SUPPORT COURSE FOR ELEMENTARY STATISTICS

Larry Green
Lake Tahoe Community College

LibreTexts™

Lake Tahoe Community College
Support Course for Elementary Statistics

Larry Green

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

The LibreTexts mission is to unite students, faculty and scholars in a cooperative effort to develop an easy-to-use online platform for the construction, customization, and dissemination of OER content to reduce the burdens of unreasonable textbook costs to our students and society. The LibreTexts project is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning by developing an Open Access Resource environment. The project currently consists of 14 independently operating and interconnected libraries that are constantly being optimized by students, faculty, and outside experts to supplant conventional paper-based books. These free textbook alternatives are organized within a central environment that is both vertically (from advance to basic level) and horizontally (across different fields) integrated.

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, 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. More information on our activities can be found via Facebook (https://facebook.com/Libretexts), Twitter (https://twitter.com/libretexts), or our blog (http://Blog.Libretexts.org).

This text was compiled on 03/18/2025

TABLE OF CONTENTS

7: Geometry

- 7.1: Angles
- 7.2: The Area of a Rectangle and Square
- 7.3: The Area of a Triangle
- 7.4: Pythagorean Theorem

8: Sampling and Data

- 8.1: Introduction
- 8.2: Definitions of Statistics, Probability, and Key Terms
- 8.3: Data, Sampling, and Variation in Data and Sampling
- 8.4: Frequency, Frequency Tables, and Levels of Measurement
- 8.5: Experimental Design and Ethics
- 8.6: Data Collection Experiment (Worksheet)
- 8.7: Sampling Experiment (Worksheet)
- 8.E: Sampling and Data (Exercises)

9: Descriptive Statistics

- 9.1: Prelude to Descriptive Statistics
- 9.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs
- 9.3: Histograms, Frequency Polygons, and Time Series Graphs
- 9.4: Measures of the Location of the Data
  - 9.4E: Measures of the Location of the Data (Exercises)
- 9.5: Box Plots
- 9.6: Measures of the Center of the Data
- 9.7: Skewness and the Mean, Median, and Mode
- 9.8: Measures of the Spread of the Data
- 9.9: Descriptive Statistics (Worksheet)
- 9.E: Descriptive Statistics (Exercises)

10: Probability Topics

- 10.1: Introduction
- 10.2: Terminology
- 10.3: Independent and Mutually Exclusive Events
- 10.4: Two Basic Rules of Probability
- 10.5: Contingency Tables
- 10.6: Tree and Venn Diagrams
- 10.7: Probability Topics (Worksheet)
- 10.E: Probability Topics (Exericses)

11: The Normal Distribution

- 11.1: Prelude to The Normal Distribution
- 11.2: The Standard Normal Distribution
  - 11.2E: The Standard Normal Distribution (Exercises)
- 11.3: Using the Normal Distribution
- 11.4: Normal Distribution - Lap Times (Worksheet)
- 11.5: Normal Distribution - Pinkie Length (Worksheet)
- 11.E: The Normal Distribution (Exercises)

Licensing

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

CHAPTER OVERVIEW

1: Decimals Fractions and Percents

This page titled 1: Decimals Fractions and Percents is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

1.1: Comparing Fractions, Decimals, and Percents

Learning Outcomes

1. Compare two fractions
2. Compare two numbers given in different forms

In this section, we will go over techniques to compare two numbers. These numbers could be presented as fractions, decimals or percents and may not be in the same form. For example, when we look at a histogram, we can compute the fraction of the group that occurs the most frequently. We might be interested in whether that fraction is greater than 25% of the population. By the end of this section we will know how to make this comparison.

Comparing Two Fractions

Whether you like fractions or not, they come up frequently in statistics. For example, a probability is defined as the number of ways a sought after event can occur over the total number of possible outcomes. It is commonly asked to compare two such probabilities to see if they are equal, and if not, which is larger. There are two main approaches to comparing fractions.

Approach 1: Change the fractions to equivalent fractions with a common denominator and then compare the numerators

The procedure of approach 1 is to first find the common denominator and then multiply the numerator and the denominator by the same whole number to make the denominators common.

Example 1.1.1

Compare: $\frac{2}{3}$ and $\frac{5}{7}$

Solution

A common denominator is the product of the two: $3 \times 7 = 21$. We convert:

$$\frac{2}{3}\frac{7}{7} = \frac{14}{21}$$

and

$$\frac{5}{7}\frac{3}{3} = \frac{15}{21}$$

Next we compare the numerators and see that $14 < 15$, hence

$\frac{2}{3} < \frac{5}{7}$

Example 1.1.2

In statistics, we say that two events are independent if the probability of the second occurring is equal to the probability of the second occurring given that the first occurs. The probability of rolling two dice and having the sum equal to 7 is $\frac{6}{36}$. If you know that the first die lands on a 4, then the probability that the sum of the two dice is a 7 is $\frac{1}{6}$. Are these events independent?

Solution

We need to compare $\frac{6}{36}$ and $\frac{1}{6}$. The common denominator is 36. We convert the second fraction to

$$\frac{1}{6}\frac{6}{6} = \frac{6}{36}$$

Now we can see that the two fractions are equal, so the events are independent.

Approach 2: Use a calculator or computer to convert the fractions to decimals and then compare the decimals

If it is easy to build up the fractions so that we have a common denominator, then Approach 1 works well, but often the fractions are not simple, so it is easier to make use of the calculator or computer.

Example 1.1.3

In computing probabilities for a uniform distribution, fractions come up. Given that the number of ounces in a medium sized drink is uniformly distributed between 15 and 26 ounces, the probability that a randomly selected medium sized drink is less than 22 ounces is $\frac{7}{11}$. Given that the weight of in a medium sized American is uniformly distributed between 155 and 212 pounds, the probability that a randomly selected medium sized American is less than 195 pounds is $\frac{40}{57}$. Is it more likely to select a medium sized drink that is less than 22 ounces or to select a medium sized American who is less than 195 pounds?

Solution

We could get a common denominator and build the fractions, but it is much easier to just turn both fractions into decimal numbers and then compare. We have:

$$\frac{7}{11} \approx 0.6364$$

and

$$\frac{40}{57} \approx 0.7018$$

Notice that

$$0.6364 < 0.7018$$

Hence, we can conclude that it is less likely to pick the medium sized 22 ounce or less drink than to pick the 195 pound or lighter medium sized person.

Exercise

If you guess on 10 true or false questions, the probability of getting at least 9 correct is $\frac{11}{1024}$. If you guess on six multiple choice questions with three choices each, then the probability of getting at least five of the six correct is $\frac{7}{729}$. Which of these is more likely?

Comparing Fractions, Decimals and Percents

When you want to compare a fraction to a decimal or a percent, it is usually easiest to convert to a decimal number first, and then compare the decimal numbers.

Example 1.1.4

Compare 0.52 and $\frac{7}{13}$.

Solution

We first convert $\frac{7}{13}$ to a decimal by dividing to get 0.5385. Now notice that

$$0.52 < 0.5385$$

Thus

$$0.52 < \frac{7}{13}$$

Example 1.1.5

When we preform a hypothesis test in statistics, We have to compare a number called the p-value to another number called the level of significance. Suppose that the p-value is calculated as 0.0641 and the level of significance is 5%. Compare these two numbers.

Solution

We first convert the level of significance, 5%, to a decimal number. Recall that to convert a percent to a decimal, we move the decimal over two places to the right. This gives us 0.05. Now we can compare the two decimals:

$$0.0641 > 0.05$$

Therefore, the p-value is greater than the level of significance.

This is an application of comparing fractions to probability.

- Example: Comparing Fractions with Different Denominators using Inequality Symbols
- Ex: Compare Fractions and Decimals using Inequality Symbols
- https://youtu.be/lSzNkQjcfEU

This page titled 1.1: Comparing Fractions, Decimals, and Percents is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Comparing Fractions, Decimals, and Percents** by Larry Green is licensed CC BY 4.0.

1.2: Converting Between Fractions, Decimals and Percents

Learning Outcomes

1. Given a decimal, convert it to a percent
2. Given a percent, convert it to a decimal
3. Convert a fraction to a decimal and percent

In this section, we will convert from decimals to percents and back. We will also start with a fraction and convert it to a decimal and a percent. In statistics we are often given a number as a percent and have to do calculations on it. To do so, we must first convert it to a percent. Also, the computer or calculator shows numbers as decimals, but for presentations, percents are friendlier. It is also much easier to compare decimals than fractions, thus converting to a decimal is helpful.

For example, we often want to see if a probability is greater than 5%. A computer will display the probability as a decimal such as 0.04836. To make the comparison we will first change it to a percent and then compare it to 5%.

Transforming a Decimal to a Percent

We have all heard of percents before. "You only have a 20% chance of winning the game", "Just 38% of all Americans approve of Congress", and "I am 95% confident that my answer is correct" are just a few of the countless examples of percents as they come up in statistics.

Defintion: Percent

Percent means Parts Per Hundred

Thus if we are given a decimal and want to convert it to a percent, we multiply the decimal by 100. In practice, this means we move the decimal point two places to the right.

Example 1.2.1

Convert the number 0.1738 to a percent.

Solution

We move the decimal over two to the right as shown below.

We get: 17.38% for the answer.

Example 1.2.2

Convert 0.7 to a percent.

Solution

We want to move the decimal two places to the right, but there is only one digit to the right of the decimal place. The good news is that we can always add a 0 to the right of the last digit. We write:

$$0.7 = 0.70$$

Now move the decimal place two digits to the right to get 70%.

Example 1.2.3

In regression analysis, an important number that is calculated is called R-Squared. It helps us determine how helpful one variable is in predicting another variable. The computer and calculator always display it as a decimal, but it is more meaningful

as a percent. Suppose that the R-Squared value that relates the amount of studying students do to prepare for a final exam and the score on the exam is: $r^2 = 0.8971$. Convert this to a percent rounded to the nearest whole number percent.

Solution

We move the decimal 0.8971 two places to the right to get 89.71%

Now round to the nearest whole number percent. Note that the digit to the left of the whole number is $7 \geq 5$. Thus we add 1 to the whole number, 89. This gives us 90%.

Exercise

A standard goal in statistics is to come up with a range of values that a population proportion is likely to lie. This range is called a confidence interval. Suppose that we want to interpret a confidence interval for the percent of patients who experience side effects from an experimental cancer treatment. The computer calculates it as the decimal range: [0.023,0.029]. What is the likely range for the percent of patients who experience side effects from the experimental cancer treatment?

Transforming a Percent to a Decimal

To convert a decimal to a percent, we multiply the decimal by 100 which is equivalent to moving the decimal two places to the right. Not surprisingly, to convert a percent to a decimal, we do exactly the opposite. We divide the number by 100 which is equivalent to moving the decimal two places to the left.

Example 1.2.4

Convert the percent 89.4% to a decimal.

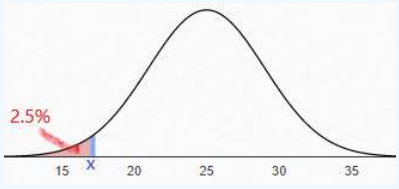
Solution

We move the decimal over two to the left as shown below.

We get: 0.894 for the answer.

Example 1.2.5

Suppose that you want to find the value of x such that 2.5% of the entire area under the Normal curve lies to the left of x. The first step will be to convert the 2.5% to a decimal. What decimal is equivalent to 2.5%?



Solution

We want to move the decimal 2.5 two places to the left, but since there is only one digit to the left of the decimal, we add a zero first: 02.5. Now move the decimal two places to the left to get 0.025.

Converting a Fraction to a Decimal and a Percent

Often in probability it is natural to represent probabilities as fractions, but it is easier to make comparisons as decimals. Thus, we need to be able to convert fractions to decimals. To do so we just divide.

Example 1.2.6

Convert the fraction $\frac{4}{7}$ to a decimal, rounding to the nearest hundredth.

Solution

We use long division:

$$\begin{array}{r} .571 \\ \hline 7)4.000 \\ \underline{35} \\ 50 \\ \underline{49} \\ 10 \end{array} \tag{1.2.1}$$

Next round to the nearest hundredth to get 0.57.

Although everyone's favorite thing to do is to perform long division by hand, in most statistics classes you will have a calculator or computer to use. Thus you just have to remember to perform the division with the calculator or computer and then round.

Example 1.2.7

In statistics we need to find basic probabilities and create a table for them. Suppose that you roll two six-sided dice, what percent of the time will the sum equal to a 4? Round to the nearest whole number percent.

Solution

First, notice that there are 36 total possibilities for rolling the dice, since there are 6 faces on the first die and for each value of the first die roll, there are 6 possibilities for the second die roll. Multiplying: 6 x 6 = 36. This will be the denominator. To find the numerator, we list all the possible outcome where the sum is 4:

(1,3), (2,2), and (3,1)

There are three possible outcomes with the sum equaling a 4. Thus:

$$P(\text{sum} = 4) = 3/36$$

Now we divide:

$$\frac{3}{36} = 0.08333...$$

Next to convert this decimal to a percent, we move the decimal two places to the right to get: 8.333...%

We are asked to round to the nearest whole number percent. The digit to the right or the whole number (8) is a 3. Since 3 < 5, we can just erase everything to the left of the 8 and leave the 8 unchanged to get 8%. Thus there is an 8% chance of getting a sum of 4 if you roll two six sided dice.

- Convert Percentages to Decimals
- Relating Fractions, Decimals, and Percents
- Statistics Application of Converting Decimals to Percents

This page titled 1.2: Converting Between Fractions, Decimals and Percents is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Converting Between Fractions, Decimals and Percents** by Larry Green is licensed CC BY 4.0.

1.3: Decimals- Rounding and Scientific Notation

Learning Outcomes

1. Understand what it means to have a number rounded to a certain number of decimal places.
2. Round a number to a fixed number of digits.
3. Convert from scientific notation to decimal notation and back.

In this section, we will go over how to round decimals to the nearest whole number, nearest tenth, nearest hundredth, etc. In most statistics applications that you will encounter, the numbers will not come out evenly, and you will need to round the decimal. We will also look at how to read scientific notation. A very common error that statistics students make is not noticing that the calculator is giving an answer in scientific notation.

For example, suppose that you used a calculator to find the probability that a randomly selected day in July will have a high temperature of over 90 degrees. Your calculator gives the answer: 0.4987230156. This is far too many digits for practical use, so it makes sense to round to just a few digits. By the end of this section you will be able to perform the rounding that is necessary to make unmanageable numbers manageable.

Brief Review of Decimal Language

Consider the decimal number: 62.5739. There is a defined way to refer to each of the digits.

- The digit 6 is in the "Tens Place"
- The digit 2 is in the "Ones Place"
- The digit 5 is in the "Tenths Place"
- The digit 7 is in the "Hundredths Place"
- The digit 3 is in the "Thousandths Place"
- The digit 9 is in the "Ten-thousandths Place"
- We also say that 62 is the "Whole Number" part.

Keeping this example in mind will help you when you are asked to round to a specific place value.

Example 1.3.1

It is reported that the mean number of classes that college students take each semester is 3.2541. Then the digit in the *hundredths place* is 5.

Rules of Rounding

Now that we have reviewed place values of numbers, we are ready to go over the process of rounding to a specified place value. When asked to round to a specified place value, the answer will erase all the digits after the specified digit. The process to deal with the other digits is best shown by examples.

Example 1.3.2: Case 1 - The Test Digit is Less Than 5

Round 3.741 to the nearest tenth.

Solution

Since the test digit (4) is less than 5, we just erase everything to the right of the tenths digit, 7. The answer is: 3.7.

Example 1.3.3: Case 2 - The Test Digit is 5 or Greater

Round 8.53792 to the nearest hundredth.

Solution

Since the test digit (6) is 5 or greater, we add one to the hundredths digit and erase everything to the right of the hundredths digit, 3. Thus the 3 becomes a 4. The answer is: 8.54.

Example 1.3.4: Case 3 - The Test Digit is 5 or Greater and the rounding position digit is a 9

Round 0.014952 to four decimal places.

Solution

The test digit is 5, so we must round up. The rounding position is a 9 and adding 1 gives 10, which is not a single digit number. Instead look at the two digits to the left of the test digit: 49. If we add 1 to 49, we get 50. Thus the answer is 0.0150.

Applications

Rounding is used in most areas of statistics, since the calculator or computer will produce numerical answers with far more digits than are useful. If you are not told how many decimal places to round to, then you often want to think about the smallest number of decimals to keep so that no important information is lost. For example suppose you conducted a sample to find the proportion of college students who receive financial aid and the calculator presented 0.568429314. You could turn this into a percent at 56.8429314%. There are no applications where keeping this many decimal places is useful. If, for example, you wanted to present this finding to the student government, you might want to round to the nearest whole number. In this case the ones digit is 6 and the test digit is 8. Since $8 \geq 5$, you add 1 to the ones digit. You can tell the student government that 57% of all college students receive financial aid.

Example 1.3.5

Suppose that you found out that the probability that a randomly selected person with who has misused prescription opioids will transition to heroin is 0.04998713. Round this number to four decimal places.

Solution

The first four decimal places are 0.0499 and the test digit is 8. Since $8 \geq 5$, we would like to add 1 to the fourth digit. Since this is a 9, we go to the next digit to the left. This is also a 9, so we go to the next one which is a 4. We can think of adding 0499 +

1 = 0500. Thus the answer is 0.0500. Note that we keep the last two 0's after the 5 to emphasize that this is accurate to the fourth decimal place.

Rounding and Arithmetic

Many times, we have to do arithmetic on numbers with several decimal places and want the answer rounded to a smaller number of decimal places. One question you might ask is should you round before you perform the arithmetic or after. For the most accurate result, you should always round after you preform the arithmetic if possible.

> *When asked to do arithmetic and present you answer rounded to a fixed number of decimal places, only round after performing the arithmetic.*

Example 1.3.6

Suppose you pick three cards from a 52 card deck with replacement and want to find the probability of the event, A, that none of the three cards will be a 2 through 7 of hearts. This probability is:

$$P(A) = (0.8846)^3$$

Round the answer to 2 decimal places.

Solution

Note that we have to first perform the arithmetic. With a computer or calculator we get:

$$0.8846^3 = 0.69221467973$$

Now we round to two decimal places. Notice that the hundredths digit is a 9 and the test digit is a 2. Thus the 9 remains unchanged and everything to the right of the 9 goes away. the result is

$$P(A) \approx 0.69$$

If we mistakenly rounded 0.8846 to two decimal places (0.88) and then cubed the answer we would have gotten 0.68 which is not the correct answer.

Scientific Notation

When a calculator presents a number in scientific notation, we must pay attention to what this represents. The standard way of writing a number in scientific notation is writing the number as a product of a number greater than or equal 1 but less than 10 followed by a power of 10. For example:

$$602,000,000,000,000,000,000,000 = 6.02 \times 10^{23}$$

The main purpose of scientific notation is to allow us to write very large numbers or numbers very close to 0 without having to use so many digits. Most calculators and computers use a different notation for scientific notation, most likely because the superscript is difficult to render on a screen. For example, with a calculator:

$$0.00000032 = 3.2E-7$$

Notice that to arrive at 3.2, the decimal needed to be moved 7 places to the right.

Example 1.3.7

A calculator displays:

$$2.0541E6$$

Write this number in decimal form.

Solution

Notice that the number following E is 6. This means move the decimal over 6 places to the right. The first 4 moves is natural, but for the last 2 moves, there are no numbers to move the decimal place past. We can always add extra zeros after the last

number to the right of the decimal place:

$$2.0541E6 = 2.054100E6$$

Now we can move the decimal place to the right 6 places to get

$$2.0541E6 = 2.054100E6 = 2,054,100$$

Example 1.3.8

If you use a calculator or computer to find the probability of flipping a coin 27 times and getting all heads, then it will display:

$$7.45E-9$$

Write this number in decimal form.

Solution

Many students will forget to look for the "E" and just write that the probability is 7.45, but probabilities can never be bigger than 1. You can not have a 745% chance of it occurring. Notice that the number following E is −9. Since the power is negative, this means move the decimal to the left, and in particular 9 places to the left. There is only one digit to the left of the decimal place, so we need to insert 8 zeros:

$$7.45E-9 = 000000007.45E-9$$

Now we can move the decimal place to the right 9 places to the left to get

$$7.45E-9 = 000000007.45E-9 = 0.00000000745$$

- Application of Rounding Decimal Numbers
- Here is a video that explains rounding.

This page titled 1.3: Decimals- Rounding and Scientific Notation is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Decimals: Rounding and Scientific Notation** by Larry Green is licensed CC BY 4.0.

1.4: Using Fractions, Decimals and Percents to Describe Charts

Learning Outcomes

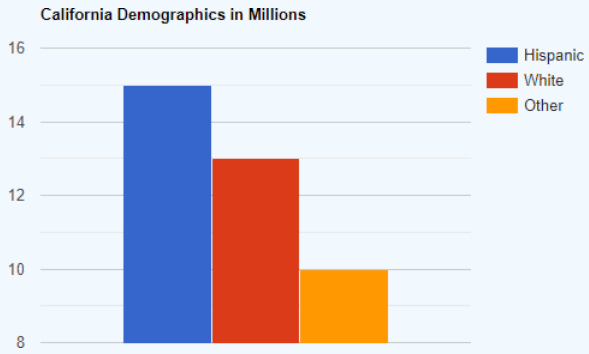
1. Interpret bar charts using fractions, decimals and percents
2. Interpret pie charts using fractions, decimals and percents

Charts, such as bar charts and pie charts are visual ways of presenting data. You can think of each slice of the pie or each bar as a part of the whole. The numerical versions of this are a list of fractions, decimals and percents. By the end of this section we will be able to look at one of these charts and produce the corresponding fractions, decimals, and percents.

Reading a Bar Chart

Bar charts occur frequently and it is definitely required to understand how to read them and interpret them in statistics. Often we want to convert the information of a bar chart to information shown numerically. We need fractions and/or percents to do this.

Example 1.4.1



The above bar chart shows the demographics of California in 2019 where the numbers represent millions of people. Here are some questions that might come up in a statistics class.

A. What fraction of Californians was Hispanic in 2019?
B. What proportion of all Californians was White in 2019? Write your answer as a decimal number rounded to four decimal places.
C. What percent of Californians who were neither Hispanic nor White in 2019? Round your answer to the nearest percent.

Solution

A. To find the fraction of California that was Hispanic in 2019, the numerator will be the total number of Hispanics and the denominator will be the total number of people in California in 2019. The height of the bar that represents Hispanics is 15. Therefore the numerator is 15. To find the total number of people in California, we add up the heights of the three bars:

$$15+13+10 = 38$$

Now we can just write down the fraction:

$$\frac{15}{38}$$

To find the proportion of Californians who were White in 2019, we start in the same way. The numerator will be the number of Whites: 13. The denominator will be the total number of Californians which we already computed as 38. Therefore the fraction of Californians who were White is:

$$\frac{13}{38}$$

To convert this to a decimal, we use a calculator to get:

$$\frac{13}{38} \approx 0.342105$$

Next round to four decimal places. Since the digit to the right of the fourth decimal place is $0 < 5$, we round down to:

$$0.3421$$

B. To find the percent of Californians who were neither Hispanic nor White in 2019, we first find the fraction who were neither. The numerator will be the number of "Other" which is: 10. The denominator will be the total which is 38. Thus the fraction is:

$$\frac{10}{38}$$

Next, use a calculator to divide these numbers to get:

$$\frac{10}{38} \approx 0.263158$$

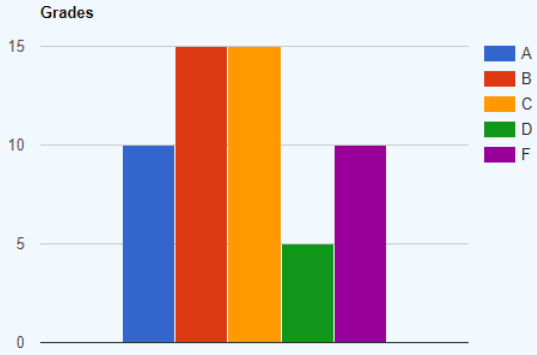
To convert this to a percent we multiply by 100% by moving the decimal two places to the right:

$$0.263158 \times 100\% = 26.3158\%$$

Finally we round to the nearest whole number. Noting that $3 < 5$, we round down to get: 26%

Exercise

The bar chart below shows the grade distribution for a math class.



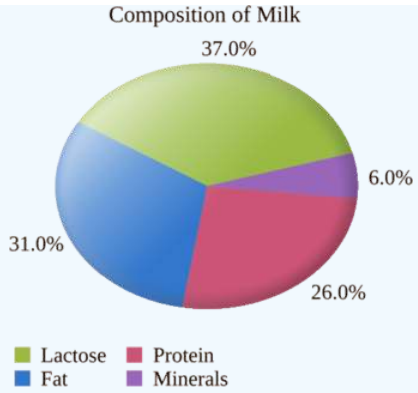
A. Find the fraction of students who received a "C" grade.
B. Find the proportion of grades below a "C". Write your answer as a decimal number rounded to the nearest hundredth.
C. What percent of the students received an "A" grade? Round your answer to the nearest whole number percent.

Reading a Pie Chart

Another important chart that is used to display the components of a whole is a pie chart. With a pie chart, it is very easy to determine the percent of each item.

Example 1.4.2

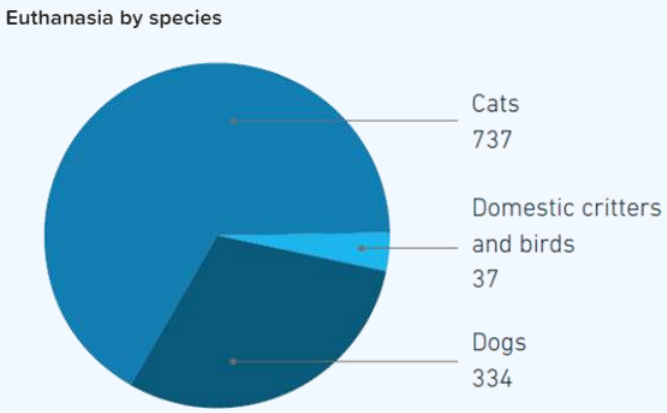
The pie chart below shows the makeup of milk. Write the proportion of fat contained in milk as a decimal.



Solution

We see that 31% of milk is fat. To convert a percent to a decimal, we just move the decimal over two places to the left. Thus, 31% becomes 0.31.

Example 1.4.3



The pie chart above shows the number of pets of each type that had to be euthanized by the humane society due to incurable illnesses.

A. What fraction of the euthanized pets were dogs?
B. What percent of the euthanized pets were cats? Round to the nearest whole number percent.

Solution

A. We take the number of dogs over the total. There were 334 euthanized dogs. To find the total we add:

$$737 + 37 + 334 = 1108$$

Therefore, the fraction of euthanized dogs is

$$\frac{334}{1108}$$

B. To find the percent of euthanized cats, we first find the fraction. There were 737 cats over a total of 1108 pets. The fraction is

$$\frac{737}{1108}$$

Next use a calculator to get the decimal number: 0.66516. Now multiply by 100% by moving the decimal place two digits to the right to get: 66.516%. Finally, we need to round to the nearest whole number percent. Since $5 \geq 5$, we round up.

> Thus the percent of euthanized cats is 67%.

- Finding Fractions, Decimals and Percents from a Bar Chart
- Ex: Find the a Percent of a Total Using an Amount in Pie Chart

---

This page titled 1.4: Using Fractions, Decimals and Percents to Describe Charts is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Using Fractions, Decimals and Percents to Describe Charts** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

2: The Number Line

Thumbnail: Demonstration the addition on the line number. (CC BY 3.0 unported; Stephan Kulla).

This page titled 2: The Number Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

2.1: Distance between Two Points on a Number Line

Learning Outcomes

1. Calculate the distance between two points on a number line when both are non-negative.
2. Calculate the distance between two points on a number line when at least one is negative.

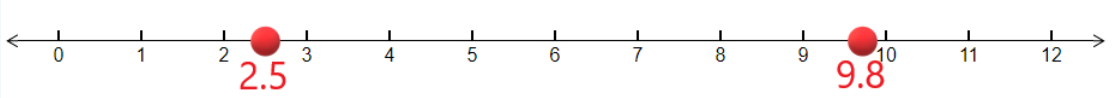
The number line is the main visual base in statistics and we often want to look at two points on the number line and determine the distance between them. This is used to find the base of a rectangle or another figure that lies above the number line. By the end of this section, you will be able to determine the distance between any two points on a number line that comes from a statistics application.

Finding the Distance Between Two Points with Positive Coordinates on a Number Line

The key to finding the distance between two points is to remember that the geometric definition of subtraction is the distance between the two numbers as long as we subtract the smaller number from the larger.

Example 2.1.1

Find the distance between the points 2.5 and 9.8 as shown below on the number line.



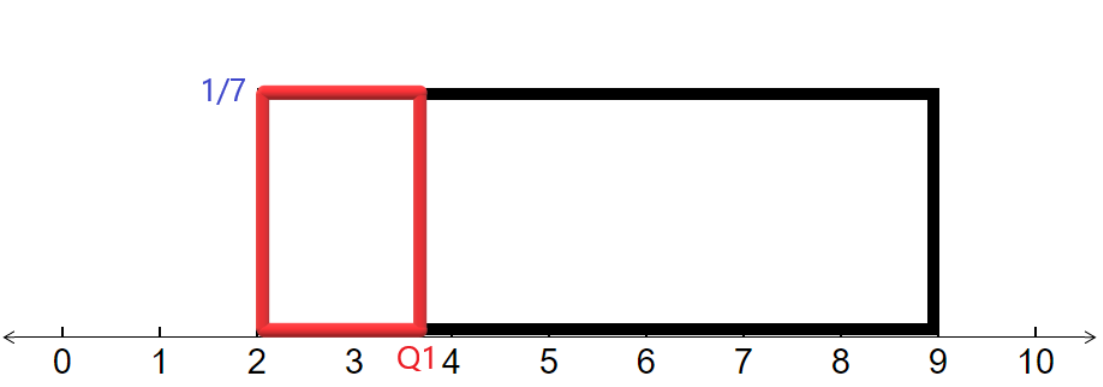
Solution

To find the distance, we just subtract:

$$9.8 - 2.5 = 7.3$$

Example 2.1.2

When finding probabilities involving a uniform distribution, we have to find the base of a rectangle that lies on a number line. Find the base of the rectangle shown below that represents a uniform distribution from 2 to 9.



Solution

We just subtract:

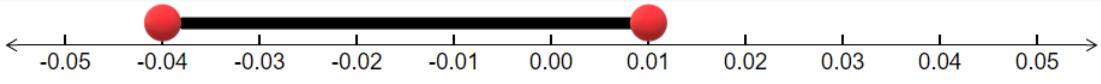
$$9 - 2 = 7$$

Finding the Distance Between Two Points on a Number Line When the Coordinates Are Not Both Positive

In statistics, it is common to have points on a number line where the points are not both positive and we need to find the distance between them.

Example 2.1.3

The diagram below shows the confidence interval for the difference between the proportion of men who are planning on going into the health care profession and the proportion of women. What is the width of the confidence interval?



Solution

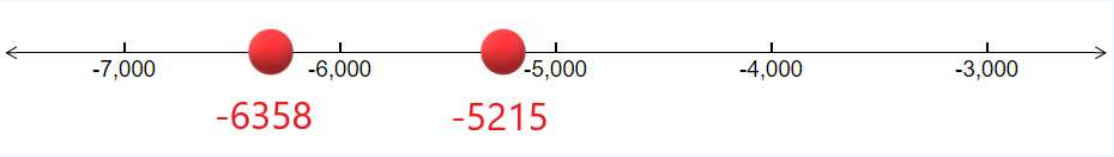
Whenever we want want to find the distance between two numbers, we always subtract. Recall that subtracting a negative number is adding.

$$0.01 - (-0.04) = 0.01 + 0.04 = 0.05$$

Therefore the width of the confidence interval is 0.05.

Example 2.1.4

The mean value of credit card accounts is -6358 dollars. A study was done of recent college graduates and found their mean value for their credit card accounts was -5215 dollars. The number line below shows this situation. How far apart are these values?



Solution

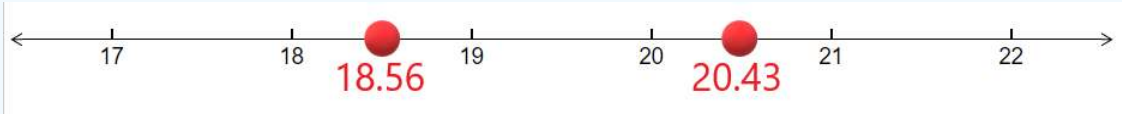
We subtract the two numbers and recall that when we subtract two negative numbers when we are looking at the right minus the left, we make them positive and subtract the positive numbers.

$$-5215 - (-6358) = 6358 - 5215 = 1143$$

Thus the mean credit card balances are $1143 apart.

Exercise

In statistics, we are asked to find a z-score, which tells us how unusual an event is. The first step in finding a z-score is to calculate the distance a value is from the mean. The number line below depicts the mean of 18.56 and the value of 20.43. Find the distance between these two points.



- Finding the Distance Between Points on a Number Line
- Integer Subtracton Using the Number Line

This page titled 2.1: Distance between Two Points on a Number Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Distance between Two Points on a Number Line** by Larry Green is licensed CC BY 4.0.

2.2: Plotting Points and Intervals on the Number Line

Learning Outcomes

1. Plot a point on the number line
2. Plot an interval on the number line

The number line is of fundamental importance and is used repeatedly in statistics. It is a tool to visualize all of the possible outcomes of a study and to organize the results of the study. Often a diagram is placed above the number line to provide us with a picture of the results. By the end of this section, you will be able to plot points and intervals on a number line and use these plots to understand the possible outcomes and actual outcomes of studies.

Drawing Points on a Number Line

A number line is just a horizontal line that is used to display all the possible outcomes. It is similar to a ruler in that it helps us describe and compare numbers. Similar to a ruler that can be marked with many different scales such as inches or centimeters, we get to choose the scale of the number line and where the center is.

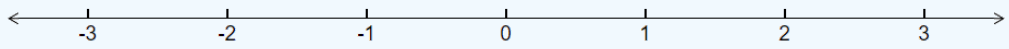
Example 2.2.1

The standard normal distribution is plotted above a number line. The most important values are the integers between -3 and 3. The number 0 is both the mean (average) and median (center).

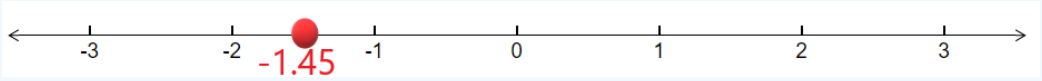
1. Plot the number line that best displays this information.
2. Plot the value -1.45 on this number line.

Solution

1. We sketch a line, mark 0 as the center, and label the numbers -3, -2, -1, 0, 1, 2, 3 from left to right.



2. To plot the point -1.45, we first have to understand that this number is between -1 and -2. It is close to half way between -1 and -2. We put a circle on the number line that is close to halfway between these values as shown below.

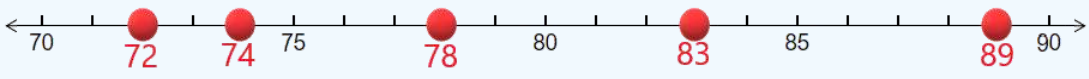


Example 2.2.2

When working with box plots, we need to first set up a number line that labels what is called the five point summary: Minimum, First Quartile, Median, Third Quartile, and Maximum. Suppose the five point summary for height in inches for a basketball team is: 72,74,78,83,89. Plot these points on a number line

Solution

When plotting points on a number line, we first have to decide what range of the line we want to show in order to best display the points that appear. Technically all numbers are on every number line, but that does not mean we show all numbers. In this example, the numbers are all between 70 and 90, so we certainly don't need to display the number 0. A good idea is to let 70 be on the far left and 90 be on the far right and then plot the points between them. We also have to decide on the spacing of the tick marks. Since the range from 70 to 90 is 20, this may be too many numbers to display. Instead we might want to count by 5's. Below is the number line that shows the numbers 70 to 90 and counts by 5's. The five point summary is plotted on this line.



Exercise

A histogram will be drawn to display the annual income that experienced registered nurses make. The boundaries of the bars of the histogram are: $81,000, $108,000, $135,000, $162,000, and $189,000. Plot these points on a number line.

Plotting an Interval on a Number Line

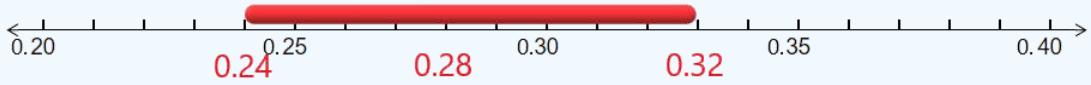
Often in statistics, instead of just having to plot a few points on a number line, we need to instead plot a whole interval on the number line. This is especially useful when we want to exhibit a range of values between two numbers, to the left of a number or to the right of a number.

Example 2.2.3

A 95% confidence interval for the proportion of Americans who work on weekends is found to be 0.24 to 0.32, with the center at 0.28. Use a number line to display this information.

Solution

We just draw a number line, include the three key numbers: 0.24, 0.32, and 0.28 and highlight the part of the interval between 0.23 and 0.31.

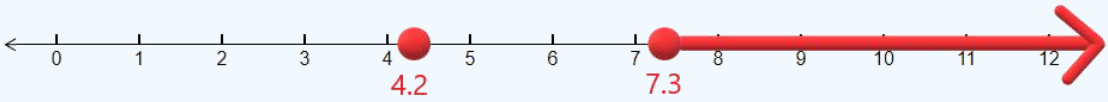


Example 2.2.4: rejection region

In Hypothesis testing, we sketch something called the rejection region which is an interval that goes off to infinity or to negative infinity. Suppose that the mean number of hours to work on the week's homework is 4.2. The rejection region for the hypothesis test is all numbers larger than 7.3 hours. Plot the mean and sketch the rejection region on a number line.

Solution

We plot the point 4.2 on the number line and shade everything to the right of 7.3 on the number line.



- Plot Integers on the Number Line
- Intervals: Given an Inequality, Graph the Interval and State Using Interval Notation
- Plotting Points on a Number Line Application

This page titled 2.2: Plotting Points and Intervals on the Number Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Plotting Points and Intervals on the Number Line** by Larry Green is licensed CC BY 4.0.

2.3: Represent an Inequality as an Interval on a Number Line

Learning Outcomes

1. Graph and inequality on a number line.
2. Graph the complement on a number line for both continuous and discrete variables.

Inequalities come up frequently in statistics and it is often helpful to plot the inequality on the number line in order to visualize the inequality. This helps both for inequalities that involve real numbers and for inequalities that refer to just integer values. As an extension of this idea, we often want to look at the complement of an inequality, that is all numbers that make the inequality false. In this section we will look at examples that accomplish this task.

Sketching an Inequality on a number line where the possible values are real numbers.

There are four different inequalities: $<$, $\leq$, $>$, $\geq$. What makes this the most challenging is when they are expressed in words. Here are some of the words that are used for each:

- $<$: "Less Than", "Smaller", "Lower", "Younger"
- $\leq$: "Less Than or Equal to", "At Most", "No More Than", "Not to Exceed"
- $>$: "Greater Than", "Larger", "Higher", "Bigger", "Older", "More Than"
- $\geq$: "Greater Than or Equal to", "At Least", "No Less than"

These are the most common words that correspond to the inequalities, but there are others that come up less frequently.

Example 2.3.1

Graph the inequality: $3 < x \leq 5$ on a number line

Solution

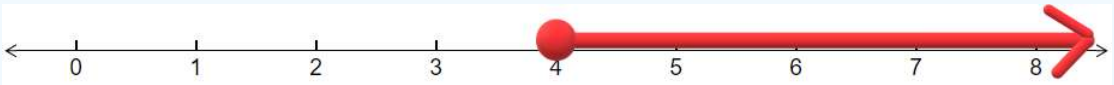
First notice that the interval does not include the number 3, but does include the number 5. We can represent not including a number with an open circle and including a number with a closed circle. The number line representation of the inequality is shown below.

Example 2.3.2

In statistics, we often want to find probabilities of an event being at least as large or no more than a given value. It helps to first plot the interval on a number line. Suppose you want to find the probability that you will have to wait in line for at least 4minutes. Sketch this inequality on a number line.

Solution

First, notice that "At Least" has the symbol $\geq$. Thus, we have a closed circle on the number 4. There is no upper bound, so we draw a long arrow from 4 to the right of 4. The solution is shown below

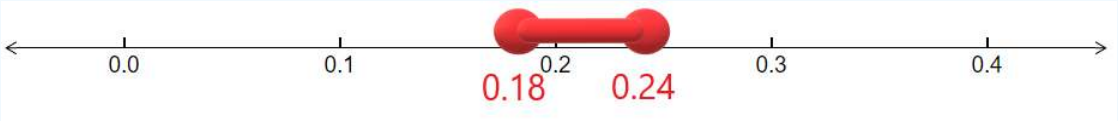


Example 2.3.3

Another main topic that comes up in statistics is confidence intervals. For example in recent poll to see the percent of Americans who think that Congress is doing a good job found that a 95% confidence interval had lower bound of 0.18 and an upper bound of 0.24. This can be written as [0.18,0,24]. Sketch this interval on the number line.

Solution

The first thing we need to do is decide on the tick marks to put on the number line. If we counted by 1's, then the interval of interest would be too small to stand out. Instead we will count by 0.1's. The number line is shown below.

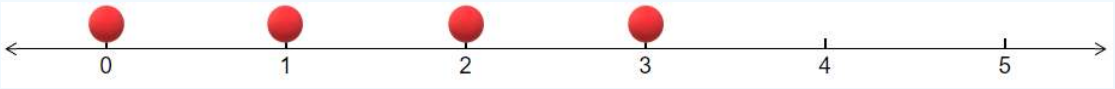


Example 2.3.4

Often in statistics, we deal with discrete variables. Most of the time this will mean that only whole number values can occur. For example, you want to find out the probability that a college student is taking at most three classes. Graph this on a number line.

Solution

First note that the outcomes can only be whole numbers. Second, note that "at most" means $\leq$. Thus the possible outcomes are: 0, 1, 2, and 3. The number line below displays these outcomes.



Graphing the Complement

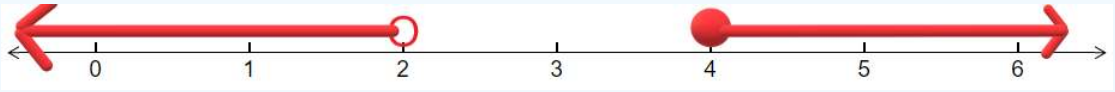
In statistics, we often want to graph the complement of an interval. The complement means everything that is not in the interval.

Example 2.3.5

Graph the complement of the interval [2,4).

Solution

Notice that the complement of numbers inside the interval between 2 and 4 is the numbers outside that interval. This will consist of the numbers to the left of 2 and to the right of 4. Since the number 2 is included in the original interval, it will not be included in the complement. Since the number 4 is not included in the original interval, it will be included in the complement. The complement is shown on the number line below.

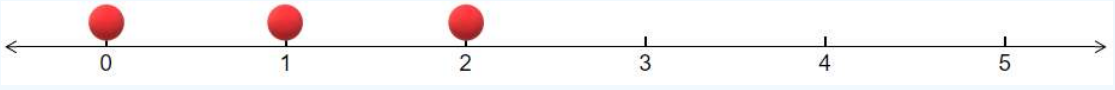


Example 2.3.6

Some calculators can only find probabilities for values less than a certain number. If we want the probability of an interval greater than a number, we need to use the complement. Suppose that you want to find the probability that a person will have traveled to more than two foreign countries in the last twelve months. Find the complement of this and graph it on a number line.

Solution

First notice that only whole numbers are possible since it does not make sense to go to a fractional number of countries. Second note that the lowest number that is more than 2 is 3. If 3 is included in the original list, then 3 will not be included in the complement. Thus, the highest number that is in the complement of "more than 2" is 2. The number line below shows the complement of more than 2.



Exercise

Suppose you want to find the probability that at least 4 people in your class have a last name that contains the letter "W". To make this calculation you will need to first find the complement of "at least 4". Sketch this complement on the number line.

- Intervals: Given an Inequality, Graph the Interval and State Using Interval Notation
- Express Inequalities as a Graph and Interval Notation
- Sketching the Complement of an Interval on a Number Line

This page titled 2.3: Represent an Inequality as an Interval on a Number Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Represent an Inequality as an Interval on a Number Line** by Larry Green is licensed CC BY 4.0.

2.4: The Midpoint

Learning Outcomes

1. Find the midpoint between two numbers.
2. Sketch the midpoint of two numbers on a number line.

As the word sounds, "midpoint" means "the point in the middle". Finding a midpoint is not too difficult and has applications in many areas of statistics, from confidence intervals to sketching distributions, to means.

Finding the Midpoint Between Two Numbers

If we are given two numbers, then the midpoint is just the average of the two numbers. To calculate the midpoint, we add them up and then divide the result by 2. The formula is as follows:

Definition: the Midpoint

Let a and b be two numbers. Then the midpoint, M of these two numbers is

$$M = \frac{a+b}{2} \tag{2.4.1}$$

Example 2.4.1

Find the midpoint of the numbers 3.5 and 7.2.

Solution

The most important thing about finding the midpoint is that the addition of the two numbers must occur before the division by 2. We can either do this one step at a time in our calculator or we can enclose the sum in parentheses. In this example we will perform the addition first:

$$3.5+7.2 = 10.7$$

Now we are ready to divide by 2:

$$\frac{10.7}{2} = 5.35$$

Thus the midpoint of 3.5 and 7.2 is 5.35.

Example 2.4.2

A major topic in statistics is the confidence interval which tells us the most likely interval that the mean or the proportion will lie in. Often the lower and upper bound of the confidence interval are given, but the midpoint of these two numbers is the best guess for what we are looking for. Suppose a 95% confidence interval for the difference between two means is -1.34 and 2.79. Find the midpoint of these numbers, which is the best guess for the difference between the two means.

Solution

We use the formula for the midpoint (Equation 2.4.1):

$$M = \frac{a+b}{2} = \frac{-1.34+2.79}{2}$$

Now let's use a calculator. We will need parentheses around the numerator:

$$(-1.34+2.79) \div 2 = 0.725$$

Thus, the midpoint of the numbers -1.34 and 2.79 is 0.725.

Sketching the Midpoint on a Number Line

Visualizing the midpoint can often reveal it much better than just writing down its value. The diagrams are of fundamental importance in statistics.

Example 2.4.3

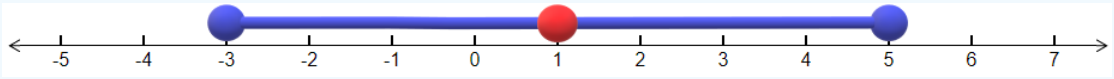
Sketch the points -3, 5 and the midpoint of these two numbers on a number line.

Solution

We start by finding the midpoint using the midpoint formula (Equation 2.4.1):

$$M = \frac{-3+5}{2} = (-3+5) \div 2 = 1$$

Now we sketch these three points on the number line:

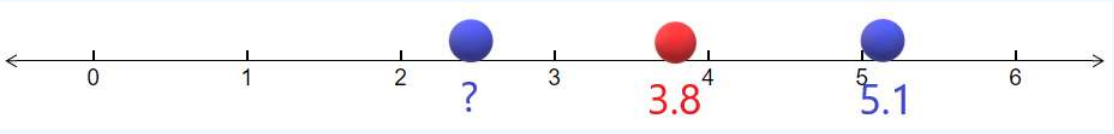


Example 2.4.4: hypothesis testing

Another application of the midpoint involves hypothesis testing. Sometimes we are given the hypothesized mean, which is the midpoint. We are also given the sample mean, which is either the left or right endpoint. The goal is to find the other endpoint. Suppose that the midpoint (hypothesized mean) is at 3.8 and the right endpoint (sample mean) is at 5.1. Find the value of the left endpoint.

Solution

It helps to sketch the diagram on the number line as shown below.



Now since 3.8 is the midpoint, the distance from the left endpoint to the midpoint is equal to the distance from 3.8 to 5.1. The distance from 3.8 to 5.1 is:

$$5.1 - 3.8 = 1.3$$

Therefore the left endpoint is 1.3 to the left of 3.8. This can be found by subtracting the two numbers:

$$3.8 - 1.3 = 2.5$$

Therefore the left endpoint is at 2.5.

Exercise

Suppose that the midpoint (hypothesized proportion) is at 0.31 and the left endpoint (sample proportion) is at 0.28. Find the value of the right endpoint.

- Midpoint on the Number line
- Finding the Right Endpoint Given the Left Endpoint and Midpoint

---

This page titled 2.4: The Midpoint is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **The Midpoint** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

3: Operations on Numbers

This page titled 3: Operations on Numbers is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

3.1: Area of a Rectangle

Learning Outcomes

- Find the area of a rectangle.
- Find the height of a rectangle given that the area is equal to 1.

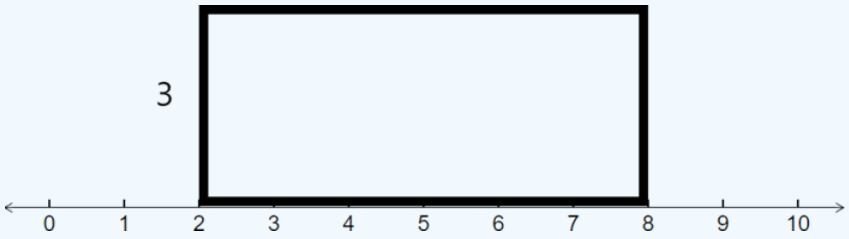
Rectangles are of fundamental importance in the portion of statistics that involves the uniform distribution. Every rectangle has a base and a height and an area. The formula for the area of a rectangle is:

$$\text{Area} = \text{Base} \times \text{Height} \tag{3.1.1}$$

When working with the uniform distribution, the area represents the probability of an event being within the bounds of the base.

Example 3.1.1

Consider the rectangle shown below.



Find the area of this rectangle.

Solution

We use the Area formula (Equation 3.1.1). To find the base, we notice that it runs from 2 to 8, so we subtract these numbers to get the base:

$$Base = 8 - 2 = 6$$

Next multiply by the height, 3, to get

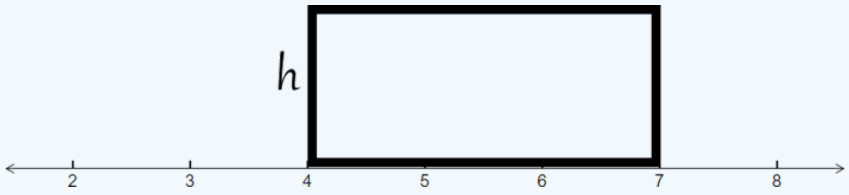
$$Area = Base \times Height = 6 \times 3 = 18$$

Example 3.1.2

It turns out that the area of the rectangles that equal to 1 will occur the most often for a uniform distribution. Suppose that we know that the area of a rectangle that depicts a uniform distribution is equal to 1 and that the base of the rectangle goes from 4 to 7. Find the height of the rectangle.

Solution

First sketch the rectangle below, labeling the height as h.



Next, find the base of the rectangle that goes from 4 to 7 by subtracting:

$$Base = 7 - 4 = 3$$

Next, plug in what we know into the area equation:

$$1 = Area = Base \times Height = 3 \times h$$

This tell us that 3 times a number is equal to 1. To find out what the number is, we just divide both sides by 3 to get:

$$h = \frac{1}{3}$$

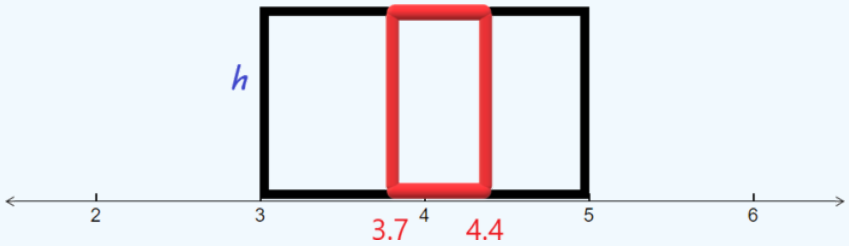
Therefore the height of an area 1 rectangle with base from 4 to 7 is $\frac{1}{3}$.

Example 3.1.3

Suppose that we know that the area of a rectangle that depicts a uniform distribution is equal to 1 and that the base of the rectangle goes from 3 to 5. There is a smaller rectangle within the larger one with the same height, but whose base goes from 3.7 to 4.4. Find the area of the smaller rectangle.

Solution

First, sketch the larger rectangle with the smaller rectangle shaded in.



Next, we find the height of the rectangle. We know that the area of the larger rectangle is 1. The base goes from 3 to 5, so the base is $5 - 3 = 2$ Hence:

$$1 = Area = Base \times Height = 2h$$

Dividing by 2, gives us that the height is $\frac{1}{2}$ or 0.5. Now we are ready to find the area of the smaller rectangle. We first find the base by subtracting:

$$\text{Base} = 4.4 - 3.7 = 0.7$$

Next, use the area formula:

$$Area = Base \times Height = 0.7 \times 0.5 = 0.35$$

Exercise 3.1.1

Suppose that elementary students' ages are uniformly distributed from 5 to 11 years old. The rectangle that depicts this has base from 5 to 11 and area 1. The rectangle that depicts the probability that a randomly selected child will be between 6.5 and 8.6 years old has base from 6.5 to 8.6 and the same height as the larger rectangle. Find the area of the smaller rectangle

- Ex: Determine the Area of a Rectangle Involving Whole Numbers
- Area of a Rectangle and the Uniform Distribution

This page titled 3.1: Area of a Rectangle is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Area of a Rectangle** by Larry Green is licensed CC BY 4.0.

3.2: Factorials and Combination Notation

Learning Outcomes

1. Evaluate a factorial.
2. Use combination notation for statistics applications.

When we need to compute probabilities, we often need to multiple descending numbers. For example, if there is a deck of 52 cards and we want to pick five of them without replacement, then there are 52 choices for the first pick, 51 choices for the second pick since one card has already been picked, 50 choices for the third, 49 choices for the fourth, and 48 for the fifth. If we want to find out how many different outcomes there are, we can use what we call the multiplication principle and multiple them: $52 \times 51 \times 50 \times 49 \times 48$. If we wanted to pick all 52 of the cards one at a time, then this list would be excessively long. Instead there is a notation that describes multiplying all the way down to 1, called the factorial. It must be exciting, since we use the symbol "!" for the factorial.

Example 3.2.1

Calculate $4!$

Solution

We use the definition which says start at 4 and multiply until we get to 1:

$$4! = 4 \times 3 \times 2 \times 1 = 24$$

Example 3.2.2

If we pick 5 cards from a 52 card deck without replacement and the same two sets of 5 cards, but in different orders, are considered different, how many sets of 5 cards are there?

Solution

From the introduction, the number of sets is just:

$$52 \times 51 \times 50 \times 49 \times 48$$

This is not quite a factorial since it stops at 48; however, we can think of this as $52!$ with $47!$ removed from it. In other words we need to find

$$\frac{52!}{47!}$$

We could just multiply the numbers from the original list, but it is a good idea to practice with your calculator or computer to find this using the ! symbol. When you do use technology, you should get:

$$\frac{52!}{47!} = 311,875,200$$

Combinations

One of the most important applications of factorials is combinations which count the number of ways of selecting a smaller collection from a larger collection when order is not important. For example if there are 12 people in a room and you want to select a team of 4 of them, then the number of possibilities uses combinations. Here is the definition:

Definition: Combinations

The number of ways of selecting k items without replacement from a collection of n items when order does not matter is:

$$\binom{n}{r} = {}_nC_r = \frac{n!}{r!\,(n-r)!} \tag{3.2.1}$$

Notice that there are a few notations. The first is more of a mathematical notation while the second is the notation that a calculator uses. For example, in the TI 84+ calculator, the notation for the number of combinations when selecting 4 from a collection of 12 is:

$$12\ {}_nC_r\ 4$$

There are many internet sites that will perform combinations. For example the math is fun site asks you to put in n and r and also state whether order is important and repetition is allowed. If you click to make both "no" then you will get the combinations.

Example 3.2.3

Calculate

$$\binom{15}{11} = {}_{15}C_{11}$$

Solution

Whether you use a hand calculator or a computer you should get the number: 1365

Example 3.2.4

The probability of winning the Powerball lottery if you buy one ticket is:

$$P(win) = \frac{1}{{}_{69}C_5 \times 26}$$

Calculate this probability.

Solution

First, let's calculate ${}_{69}C_5$. Using a calculator or computer, you should get 11,238,513. Next, multiply by 26 to get

$$11,238,513 \times 26 = 292,201,338$$

Thus, there is a one in 292,201,338 chance of winning the Powerball lottery if you buy a ticket. We can also write this as a decimal by dividing:

$$P\left(win\right) = \frac{1}{292,201,338} = 0.000000003422$$

As you can see, your chances of winning the Powerball are very small.

Exercise

A classroom is full of 28 students and there will be one president of the class and a "Congress" of 4 others selected. The number of different leadership group possibilities is:

$$28 \times {}_{27}C_4$$

Calculate this number to find out how many different leadership group possibilities there are.

Ex 1: Simplify Expressions with Factorials

Combinations

Combinations

This page titled 3.2: Factorials and Combination Notation is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Factorials and Combination Notation** by Larry Green is licensed CC BY 4.0.

3.3: Order of Operations

Learning Outcomes

1. Use the order of operations to correctly perform multi-step arithmetic
2. Apply the order of operations to statistics related complex questions.

When we are given multiple arithmetic operations within a calculation, there is a, established order that we must do them in based on how the expression is written. Understanding these rules is especially important when using a calculator, since calculators are programmed to strictly follow the order of operations. This comes up in every topic in statistics, so knowing the order of operations is an essential skill for all successful statistics students to have.

PEMDAS

The order of operations are as follows:

1. **P**arentheses
2. **E**xponents
3. **M**ultiplication and **D**ivision
4. **A**ddition and **S**ubtraction

When there is a tie, the rule is to go from left to right.

Notice that Multiplication and division are listed together as item 3. If you see multiplication and division in the same expression the rule is to go from left to right. Similarly, if you see addition and subtraction in the same expression the rule is to from go left to right. The same goes for two of the same arithmetic operators.

Example 3.3.1

Evaluate: $20-6\div 3+\left(2\times 3^2\right)$

Solution

We start with what is inside the parentheses: $2+3^2$. Since exponents comes before addition, we find $3^2=9$ first. We now have

$$20-6\div 3+(2\times 9)$$

We continue inside the parentheses and perform the multiplication: $2\times 9=18$.

This gives

$$20-6\div 3+18$$

Since division comes before addition and subtraction, we next calculate $6\div 3=2$ to get

$$20-2+18$$

Since subtraction and addition are tied, we go from left to right. We calculate: $20-2=18$ to get

$$18+18\ =36$$

The key to arriving at the correct answer is to go slow and write down each step in the arithmetic.

Hidden Parentheses

You may think that since you always have a calculator or computer at hand, that you don't need to worry about order of operations. Unfortunately, the way that expressions are written is not the same as the way that they are entered into a computer or calculator. In particular, exponents need to be treated with care as do fractions bars.

Example 3.3.3

Evaluate 2.1^{6-2}

Solution

First, note that we use the symbol "^" to tell a computer or calculator to exponentiate. If you were to enter 2.1^6-2 into a computer, it would give you the answer of 83.766121 which is not correct, since the computer will first expontiate and then subtract. Since the subtraction is within the exponent, it must be performed first. To tell a calculator or computer to perform the subtraction first, we use parentheses:

2.1^(6 - 2) = 19.4481

Example 3.3.4: z-scores

The "z-score" is defined by:

$$z = \frac{x - \mu}{\sigma}$$

Find the z-score rounded to one decimal place if:

$$x = 2.323,\ \mu = 1.297,\ \sigma = 0.241$$

Solution

Once again, if we put these numbers into the z-score formula and use a computer or calculator by entering $3.323 - 1.297 \div 0.241$ we will get -0.259 which is the wrong answer. Instead, we need to know that the fraction bar separates the numerator and the denominator, so the subtraction must be done first. We compute

$$\frac{2.323 - 1.297}{0.241} = (2.323 - 1.297) \div 0.241 = 4.25726141$$

Now round to one decimal place to get 4.3. Notice that if you rounded before you did the arithmetic, you would get exactly 5 which is very different. 4.3 is more accurate.

Exercise

Suppose the equation of the regression line for the number of pairs of socks a person owns, y, based on the number of pairs of shoes, x, the person owns is

$$\hat{y} = 6 + 2x$$

Use this regression line to predict the number of pairs of socks a person owns for a person who owns 4 pairs of shoes.

- Order of Operations - The Basics
- Order of Operations

This page titled 3.3: Order of Operations is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Order of Operations** by Larry Green is licensed CC BY 4.0.

3.4: Order of Operations in Expressions and Formulas

Learning Outcomes

- Use Order of Operations in Statistics Formulas.

We have already encountered the order of operations: Parentheses, Exponents, Multiplication and Division, Addition and Subtraction. In this section, we will give some additional examples where the order of operations must be used properly to evaluate statistics.

Example 3.4.1

The sample standard deviation asks us to add up the squared deviations, take the square root and divide by one less than the sample size. For example, suppose that there are three data values: 3, 5, 10. The mean of these values is 6. Then the standard deviation is:

$$s = \sqrt{\frac{(3-6)^2 + (5-6)^2 + (10-6)^2}{3-1}}$$

Evaluate this number rounded to the nearest hundredth.

Solution

The first thing in the order of operations is to do what is in the parentheses. We must subtract:

$$3-6 = -3, \quad 5-6 = -1, \quad 10-6 = 4$$

We can substitute the numbers in to get:

$$= \sqrt{\frac{(-3)^2 + (-1)^2 + (4)^2}{3-1}}$$

Next, we exponentiate:

$$(-3)^2 = 9, \quad (-1)^2 = 1, \quad 4^2 = 16$$

Substitute these in to get:

$$\sqrt{\frac{9+1+16}{3-1}}$$

We can now perform the addition inside the square root to get:

$$\sqrt{\frac{26}{3-1}}$$

Next, perform the subtraction of the denominator to get:

$$\sqrt{\frac{26}{2}}$$

We can divide to get:

$$\sqrt{13}$$

We don't want to do this by hand, so in a calculator or computer type in:

$$13^{0.5} = 3.61$$

Example 3.4.2

When calculating the probability that a value will be less than 4.6 if the value is taken randomly from a uniform distribution between 3 and 7, we have to calculate:

$$(4.6-3)\times\frac{1}{7-3}$$

Find this probability.

Solution

We can use a calculator or computer, but we must be very careful about the order of operations. Notice that there are implied parentheses due to the fraction bar. The answer is:

$$\frac{(4.6-3)\times 1}{7-3}$$

Using technology, we get:

$$(4.6-3)\times\frac{1}{7-3}=0.4$$

Exercise

When finding the upper bound, U, of a confidence interval given the lower bound, L, and the margin of error, E, we use the formula

$$U=L+2E$$

Find the upper bound of the confidence interval for the proportion of babies that are born preterm if the lower bound is 0.085 and the margin of error is 0.03.

- Ex: Evaluate an Expression Using the Order of Operations
- Order of Operations and Confidence Intervals

This page titled 3.4: Order of Operations in Expressions and Formulas is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Order of Operations in Expressions and Formulas** by Larry Green is licensed CC BY 4.0.

3.5: Perform Signed Number Arithmetic

Learning Outcomes

1. Add signed numbers.
2. Subtract signed numbers.
3. Multiply signed numbers.
4. Divide signed numbers.

Even though negative numbers seem not that common in the real world, they do come up often when doing comparisons. For example, a common question is how much bigger is one number than another, which involves subtraction. In statistics we don't know the means until we collect the data and do the calculations. This often results in subtracting a larger number from a smaller number which yields a negative number. Because of this and for many other reasons, we need to be able to perform arithmetic on both positive and negative numbers.

Adding Signed Numbers

We will assume that you are very familiar with adding positive numbers, but when there are negative numbers involved, there are some rules to follow:

1. When adding two negative numbers, ignore the negative signs, add the positive numbers and then make the result negative.
2. When adding two numbers such that one is positive and the other is negative, ignore the sign, subtract the smaller from the larger. If the larger of the positive numbers was originally negative, then make the result negative. Otherwise keep the result positive.

Example 3.5.1

Add:

$$-4+(-3)$$

Solution

First we ignore the signs and add the positive numbers.

$$4+3=7$$

Next we make the result negative.

$$-4+(-3)=-7$$

Example 3.5.2

Add:

$$-2+5$$

Solution

Since one of the numbers is positive and the other is negative, we subtract:

$$5-2=3$$

Of the two numbers, 2 and 5, 5 is the larger one and started positive. Hence we keep the result positive:

$$-2+5=3$$

Subtracting Numbers

Subtraction comes up often when we want to find the width of an interval in statistics. Here are the cases for subtracting: $a-b$:

1. If $a \geq b \geq 0$, then this is just ordinary subtraction.
2. If $b \geq a \geq 0$, then find $b - a$ and make the result negative.
3. If $a < 0,\ b \geq 0$, then make both positive, add the two positive numbers and make the result negative.
4. If $b < 0$ then you use the rule that subtracting a negative number is the same as adding the positive number.

Example 3.5.3

Evaluate $5 - 9$

Solution

Since 9 is bigger than 5, we subtract:

$$9 - 5 = 4$$

Next, we make the result negative to get:

$$5 - 9 = -4$$

Example 3.5.4

Evaluate $-9 - 4$

Solution

We are in the case $a < 0,\ b \geq 0$. Therefore, we first make both positive and add the positive numbers.

$$9 + 4 = 13$$

The final step is to make the answer negative to get

$$-9 - 4 = -13$$

Example 3.5.5: Uniform distributions

In statistics, we call a *distribution Uniform* if an event is just as likely to be in any given interval within the bounds as any other interval within the bounds as long as the intervals are both of the same width. Finding the width of a given interval is usually the first step in solving a question involving uniform distributions. Suppose that the temperature on a winter day has a Uniform distribution on [-8,4]. Find the width of this interval

Solution

To find the width of an interval, we subtract the left endpoint from the right endpoint:

$$4 - (-8)$$

Since we are subtracting a negative number, the "-" signs become addition:

$$4 - (-8) = 4 + 8 = 12$$

Thus the width of the interval is 12.

Multiplying and Dividing Signed Numbers

When we have a multiplication or division problem, we just remember that two negatives make a positive. So if there are an even number of negative numbers that are multiplied or divided, the result is negative. If there are an odd number of negative numbers that are multiplied or divided, the result is positive.

Example 3.5.6

Perform the arithmetic:

$$\frac{(-6)(-10)}{(-4)(-5)}$$

Solution

First, just ignore all of the negative signs and multiply the numerator and denominator separately:

$$\frac{(6)(10)}{(4)(5)} = \frac{60}{20}$$

Now divide:

$$\frac{60}{20} = \frac{6}{2} = 3$$

Finally, notice that there are four negative numbers in the original multiplication and division problem. Four is an even number, so the answer is positive:

$$\frac{(-6)(-10)}{(-4)(-5)} = 3$$

Example 3.5.7

A confidence interval for the population mean difference in books read per year by men and women was was found to be [-4,1]. Find the midpoint of this interval.

Solution

First recall that to find the midpoint of two numbers, we add then and then divide by 2. Hence, our first step is to add -4 and 1. Since 1 is positive and -4 is negative, we first subtract the two numbers:

$$4 - 1 = 3$$

Of the two numbers, 4 and 1, 4 is the larger one and started negative. Hence we change the sign to negative::

$$-4 + 1 = -3$$

The final step in finding the midpoint is to divide by 2. First we divide them as positive numbers:

$$\frac{3}{2} = 1.5$$

Since the original quotient has a single negative number (an odd number of negative numbers), the answer is negative. Thus the midpoint of -4 and 1 is -1.5.

Exercise

The difference between the observed value and the expected value in linear regression is called the residual. Suppose that the three observed values are: -4, 2, and 5. The expected values are -3, 7, and -1. First find the residuals and then find the sum of the residuals.

- Signed Number Operations (L1.4)
- signed arithmetic

This page titled 3.5: Perform Signed Number Arithmetic is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Perform Signed Number Arithmetic** by Larry Green is licensed CC BY 4.0.

3.6: Powers and Roots

Learning Outcomes

1. Raise a number to a power using technology.
2. Take the square root of a number using technology.
3. Apply the order of operations when there is root or a power.

It can be a challenge when we first try to use technology to raise a number to a power or take a square root of a number. In this section, we will go over some pointers on how to successfully take powers and roots of a number. We will also continue our practice with the order of operations, remembering that as long as there are no parentheses, exponents always come before all other operations. We will see that taking a power of a number comes up in probability and taking a root comes up in finding standard deviations.

Powers

Just about every calculator, computer, and smartphone can take powers of a number. We just need to remember that the symbol "^" is used to mean "to the power of". We also need to remember to use parentheses if we need to force other arithmetic to come before the exponentiation.

Example 3.6.1

Evaluate: 1.04^5 and round to two decimal places.

Solution

This definitely calls for the use of technology. Most calculators, whether hand calculators or computer calculators, use the symbol "^" (shift 6 on the keyboard) for exponentiation. We type in:

$$1.04^5 = 1.2166529$$

We are asked to round to two decimal places. Since the third decimal place is a 6 which is 5 or greater, we round up to get:

$$1.04^5 \approx 1.22$$

Example 3.6.2

Evaluate: $2.8^{5.3\times 0.17}$ and round to two decimal places.

Solution

First note that on a computer we use "*" (shift 8) to represent multiplication. If we were to put in 2.8 ^ 5.3 * 0.17 into the calculator, we would get the wrong answer, since it will perform the exponentiation before the multiplication. Since the original question has the multiplication inside the exponent, we have to force the calculator to perform the multiplication first. We can ensure that multiplication occurs first by including parentheses:

$$2.8^{5.3\times 0.17} = 2.52865$$

Now round to decimal places to get:

$$2.8^{5.3\times 0.17} \approx 2.53$$

Example 3.6.3

If we want to find the probability that if we toss a six sided die five times that the first two rolls will each be a 1 or a 2 and the last three die rolls will be even, then the probability is:

$$\left(\frac{1}{3}\right)^2 \times \left(\frac{1}{2}\right)^3$$

What is this probability rounded to three decimal places?

Solution

We find:

$$(1/3)^2(1/2)^3 \approx 0.013888889$$

Now round to three decimal places to get

$$\left(\frac{1}{3}\right)^2 \times \left(\frac{1}{2}\right)^3 \approx 0.014$$

Square Roots

Square roots come up often in statistics, especially when we are looking at standard deviations. We need to be able to use a calculator or computer to compute a square root of a number. There are two approaches that usually work. The first approach is to use the $\sqrt{\ }$ symbol on the calculator if there is one. For a computer, using sqrt() usually works. For example if you put 10*sqrt(2) in the Google search bar, it will show you 14.1421356. A second way that works for pretty much any calculator, whether it is a hand held calculator or a computer calculator, is to realize that the square root of a number is the same thing as the number to the 1/2 power. In order to not have to wrap 1/2 in parentheses, it is easier to type in the number to the 0.5 power.

Example 3.6.3

Evaluate $\sqrt{42}$ and round your answer to two decimal places.

Solution

Depending on the technology you are using you will either enter the square root symbol and then the number 42 and then close the parentheses if they are presented and then hit enter. If you are using a computer, you can use sqrt(42). The third way that will work for both is to enter:

$$42^{0.5} \approx 6.4807407$$

You must then round to two decimal places. Since 0 is less than 5, we round down to get:

$$\sqrt{42} \approx 6.48$$

Example 3.6.4

The "z-score" is for the value of 28 for a sampling distribution with sample size 60 coming from a population with mean 28.3 and standard deviation 5 is defined by:

$$z = \frac{28 - 28.3}{\frac{5}{\sqrt{60}}}$$

Find the z-score rounded to two decimal places.

Solution

We have to be careful about the order of operations when putting it into the calculator. We enter:

$$(28 - 28.3)/(5/60\text{^}0.5) = -0.464758$$

Finally, we round to 2 decimal places. Since 4 is smaller than 5, we round down to get:

$$z = \frac{28 - 28.3}{\frac{5}{\sqrt{60}}} = -0.46$$

Exercise

The standard error, which is an average of how far sample means are from the population mean is defined by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is the standard error, σ is the standard deviation, and n is the sample size. Find the standard error if the population standard deviation, σ, is 14 and the sample size, n, is 11.

- Square Root on the TI-83plus and TI-84 family of Calculators
- Square Roots with a Computer

This page titled 3.6: Powers and Roots is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Powers and Roots** by Larry Green is licensed CC BY 4.0.

3.7: Using Summation Notation

Learning Outcomes

1. Evaluate an expression that includes summation notation.
2. Apply summation notation to calculate statistics.

This notation is called summation notation and appears as:

$$\sum_{i=1}^{n} a_i$$

In this notation, the a_i is an expression that contains the index i and you plug in 1 and then 2 and then 3 all the way to the last number n and then add up all of the results.

Example 3.7.1

Calculate

$$\sum_{i=1}^{4} 3i$$

Solution

First notice that i = 1, then 2, then 3 and finally 4. We are supposed to multiply each of these by 3 and add them up:

$$\begin{aligned}\sum_{i=1}^{4} 3i &= 3(1)+3(2)+3(3)+3(4)\\ &= 3+6+9+12 = 30\end{aligned}$$

Example 3.7.2

The formula for the sample mean, sometimes called the average, is

$$\bar{x} = \frac{\sum_{i-1}^{n} x_i}{n}$$

A survey was conducted asking 8 older adults how many sexual partners they have had in their lifetime. Their answers were {4,12,1,3,4,9,24,7}. Use the formula to find the sample mean.

Solution

Notice that the numerator of the formula just tells us to add the numbers up. Computing the numerator first gives:

$$\sum_{i=1}^{8} x_i = 4+12+1+3+4+9+24+7 = 64$$

Now that we have the numerator calculated, the formula tells us to divide by *n*, which is just 8. We have:

$$\bar{x} = \frac{64}{8} = 8$$

Thus, the sample mean number of sexual partners this group had in their lifetimes is 8.

Example 3.7.3

The next most important statistic is the standard deviation. The formula for the sample standard deviation is:

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

Let's consider the data in the previous example. Find the standard deviation.

Solution

The formula is quite complicated, but if tackle it one piece at a time using the order of operations properly, we can succeed in finding the sample standard deviation for the data. Notice that there are parentheses, so based on the order of operations, we must do the subtraction within the parentheses first. Since this is all part of the sum, we have eight different subtractions to do. From our calculations in the previous example, the sample mean was $\bar{x} = 8$. We compute the 8 subtractions:

$$4-8 = -4,\ 12-8 = 4,\ 1-8 = -7,\ 3-8 = -5,$$

$$4-8 = -4,\ 9-8 = 1,\ 24-8 = 16,\ 7-8 = -1$$

The next arithmetic to do is to square each of the differences to get:

$$(-4)^2 = 16,\ (4)^2 = 16, (-7)^2 = 49,\ (-5)^2 = 25,$$

$$(-4)^2 = 16,\ 1^2 = 1,\ 16^2 = 256,\ (-1)^2 = 1$$

Now we have all the entries in the summation, so we add them all up:

$$16+16+49+25+16+1+256+1 = 380$$

Now we can write

$$s = \sqrt{\frac{380}{8-1}} = \sqrt{\frac{380}{7}}$$

We can put this into the calculator or computer to get:

$$s = \sqrt{\frac{380}{7}} = 7.3679$$

Exercise: expected value

The expected value, EV, is defined by the formula

$$EV = \sum_{i=1}^{n} x_i\, P(x_i)$$

Where x_i are the possible outcomes and $P(x_i)$ are the probabilities of the outcomes occurring. Suppose the table below shows the number of eggs in a bald eagle clutch and the probabilities of that number occurring.

Probability Distribution Table with Outcomes, x, and probabilities, P(x)

x	1	2	3	4
P(x)	0.2	0.4	0.3	0.1

Find the expected value.

Ex 1: Find a Sum Written in Summation / Sigma Notation

Summation Notation and Expected Value

This page titled 3.7: Using Summation Notation is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Using Summation Notation** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

4: Sets

This page titled 4: Sets is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

4.1: Set Notation

Learning Outcomes

1. Read set notation.
2. Describe sets using set notation.

A set is just a collection of items and there are different ways of representing a set. We want to be able to both read the various ways and be able to write down the representation ourselves in order to best display the set. We have already seen how to represent a set on a number line, but that can be cumbersome, especially if we want to just use a keyboard. Imagine how difficult it would be to text a friend about a cool set if the only way to do this was with a number line. Fortunately, mathematicians have agreed on notation to describe a set.

Example 4.1.1

If we just have a few items to list, we enclose them in curly brackets "{" and "}" and separate the items with commas. For example,

$$\{\text{Miguel}, \text{Kristin}, \text{Leo}, \text{Shanice}\}$$

means the set the contains these four names.

Example 4.1.2

If we just have a long collection of numbers that have a clear pattern, we use the "..." notation to mean "start here, keep going, and end there". For example,

$$\{3, 6, 9, 12, \ldots, 90\}$$

This set contains more than just the five numbers that are shown. It is clear that the numbers are separated by three each. After the 12, even though it is not explicitly shown, is a 15 which is part of this set. It also contains 18, 21 and keeps going including all the multiples of 3 until it gets to its largest number 90.

Example 4.1.3

If we just have a collection of numbers that have a clear pattern, but never ends, we use the "..." without a number at the end. For example,

$$\left\{\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \cdots\right\}$$

This set contains an infinite number of fractions, since there is no number followed by the "...".

Example 4.1.4

Sometimes we have a set that it best described by stating a rule. For example, if you want to describe the set of all people who are over 18 years old but not 30 years old, you announce the conditions by putting them to the left of a vertical line segment. We read the line segment as "such that".

$$\{x \mid x > 18 \; and \; x \neq 30\}$$

This can be read as "the set of all numbers x such that x is greater than 18 and x is not equal to 30".

Exercise

Describe using set notation the collection of all positive even whole numbers that are not equal to 20 or 50.

- Set-Builder Notation

- https://youtu.be/VGphtczN0-c

This page titled 4.1: Set Notation is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Set Notation** by Larry Green is licensed CC BY 4.0.

4.2: The Complement of a Set

Learning Outcomes

1. Determine the complement of a set.
2. Write the complement of a set using set notation.

We saw in the section "Represent an Inequality as an Interval on a Number Line" how to graph the complement for a set defined by an inequality. Complements come up very often in statistics, so it is worth revisiting this, but instead of graphically we will focus on set notation. Recall that the complement of a set is everything that is not in that set. Sometimes it is much easier to find the probability of a complement than of the original set, and there is an easy relationship between the probability of an event happening and the probability of the complement of that event happening.

$$P(A) = 1 - P(not\ A)$$

Example 4.2.1

Find the complement of the set:

$$A = \{x \mid x < 4\}$$

Solution

The complement of the set of all numbers that are less than 4 is the set of all numbers that are at least as big as 4. Notice that the number 4 is not in the set A, since the inequality is strict (does not have an "="). Therefore the number 4 is in the complement of the set A. In set notation:

$$A^c = \{x \mid x \geq 4\}$$

Example 4.2.2

When computing probabilities the complement is sometimes much easier than the original set. For example suppose you roll a die 6 times and want to find the probability that the number 3 comes up at least once. Find the complement of this event.

Solution

First note that the event of at least once means that there could be one 3, two 3's, three 3's, four 3's, five 3's, or six 3's. It turns out that this would be a burden to deal with each of these possibilities. However the complement is quite easy. The complement of getting at least one 3 is that you go no 3's.

Example 4.2.3

Suppose that we want to find the probability that at least 20 people in the class have done their homework. Find the complement of this event.

Solution

Sometimes it is easiest to list nearby outcomes and then determine the outcomes that satisfy the event. Finally, to find the complement, you select the rest. First list numbers near 20:

$$\ldots, 17, 18, 19, 20, 21, 22, \ldots$$

Now, the ones that are at least 20 are all the ones including 20 and to the right of 20:

$$20, 21, 22, \ldots$$

These are the large numbers. The complement includes all the small numbers.

$$\ldots, 17, 18, 19$$

We can write this in set notation as:

$$\{x \mid x \leq 19\}$$

or equivalently

$$\{x \mid x < 20\}$$

Example 4.2.4

Suppose a number is picked at random from the whole numbers from 1 to 10. Let A be the event that a number is both even and less than 8. Find the complement of A.

Solution

First, the set of numbers that are both even and less than 8 is:

$$A = \{2, 4, 6\}$$

The complement of this set is all the numbers from 1 to 10 that are not in A:

$$A^c = \{1, 3, 5, 7, 8, 9, 10\}$$

Exercise

Suppose that two six sided dice are rolled. Let the A be the event that either the first die is even or the sum of the dice is greater than 5 or both have occurred. Find the complement of A.

- Ex: Find the Intersection of a Set and A Complement Using a Venn Diagram
- https://youtu.be/ek3QwY2gw4w

This page titled 4.2: The Complement of a Set is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **The Complement of a Set** by Larry Green is licensed CC BY 4.0.

4.3: The Union and Intersection of Two Sets

Learning Outcomes

1. Find the union of two sets.
2. Find the intersection of two sets.
3. Combine unions intersections and complements.

All statistics classes include questions about probabilities involving the union and intersections of sets. In English, we use the words "Or", and "And" to describe these concepts. For example, "Find the probability that a student is taking a mathematics class or a science class." That is expressing the union of the two sets in words. "What is the probability that a nurse has a bachelor's degree and more than five years of experience working in a hospital." That is expressing the intersection of two sets. In this section we will learn how to decipher these types of sentences and will learn about the meaning of unions and intersections.

Unions

An element is in the union of two sets if it is in the first set, the second set, or both. The symbol we use for the union is $\cup$. The word that you will often see that indicates a union is "or".

Example 4.3.1: Union of Two sets

Let:

$$A = \{2, 5, 7, 8\}$$

and

$$B = \{1, 4, 5, 7, 9\}$$

Find $A \cup B$

Solution

We include in the union every number that is in A or is in B:

$$A \cup B = \{1, 2, 4, 5, 7, 8, 9\}$$

Example 4.3.2: Union of Two sets

Consider the following sentence, "Find the probability that a household has fewer than 6 windows or has a dozen windows." Write this in set notation as the union of two sets and then write out this union.

Solution

First, let A be the set of the number of windows that represents "fewer than 6 windows". This set includes all the numbers from 0 through 5:

$$A = \{0, 1, 2, 3, 4, 5\}$$

Next, let B be the set of the number of windows that represents "has a dozen windows". This is just the set that contains the single number 12:

$$B = \{12\}$$

We can now find the union of these two sets:

$$A \cup B = \{0, 1, 2, 3, 4, 5, 12\}$$

Intersections

An element is in the intersection of two sets if it is in the first set and it is in the second set. The symbol we use for the intersection is $\cap$. The word that you will often see that indicates an intersection is "and".

Example 4.3.3: Intersection of Two sets

Let:

$$A = \{3, 4, 5, 8, 9, 10, 11, 12\}$$

and

$$B = \{5, 6, 7, 8, 9\}$$

Find $A \cap B$.

Solution

We only include in the intersection that numbers that are in both A and B:

$$A \cap B = \{5, 8, 9\}$$

Example 4.3.4: Intersection of Two sets

Consider the following sentence, "Find the probability that the number of units that a student is taking is more than 12 units and less than 18 units." Assuming that students only take a whole number of units, write this in set notation as the intersection of two sets and then write out this intersection.

Solution

First, let A be the set of numbers of units that represents "more than 12 units". This set includes all the numbers starting at 13 and continuing forever:

$$A = \{13,\ 14,\ 15,\ \ldots\}$$

Next, let B be the set of the number of units that represents "less than 18 units". This is the set that contains the numbers from 1 through 17:

$$B = \{1,\ 2,\ 3,\ \ldots,\ 17\}$$

We can now find the intersection of these two sets:

$$A \cap B = \{13,\ 14,\ 15,\ 16,\ 17\}$$

Combining Unions, Intersections, and Complements

One of the biggest challenges in statistics is deciphering a sentence and turning it into symbols. This can be particularly difficult when there is a sentence that does not have the words "union", "intersection", or "complement", but it does implicitly refer to these words. The best way to become proficient in this skill is to practice, practice, and practice more.

Example 4.3.5

Consider the following sentence, "If you roll a six sided die, find the probability that it is not even and it is not a 3." Write this in set notation.

Solution

First, let A be the set of even numbers and B be the set that contains just 3. We can write:

$$A = \{2, 4, 6\}, \quad B = \{3\}$$

Next, since we want "not even" we need to consider the complement of A:

$$A^c = \{1, 3, 5\}$$

Similarly since we want "not a 3", we need to consider the complement of B:

$$B^c = \{1, 2, 4, 5, 6\}$$

Finally, we notice the key word "and". Thus, we are asked to find:

$$A^c \cap B^c = \{1, 3, 5\} \cap \{1, 2, 4, 5, 6\} = \{1, 5\}$$

Example 4.3.6

Consider the following sentence, "If you randomly select a person, find the probability that the person is older than 8 or is both younger than 6 and is not younger than 3." Write this in set notation.

Solution

First, let A be the set of people older than 8, B be the set of people younger than 6, and C be the set of people younger than 3. We can write:

$$A = \{x \mid x > 8\}, \quad B = \{x \mid x < 6\}, \; C = \{x \mid x < 3\}$$

We are asked to find

$$A \cup (B \cap C^c)$$

Notice that the complement of "$<$" is "$\geq$". Thus:

$$C^c = \{x \mid x \geq 3\}$$

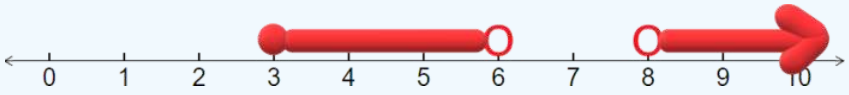
Next we find:

$$B \cap C^c = \{x \mid x < 6\} \cap \{x \mid x \geq 3\} = \{x \mid 3 \leq x < 6\}$$

Finally, we find:

$$A \cup (B \cap C^c) = \{x \mid x > 8\} \cup \{x \mid 3 \leq x < 6\}$$

The clearest way to display this union is on a number line. The number line below displays the answer:



Exercise

Suppose that we pick a person at random and are interested in finding the probability that the person's birth month came after July and did not come after September. Write this event using set notation.

- Ex: Find the Intersection of a Set and A Complement Using a Venn Diagram
- Intersection and Complements of Sets

This page titled 4.3: The Union and Intersection of Two Sets is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **The Union and Intersection of Two Sets** by Larry Green is licensed CC BY 4.0.

4.4: Venn Diagrams

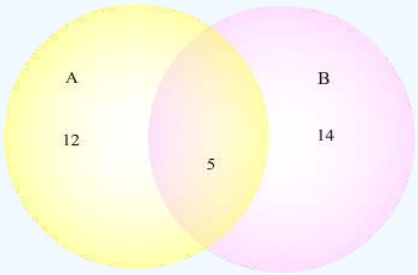
Learning Outcomes

1. Read a Venn Diagram to extract information.
2. Draw a Venn Diagram.

Venn Diagrams are a simple way of visualizing how sets interact. Many times we will see a long wordy sentence that describes a numerical situation, but it is a challenge to understand. As the saying goes, "A picture is worth a thousand words." In particular, a Venn Diagram describes how many elements are in each set displayed and how many elements are in their intersections and complements.

Example 4.4.1

Consider the Venn Diagram shown below.



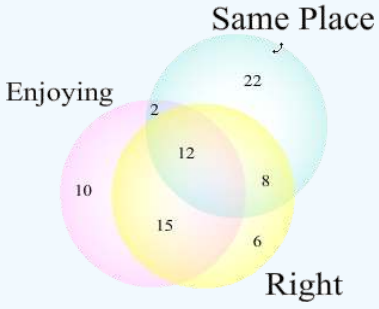
Describe how many elements are in each of the sets.

Solution

Once we understand how to read the Venn Diagram we can use it in many applications. For the Venn Diagram above, there are 12 from A that are not in B, there are 5 in both A and B, and there are 14 in B that are not in A. If we wanted to find the total in A, we would just add 12 and 5 to get 17 total in A. Similarly, there are 19 total in B.

Example 4.4.2

Consider the Venn Diagram below that shows the results of a study asking students whether their first college class was at the same place they are at now, whether they are right handed, and whether they are enjoying their experience at their college.



Determine how many students are:

1. Right handed and enjoy college.
2. At the same place but not right handed.
3. Enjoy college.

Solution

1. To be right handed and enjoy college they must be in both the Right circle and the Enjoying circle. Notice that the numbers 12 and 15 are in both these circles. Thus, there are 12 + 15 = 27 total students who are right handed and enjoy college.

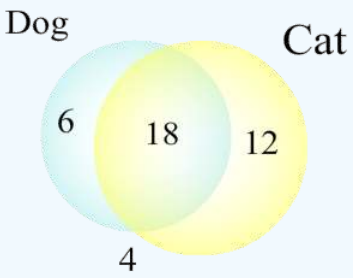
2. To be in the same place and not be right handed, the number must be in the same place circle but not in the right circle. We see that 2 and 22 are the numbers in the same place circle but not in the right circle. Adding these gives 2 + 22 = 24 total students who are at the same place but not right handed.
3. We must count all the numbers in the Enjoying circle. These are 2, 10, 12, and 15. Adding these up gives: 2 + 10 + 12 + 15 = 39. Thus, 39 students enjoy college.

Example 4.4.3

Suppose that a group of 40 households was looked at. 24 of them housed dogs, 30 of them housed cats, and 18 of them housed both cats and dogs. Sketch a Venn Diagram that displays this information.

Solution

To get ready to sketch the Venn Diagram, we first plan on what it will look like. There are two main groups here: houses with dogs and houses with cats. Therefore we will have two circles. The intersection will have the number 18. Since there are 24 houses with dogs and 18 also have cats, we subtract 24 - 18 = 6 to find the houses with dogs but no cats. Similarly, we subtract 30 - 18 = 12 houses with cats and no dogs. If we add 18 + 6 + 12 = 36, we find the total number of houses with a dog, cat or both. Therefore there are 40 - 36 = 4 houses without any pets. Now we are ready to put in the numbers into the Venn Diagram. It is shown below.



Exercise

Suppose that a group of 55 businesses was researched. 29 of them were open on the weekends, 25 of them paid more than minimum wage for everyone , 17 of them were both open on the weekends and paid more than minimum wage for everyone, and 4 of them were government consulting businesses. None of the government consulting businesses were open on the weekend nor did they pay more than minimum wage for everyone. Sketch a Venn Diagram that displays this information.

- **Solving Problems with Venn Diagrams**
- https://youtu.be/t67RMAWGMdY

This page titled 4.4: Venn Diagrams is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Venn Diagrams** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

5: Expressions, Equations and Inequalities

This page titled 5: Expressions, Equations and Inequalities is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

5.1: Evaluate Algebraic Expressions

Learning Outcomes

1. Evaluate an algebraic expression given values for the variables.
2. Recognize given values in a word problem and evaluate an expression using these values.

There are many formulas that are encountered in a statistics class and the values of each variable will be given. It will be your task to carefully evaluate the expression after plugging in each of the given values into the formula. In order to be successful you should not rush through the process and you need to be aware of the order of operations and use parentheses when necessary.

Example 5.1.1

Suppose that equation of the regression line for the number of days a week, x, a person exercises and the number of days, $\hat{y}$, a year a person is sick is:

$$\hat{y} = 12.5 - 1.6x$$

We use $\hat{y}$ instead of y since this is a prediction instead of an actual data value's y-coordinate. Use this regression line to predict the number of times a person who exercises 4 days a week will be sick this year.

Solution

The first step is always to identify the variable or variables that are given. In this case, we have 4 days of exercise a week, so:

$$x = 4$$

Next, we plug in to get:

$$\hat{y} = 12.5 - 1.6(4) = 6.1$$

Since we are predicting the number of days a year being sick, it is a good idea to round to the nearest whole number. We get that the best prediction for the number of sick days for a person who exercises 4 days per week is that they will be sick 6 days this year.

Example 5.1.2

For a yes/no question, a sample size is considered large enough to use a Normal distribution if

$np > 5$ and $nq > 5$

where n is the sample size, p is the proportion of Yes answers, and q is the proportion of No answers. A survey was given to 59 American adults asking them if they were food insecure today. 6.8% of them said they were food insecure today. Was the sample size large enough to use the Normal distribution?

Solution

Our first task is to list out each of the needed variables. Let's start with n, the sample size. We are given that 59 Americans were surveyed. Thus

$$n = 59$$

Next, we will find p, the proportion of Yes answers. We are given that 6.8% said Yes. Since this is a percent and not a proportion, we must convert the percent to a proportion by moving the decimal place two places to the right. It helps to place a 0 to the left of the 6, so that the decimal point has a place to go. A common error is to rush through this and wrongly write down 0.68. Instead, the proportion is:

$$p = 0.068$$

Our next task is to find q, the proportion of No answers. For a Yes/No question, the proportion of Yes answers and the proportion of No answers must always add up to 1. Thus:

$$q = 1 - 0.068 \ = \ 0.932$$

Now we are ready to plug into the two inequalities:

$$np = 59 \times 0.068 = 4.012$$

and

$$nq = 59 \times 0.932 = 54.988$$

Although $nq \ = \ 54.988 > 5$, we have $np \ = \ 4.012 < 5$, so the sample size was not large enough to use the Normal distribution.

Example 5.1.3

For a quantitative study, the sample size, n, needed in order to produce a confidence interval with a margin of error no more than $\pm E$, is

$$n = \left(\frac{z\sigma}{E}\right)^2$$

where z is a value that is determined from the confidence level and σ is the population standard deviation. You want to conduct a survey to estimate the population mean amount of years it takes psychologists to get through college and you require a margin of error of no more than ± 0.1 years. Suppose that you know that the population standard deviation is 1.3 years. If you want a 95% confidence interval that comes with a $z = 1.96$, at least how many psychologists must you survey? Round your answer up.

Solution

We start out by identifying the given values for each variable. Since we want a margin of error of no more than ± 0.1, we have:

$$E \ = \ 0.1$$

We are told that the population standard is 1.3, so:

$$\sigma = 1.3$$

We are also given the value of z:

$$z = 1.96$$

Now put this into the formula to get:

$$n = \left(\frac{1.96 \times 1.3}{0.1}\right)^2$$

We put this into a calculator or computer to get:

$$(1.96 \times 1.3 \div 0.1)^2 = 649.2304$$

We round up and can conclude that we need to survey 650 psychologists.

Example 5.1.4

Based on the Central Limit Theorem, the standard deviation of the sampling distribution when samples of size n are taken from a population with standard deviation, σ, is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation for the number of customers who walk into a fast food restaurant is 12, what is the standard deviation of the sampling distribution for samples of size 35? Round your answer to two decimal places.

Solution

First we identify each of the given variables. Since the population standard deviation was 12, we have:

$$\sigma = 12$$

We are told that the sample size is 35, so:

$$n = 35$$

Now we put these numbers into the formula for the standard deviation of the sampling distribution to get:

$$\sigma_{\bar{x}} = \frac{12}{\sqrt{35}}$$

We are now ready to put this into our calculator or computer. We put in:

$$\sigma_x = \frac{12}{\sqrt{35}} = 12 \div (35 \text{^} 0.5) = 2.02837$$

Rounded to two decimal places, we can say that the standard deviation of the sampling distribution is 2.03.

Example 5.1.5: Z score

The z-score for a given sample mean $\bar{x}$ for a sampling distribution with population mean μ, population standard deviation σ, and sample size n is given by:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

An environmental scientist collected data on the amount of glacier retreat. She measured 45 glaciers. The population mean retreat is 22 meters and the population standard deviation is 16 meters. The sample mean for her data was 27 meters and the sample standard deviation for her data was 18 meters. What was the z-score?

Solution

First we identify each of the given variables. Since the sample mean was 27, we have:

$$\bar{x} = 27$$

We are told that the population mean is 22 meters, so:

$$\mu = 22$$

We are also given that the population standard deviation is 16 meters, hence:

$$\sigma = 16$$

Finally, since she measured 45 glaciers, we have:

$$n = 45$$

Now we put the numbers into the formula for the z-score to get:

$$z = \frac{27 - 22}{\frac{16}{\sqrt{45}}}$$

We are now ready to put this into our calculator or computer. We must pay attention to the order of operations and put parentheses around the numerator, since the subtraction happens for this expression before the division. We also must put parentheses around the denominator. We put in:

$$z = (27 - 22) \div \left(16 \div \sqrt{45}\right) = 2.0963$$

Exercise

You want to come up with a 90% confidence interval for the proportion of people in your community who are obese and require a margin of error of no more than $\pm 3\%$. According to the Journal of the American Medical Association (JAMA) 34% of all Americans are obese. The equation to find the sample size, n, needed in order to come up with a confidence interval is:

$n = p\,(1-p)\left(\frac{z}{E}\right)^2$

where p is the preliminary estimate for the population proportion. Based on calculations, $z = 1.645$. How many people in your community must you survey?

Evaluating Algebraic Expressions (L2.1)

https://youtu.be/HLjUT8Kvc5U

This page titled 5.1: Evaluate Algebraic Expressions is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Evaluate Algebraic Expressions** by Larry Green is licensed CC BY 4.0.

5.2: Inequalities and Midpoints

Learning Objectives

- Write out an inequality from words.
- Go from a midpoint and error to an inequality.
- Go from inequality to a midpoint and error.

Inequalities are an essential component of statistics. One very important use of inequalities is when we have found a mean or proportion from a sample and want to write out an inequality that gives where the population mean or proportion is likely to lie. Another application is in probability where we want to find the probability of a value being more than a number, less than a number, or between two numbers.

Converting Words to Inequalities

Example 5.2.1

You want to find the probability that it will a patient will "take at least three hours to wake up after surgery". Write an inequality for this situation.

Solution

The key words here are "at least". These words can be written symbolically as "≤". Therefore we can write "take at least three hours to wake up after surgery" as:

$$x \leq 3$$

Example 5.2.2

Suppose you want to find the probability that a relationship will last "more than 1 week and at most 8 weeks". Write an inequality for this situation.

Solution

Let's first translate the words "more than". This is equivalent to ">". Next translate the words "at most". This is equivalent to "≤". Now we can put this together to get:

$$1 < x \leq 8$$

Midpoints and Inequalities

There are two ways of thinking about an interval. The first is that x is greater than the lower bound and less than the upper bound. The second is that the center or midpoint of the interval is a given value and the interval goes no more than a certain distance from that value. In statistics, this is important when we look at confidence intervals. Both ways of presenting the interval are commonly used, so we need to be able to go from one way to the other.

Example 5.2.3

A researcher observed 45 startup companies to find a 95% confidence interval for the population mean amount of time it takes to make a profit. The sample mean was 14 months and the margin of error was plus or minus 8 months. In symbols the confidence interval can be written as:

$$14 \pm 8$$

Express this as a trilinear inequality.

Solution

We first find the lower bound by subtracting:

$$14-8=6$$

Next, we find the upper bound by adding:

$$14+8=22$$

We can now put this together as a trilinear inequality:

$$6 \leq x \leq 22$$

Example 5.2.4

A researcher interviewed 1000 Americans to asking them if they thought abortion should be against the law. The following 95% confidence interval was given for the population proportion of all Americans who are against abortion:

$$(0.41, 0.47)$$

Find the midpoint and the margin or error. That is write this interval in the form:

$$a \pm b \tag{5.2.1}$$

Solution

Let's first find the midpoint. This is the average of the left and right endpoints:

$$a = \frac{0.41+0.47}{2} = 0.44$$

Next, find the distance from the midpoint to either boundary:

$$b = 0.47 - 0.44 = 0.3$$

Finally we can put these two together to get:

$$0.44 \pm 0.03$$

Exercise 5.2.1

A study was done to see how many years longer it takes low income students to finish college compared to high income students. The confidence interval for the population mean difference was found to be:

$$[0.67, 0.84]$$

Find the midpoint and the margin of error. That is write this interval as in the form:

$$a \pm b$$

Converting an Inequality from Interval Notation to Midpoint and Error Notation (Links to an external site.)

Writing Equations and Inequalities for Scenarios

5.2: Inequalities and Midpoints is shared under a not declared license and was authored, remixed, and/or curated by LibreTexts.

- **Inequalities and Midpoints** has no license indicated.

5.3: Solve Equations with Roots

Learning Outcomes

- Solve equations that include square roots.

Square roots occur frequently in a statistics course, especially when dealing with standard deviations and sample sizes. In this section we will learn how to solve for a variable when that variable lies under the square root sign. The key thing to remember is that the square of a square root is what lies inside. In other words, squaring a square root cancels the square root.

Example 5.3.1

Solve the following equation for x.

$$2+\sqrt{x-3} = 6$$

Solution

What makes this a challenge is the square root. The strategy for solving is to isolate the square root on the left side of the equation and then square both sides. First subtract 2 from both sides:

$$\sqrt{x-3}=4$$

Now that the square root is isolated, we can square both sides of the equation:

$$\left(\sqrt{x-3}\right)^2=4^2$$

Since the square and the square root cancel we get:

$$x-3=16$$

Finally add 3 to both sides to arrive at:

$$x=19$$

It's always a good idea to check your work. We do this by plugging the answer back in and seeing if it works. We plug in $x=19$ to get

$$\begin{aligned} 2+\sqrt{19-3} &= 2+\sqrt{16} \\ &= 2+4 \\ &= 6 \end{aligned}$$

Yes, the solution is correct.

Example 5.3.2

The standard deviation, $\sigma_{\hat{p}}$, of the sampling distribution for a proportion follows the formula:

$$\sigma_{\hat{p}}=\sqrt{\frac{p\left(1-p\right)}{n}}$$

Where p is the population proportion and n is the sample size. If the population proportion is 0.24 and you need the standard deviation of the sampling distribution to be 0.03, how large a sample do you need?

Solution

We are given that $p=0.24$ and $\sigma_{\hat{p}}=0.03$

Plug in to get:

$$0.03=\sqrt{\frac{0.24\left(1-0.24\right)}{n}}$$

We want to solve for n, so we want n on the left hand side of the equation. Just switch to get:

$$\sqrt{\frac{0.24\,(1-0.24)}{n}} = 0.03$$

Next, we subtract:

$$1-0.24 = 0.76$$

And them multiply:

$$0.24\,(0.76) = 0.1824$$

This gives us

$$\sqrt{\frac{0.1824}{n}} = 0.03$$

To get rid of the square root, square both sides:

$$\left(\sqrt{\frac{0.1824}{n}}\right)^2 = 0.03^2$$

The square cancels the square root, and squaring the right hand side gives:

$$\frac{0.1824}{n} = 0.0009$$

We can write:

$$\frac{0.1824}{n} = \frac{0.0009}{1}$$

Cross multiply to get:

$$0.0009\,n = 0.1824$$

Finally, divide both sides by 0.0009:

$$n = \frac{0.1824}{0.0009} = 202.66667$$

Round up and we can conclude that we need a sample size of 203 to get a standard error that is 0.03. We can check to see if this is reasonable by plugging $n = 203$ back into the equation. We use a calculator to get:

$$\sqrt{\frac{0.24\,(1-0.24)}{203}} = 0.029975$$

Since this is very close to 0.03, the answer is reasonable.

Exercise

The standard deviation, $\sigma_{\bar{x}}$, of the sampling distribution for a mean follows the formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where σ is the population standard deviation and n is the sample size. If the population standard deviation is 3.8 and you need the standard deviation of the sampling distribution to be 0.5, how large a sample do you need?

- Ex 1: Solve a Basic Radical Equation - Square Roots
- https://youtu.be/u1aGMkJIlMI

This page titled 5.3: Solve Equations with Roots is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Solve Equations with Roots** by Larry Green is licensed CC BY 4.0.

5.4: Solving Linear Equations in One Variable

Learning Outcomes

- Solve linear equations for the variable.

It is a common task in algebra to solve an equation for a variable. The goal will be to get the variable on one side of the equation all by itself and have the other side of the equation just be a number. The process will involve identifying the operations that are done on the variable and apply the inverse operation to both sides of the equation. This will be managed in the reverse of the order of operations.

Example 5.4.1

Solve the following equation for x.

$$3x+4=11 \tag{5.4.1}$$

Solution

We begin by looking at the operations that are done to x, keeping track the order. The first operation is "multiply by 3" and the second is "add 4". We now do everything backwards. Since the last operation is "add 4", our first step is to subtract 4 from both sides of Equation 5.4.1.

$$3x+\cancel{4}\cancel{-4}=11-4$$

which simplifies the equation

$$3x=7$$

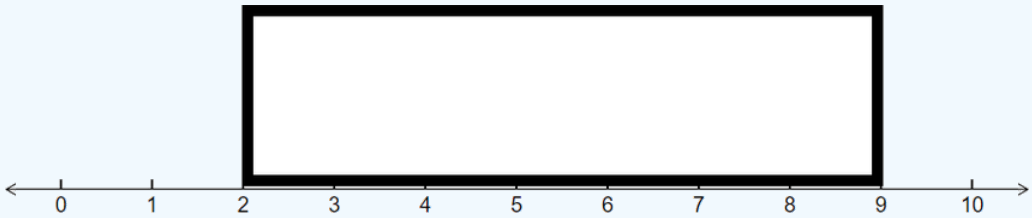
Next, the way to undo "multiply by 3" is to divide both sides by 3. We get

$$\frac{\cancel{3}x}{\cancel{3}}=\frac{7}{3}$$

or

$$x=\frac{7}{3}$$

Example 5.4.2



The rectangle above is a diagram for a uniform distribution from 2 to 9 that asks for the first quartile. The area of the smaller red rectangle that has base from 2 to Q1 and height 1/7 is 1/4. Find Q1.

Solution

We start by using the area formula for a rectangle:

$$\text{Area}=\text{Base}\times\text{Height} \tag{5.4.2}$$

We have:

- Area = $\frac{1}{4}$
- Base = $Q1-2$

- Height = $\frac{1}{7}$

Plug this into Equation 5.4.2 to get:

$$\frac{1}{4} = (Q1 - 2)\left(\frac{1}{7}\right) \tag{5.4.3}$$

We need to solve for $Q1$. First multiple both sides of Equation 5.4.3 by 7 to get:

$$7\left(\frac{1}{4}\right) = 7(Q1 - 2)\left(\frac{1}{7}\right)$$

$$\frac{7}{4} = Q1 - 2 \tag{5.4.4}$$

Now add 2 to both sides of Equation 5.4.4 to get:

$$\frac{7}{4} + 2 = Q1 - 2 + 2$$

$$\frac{7}{4} + 2 = Q1$$

or

$$Q1 = \frac{7}{4} + 2$$

Putting this into a calculator gives:

$$Q1 = 3.75$$

Example 5.4.3: z-score

The z-score for a given value x for a distribution with population mean μ and population standard deviation σ is given by:

$$z = \frac{x - \mu}{\sigma}$$

An online retailer has found that the population mean sales per day is \$2,841 and the population standard deviation is \$895. A value of x is considered an outlier if the z-score is less than -2 or greater than 2. How many sales must be made to have a z-score of 2?

Solution

First we identify each of the given variables. Since the population mean is 2,841, we have:

$$\mu = 2841$$

We are told that the population standard deviation is 895 meters, so:

$$\sigma = 895$$

We are also given that the z-score is 2, hence:

$$z = 2$$

Now we put the numbers into the formula for the z-score to get:

$$2 = \frac{x - 2841}{895}$$

We can next switch the order of the equation so that the x is on the left hand side of the equation:

$$\frac{x - 2841}{895} = 2$$

Next, we solve for x. First multiply both sides of the equation by 895 to get

$$x - 2841 = 2\,(895) = 1790$$

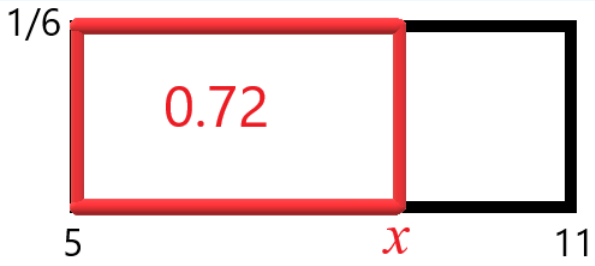
Finally, we can add 2841 to both sides of the equation to get x by itself:

$$x = 1790 + 2841 = 4631$$

We can conclude that if the day's sales is at $4631, the z-score is 2.

Exercise

The rectangle below is a diagram for a uniform distribution from 5 to 11 that asks for the 72nd percentile. The area of the smaller red rectangle that has base from 5 to the 72nd percentile, x, and height 1/6 is 0.72. Find x.



- Solving Two Step Equations: The Basics
- Solving Linear Equations

This page titled 5.4: Solving Linear Equations in One Variable is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Solving Linear Equations in One Variable** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

6: Graphing Points and Lines in Two Dimensions

This page titled 6: Graphing Points and Lines in Two Dimensions is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

6.1: Finding Residuals

Learning Outcomes

- Given a Regression line and a data point, find the residual

In the linear regression part of statistics we are often asked to find the residuals. Given a data point and the regression line, the residual is defined by the vertical difference between the observed value of y and the computed value of $\hat{y}$ based on the equation of the regression line:

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

Example 6.1.1

A study was conducted asking female college students how tall they are and how tall their mother is. The results are show in the table below:

Table of Mother and Daughter Heights

Mother's Height	63	67	64	60	65	67	59	60
Daughter's Height	58	64	65	61	65	67	61	64

The equation of the regression line is

$$\hat{y} = 30.28 \ + 0.52x$$

Find the residual for the mother who is 59 inches tall.

Solution

First note that the Daughter's Height associated with the mother who is 59 inches tall is 61 inches. This is y. Next we use the equation of the regression line to find $\hat{y}$. Since $x = 59$, we have

$$\hat{y} = 30.28 \ + 0.52(59)$$

We can use a calculator to get:

$$\hat{y} = 60.96$$

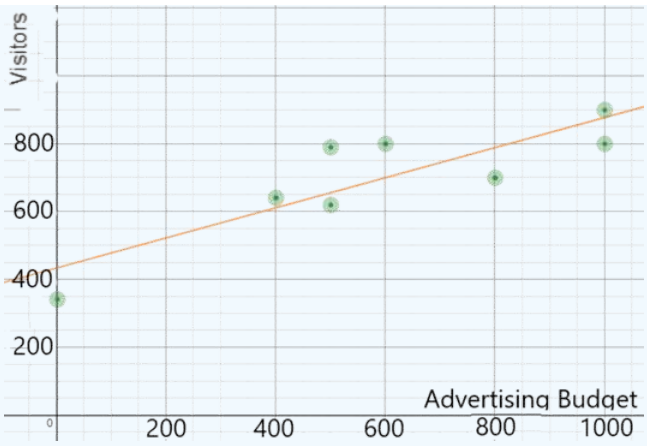
Now we are ready to put the values into the residual formula:

$$\text{Residual} = y - \hat{y} = 61 - 60.96 = 0.04$$

Therefore the residual for the 59 inch tall mother is 0.04. Since this residual is very close to 0, this means that the regression line was an accurate predictor of the daughter's height.

Example 6.1.2

An online retailer wanted to see how much bang for the buck was obtained from online advertising. The retailer experimented with different weekly advertising budgets and logged the number of visitors who came to the retailer's online site. The regression line for this is shown below.



Find the residual for the week when the retailer spent $600 on advertising.

Solution

First notice that the point of the scatterplot with x-coordinate of 600 has y-coordinate 800. Thus $y = 800$. Next note that the point on the line with x-coordinate 600 has y-coordinate 700. Thus $\hat{y} = 700$. Now we are ready to put the values into the residual formula:

$$\text{Residual} = y - \hat{y} = 800 - 700 = 100$$

Therefore the residual for the $600 advertising budget is -100.

Exercise

Data was taken from the recent Olympics on the GDP in trillions of dollars of 8 of the countries that competed and the number of gold medals that they won. The equation of the regression line is:

$$\hat{y} = 7.55 + 1.57x$$

The table below shows the data:

GDP	21	1.6	16	1.8	4	5.4	3.1	2.3
Medals	46	8	26	19	17	12	10	9

Find the residual for the country with a GDP of 4 trillion dollars.

- Calculating residual example | Exploring bivariate numerical data | AP Statistics | Khan Academy
- Finding a Residual

This page titled 6.1: Finding Residuals is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Finding Residuals** by Larry Green is licensed CC BY 4.0.

6.2: Find the Equation of a Line given its Graph

Learning Outcomes

1. Find the slope of a line given its graph.
2. Find the y-intercept of a line given its graph.
3. Find the equation of a line given its graph.

There are two main ways of representing a line: the first is with its graph, and the second is with its equation. In this section, we will practice how to find the equation of the line if we are given the graph of the line. The two key numbers in the equation of a line are the slope and the y-intercept. Thus the main steps in finding the equation of a line are finding the slope and finding the y-intercept. In statistics we are often presented with a **scatterplot** where we can eyeball the line. Once we have the graph of the line, getting the equation is helpful for making predictions based on the line.

Finding the Slope of a Line Given Its Graph

The steps to follow to fine the slope of the line given its graph are the following.

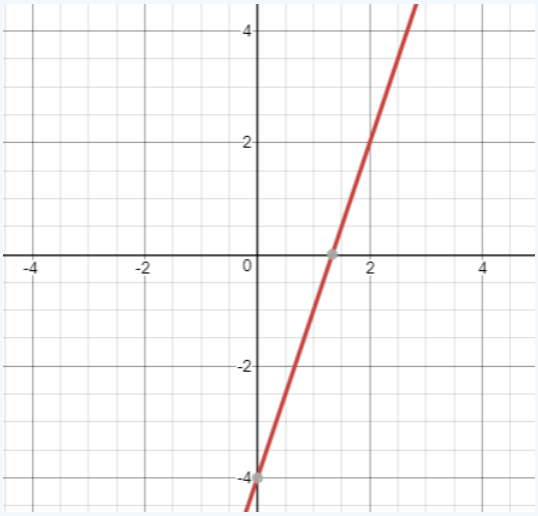
Step 1: Identify two points on the line. Any two points will do, but it is recommended to find points with nice x and y coordinates.

Step 2: The slope is the rise over the run. Thus if the points have coordinates (x_1, y_1) and $(x_2,\ y_2)$, then the slope is:

$$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1}$$

Example 6.2.1

Find the slope of the line shown below.



Solution

First, we locate points on the line that are as easy as possible to work with. The points with integer coordinates are (0,-4) and (2,2).

Next, we use the rise over run formula to find the slope of the line.

$$Slope = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2 - (-4)}{2 - 0} = \frac{6}{2} = 3$$

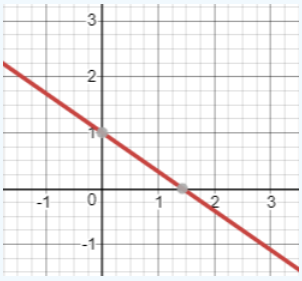
Finding the y-intercept from the graph

If the portion of the graph that is in view includes the y-axis, then the y-intercept is very easy to spot. You just see where it crosses the y-axis. On the other hand, if the portion of the graph in view does not contain the y-axis, then it is best to first find the equation

of the line and then use the equation to find the y-intercept.

Example 6.2.2

Find the y-intercept of the line shown below.



Solution

We just look at the line and notice that it crosses the y-axis at $y = 1$. Therefore, the y-intercept is 1 or (0,1).

Finding the equation of the line given its graph

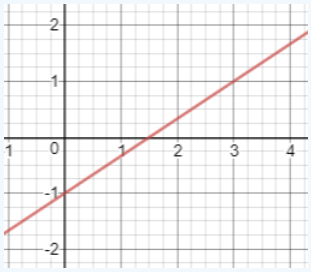
If you are given the graph of a line and want to find its equation, then you first find the slope as in Example 6.2.1. Then you use one of the points you found $(x_1,\ y_1)$ when you computed the slope, m, and put it into the **point slope equation**:

$$y - y_1 = m\,(x - x_1)$$

Then you multiply the slope through and add y_1 to both sides to get y by itself.

Example 6.2.3

Find the equation of the line shown below.



Solution

First we find the slope by identifying two nice points. Notice that the line passes through (0,-1) and (3,1). Now compute the slope using the rise over run formula:

$$Slope = \frac{rise}{run} = \frac{1-(-1)}{3-0} = \frac{2}{3}$$

Next use the point slope equation with the point (0,-1).

$$y-(-1) = \frac{2}{3}(x-0)$$

Now simplify:

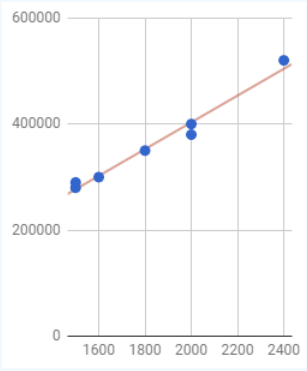
$$y+1 = \frac{2}{3}x$$

Finally subtract 1 from both sides to get:

$$y = \frac{2}{3}x - 1$$

Example 6.2.4

A study was done to look at the relationship between the square footage of a house and the price of the house. The scatter plot and regression line are shown below. Find the equation of the regression line.



Solution

First we find the slope by identifying two nice points. You will have to eyeball it and notice that the line passes through (1600, 300000) and (2000,400000). Now compute the slope using the rise over run formula:

$$\frac{rise}{run} = \frac{400000 - 300000}{2000 - 1600} = \frac{100000}{400} = 250$$

Next use the point slope equation with the point (2000,400000).

$$y - (400000) = 250\,(x - 2000)$$

Now simplify:

$$y - 400000 = 250x - 500000$$

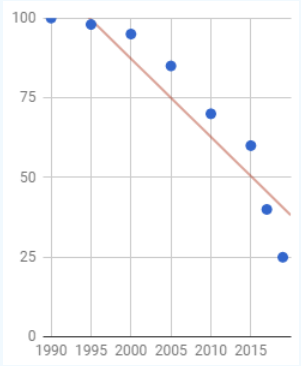
Finally add 400000 to both sides to get:

$$y = 250x - 100000$$

Notice that although the y-intercept is not visible from the graph of the line, we can see from the equation of the line that the y-intercept is -100000 or (0,-100000).

Exercise

The regression line and scatterplot below show the result of surveys that were taken in multiple years to find out the percent of households that had a landline telephone.



Find the equation of this regression line.

This page titled 6.2: Find the Equation of a Line given its Graph is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Find the Equation of a Line given its Graph** by Larry Green is licensed CC BY 4.0.

6.3: Find y given x and the Equation of a Line

Learning Outcomes

1. Find the value of y given x and the equation of a line.
2. Use a line to make predictions.

A line can be thought of as a function, which means that if a value of x is given, the equation of the line produces exactly one value of y; This is particularly useful in regression analysis where the line is used to make a prediction of one variable given the value of the other variable.

Example 6.3.1

Consider the line with equation:

$$y = 3x - 4$$

Find the value of y when x is 5.

Solution

Just replace the variable x with the number 5 in the equation and perform the arithmetic:

$$y = 3(5) - 4 = 15 - 4 = 11$$

Example 6.3.2

A survey was done to look at the relationship between a woman's height, x and the woman's weight, y. The equation of the regression line was found to be:

$$y = -220 + 5.5x$$

Use this equation to estimate the weight in pounds of a woman who is 5' 2" (62 inches) tall.

Solution

Just replace the variable x with the number 62 in the equation and perform the arithmetic:

$$y = -220 + 5.5(62)$$

We can put this into a calculator or computer to get:

$$y = 121$$

Therefore, our best prediction for the weight of a woman who is 5' 2" tall is that she is 121 lbs.

Exercise

A biologist has collected data on the girth (how far around) of pine trees and the pine tree's height. She found the equation of the regression line to be:

$$y = 1.3 + 2.7x$$

Where the girth, x, is measured in inches and the height, y, is measured in feet. Use the regression line to predict the height of a tree with girth 28 inches.

https://youtu.be/cS95PlUKZ6I

This page titled 6.3: Find y given x and the Equation of a Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Find y given x and the Equation of a Line** by Larry Green is licensed CC BY 4.0.

6.4: Graph a Line given its Equation

Learning Outcomes

1. Identify the slope and y-intercept from the equation of a line.
2. Plot the y-intercept of a line given its equation.
3. Plot a second point on a line given the y-intercept and the slope.
4. Graph a line given its equation in slope y-intercept form.

Often we are given an equation of a line and we want to visualize it. For this reason, it is important to be able to graph a line given its equation. We will look at lines that are in slope intercept form: $y = a + bx$ where a is the y-intercept of the line and b is the slope of the line. The y-intercept is the value of y where the line crosses the y-axis. The slope is the rise over run. If we write the slope as a fraction, then the numerator tells us how far to move up (or down if it is negative) and the denominator tells us how far to the right we need to go. the main application to statistics is in regression analysis which is the study of how to use a line to make a prediction about one variable based on the value of the other variable.

Example 6.4.1

Graph the line given by the equation:

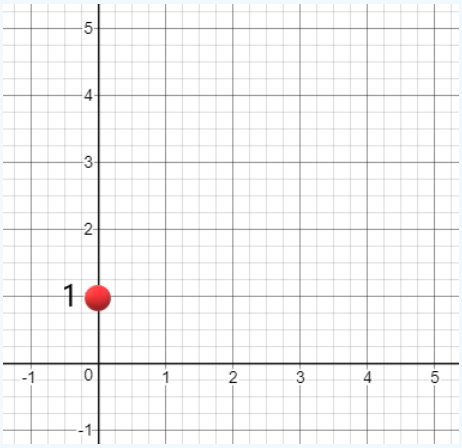
$$y = 1 + \frac{3}{2}x$$

Solution

We follow the three step process:

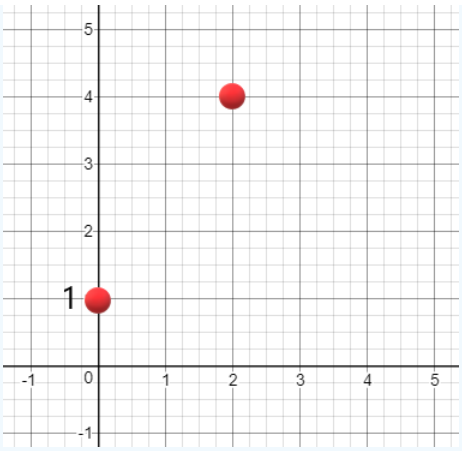
Step 1: Plot the y-intercept

The y-intercept is the number that is not associated with the x. For this example, it is 1. The x-coordinate of the y-intercept is always 0. So the coordinates of the y-intercept are (0,1). Thus start at the origin and move up 1:



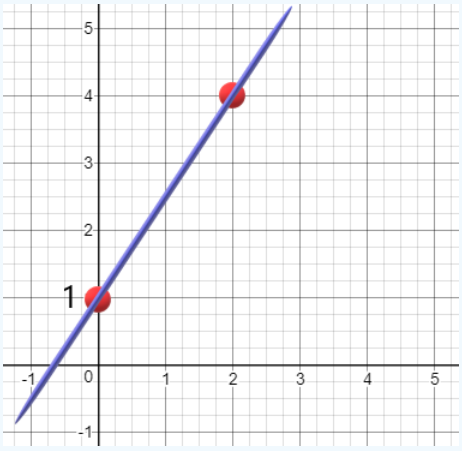
Step 2: Plot the Slope.

The slope of a line is the coefficient of the x term. Here it is $\frac{3}{2}$. What this means is that we rise 3 and run to the right 2. Rising 3 from an original y-coordinate of 1 gives a new y-coordinate of 4. Running 2 to the right from an initial x-coordinate of 0gives a new x-coordinate of 2. Thus we next plot the point (2,4).



Step 3: Connect the Dots

The last thing we need to do is connect the dots with a line:



Example 6.4.2

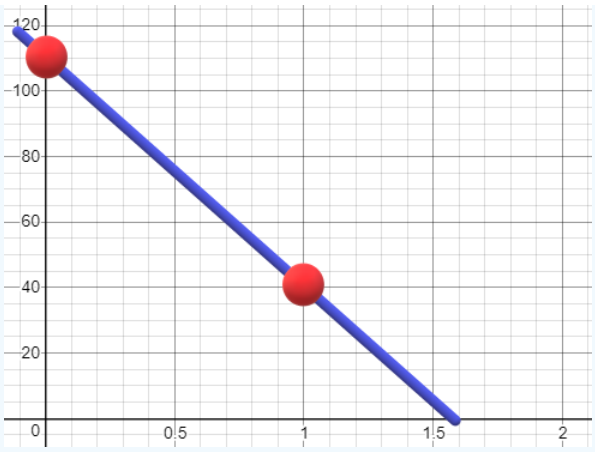
A study was done to look at the relationship between the weight of a car, x, in tons and its gas mileage in mpg, y. The equation of the regression line was found to be:

$$y = 110 - 70x \tag{6.4.1}$$

Graph this line.

Solution

The fist step is to note that the y-intercept is 110, hence the graph goes through the point (0,110). The next step is to see that the slope is -70. We can always put a number over 1 in order to make it a fraction. The slope of $-\frac{70}{1}$ tells us that y goes down by 70 if x goes up by 1. We use this to find the second point. The y-coordinate is: $110 - 70 = 40$. The x-coordinate is 1. Thus, a second point is (1,40). We can now plot the two points and connect the dots with a line.



Exercise

The regression line that relates the ounces of beer consumed just before a test, x, and the score on the test, y, is given by

$y = 93 - 1.2x$

Graph this line.

Graphing a Line in Slope-Intercept Form

https://youtu.be/z3rM-ZidXaw

This page titled 6.4: Graph a Line given its Equation is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Graph a Line given its Equation** by Larry Green is licensed CC BY 4.0.

6.5: Interpreting the Slope of a Line

Learning Outcomes

1. Interpret the slope of a line as the change in y when x changes by 1.

Template for Interpreting the Slope of a Line

For every increase in the x-variable by 1, the y-variable tends to change by (xxx the slope).

A common issue when we learn about the equation of a line in algebra is to state the slope as a number, but have no idea what it represents in the real world. The slope of a line is the rise over the run. If the slope is given by an integer or decimal value we can always put it over the number 1. In this case, the line rises by the slope when it runs 1. "Runs 1" means that the x value increases by 1 unit. Therefore the slope represents how much the y value changes when the x value changes by 1 unit. In statistics, especially regression analysis, the x value has real life meaning and so does the y value.

Example 6.5.1

A study was done to see the relationship between the time it takes, x, to complete a college degree and the student loan debt incurred, y. The equation of the regression line was found to be:

$$y = 25142 + 14329x \tag{6.5.1}$$

Interpret the slope of the regression line in the context of the study.

Solution

First, note that the slope is the coefficient in front of the x. Thus, the slope is 14,329. Next, the slope is the rise over the run, so it helps to write the slope as a fraction:

$$Slope = \frac{rise}{run} = \frac{14,329}{1} \tag{6.5.2}$$

The rise is the change in y and y represents student loan debt. Thus, the numerator represents an increase of $14,329 of student loan debt. The run is the change in x and x represents the time it takes to complete a college degree. Thus, the denominator represents an increase of 1 year to complete a college degree. We can put this all together and interpret the slope as telling us that

For every additional year it takes to complete a college degree, on average the student loan debt tends to increase by $14,329.

Example 6.5.2

Suppose that a research group tested the cholesterol level of a sample of 40 year old women and then waited many years to see the relationship between a woman's HDL cholesterol level in mg/dl, x, and her age of death, y. The equation of the regression line was found to be:

$$y = 103 - 0.3x \tag{6.5.3}$$

Interpret the slope of the regression line in the context of the study.

Solution

The slope of the regression line is -0.3. The slope as a fraction is:

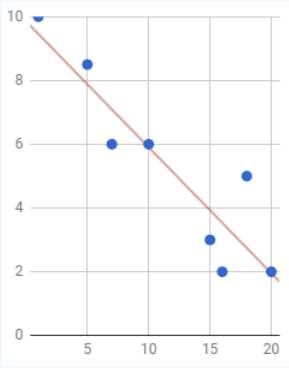
$Slope = \frac{rise}{run} = \frac{-0.3}{1}$ ” $width =$” 233

The rise is the change in y and y represents age of death. Since the slope is negative, the numerator indicates a decrease in lifespan. Thus, the numerator represents a decrease in lifespan of 0.3 years. The run is the change in x and x represents the HDL cholesterol level. Thus, the denominator represents an HDL cholesterol level increase of 1 mg/dl. Now, put this all together and interpret the slope as telling us that

For every additional 1 mg/dl of HDL cholesterol, on average women are predicted to die 0.3 years younger.

Example 6.5.3

A researcher asked several employees who worked overtime "How many hours of overtime did you work last week?" and "On a scale from 1 to 10 how satisfied are you with your job?". The scatterplot and the regression line from this study are shown below.



Interpret the slope of the regression line in the context of the study.

Solution

We first need to determine the slope of the regression line. To find the slope, we get two points that have as nice coordinates as possible. From the graph, we see that the line goes through the points (10,6) and (15,4). The slope of the regression line can now be found using the rise over the run formula:

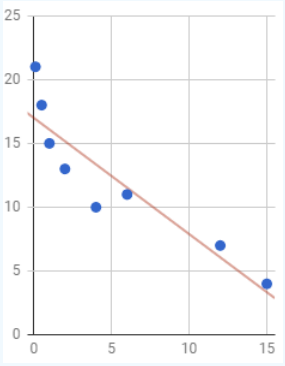
$$Slope = \frac{rise}{run} = \frac{4-6}{15-10} = \frac{-2}{5} \tag{6.5.4}$$

The rise is the change in y and y represents job satisfaction rating. Since the slope is negative, the numerator indicates a decrease in job satisfaction. Thus, the numerator represents a decrease in job satisfaction of 2 on the scale from 1 to 10. The run is the change in x and x represents the overtime work hours. Thus, the denominator represents an increase of 5 hours of overtime work. Now, put this all together and interpret the slope as telling us that

For every additional 5 hours of overtime work that employees are asked to do, their job satisfaction tends to go down an average of 2 points.

Exercise

The scatterplot and regression line below are from a study that collected data on the population (in hundred thousands) of cities and the average number of hours per week the city's residents spend outdoors.



Interpret the slope of this regression line in the context of the study.

Interpret the Meaning of the Slope of a Linear Equation - Smokers

Interpreting the Slope of a Regression Line

This page titled 6.5: Interpreting the Slope of a Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Interpreting the Slope of a Line** by Larry Green is licensed CC BY 4.0.

6.6: Interpreting the y-intercept of a Line

Learning Outcomes

1. Interpret the y-intercept of a line as the value of y when x equals to 0.
2. Determine whether the y-intercept is useful for interpreting the relationship between x and y

Just like the slope of a line, many algebra classes go over the y-intercept of a line without explaining how to use it in the real world. The y-intercept of a line is the value of y where the line crosses the y-axis. In other words, it is the value of y when the value of x is equal to 0. Sometimes this has true meaning for the model that the line provides, but other times it is meaningless. We will encounter examples of both types in this section.

Template for the y-Intercept Interpretation

When the value for the x-variable is 0, the best prediction for the value of the y-variable is (xxx the y-intercept).

Example 6.6.1

A study was done to see the relationship between the ounces of meat, x, that people eat each day on average and the hours per week, y they watch sports. The equation of the regression line was found to be:

$$y = 1.3 + 0.4x$$

Interpret the y-intercept of the regression line in the context of the study or explain why it has no practical meaning.

Solution

First, note that the y-intercept is the number that is not in front of the x. Thus, the y-intercept is 1.3. Next, the y-intercept is the value of y when x equals zero. For this example, x represents the ounces of meat consumed each day.

When the consumption of meat is 0, the best prediction for the value of the hours of sports each week is 1.3.

If x is equal to 0, this means the person does not consume any meat. Since there are people, called vegetarians, who consume no meat, it is meaningful to have an x-value of 0. The y-value of 1.3 represents the hours of sports the person watches. Putting this all together we can state:

A vegetarian is predicted to watch 1.3 hours of sports each week.

Example 6.6.2

A neonatal nurse at Children's Hospital has collected data on the birth weight, x, in pounds the number of days, y, that the newborns stay in the hospital. The equation of the regression line was found to be

$$y = 45 - 3.9x$$

Interpret the y-intercept of the regression line in the context of the study or explain why it has no practical meaning.

Solution

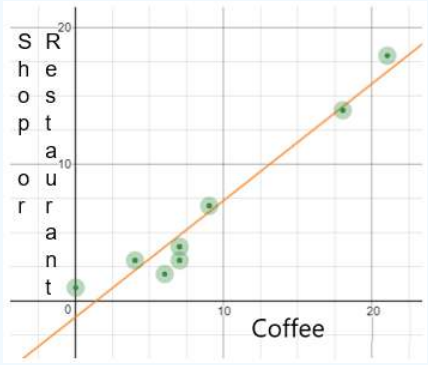
Again, we note that the y-intercept is the number that is not in front of the x. Thus, the y-intercept is 45. Next, the y-intercept is the value of y when x equals zero.

When the birth weight in pounds is 0, the best prediction for the value of the number of days the newborn is predicted to stay in the hospital is 45 days.

For this example, x represents the new born baby's birth weight in pounds. If x is equal to 0, this means the baby was born with a weight of 0 pounds. Since it makes no sense for a baby to weigh 0 pounds, we can say that the y-intercept of this regression line has no practical meaning.

Example 6.6.3

A researcher asked several people "How many cups of coffee did you drink last week?" and "How many times did you go to a shop or restaurant for a meal or a drink last week?" The scatterplot and the regression line from this study are shown below.



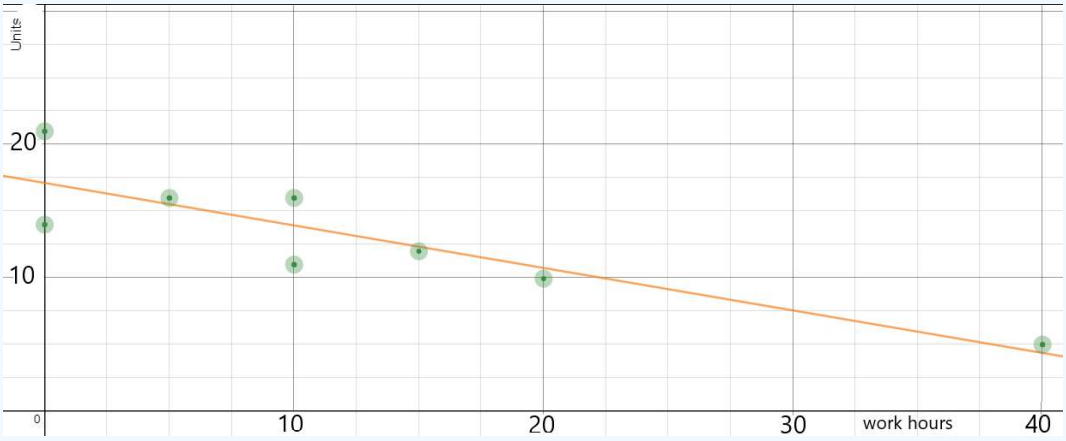
Interpret the y-intercept of the regression line in the context of the study or explain why it has no practical meaning.

Solution

The y-intercept of a line is where it crosses the y-axis. In this case, the line crosses at around y = -1. The value of x, by definition is 0 and the x-axis represents the number of cups of coffee a person drank last week. Since there are people who don't drink coffee, it does male sense to have an x-value of 0. The y-axis represents the number of times the person went to a shop or restaurant last week to purchase a meal or a drink. It makes no sense to say that a person went -1 times to a shop or restaurant last week to purchase a meal or a drink. Therefore the y-intercept of this regression line has no practical meaning.

Exercise

The scatterplot and regression line below are from a study that collected data from a group of college students on the number of hours per week during the school year they work at a paid job and the number of units they are taking. Interpret the y-intercept of the regression line or explain why it has no practical meaning.



- Interpret the Meaning of the y-intercept Given a Linear Equation
- Interpreting the y-Intercept

This page titled 6.6: Interpreting the y-intercept of a Line is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Interpreting the y-intercept of a Line** by Larry Green is licensed CC BY 4.0.

6.7: Plot an Ordered Pair

Learning Outcomes

1. Draw x and y axes.
2. Plot a point in the xy-plane

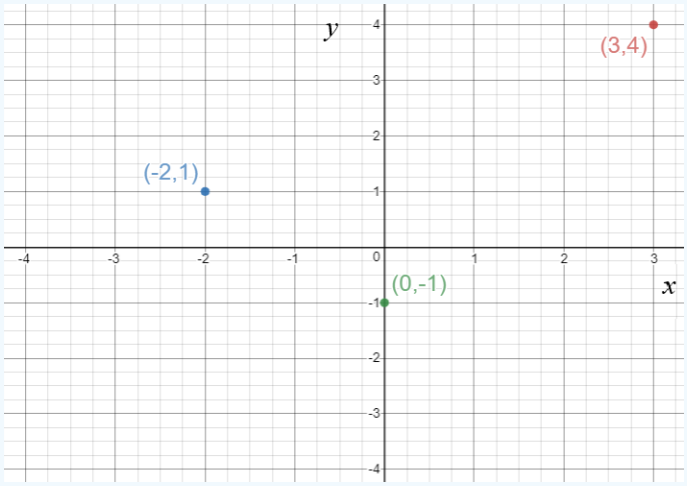
We have already gone into detail about how to plot points on a number line, and that is very useful for single variable presentations. Now we will move to questions that involve comparing two variables. Working with two variables is frequently encountered in statistical studies and we would like to be able to display the results graphically. This is best done by plotting points in the xy-plane.

Example 6.7.1

Plot the points: $(3,4)$,$(-2,1)$, and $(0,-1)$

Solution

The first thing to do when plotting points is to sketch the x-axis and y-axis and decide on the tick marks. Here the numbers are all less than 5, so it is reasonable to count by 1's. Next, we plot the first point, $(3,4)$. This means to start at the origin, where the axes intersect. Then move 3 units to the right and 4 units up. After arriving there, we just draw a dot. For the next point, $(-2,1)$, we start at the origin, move 2 units to the left and 1 unit up and draw the dot. For the third point, $(0,-1)$, we don't move left or right at all since the x-coordinate is 0, but we do move 1 unit down and draw the dot. The plot is shown below.



Example 6.7.2

A survey was done to look at the relationship between a person's age and their income. The first three answers are shown in the table below:

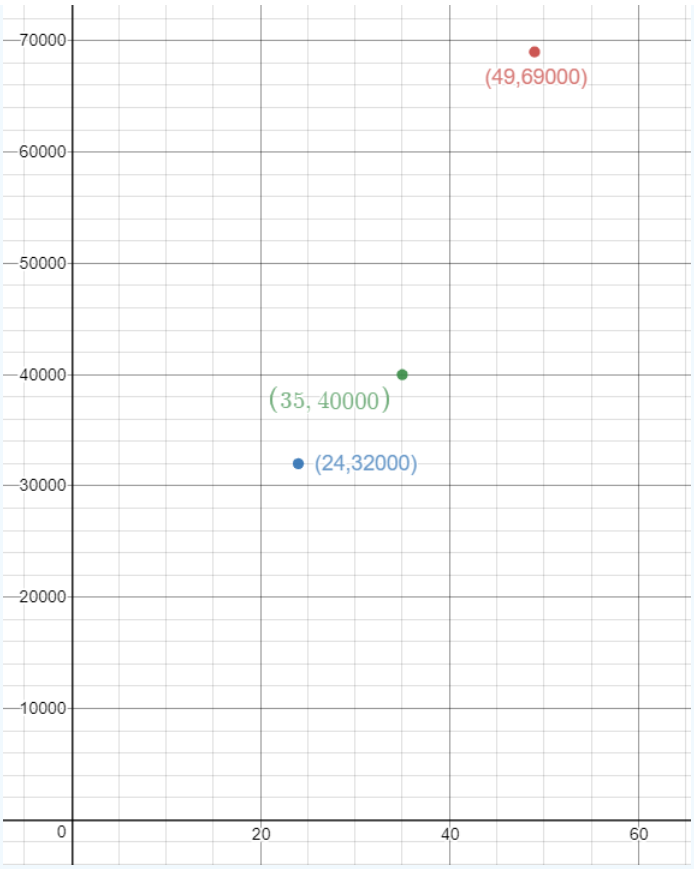
Table of ages and income

Age	49	24	35
Income	69,000	32,000	40,000

Graph the three points on the xy-plane.

Solution

Notice that the numbers are all relatively large. Therefore counting by 1's would not make sense. Instead, it makes better sense to count the Age axis, x, by 10's and the Income axis, y, by 1000's. The points are plotted below.



Exercise

A hotel manager was interested in seeing the relationship between the price per night, x, that the hotel charged and the number of occupied rooms, y. The results were (75,83), (100,60), (110,55), and (125,40). Plot these points in the xy-plane.

Ex: Plotting Points on the Coordinate Plane

Plotting Points

This page titled 6.7: Plot an Ordered Pair is shared under a CC BY license and was authored, remixed, and/or curated by Larry Green.

- **Plot an Ordered Pair** by Larry Green is licensed CC BY 4.0.

CHAPTER OVERVIEW

7: Geometry

Thumbnail: Similar Triangles. (CC BY-SA 3.0; Nguyenthephuc via Wikipedia).

This page titled 7: Geometry is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Henry Africk (New York City College of Technology at CUNY Academic Works) .

7.1: Angles

An *angle* is the figure formed by two rays with a common end point, The two rays are called the sides of the angle and the common end point is called the *vertex* of the angle, The symbol for angle is $\angle$

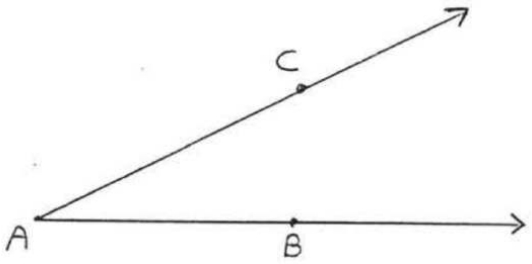


Figure 7.1.1: Angle BAC has vertex A and sides $\overrightarrow{AB}$ and $\overrightarrow{AC}$

The angle in Figure 7.1.1 has vertex A and sides AB and AC, It is denoted by $\angle BAC$ or $\angle CAB$ or simply $\angle A$. When three letters are used, the middle letter is always the vertex, In Figure 7.1.2 we would not use the notation $\angle A$ as an abbreviation for $\angle BAC$ because it could also mean $\angle CAD$ or $\angle BAD$, We could however use the simpler name $\angle x$ for $\angle BAC$ if "x" is marked in as shown,

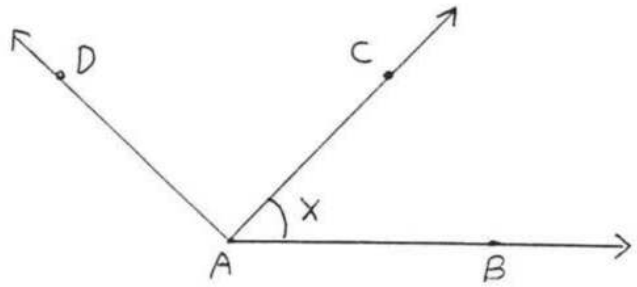


Figure 7.1.2 $\angle BAC$ may also be denoted by $\angle x$.

Angles can be measured with an instrument called a *protractor.* The unit of measurement is called a *degree* and the symbol for degree is $^\circ$.

To measure an angle, place the center of the protractor (often marked with a cross or a small circle) on the vertex of the angle, Position the protractor so that one side of the angle cuts across 0, at the beginning of the scale, and so that the other side cuts across a point further up on the scale, We use either the upper scale or the lower scale, whichever is more convenient, For example, in Figure 7.1.3, one side of $\angle BAC$ crosses 0 on the lower scale and the other side crosses 50 on the lower scale. The measure of $\angle BAC$ is therefore 50° and we write $\angle BAC = 50^\circ$.

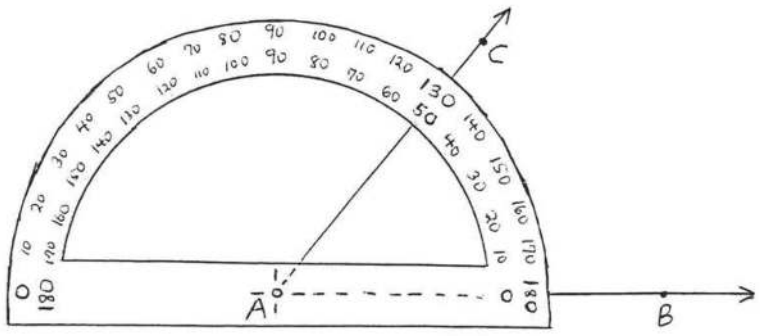


Figure 7.1.3 The protractor shows $\angle BAC = 50^\circ$

In Figure 7.1.4, side $\overrightarrow{AD}$ of $\angle DAC$ crosses 0 on the upper scale. Therefore we look on the upper scale for the point at which $\overrightarrow{AC}$ crosses and conclude that $\angle DAC = 130^\circ$.

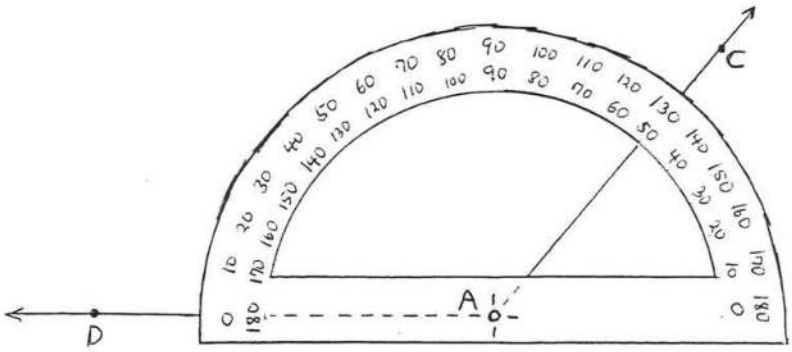


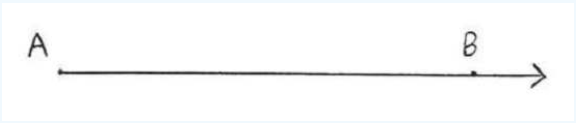
Figure 7.1.4 $\angle DAC = 130^\circ$.

Example 7.1.1

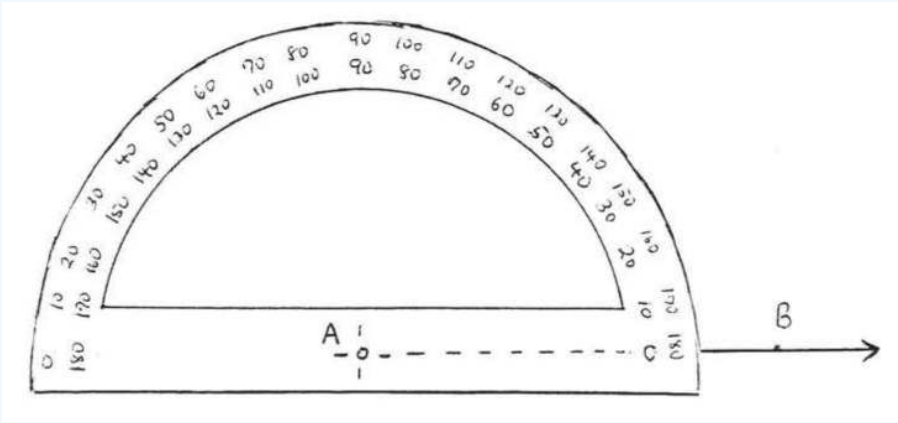
Draw an angle of 40° and label it $\angle BAC$.

Solution

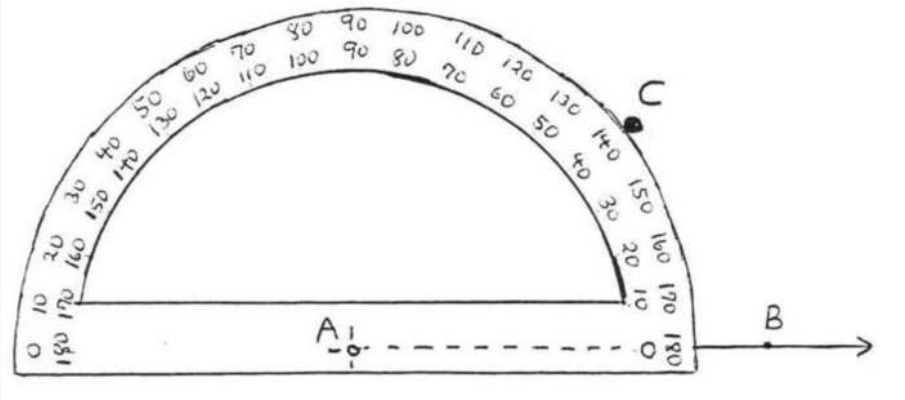
Draw ray $\overrightarrow{AB}$ using a straight edge:



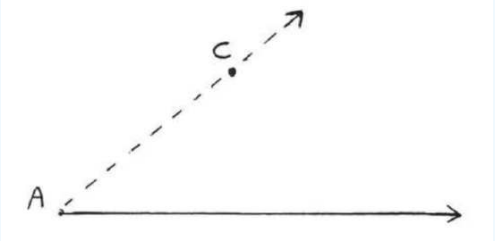
Place the protractor so that its center coincides with A and $\overrightarrow{AB}$ crosses the scale at 0:



Mark the place on the protractor corresponding to 40°. Label this point C:



Connect A with C:



Two angles are said to be equal if they have the same measure in degrees. We often indicate two angles are equal by marking them in the same way. In Figure 7.1.5, $\angle A = \angle B$.

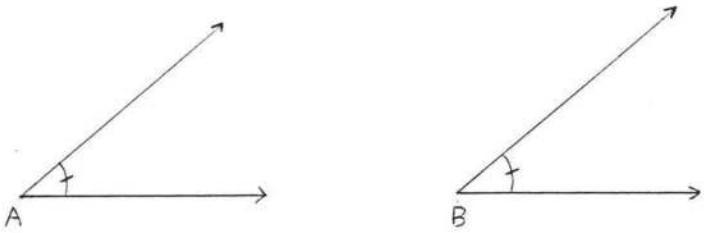


Figure 7.1.5 Equal angles.

An angle bisector is a ray which divides an angle into two equal angles. In Figure 7.1.6, $\overrightarrow{AC}$ is an angle bisector of $\angle BAD$. We also say $\overrightarrow{AC}$ bisects $\angle BAD$.

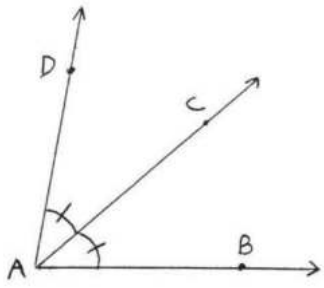
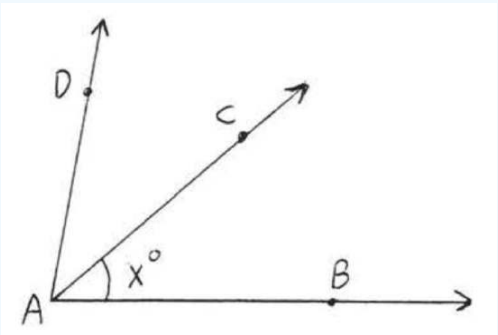


Figure 7.1.6 $\overrightarrow{AC}$ bisects $\angle BAD$.

Example 7.1.2

Find x if $\overrightarrow{AC}$ bisects $\angle BAD$ and $\angle BAD = 80^\circ$:



Solution

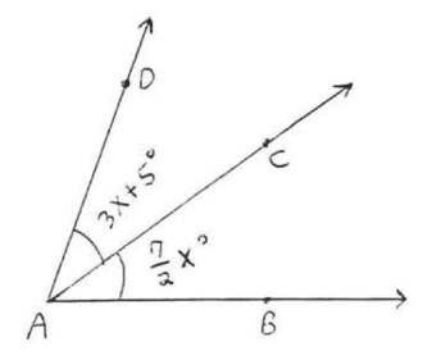
$x^\circ = \frac{1}{2}\angle BAD = \frac{1}{2}(80^\circ) = 40^\circ$

Answer: $x = 40$.

https://stats.libretexts.org/@go/page/24853

Example 7.1.3

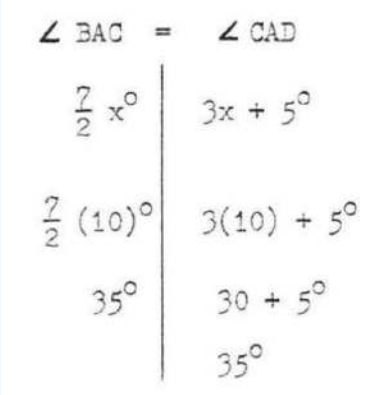
Find x if $\overrightarrow{AC}$ bisects $\angle BAD$:



Solution

$$\begin{aligned} \angle BAC &= \angle CAD \\ \frac{7}{2}x &= 3x+5 \\ (2)\frac{7}{2}x &= (2)(3x+5) \\ 7x &= 6x+10 \\ 7x-6x &= 10 \\ x &= 10 \end{aligned} \tag{7.1.1}$$

Check:

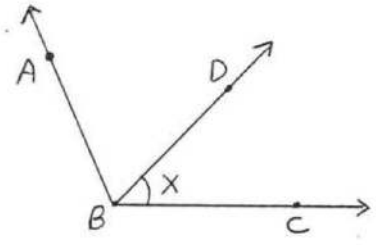


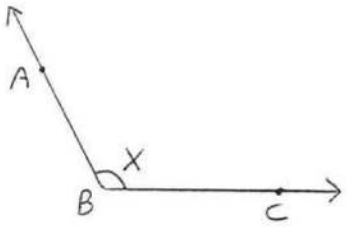
Answer: $x = 10$.

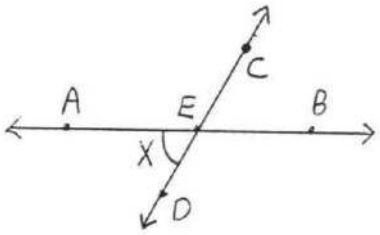
Problems

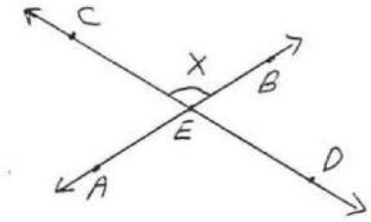
1 - 6. For each figure, give another name for $\angle x$:

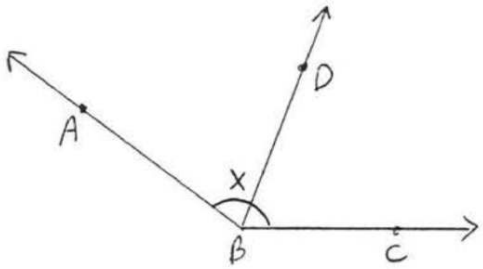
1.

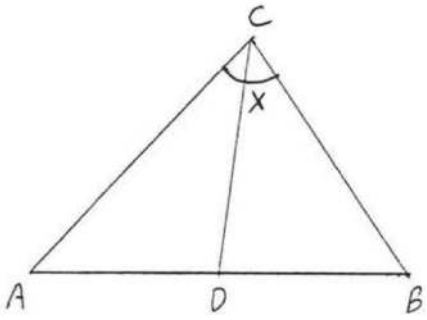


2. 

3. 

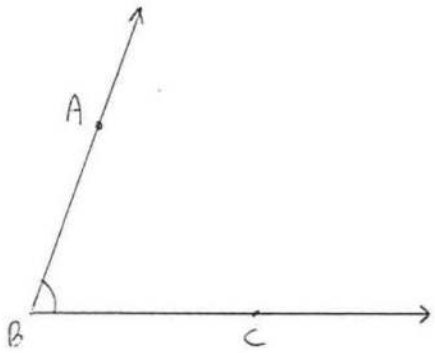
4. 

5. 

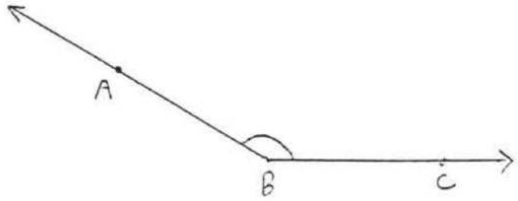
6. 

7 - 16, Measure each of the indicated angles:

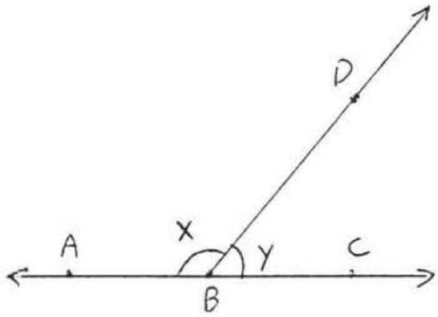
7.



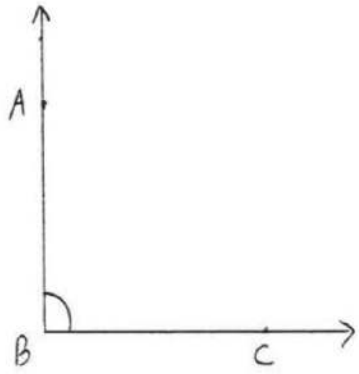
8.



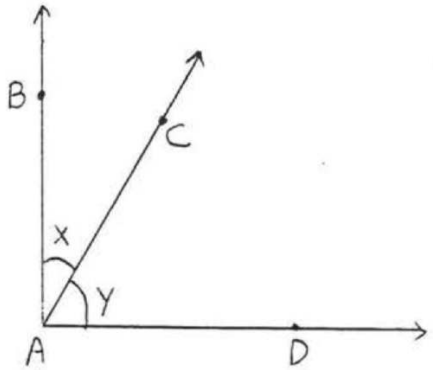
9.

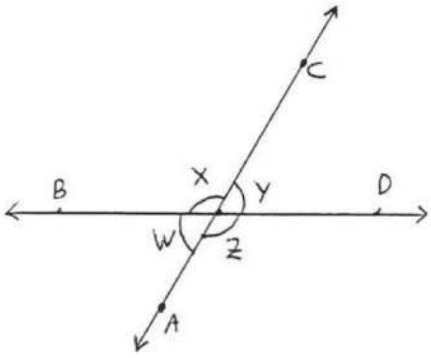


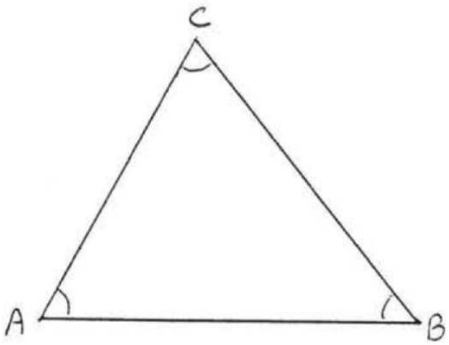
10.

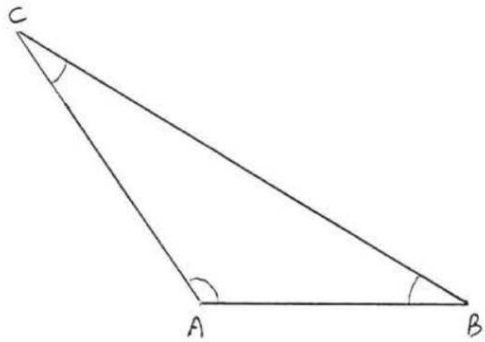


https://stats.libretexts.org/@go/page/24853

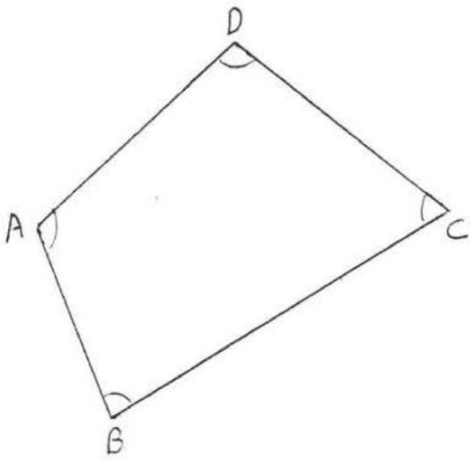
11. 

12. 

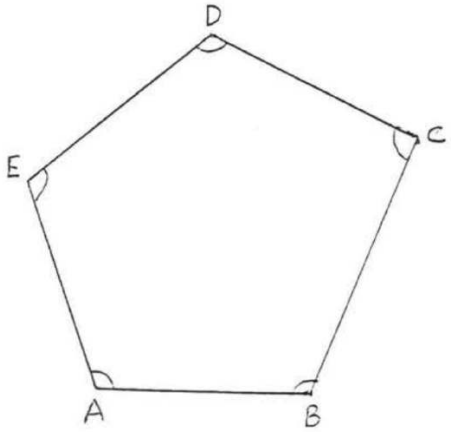
13. 

14. 

15.



16.



17 - 24. Draw and label each angle:

17. $\angle BAC = 30^\circ$

18. $\angle BAC = 40^\circ$

19. $\angle ABC = 45^\circ$

20. $\angle EFG = 60^\circ$

21. $\angle RST = 72^\circ$

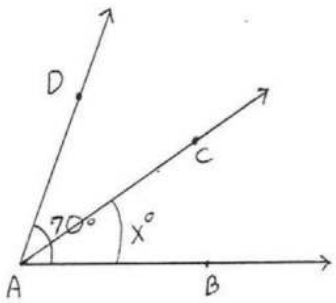
22. $\angle XYZ = 90^\circ$

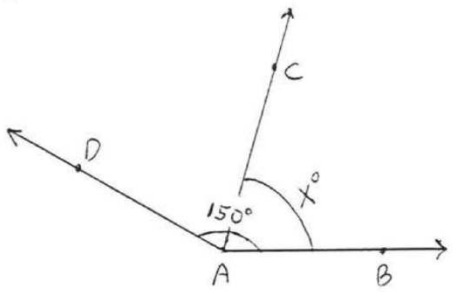
23. $\angle PQR = 135^\circ$

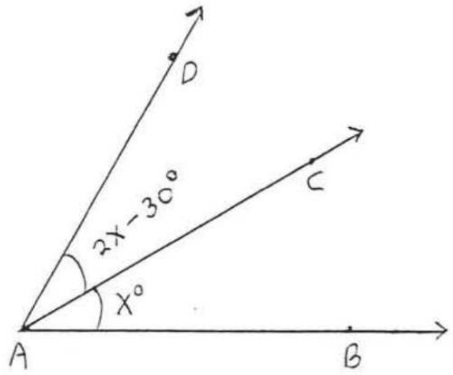
24. $\angle JKL = 164^\circ$

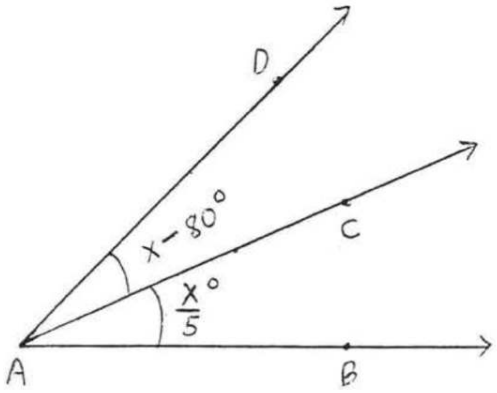
25 - 28. Find x if $\overrightarrow{AC}$ bisects $\angle BAD$:

25.



26. 

27. 

28. 

This page titled 7.1: Angles is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Henry Africk (New York City College of Technology at CUNY Academic Works) .

- **1.2: Angles** by Henry Africk is licensed CC BY-NC-SA 4.0. Original source: https://academicworks.cuny.edu/ny_oers/44.

7.2: The Area of a Rectangle and Square

The measurement of the area of geometric figures is one of the most familitax ways mathematics is used in our daily lives. The floor space of a building, the stae of a picture, the amount of paper in a roll of paper towels are all examples of Items often measured in terms of area. In this chapter we will derive formulas for the areas of the geometric objects which we have studied.

Area is measured in square inches, square feet, square centimeters, etc. The bastc unit of measurement is the unit square, the square whose sides are of length 1 (Figure 7.2.1). Its area Is 1 square inch, 1 square foot, 1 square centimeter, etc., depending on which measurement of Iength is chosen. The area of any closed figure is defined to be the number of unit squares it contains.

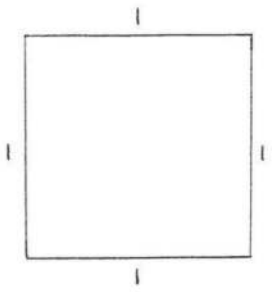


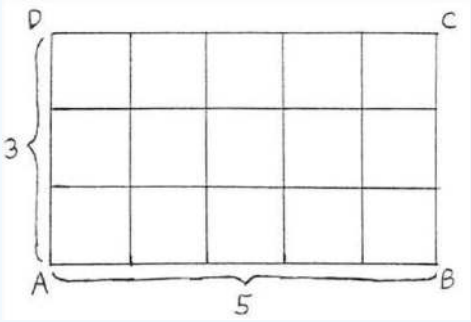
Figure 7.2.1: The unit square.

Example 7.2.1

Find the area of a rectangle with length 5 and width 3.

Solution

We see from the diagram that the area is $(5)(3) = 15$



Answer: 15.

This suggests the following theorem:

Theorem 7.2.1

The area of a rectangle is the length times its width.

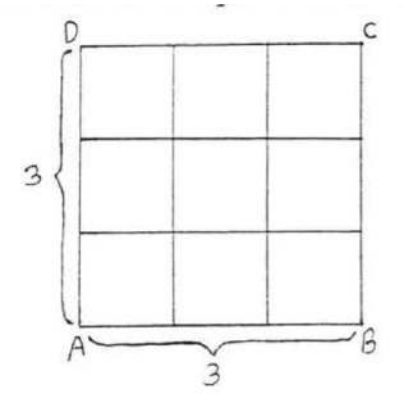
$$A = lw$$

Example 7.2.2

Find the area of a square with side 3.

Solution

Area = $(3)(3) = 3^2 = 9$.



Answer: 9.

The formula for a squa:re is now self-evident:

Theorem 7.2.2

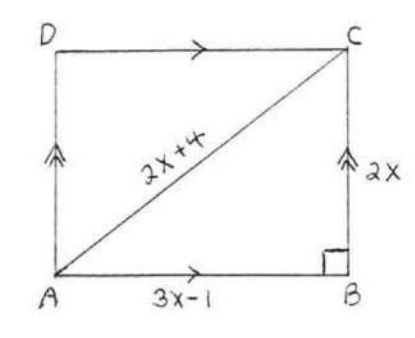
The area of a square is the square of one of its sides.

$$A = s^2$$

The perimeter of a polygon is the su.m of the lengths of its sides. For instance the perimeter of the rectangle of Example 7.2.1 would be 5 + 5 + 3 + 3 = 16.

Example 7.2.3

Find the area and perimeter of rectangle $ABCD$:



Solution

We first use the Pythagorean Theorem to find x:

$$\begin{aligned} AB^2 + BC^2 &= AC^2 \\ (3x-1)^2 + (2x)^2 &= (2x+4)^2 \\ 9x^2 - 6x + 1 + 4x^2 &= 4x^2 + 16x + 16 \\ 9x^2 - 22x - 15 &= 0 \\ (9x+5)(x-3) &= 0 \end{aligned}$$

$$\begin{aligned} 9x+5 &= 0 & \qquad x-3 &= 0 \\ x &= -\frac{5}{9} & \qquad x &= 3 \end{aligned}$$

We reject the answer $x = -\frac{5}{9}$ because $BC = 2x = 2(-\frac{5}{9}) = -\frac{10}{9}$ would have negative length. Therefore $x = 3$.

$AB = 3x - 1 = 3(3) - 1 = 9 - 1 = 8$. $BC = 2x = 2(3) = 6$.

$AC = 2x + 4 = 2(3) + 4 = 6 + 4 = 10$.

Check:

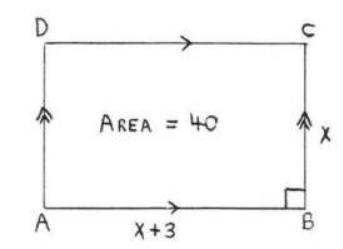
$$\begin{array}{r|l} AB^2 + BC^2 = AC^2 \\ 8^2 + 6^2 & 10^2 \\ 64 + 36 & 100 \\ 100 & \end{array}$$

Area = $lw = (8)(6) = 48$. Perimeter = 8 + 8 + 6 + 6 = 28.

Answer: Area = 48, Perimeter = 28.

Example 7.2.4

Find x:



Solution

$$\begin{array}{rcl} A & = & lw \\ 40 & = & (x+3)(x) \\ 40 & = & x^2+3x \\ 0 & = & x^2+3x-40 \\ 0 & = & (x-5)(x+8) \\ x & = & 5 \quad x=-8 \end{array}$$

We reject $x = -8$ because side $BC = x$ of the rectangle would be negative.

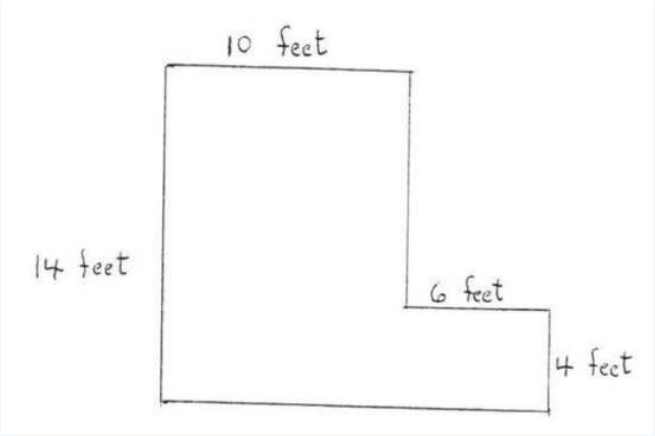
Check, $x = 5$:

$$\begin{array}{r|l} A = lw \\ 40 & (x+3)(x) \\ & (5+3)(5) \\ & (8)(5) \\ & 40 \end{array}$$

Answer: $x = 5$.

Example 7.2.5

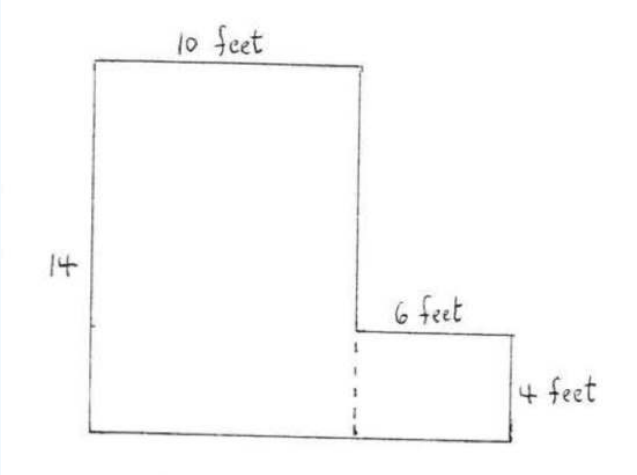
An L-shaped room has the dimensions indicated in the diagram, How many one by one foot tiles are needed to tile the floor?



Solution

Divide the room into two rectangles as shown.

$$\begin{aligned} \text{Area of room} &= \text{Area of large rectangle} \\ &+ \text{Area of small rectangle} \\ &= (14)(10)+(6)(4) \\ &= 140+24 \\ &= 164 \text{ square feet.} \end{aligned}$$



Answer: 164

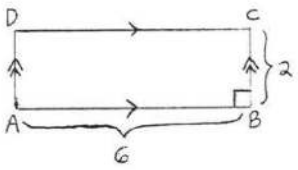
Historical Note

The need to measu:re land areas was one of the ancient problems which led to the development of geometry. Both the early Egyptians and Babylonians had formulas for the areas of rectangles, triangles, and trapezoids, but some of their formulas were not entirely accurate. The formulas in this chapter were known to the Greeks and are found in Euclid's Elements.

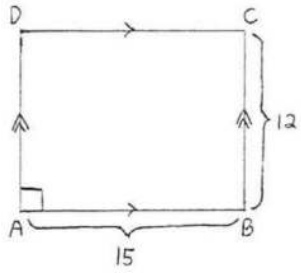
Problems

1 - 14. Find the area and perimeter of $ABCD$:

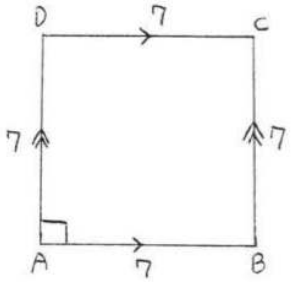
1.



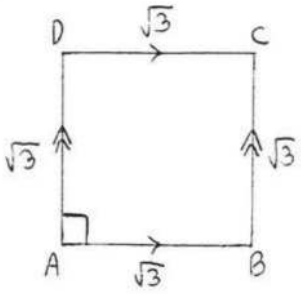
2.



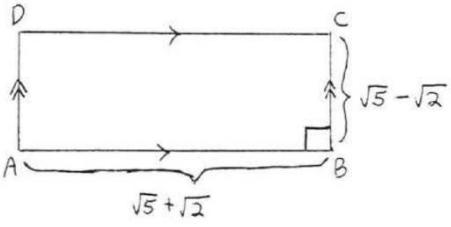
3.



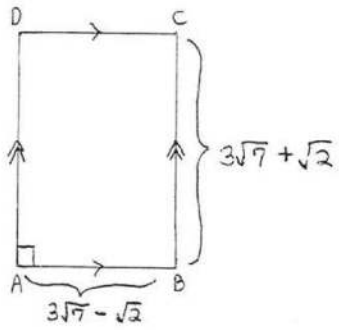
4.



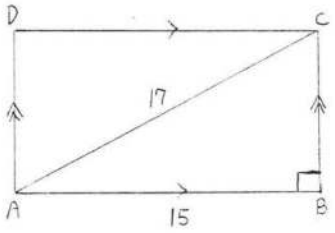
5.



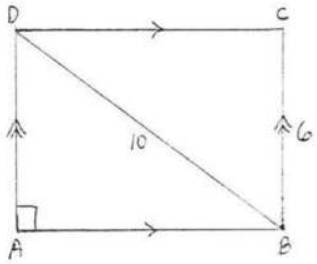
6.



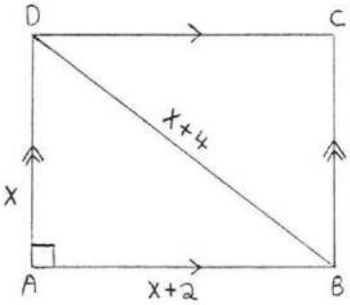
7.



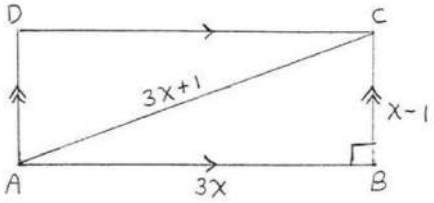
8.



9.

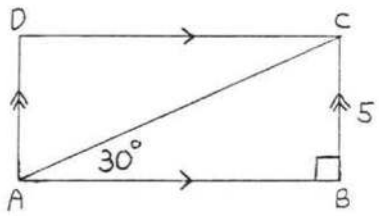


10.

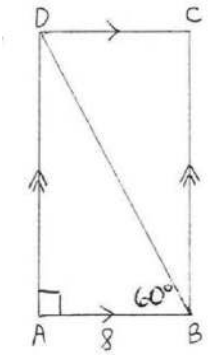


https://stats.libretexts.org/@go/page/24854

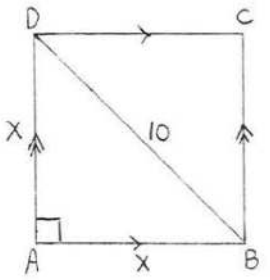
11.



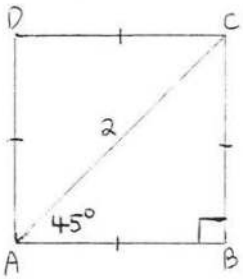
12.



13.

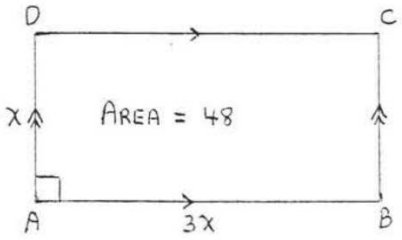


14.

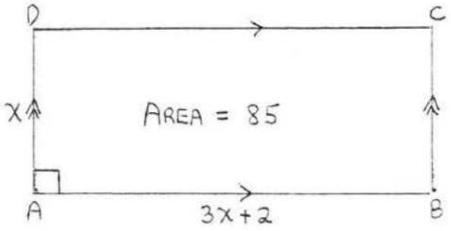


15 - 18. Find x:

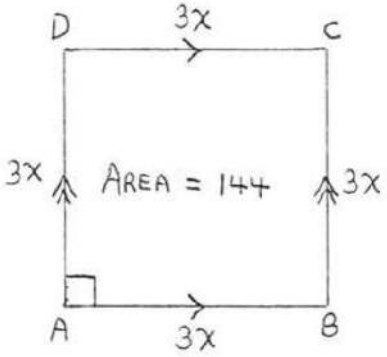
15.



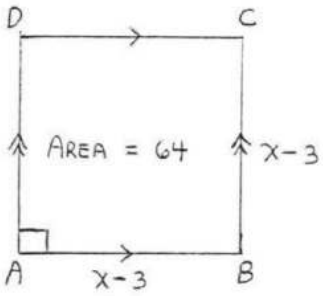
16.



17.



18.

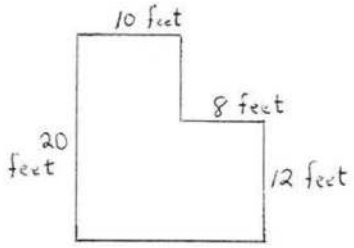


19. A football field has length 300 feet and width 160 feet. What is the area?

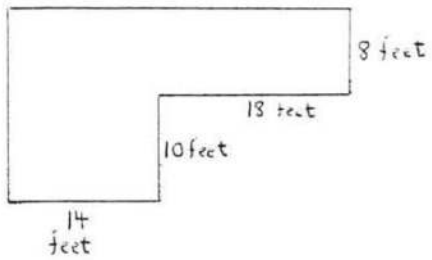
20. A tennis court is 78 feet long and 36 feet wide, What is the area?

21 - 24. How many one by one foot tiles are needed to tile each of the following rooms?

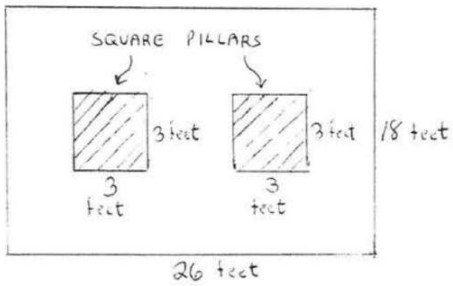
21.



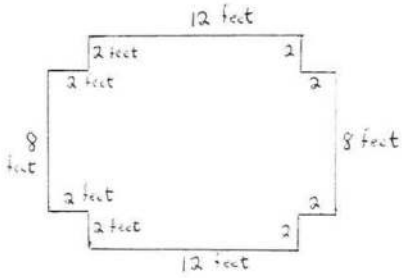
22.



23.



24.



25. A concrete slab weighs 60 pounds per square foot, What is the total weight of a rectangular slab 10 feet long and 3 feet wide?

26. A rectangular piece of plywood is 8 ey 10 feet, If the plywood weighs 3 pounds per square foot, what is the weight of the whole piece?

This page titled 7.2: The Area of a Rectangle and Square is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Henry Africk (New York City College of Technology at CUNY Academic Works) .

- **6.1: The Area of a Rectangle and Square** by Henry Africk is licensed CC BY-NC-SA 4.0. Original source: https://academicworks.cuny.edu/ny_oers/44.

7.3: The Area of a Triangle

For each of the triangles in Figure 7.3.1, side AB is called the base and CD is called the **height** or **altitude** drawn to this base. The base can be any state of the triangle though it is usually chosen to be the side on which the triangle appears to be resting. The height is the line drawn perpendicular to the base from the opposite vertex. Note that the height may fall outside the triangle, If the triangle is obtuse, and that the height may be one of the legs, if the triangle is a right triangle.

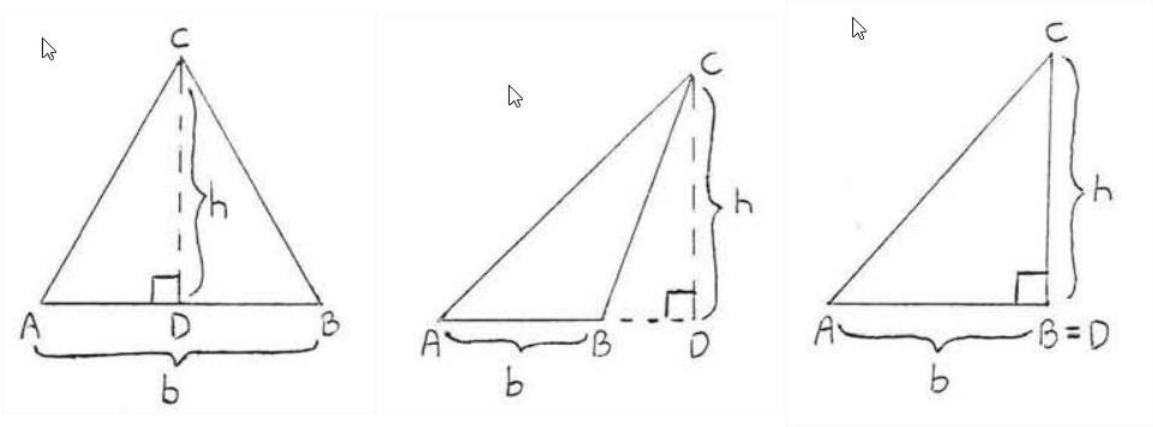


Figure 7.3.1: Triangles with base b and height h.

Theorem 7.3.1

The area of a triangle is equal to one-half of its base times its height.

$$A = \frac{1}{2}bh \tag{7.3.1}$$

Proof

For each of the triangles illustrated in Figure 7.3.1, draw AE and CE so that $ABCE$ is a parallelogram (Figure $PageIndex2$). $\triangle ABC \cong \triangle CEA$ so area of $\triangle ABC =$ area of $\triangle CEA$. Therefore area of $\triangle ABC = \frac{1}{2}$ area of parallelogram $ABCE = \frac{1}{2}bh$.

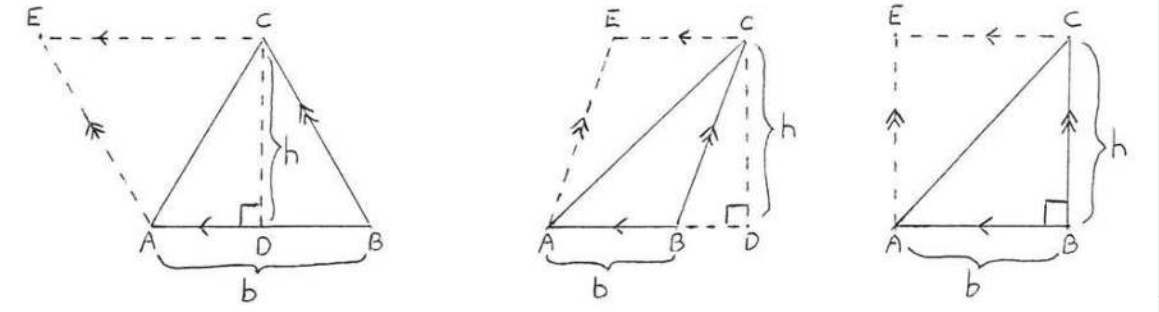
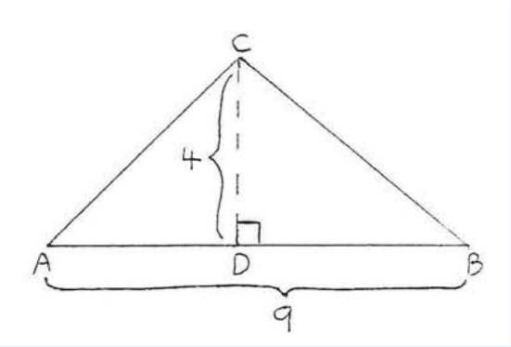


Figure 7.3.2 Draw AE and CE so that $ABCE$ is a parallelogram.

Example 7.3.1

Find the area:



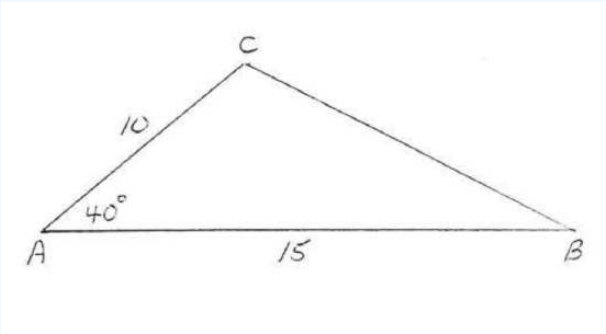
Solution

$A = \frac{1}{2}bh = \frac{1}{2}(9)(4) = \frac{1}{2}(36) = 18.$

Answer: 18.

Example 7.3.2

Find the area to the nearest tenth:



Solution

Draw the height h as shown in Figure 7.3.3

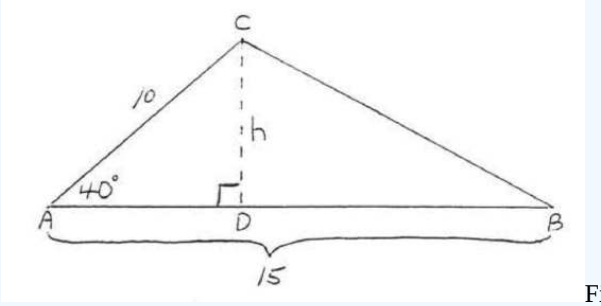


Figure 7.3.3 Draw height h.

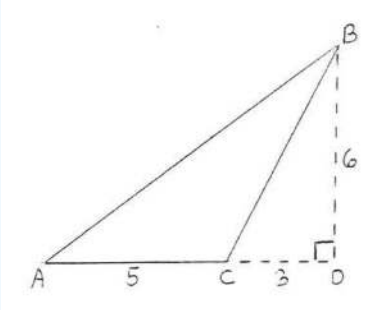
$$\begin{aligned} \sin 40^\circ &= \frac{h}{10} \\ .6428 &= \frac{h}{10} \\ (10)(.6428) &= \frac{h}{10}(10) \\ 6.428 &= h \end{aligned} \tag{7.3.2}$$

Area = $\frac{1}{2}bh = \frac{1}{2}(15)(6.428) = \frac{1}{2}(96.420) = 48.21 = 48.2$

Answer: $A = 48.2$

Example 7.3.3

Find the area and perimeter:



Solution

$A = \frac{1}{2}bh = \frac{1}{2}(5)(6) = \frac{1}{2}(30) = 15.$

To find AB and BC we use the Pythagorean theorem:

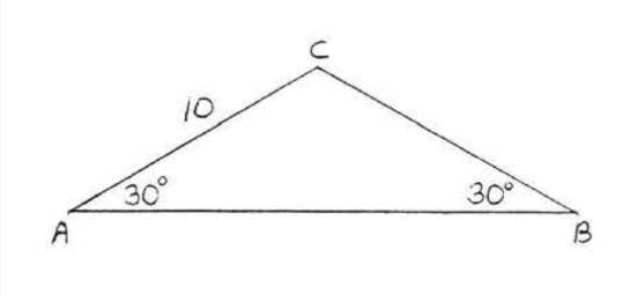
$$\begin{aligned} \mathrm{AD}^2 + \mathrm{BD}^2 &= \mathrm{AB}^2 & \mathrm{CD}^2 + \mathrm{BD}^2 &= \mathrm{BC}^2 \\ 8^2 + 6^2 &= \mathrm{AB}^2 & 3^2 + 6^2 &= \mathrm{BC}^2 \\ 64 + 36 &= \mathrm{AB}^2 & 9 + 36 &= \mathrm{BC}^2 \\ 100 &= \mathrm{AB}^2 & 45 &= \mathrm{BC}^2 \\ 10 &= \mathrm{AB} & \mathrm{BC} = \sqrt{45} &= \sqrt{9}\sqrt{5} = 3\sqrt{5} \end{aligned}$$

Perimeter = $AB + AC + BC = 10 + 5 + 3\sqrt{5} = 15 + 3\sqrt{5}$

Answer: $A = 15, P = 15 + 3\sqrt{5}$.

Example 7.3.4

Find the area and perimeter:



Solution

$\angle A = \angle B = 30^\circ$ so $\triangle ABC$ is isosceles with $BC = AC = 10$. Draw height h as in Figure 7.3.4.

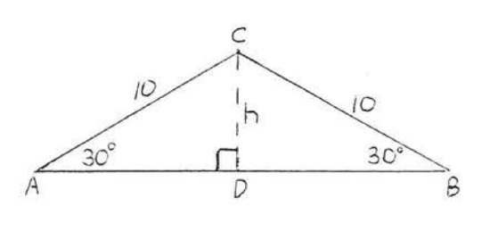


Figure 7.3.4 Draw height h.

$\triangle ACD$ is a $30^\circ - 60^\circ - 90^\circ$ triangle hence

$$\begin{aligned} \text{hypotenuse} &= 2(\text{short leg}) \\ 10 &= 2h \\ 5 &= h \\ \text{long leg} &= (\text{short leg})(\sqrt{3}) \\ AD &= h\sqrt{3} = 5\sqrt{3}. \end{aligned} \tag{7.3.3}$$

Similarly $BD = 5\sqrt{3}$.

Area = $\frac{1}{2}bh = \frac{1}{2}(5\sqrt{3}+5\sqrt{3})(5) = \frac{1}{2}(10\sqrt{3})(5) = \frac{1}{2}(50\sqrt{3}) = 25\sqrt{3}$.

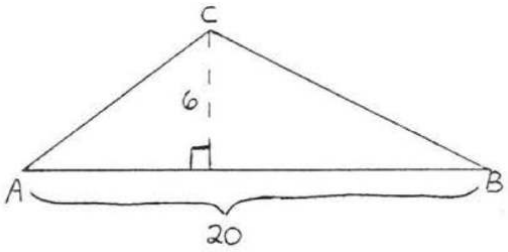
Perimeter = $10+10+5\sqrt{3}+5\sqrt{3} = 20+10\sqrt{3}$.

Answer: $A = 25\sqrt{3}, P = 20+10\sqrt{3}$.

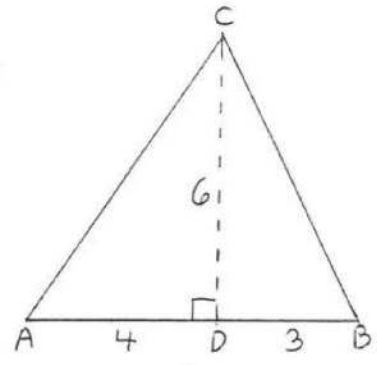
Problems

1 - 4. Find the area of $\triangle ABC$:

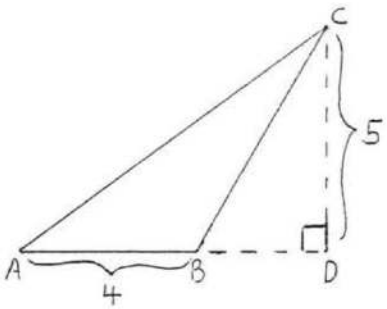
1.



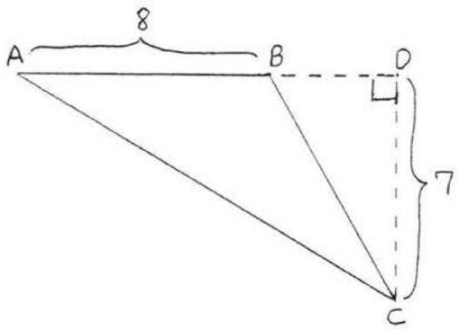
2.



3.

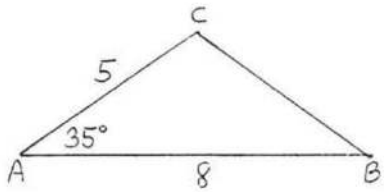


4.

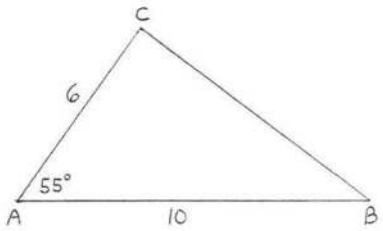


5 - 6. Find the area to the nearest tenth:

5.

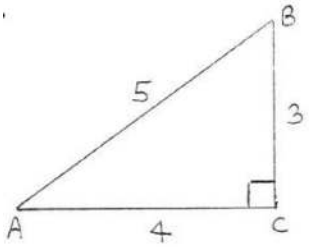


6.

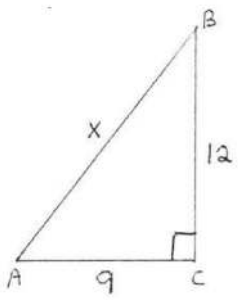


7 - 20. Find the area and perimeter of $\triangle ABC$:

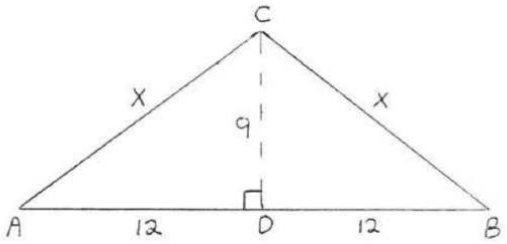
7.



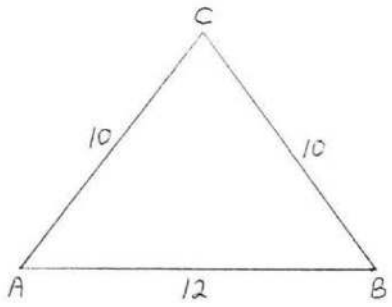
8.



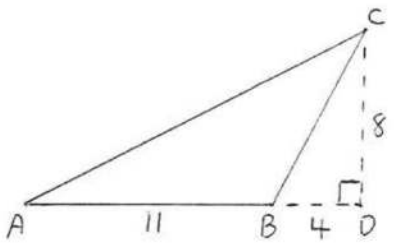
9.



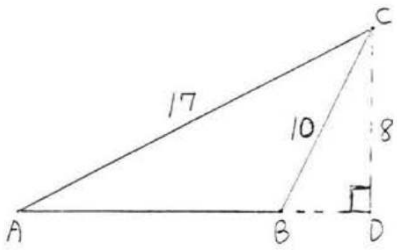
10.



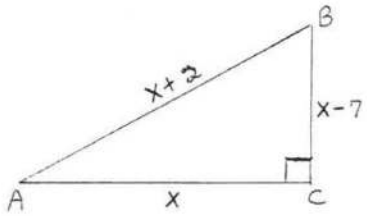
11.



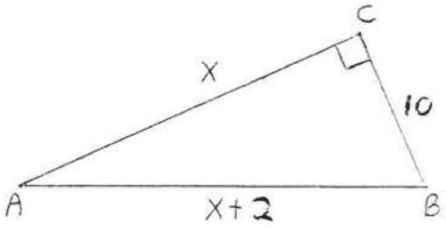
12.



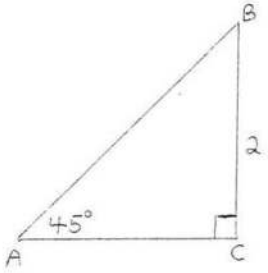
13.



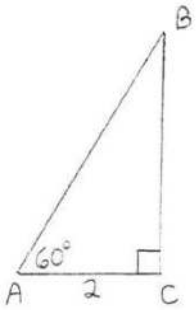
14.



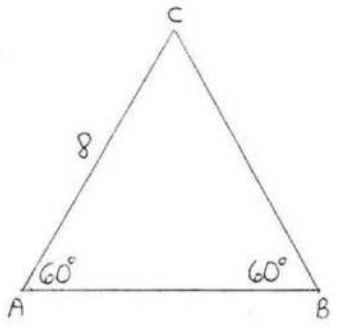
15.



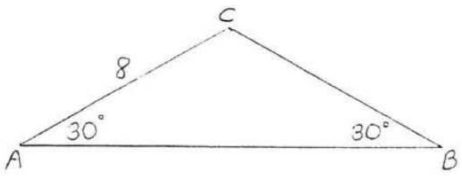
16.



17.

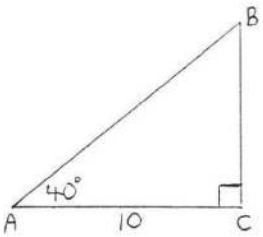


18.

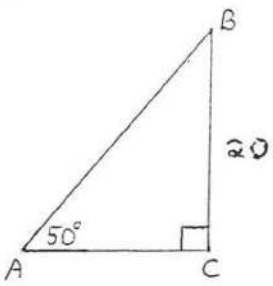


19 - 20. Find the area and perimeter to the nearest tenth:

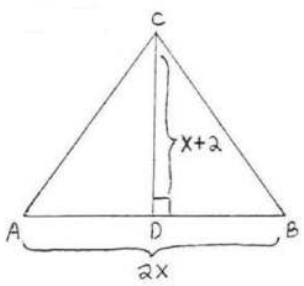
19.



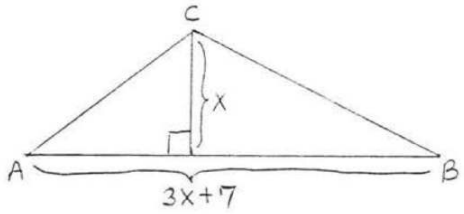
20.



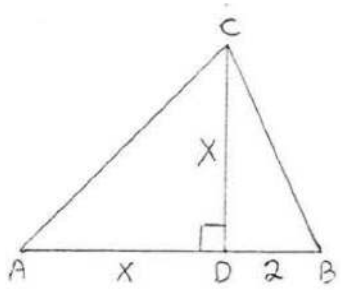
21. Find x if the area of $\triangle ABC$ is 35:



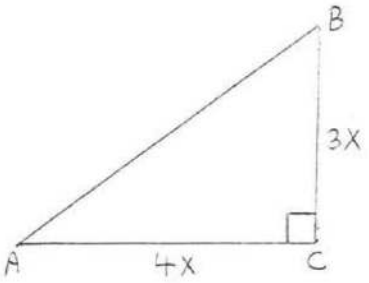
22. Find x if the area of $\triangle ABC$ is 24.



23. Find x if the area of $\triangle ABC$ is 12:



24. Find x if the area of $\triangle ABC$ is 108:



This page titled 7.3: The Area of a Triangle is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Henry Africk (New York City College of Technology at CUNY Academic Works) .

- **6.3: The Area of a Triangle** by Henry Africk is licensed CC BY-NC-SA 4.0. Original source: https://academicworks.cuny.edu/ny_oers/44.

7.4: Pythagorean Theorem

In a right triangle, the sides of the right angle are called the **legs** of the triangle and the remaining side is called the **hypotenuse**. In Figure 7.4.1, side AC and BC are the legs and side AB is the hypotenuse.

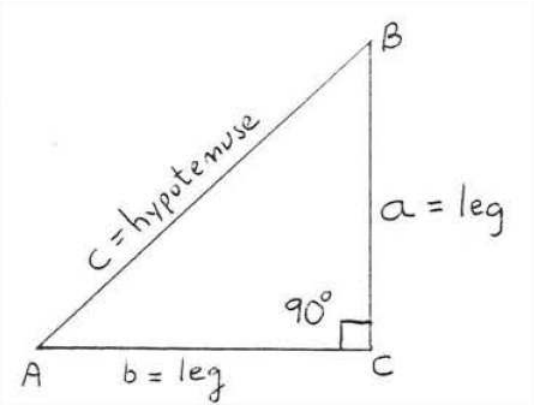


Figure 7.4.1: A right triangle.

The following is one of the most famous theorems in mathematics.

Theorem 7.4.1: Pythagorean Theorem

In a right triangle, the square of the hypotenuse is equal to the sum of the squares of the legs. That is,

$$\text{leg}^2 + \text{leg}^2 = \text{hypotenuse}^2 \tag{7.4.1}$$

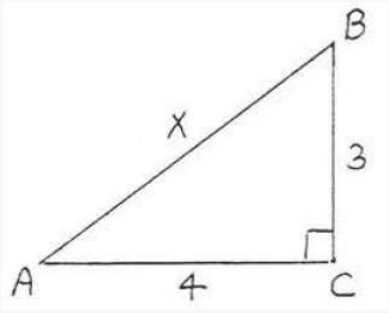
Thus, for the sides of the triangle in Figure 7.4.1,

$$a^2 + b^2 = c^2$$

Before we prove Theorem 7.4.1, we will give several examples.

Example 7.4.1

Find x



Solution

$$\begin{aligned} \text{leg}^2 + \text{leg}^2 &= \text{hyp}^2 \\ 3^2 + 4^2 &= x^2 \\ 9 + 16 &= x^2 \\ 25 &= x^2 \\ 5 &= x \end{aligned}$$

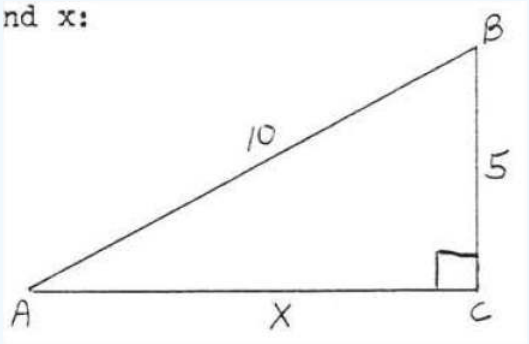
Check:

$$\begin{array}{rcl} \text{leg}^2 + \text{leg}^2 & = & \text{hyp}^2 \\ 3^2 + 4^2 & | & x^2 \\ 9 + 16 & | & 5^2 \\ 25 & | & 25 \end{array}$$

Answer: $x = 5$.

Example 7.4.2

Find x:



Solution

$$\begin{aligned} \text{leg}^2 + \text{leg}^2 &= \text{hyp}^2 \\ 5^2 + x^2 &= 10^2 \\ 25 + x^2 &= 100 \\ x^2 &= 75 \\ x &= \sqrt{75} = \sqrt{25}\sqrt{3} = 5\sqrt{3} \end{aligned}$$

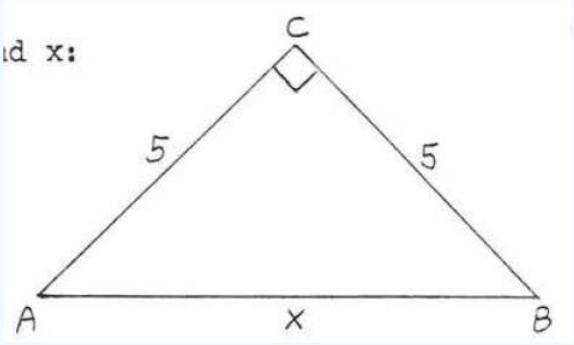
Check:

$$\begin{array}{rcl} \text{leg}^2 + \text{leg}^2 & = & \text{hyp}^2 \\ 5^2 + x^2 & | & 10^2 \\ 25 + (5\sqrt{3})^2 & | & 100 \\ 25 + 25\sqrt{9} & | & \\ 25 + 25(3) & | & \\ 25 + 75 & | & \\ 100 & | & \end{array}$$

Answer: $x = 5\sqrt{3}$.

Example 7.4.3

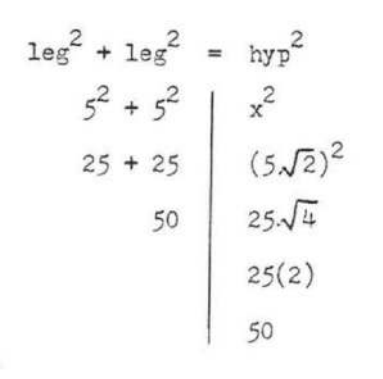
Find x:



Solution

$$\begin{aligned} \text{leg}^2 + \text{leg}^2 &= \text{hyp}^2 \\ 5^2 + 5^2 &= x^2 \\ 25 + 25 &= x^2 \\ 50 &= x^2 \\ x &= \sqrt{50} = \sqrt{25}\sqrt{2} = 5\sqrt{2} \end{aligned}$$

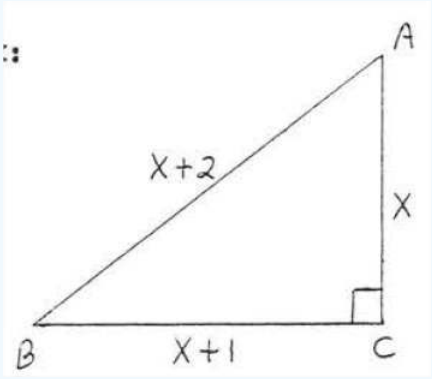
Check:



Answer: $x = 5\sqrt{2}$.

Example 7.4.4

Find x



Solution

$$\begin{aligned}
\text{leg}^2+\text{leg}^2 &= \text{hyp}^2\\
x^2+(x+1)^2 &= (x+2)^2\\
x^2+x^2+2x+1 &= x^2+4x+4\\
x^2+x^2+2x+1-x^2-4x-4 &= 0\\
x^2-2x-3 &= 0\\
(x-3)(x+1) &= 0\\
x-3=0 \quad & \quad x+1=0\\
x=3 \quad & \quad x=-1
\end{aligned}$$

We reject $x=-1$ because $AC=x$ cannot be negative.

Check, $x=3$:

$$\begin{array}{l|l}
\text{leg}^2+\text{leg}^2=\text{hyp}^2 & \\
x^2+(x+1)^2 & (x+2)^2\\
3^2+(3+1)^2 & (3+2)^2\\
9+4^2 & 5^2\\
9+16 & 25\\
25 &
\end{array}$$

Answer: $x=3$.

We will now restate and prove Theorem 7.4.1:

Theorem 7.4.1 Pythagorean Theorem

In a right triangle, the square of the hypotenuse is equal to the sum of the squares of the legs. That is,

$\text{leg}^2+\text{leg}^2=\text{hypotenuse}^2$.

In Figure 7.4.1,

$a^2+b^2=c^2$.

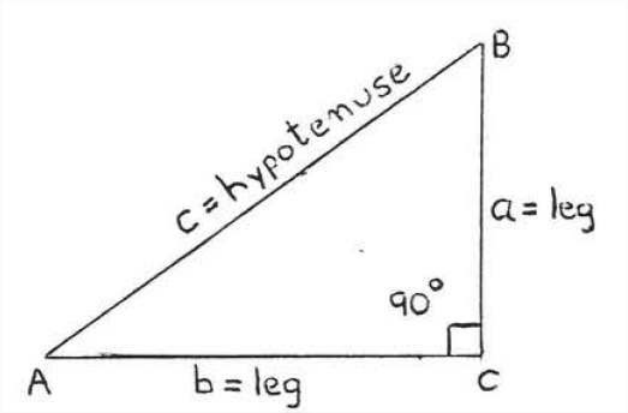


Figure 7.4.1 A right triangle.

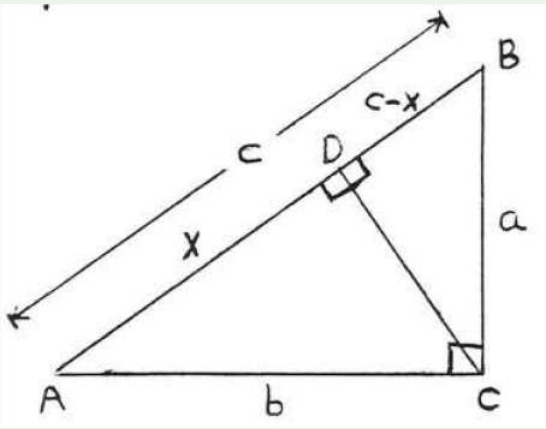
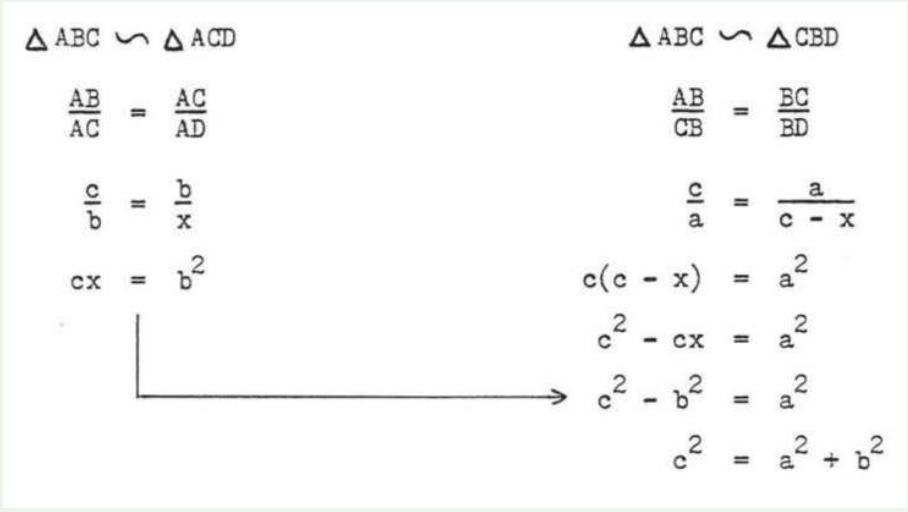


Figure 7.4.2 Draw CD perpendicular to AB.

Proof

In Figure 7.4.1, draw CD perpendicular to AB. Let $x = AD$. Then $BD = c - x$ (Figure 7.4.2). As in Example 7.4.3, section 4.2, $\triangle ABC \sim \triangle ACD$ and $\triangle ABC \sim \triangle CBD$. From these two similarities we obtain two proportions:



The converse of the Pythagorean Theorem also holds:

Theorem 7.4.2 (converse of the Pythagorean Theorem).

In a triangle, if the square of one side is equal to the sun of the squares of the other two sides then the triangle is a right triangle.

In Figure 7.4.3, if $c^2 = a^2 + b^2$ then $\triangle ABC$ is a right triangle with $\angle C = 90°$.

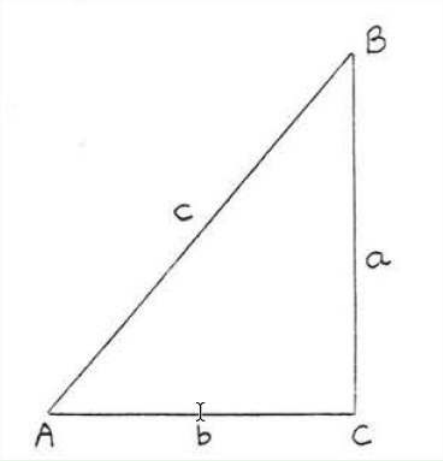


Figure 7.4.3 If $c^2 = a^2 + b^2$ then $\angle C = 90°$.

Proof

Draw a new triangle, $\triangle DEF$, so that $\angle F = 90°$, $d = a$, and $e = b$ (Figure 7.4.4). $\triangle DEF$ is a right triangle, so by Theorem 7.4.1, $f^2 = d^2 + e^2$. We have $f^2 = d^2 + e^2 = a^2 + b^2 = c^2$ and therefore $f = c$. Therefore $\triangle ABC \cong \triangle DEF$ because $SSS = SSS$. Therefore, $\angle C + \angle F = 90°$.

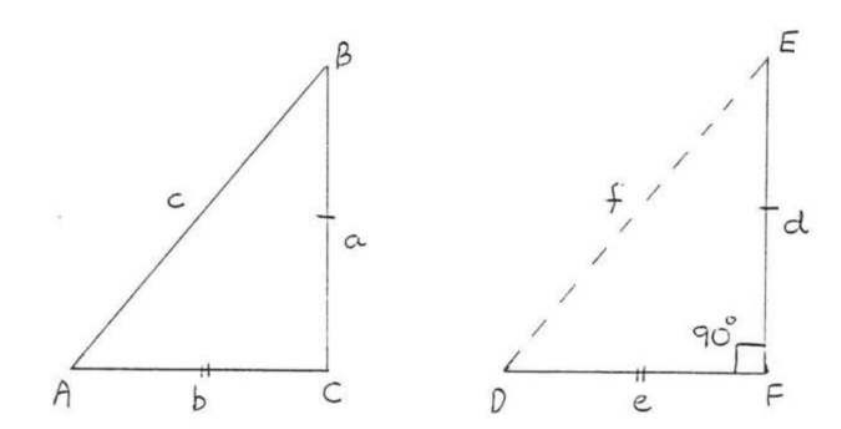
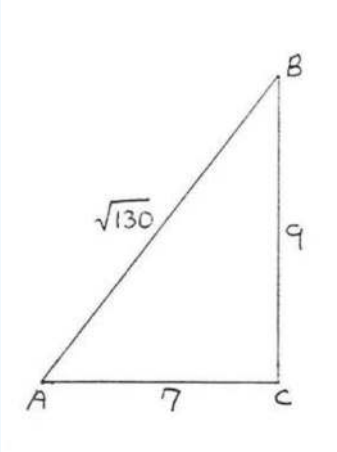


Figure 7.4.4 Given $\triangle ABC$, draw $\triangle DEF$ so that $\angle F = 90°$, $d = a$ and $e = b$.

Example 7.4.5

Is $\triangle ABC$ a right triangle?



https://stats.libretexts.org/@go/page/24856

Solution

$\mathrm{AC}^2 = 7^2 = 49$

$\mathrm{BC}^2 = 9^2 = 81$

$\mathrm{AB}^2 = (\sqrt{130})^2 = 130$

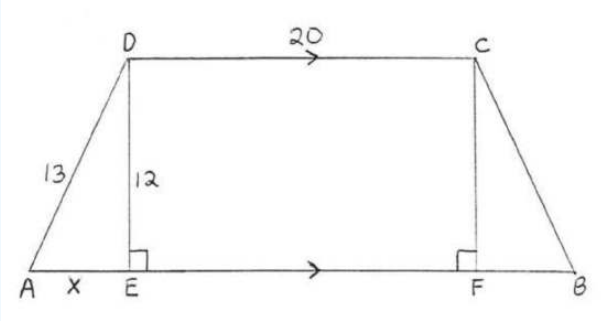
$49 + 81 = 130.$

so by Theorem $7.4.2$, $\triangle ABC$ is a right triangle.

Answer: yes.

Example 7.4.6

Find x and AB:



Solution

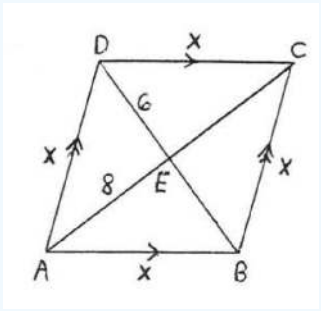
$$\begin{aligned} x^2 + 12^2 &= 13^2 \\ x^2 + 144 &= 169 \\ x^2 &= 169 - 144 \\ x^2 &= 25 \\ x &= 5 \end{aligned}$$

$CDEF$ is a rectangle so $EF = CD = 20$ and $CF = DE = 12$. Therefore $FB = 5$ and $AB = AE + EF + FB = 5 + 20 + 5 = 30$.

Answer: $x = 5$, $AB = 30$.

Example 7.4.7

Find x, AC and BD:



Solution

$ABCD$ is a rhombus. The diagonals of a rhombus are perpendicular and bisect each other.

https://stats.libretexts.org/@go/page/24856

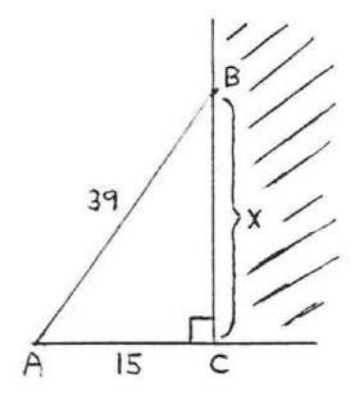
$$\begin{aligned} 6^2+8^2 &= x^2 \\ 36+64 &= x^2 \\ 100 &= x^2 \\ 10 &= x \end{aligned}$$

$AC = 8+8 = 16, BD = 6+6 = 12.$

Answer: $x = 10, AC = 16, BD = 12.$

Example 7.4.8

A ladder 39 feet long leans against a building, How far up the side of the building does the ladder reach if the foot of the ladder is 15 feet from the building?



Solution

$$\begin{aligned} \text{leg}^2+\text{leg}^2 &= \text{hyp}^2 \\ x^2+15^2 &= 39^2 \\ x^2+225 &= 1521 \\ x^2 &= 1521-225 \\ x^2 &= 1296 \\ x &= \sqrt{1296} = 36 \end{aligned}$$

Answer: 36 feet.

Historical Note

Pythagoras (c. 582 - 507 B.C.) was not the first to discover the theorem which bears his name. It was known long before his time by the Chinese, the Babylonians, and perhaps also the Egyptians and the Hindus, According to tradition, Pythagoras was the first to give a nroof of the theorem, His proof probably made use of areas, like the one suggested. In Figure 7.4.5 below, (each square contains four congruent right triangles with sides of lengths a, b, and c, In addition the square on the left contains a square with side a and a square with side b while the one on the right contains a square with side c.)

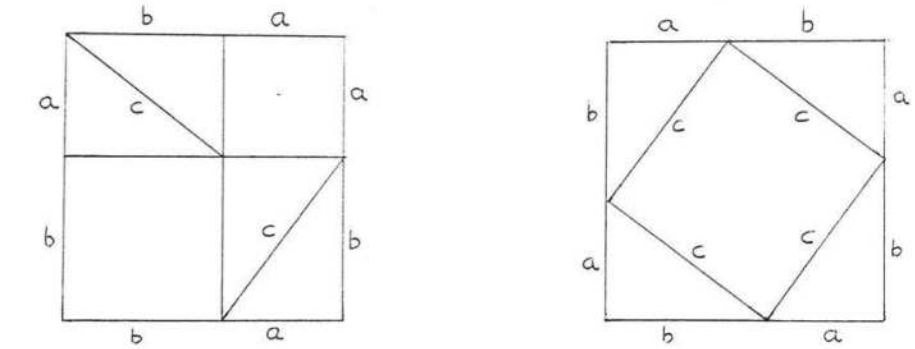


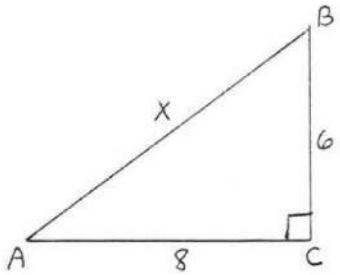
Figure 7.4.5 Pythagoras may have proved $a^2 + b^2 = c^2$ in this way. Since the time of Pythagoras, at least several hundred different proofs of the Pythagorean Theorem have been proposed, Pythagoras was the founder of the Pythagorean school, a secret religious society devoted to the study of philosophy, mathematics, and science. Its membership was a select group, which tended to keep the discoveries and practices of the society secret from outsiders. The Pythagoreans believed that numbers were the ultimate components of the universe and that all physical relationships could be expressed with whole numbers, This belief was prompted in part by their discovery that the notes of the musical scale were related by numerical ratios. The Pythagoreans made important contributions to medicine, physics, and astronomy, In geometry, they are credited with the angle s

um theorem for triangles, the properties of parallel lines, and the theory of similar triangles and proportions.

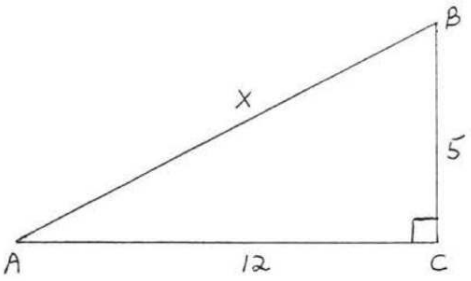
Problems

1 - 10. Find x. Leave answers in simplest radical form.

1.

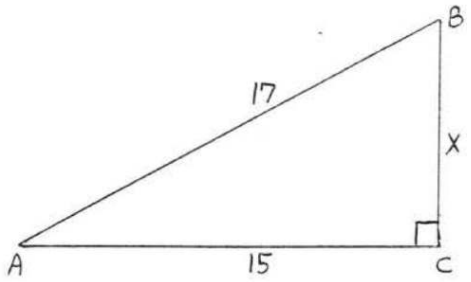


2.

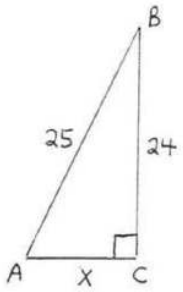


3.

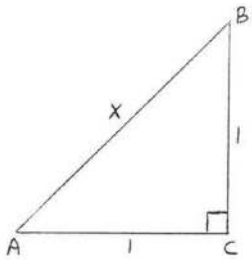
https://stats.libretexts.org/@go/page/24856



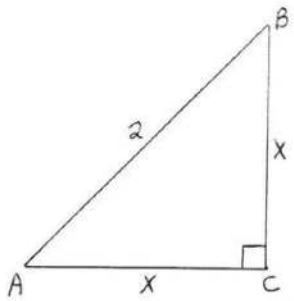
4.



5.

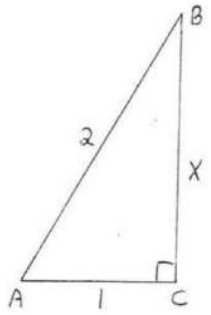


6.

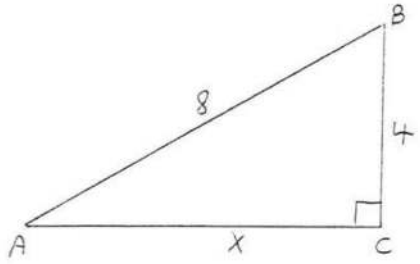


7.

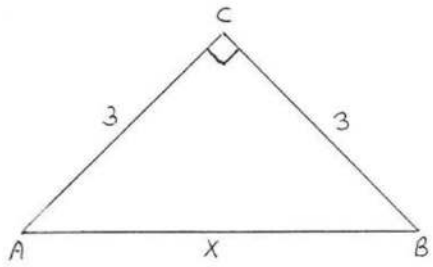
https://stats.libretexts.org/@go/page/24856



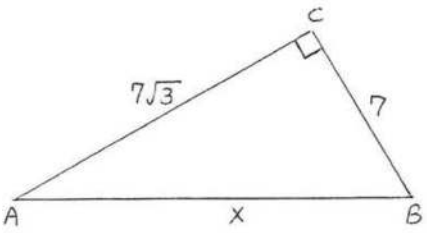
8.



9.



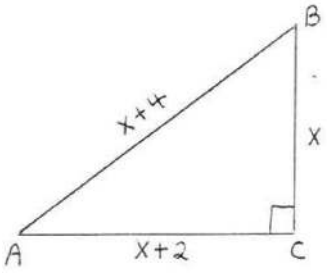
10.



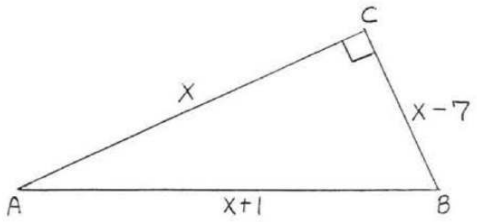
11 - 14. Find x and all sides of the triangle:

11.

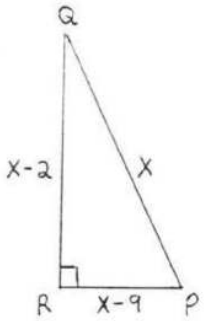
https://stats.libretexts.org/@go/page/24856



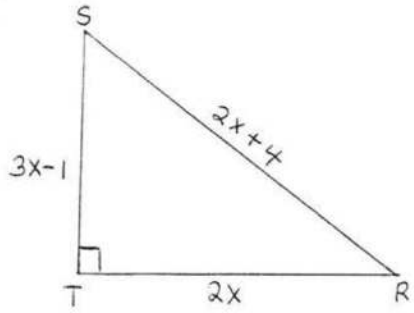
12.



13.



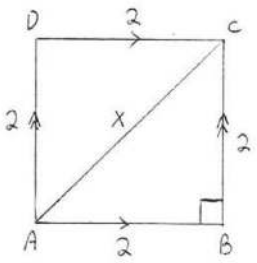
14.



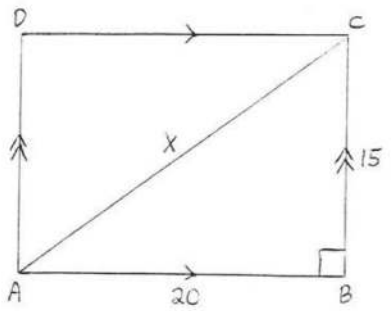
15 - 16. Find x:

15.

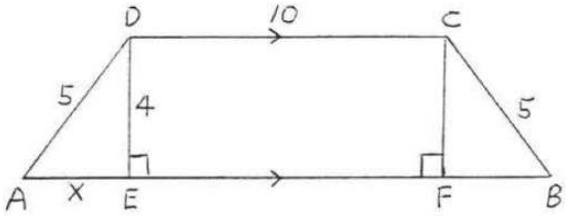
https://stats.libretexts.org/@go/page/24856



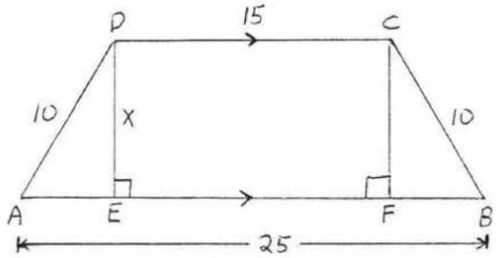
16.



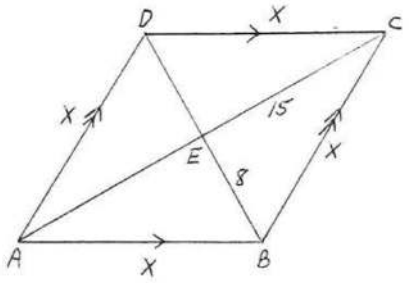
17. Find x and AB.



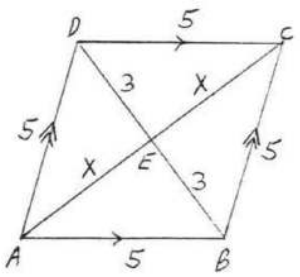
18. Find x:



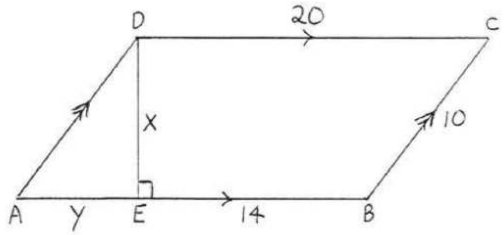
19. Find x, AC and BD:



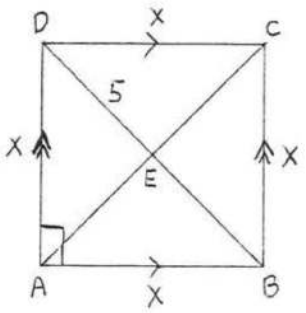
20. Find x, AC and BD:



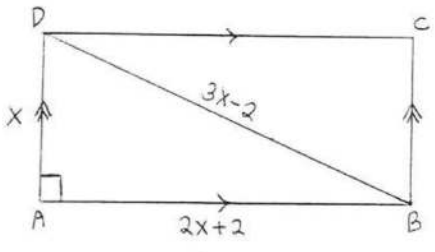
21. Find x and y:



22. Find x, AC and BD:

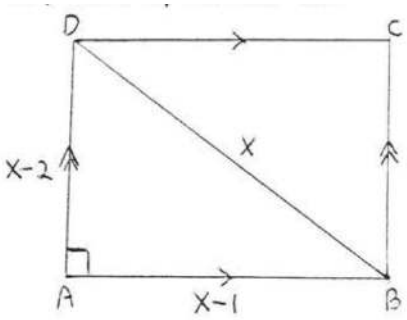


23. Find x, AB and BD:



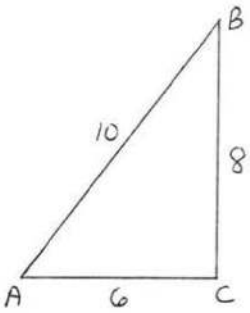
24. Find x, AB and AD:

https://stats.libretexts.org/@go/page/24856

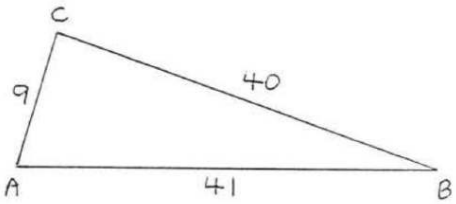


25 - 30. Is $\triangle ABC$ a right triangle?

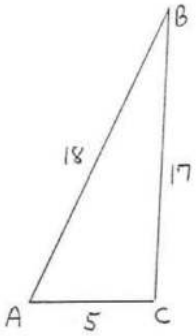
25.



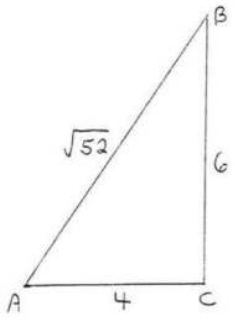
26.



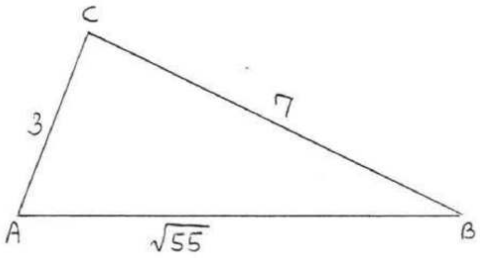
27.



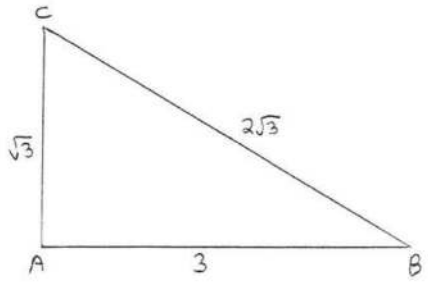
28.



29.



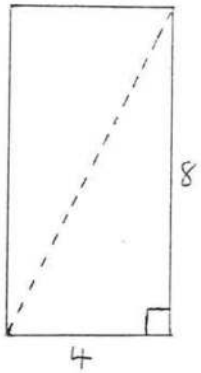
30.



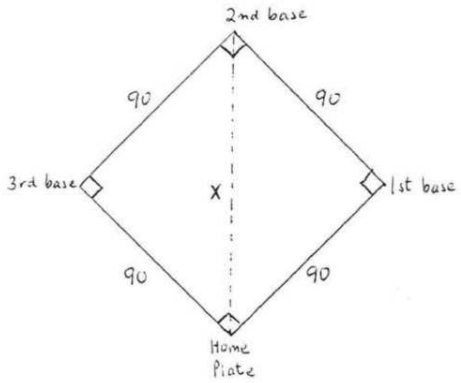
31. A ladder 25 feet long leans against a building, How far up the side of the building does the ladder reach if the foot of the ladder is 7 feet from the building?

32. A man travels 24 miles east and then 10 miles north. At the end of his journey how far is he from his starting point?

33. Can a table 9 feet wide (with its legs folded) fit through a rectangular doorway 4 feet by 8 feet?



34. A baseball diamond is a square 90 feet on each side, Find the distance from home plate to second base (leave answer in simplest radical form).



This page titled 7.4: Pythagorean Theorem is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Henry Africk (New York City College of Technology at CUNY Academic Works) .

- **4.4: Pythagorean Theorem** by Henry Africk is licensed CC BY-NC-SA 4.0. Original source: https://academicworks.cuny.edu/ny_oers/44.

CHAPTER OVERVIEW

8: Sampling and Data

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

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8: Sampling and Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.1: Introduction

Skills to Develop

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

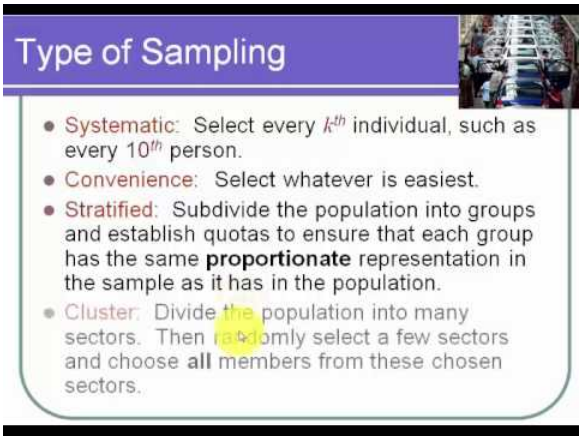
- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

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

Figure 8.1.1: We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

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

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



Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.2: Definitions of Statistics, Probability, and Key Terms

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

COLLABORATIVE EXERCISE

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5; 5.5; 6; 6; 6; 6.5; 6.5; 6.5; 6.5; 7; 7; 8; 8; 9

The dot plot for this data would be as follows:

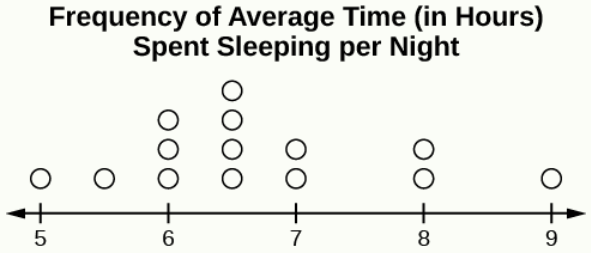


Figure 8.2.1

Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

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

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

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

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

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example 8.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Answer

- The **population** is all first year students attending ABC College this term.
- The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (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 ABC College 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 ABC College.
- The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Exercise 8.2.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Answer

- The **population** is all families with children attending Knoll Academy.
- The **sample** is a random selection of 100 families with children attending Knoll Academy.
- The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.
- The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.
- The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.
- The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Example 8.2.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1._____ Population 2._____ Statistic 3._____ Parameter 4._____ Sample 5._____ Variable 6._____ Data

a. all students who attended the college last year
b. the cumulative GPA of one student who graduated from the college last year
c. 3.65, 2.80, 1.50, 3.90
d. a group of students who graduated from the college last year, randomly selected
e. the average cumulative GPA of students who graduated from the college last year
f. all students who graduated from the college last year
g. the average cumulative GPA of students in the study who graduated from the college last year

Answer

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

Example 8.2.3

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of "drive" (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Answer

- The **population** is all cars containing dummies in the front seat.
- The **sample** is the 75 cars, selected by a simple random sample.

- The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.
- The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.
- The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.
- The **data** are either: yes, had head injury, or no, did not.

Example 8.2.4

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Answer

- The **population** is all medical doctors listed in the professional directory.
- The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.
- The **sample** is the 500 doctors selected at random from the professional directory.
- The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.
- The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.
- The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

COLLABORATIVE EXERCISE

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

1. The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories...stDummies.html (accessed May 1, 2013).

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:

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

Researcher B:

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

Exercise 1.2.2

population

Answer

AIDS patients.

Exercise 1.2.3

sample

Exercise 1.2.4

parameter

Answer

The average length of time (in months) AIDS patients live after treatment.

Exercise 1.2.5

statistic

Exercise 1.2.6

variable

Answer

$X =$ the length of time (in months) AIDS patients live after treatment

Glossary

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Average

also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative**(an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.2: Definitions of Statistics, Probability, and Key Terms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.3: Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of *counting* or *measuring* attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either *discrete* or *continuous*.

All data that are the result of counting are called *quantitative discrete data*. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are *quantitative continuous data* assuming that we can measure accurately. Measuring angles in radians might result in such numbers as $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, π, $\frac{3\pi}{4}$, and so on. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the **quantitative discrete data**.

Exercise 8.3.1

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Answer

quantitative discrete data

Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are **quantitative continuous data** because weights are measured.

Exercise 8.3.2

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Answer

quantitative continuous data

Exercise 8.3.3

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

Sample of qualitative data

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are **qualitative data**.

Exercise 8.3.4

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Answer

qualitative data

COLLABORATIVE EXERCISE 8.3.1

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. where you go on vacation
d. the distance it is from your home to the nearest grocery store
e. the number of classes you take per school year.
f. the tuition for your classes
g. the type of calculator you use
h. movie ratings
i. political party preferences
j. weights of sumo wrestlers
k. amount of money (in dollars) won playing poker
l. number of correct answers on a quiz
m. peoples' attitudes toward the government
n. IQ scores (This may cause some discussion.)

Answer

Items a, e, f, k, and l are quantitative discrete; items d, j, and n are quantitative continuous; items b, c, g, h, i, and m are qualitative.

Exercise 8.3.5

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Answer

quantitative discrete

Exercise 8.3.6

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart Figure 8.3.1. What type of data does this graph show?

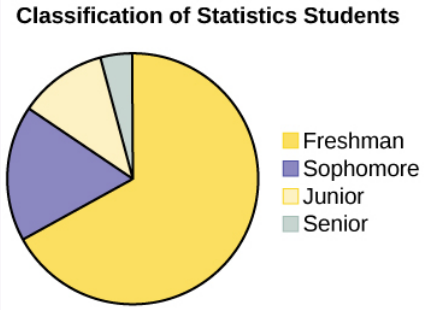


Figure 8.3.1

Answer

This pie chart shows the students in each year, which is **qualitative data**.

Exercise 8.3.7

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

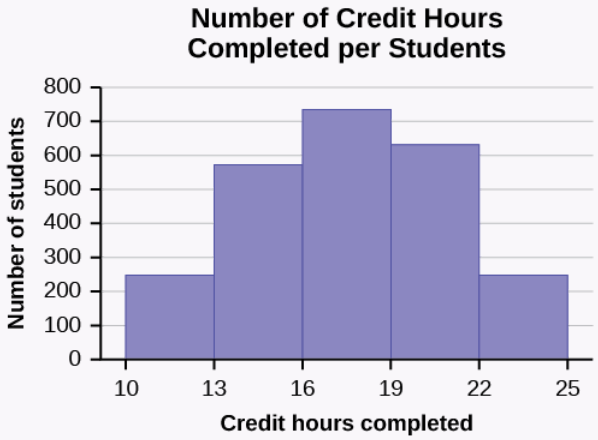


Figure 8.3.2

What type of data does this graph show?

Answer

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

Table 8.3.1*: Fall Term 2007 (Census day)*

De Anza College				Foothill College		
	Number	Percent			Number	Percent

De Anza College				Foothill College		
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

- In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.
- In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.
- A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at Figures 8.3.3 and 8.3.4 and determine which graph (pie or bar) you think displays the comparisons better.

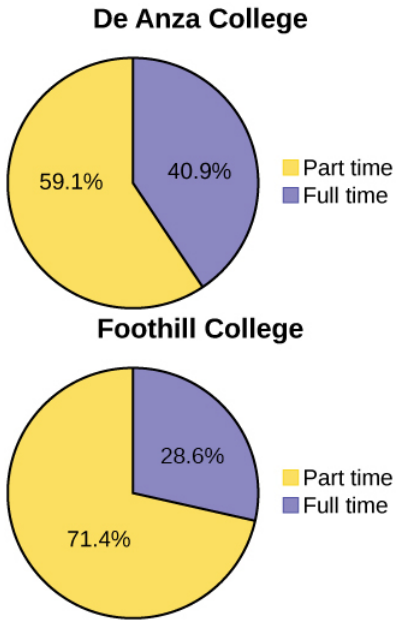


Figure 8.3.3: *Pie Charts*

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context. Our choice also depends on what we are using the data for.

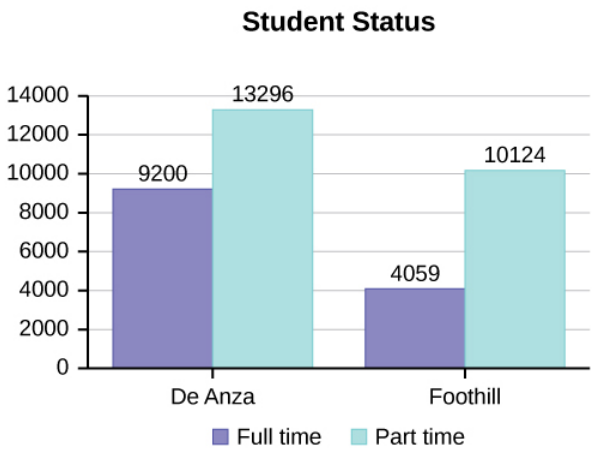


Figure 8.3.4: *Bar chart*

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Table 8.3.2: De Anza College Spring 2010

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

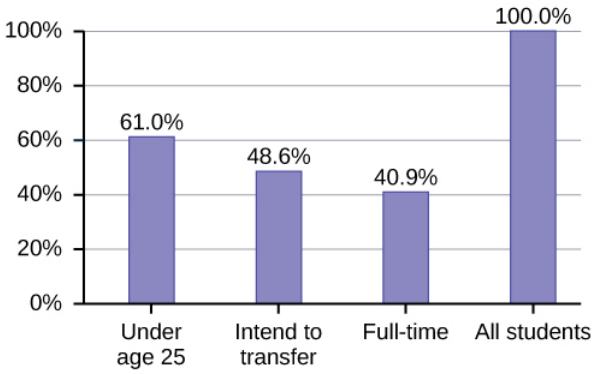


Figure 8.3.2: Bar chart of data in Table 8.3.2.

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

Table 8.3.2: 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%

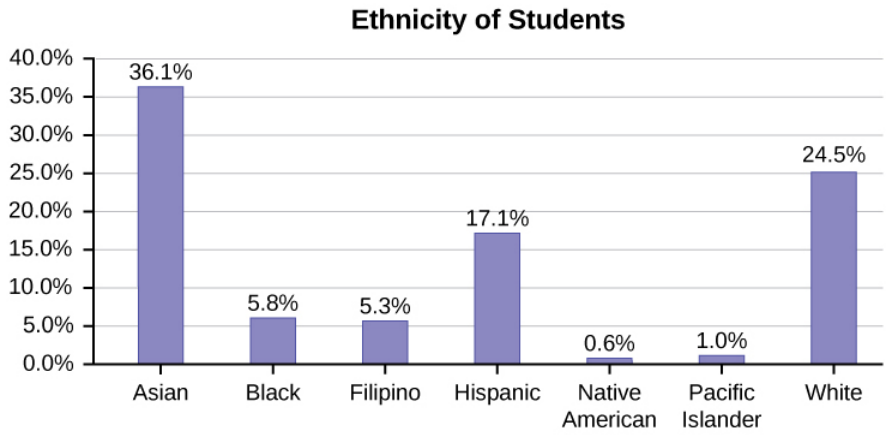


Figure 8.3.3: Enrollment of De Anza College (Spring 2010)

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 8.3.4 can be difficult to understand visually. The graph in Figure 8.3.5 is a **Pareto chart**. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

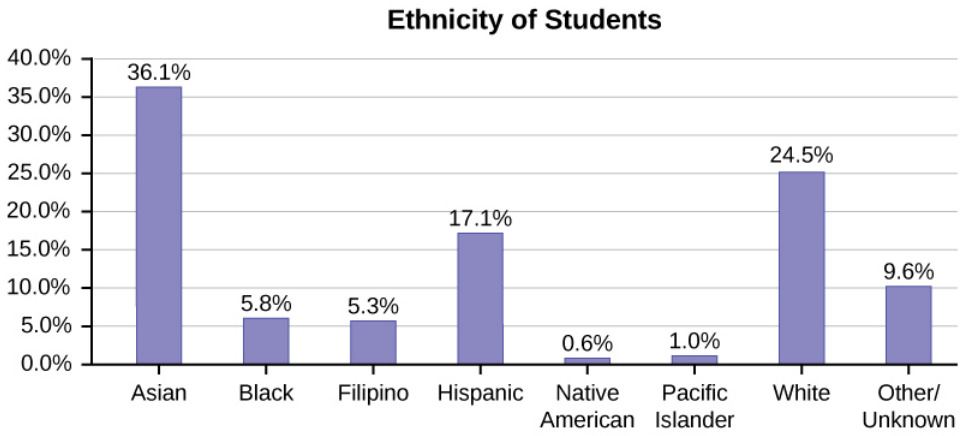


Figure* 8.3.4*: Bar Graph with Other/Unknown Category

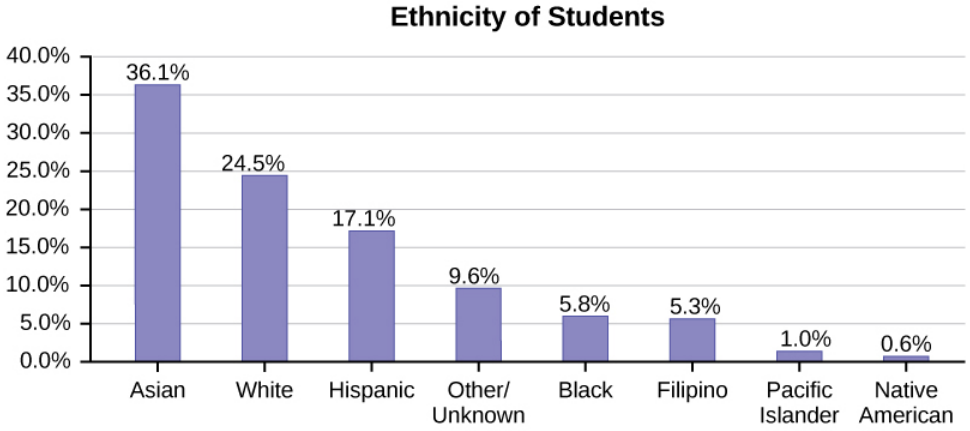


Figure* 8.3.5*: *Pareto Chart With Bars Sorted by Size*

Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The chart in Figure 8.3.6 is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in Figure 8.3.6.

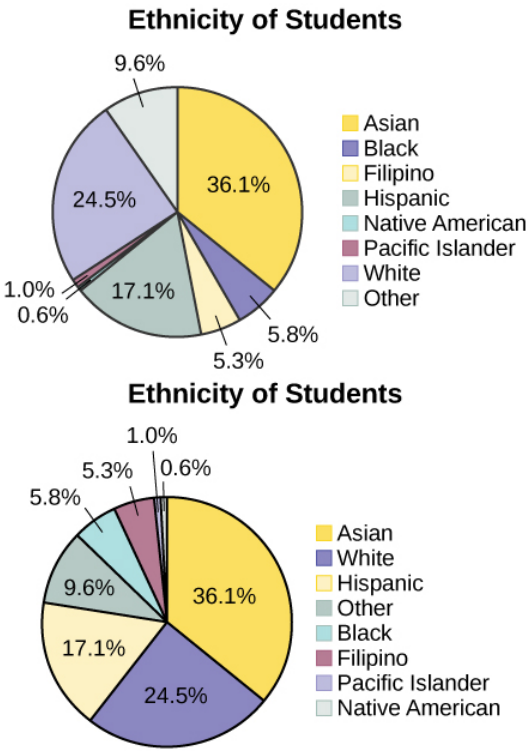


Figure 8.3.6.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen by any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in Table 1.3.2:

Table 8.3.3: Class Roster

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai

ID	Name	ID	Name	ID	Name
09	Jiao	20	Reyes	30	Wood
10	Khan				

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360; 0.99832; 0.14669; 0.51470; 0.40581; 0.73381; 0.04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, Cuningham, and Cuarismo.

The easiest way to generate a random number is to ask your smartphone, Google Assistant, or Alexa or for a random number. For example just say "Generate a random number between 4 and 15" and the technology will give you one.

Random Number Generator

Below is a simple random number generator:

Enter the lower bound and the upper bound for the random number and click in the Get Random button to generate a random number.

Low:	High:		Random Number:	
		Get Random		

To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.

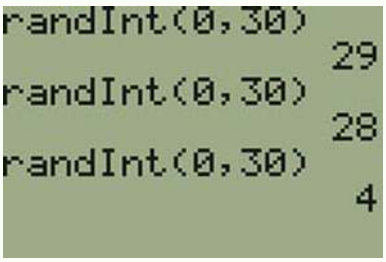


Figure 8.3.7

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Exercise 8.3.8

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. 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

Example 8.3.9: Calculator

You are going to use the random number generator to generate different types of samples from the data. This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

a. Create a stratified sample by column. Pick three quiz scores randomly from each column.

- Number each row one through ten.
- On your calculator, press Math and arrow over to PRB.
- For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
- Repeat for columns two through six.
- These 18 quiz scores are a stratified sample.

b. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.

- Press MATH and arrow over to PRB.
- Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
- The two numbers are for two of the columns.
- The quiz scores (20 of them) in these 2 columns are the cluster sample.

c. Create a simple random sample of 15 quiz scores.

- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

d. Create a systematic sample of 12 quiz scores.

- Use the numbering one through 60.
- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example 8.3.10

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Answer

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Exercise 8.3.11

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Answer

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 8.3.12: Sampling

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task. Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Answer

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Answer

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Answer

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Exercise 8.3.12

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Answer

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

COLLABORATIVE EXERCISE 8.3.8

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

a. To find the average GPA of all students in a university, use all honor students at the university as the sample.
b. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
c. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
d. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
e. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen’s and Jung’s samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This *variability in samples* cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

COLLABORATIVE EXERCISE 8.3.8

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in the following table (“frequency” is the number of times a particular face of the die occurs):

First Experiment (20 rolls)		Second Experiment (20 rolls)	
Face on Die	**Frequency**	**Face on Die**	**Frequency**

First Experiment (20 rolls)			Second Experiment (20 rolls)	
Face on Die	**Frequency**		**Face on Die**	**Frequency**
1				
2				
3				
4				
5				
6				

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

1. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).
2. Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).
3. Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/ga...questions.aspx (accessed May 1, 2013).
4. Data from http://www.bookofodds.com/Relationsh...-the-President
5. Dominic Lusinchi, "'President' Landon and the 1936 Literary Digest Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).
6. "The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).
7. "Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/ga...9362004.aspx#4 (accessed May 1, 2013).
8. The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

9. LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/f...hts.html#focus (accessed May 1, 2013).
10. Data from San Jose Mercury News

Chapter Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Footnotes

1. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
2. Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).
3. Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodol...wer-your-polls (accessed May 1, 2013).

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.3: Data, Sampling, and Variation in Data and Sampling is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.4: Frequency, Frequency Tables, and Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. Some examples are Sony, Motorola, Nokia, Samsung and Apple. This is just a list and there is no agreed upon order. Some people may favor Apple but that is a matter of opinion. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table lists the different data values in ascending order and their frequencies.

Table 8.4.1: Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Definition: relative frequency

A frequency is the number of times a value of the data occurs. According to Table Table 8.4.1, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

Definition: Relative frequencies

A *relative frequency* is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

Table 8.4.2: Frequency Table of Student Work Hours with Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

The sum of the values in the relative frequency column of Table 8.4.2 is $\frac{20}{20}$, or 1.

Definition: Cumulative relative frequency

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in Table 8.4.3.

Table 8.4.3: Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

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 each should be close to one.

Table 8.4.4 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

Table 8.4.4: Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = 0.05$	0.05
61.95–63.95	3	$\frac{3}{100} = 0.03$	$0.05 + 0.03 = 0.08$
63.95–65.95	15	$\frac{15}{100} = 0.15$	$0.08 + 0.15 = 0.23$
65.95–67.95	40	$\frac{40}{100} = 0.40$	$0.23 + 0.40 = 0.63$
67.95–69.95	17	$\frac{17}{100} = 0.17$	$0.63 + 0.17 = 0.80$
69.95–71.95	12	$\frac{12}{100} = 0.12$	$0.80 + 0.12 = 0.92$
71.95–73.95	7	$\frac{7}{100} = 0.07$	$0.92 + 0.07 = 0.99$
73.95–75.95	1	$\frac{1}{100} = 0.01$	$0.99 + 0.01 = 1.00$
	Total = 100	**Total = 1.00**	

Try filling in the blanks of the relative frequency table.

Table 8.4.4 shows a frequency table with missing values. See if you can enter in the missing values.

Table 8.4.4: Frequency Table to be Filled in

Intervals	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
First		0.25	
Second			0.45
Third			0.8

Intervals	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
Fourth			
	Total = **20**		

New Table Check Answer

This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Exercise 8.4.1

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

c.

Answer

a. If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5+3+15=23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.
b. Add the relative frequencies in the second and third rows: $0.03+0.15=0.18$ or 18%.

c.

Exercise 8.4.2

Table $8.4.5$ shows the amount, in inches, of annual rainfall in a sample of towns.

Table 8.4.5:

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50}=0.12$	0.12
4.97–6.99	7	$\frac{7}{50}=0.14$	$0.12+0.14=0.26$
6.99–9.01	15	$\frac{15}{50}=0.30$	$0.26+0.30=0.56$
9.01–11.03	8	$\frac{8}{50}=0.16$	$0.56+0.16=0.72$
11.03–13.05	9	$\frac{9}{50}=0.18$	$0.72+0.18=0.90$
13.05–15.07	5	$\frac{5}{50}=0.10$	$0.90+0.10=1.00$
	Total = 50	Total = 1.00	

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.

Answer

a. 0.56 or 56
b. $0.30 + 0.16 + 0.18 = 0.64$ or 64

Exercise 8.4.3

Use the heights of the 100 male semiprofessional soccer players in Table 8.4.4. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 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 8.4.4

From Table 8.4.5, find the number of towns that have rainfall between 2.95 and 9.01 inches.

Answer

$6 + 7 + 15 = 28$ towns

COLLABORATIVE EXERCISE 8.4.7

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

a. What percentage of the students in your class have no siblings?
b. What percentage of the students have from one to three siblings?
c. What percentage of the students have fewer than three siblings?

Example 8.4.7

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 was produced:

Table 8.4.6: Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{3}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

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. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$
d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Exercise 8.4.8

Table represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Answer

$\frac{9}{50}$

Example 8.4.9

Table contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Table 8.4.7:

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

Year	Total Number of Deaths
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 8.4.10

Table contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Table 8.4.8:

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 cumulative relative frequency for 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.

References

1. "State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).
2. "State & County QuickFacts: Quick, easy access to facts about people, business, and geography," U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).
3. "Table 5: Direct hits by mainland United States Hurricanes (1851-2004)," National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).
4. "Levels of Measurement," http://infinity.cos.edu/faculty/wood...ata_Levels.htm (accessed May 1, 2013).
5. Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/Helpa...easurement.htm (accessed May 1, 2013).
6. David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Chapter Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

Exercise 8.4.11

What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: Superior, Average, Above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box
d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store

j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65 year-old women
l. Common letter grades: A, B, C, D, and F

Answer

a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

Glossary

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency

the number of times a value of the data occurs

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.4: Frequency, Frequency Tables, and Levels of Measurement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.5: Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the response variable. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Example 8.5.1

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Answer

- The *population* is men aged 50 to 84.
- The *sample* is the 400 men who participated.
- The *experimental units* are the individual men in the study.
- The *explanatory variable* is oral medication.

- The *treatments* are aspirin and a placebo.
- The *response variable* is whether a subject had a heart attack.

Example 8.5.2

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

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

Answer

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

Example 8.5.3

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Answer

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Exercise 8.5.4

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Answer

a. Explanatory: presence of distraction from texting; response: response time measured in seconds
b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.

f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[2]

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[3] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example 8.5.5

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Answer

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
c. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Exercise 8.5.6

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.
d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

Answer

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

1. “Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritio...rce/vitamin-e/ (accessed May 1, 2013).

2. Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).
3. Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-as...s-study-300443 (accessed May 1, 2013).
4. The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories...dLearning.html (accessed May 1, 2013).
5. M.L. Jacskon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).
6. “Earthquake Information by Year,” U.S. Geological Survey. http://earthquake.usgs.gov/earthquak...archives/year/ (accessed May 1, 2013).
7. “Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).
8. Data from www.businessweek.com (accessed May 1, 2013).
9. Data from www.forbes.com (accessed May 1, 2013).
10. “America’s Best Small Companies,” http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).
11. U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.
12. “April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april...onsumer-report (accessed May 1, 2013).
13. Lori Alden, “Statistics can be Misleading,” econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).
14. Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

Chapter Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”[4] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Exercise 8.5.7

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

Exercise 8.5.7

Discuss potential violations of the rule requiring informed consent.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children’s allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Answer

a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

Footnotes

[1] McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

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

[3] "Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upl...012_UK_web.pdf (accessed May 1, 2013).

[4] Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, http://www.stat.columbia.edu/~gelman...nceEthics1.pdf (accessed May 1, 2013).

Glossary

Explanatory Variable
the independent variable in an experiment; the value controlled by researchers

Treatments
different values or components of the explanatory variable applied in an experiment

Response Variable
the dependent variable in an experiment; the value that is measured for change at the end of an experiment

Experimental Unit
any individual or object to be measured

Lurking Variable
a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Random Assignment
the act of organizing experimental units into treatment groups using random methods

Control Group
a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Informed Consent
Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board
a committee tasked with oversight of research programs that involve human subjects

Placebo
an inactive treatment that has no real effect on the explanatory variable

Blinding
not telling participants which treatment a subject is receiving

Double-blinding

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 8.5: Experimental Design and Ethics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.6: Data Collection Experiment (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey

Ask five classmates from a different class how many movies they saw at the theater last month. Do not include rented movies.

1. Record the data.
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four names on the class list. Mark that person's name. Continue doing this until you have marked 12 names. You may need to go back to the start of the list. For each marked name record the five data values. You now have a total of 60 data values.
3. For each name marked, record the data.

___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___	___

Order the Data

Complete the two relative frequency tables below using your class data.

Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0–1			

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
2–3			
4–5			
6–7+			

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

Discussion Questions

1. Is one of the tables “more correct” than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

This page titled 8.6: Data Collection Experiment (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.7: Sampling Experiment (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

Note 1.7.1

The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

Restaurants Stratified by City and Entree Cost

	Restaurants Used in Sample			
Entree Cost	Under $10	$10 to under $15	$15 to under $20	Over $20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, Panda Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

A Simple Random Sample

Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Systematic Sample

Pick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

A Stratified Sample

Pick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4. _____	9. _____	14. _____	19. _____	24. _____
5. _____	10. _____	15. _____	20. _____	25. _____

This page titled 8.7: Sampling Experiment (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

8.E: Sampling and Data (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

1.1: Introduction

1.2: Definitions of Statistics, Probability, and Key Terms

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Q 1.2.1

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Q. 1.2.2

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

S 1.2.2

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. $X =$ the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

Q 1.2.3

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Q 1.2.4

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

S 1.2.5

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. $X =$ the health costs of one client
f. values for X, such as 34, 9, 82, and so on

Q 1.2.6

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Q 1.2.7

A marriage counselor is interested in the proportion of clients she counsels who stay married.

S 1.2.7

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. $X =$ the number of couples who stay married
f. yes, no

Q 1.2.8

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Q 1.2.9

A marketing company is interested in the proportion of people who will buy a particular product.

S 1.2.9

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. $X =$ the number of people who will buy it
f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Q 1.2.10

What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

Q 1.2.11

Consider the following:

$X =$ number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

S 1.2.12

a

Q 1.2.12

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

1.3: Data, Sampling, and Variation in Data and Sampling

Practice

Exercise 1.3.11

"Number of times per week" is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Use the following information to answer the next four exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

Exercise 1.3.12

The sampling method was

a. simple random
b. systematic
c. stratified
d. cluster

Answer

b

Exercise 1.3.13

"Duration (amount of time)" is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Exercise 1.3.14

The colors of the houses around the park are what kind of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Answer

a

Exercise 1.3.15

The population is _____

Exercise 1.3.16

Table contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

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

Year	Total Number of Deaths
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Use Table to answer the following questions.

a. What is the proportion of deaths between 2007 and 2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that happened before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.

Answer

a. 0.5242
b. 0.03%
c. 6.86%
d. $\frac{823,088}{823,856}$
e. quantitative discrete
f. quantitative continuous
g. In both years, underwater earthquakes produced massive tsunamis.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

Exercise 1.3.17

A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

Exercise 1.3.18

A market researcher polls every tenth person who walks into a store.

Answer

systematic

Exercise 1.3.19

The first 50 people who walk into a sporting event are polled on their television preferences.

Exercise 1.3.20

A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Answer

simple random

Use the following information to answer the next seven exercises: Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients

live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

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

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Exercise 1.3.21

Complete the tables using the data provided:

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Researcher B

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Exercise 1.3.22

Determine what the key term data refers to in the above example for Researcher A.

Answer

values for X, such as 3, 4, 11, and so on

Exercise 1.3.23

List two reasons why the data may differ.

Exercise 1.3.24

Can you tell if one researcher is correct and the other one is incorrect? Why?

Answer

No, we do not have enough information to make such a claim.

Exercise 1.3.25

Would you expect the data to be identical? Why or why not?

Exercise 1.3.26

How might the researchers gather random data?

Answer

Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

Exercise 1.3.27

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise 1.3.28

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Answer

This would be convenience sampling and is not random.

Use the following data to answer the next five exercises: Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher B

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
8–10	11	0.07	0.97
10–12	4	0.03	1

Exercise 1.3.29

Give a reason why the data may differ.

Exercise 1.3.30

Would the sample size be large enough if the population is the students in the school?

Answer

Yes, the sample size of 150 would be large enough to reflect a population of one school.

Exercise 1.3.31

Would the sample size be large enough if the population is school-aged children and young adults in the United States?

Exercise 1.3.32

Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

Answer

Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

Exercise 1.3.33

As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in the first Table. The second study collected the data in the second Table.

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Exercise 1.3.34

Given what you know, which study is correct?

Answer

There is not enough information given to judge if either one is correct or incorrect.

Exercise 1.3.35

The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

Exercise 1.3.36

Both groups that performed the study concluded that the software works. Is this accurate?

Answer

The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

Exercise 1.3.37

The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

Exercise 1.3.38

Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from Exercise?

Answer

Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

Exercise 1.3.39

Is a sample size of 1,000 a reliable measure for a population of 5,000?

Exercise 1.3.40

Is a sample of 500 volunteers a reliable measure for a population of 2,500?

Answer

No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

Exercise 1.3.41

A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

Exercise 1.3.42

Is a sample size of two representative of a population of five?

Answer

No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

Exercise 1.3.43

Is it possible for two experiments to be well run with similar sample sizes to get different data?

Bringing It Together

Exercise 1.3.44

Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed below.

LBCC Distance Learning Survey Results

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Exercise 1.3.45

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

Answer

Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

Q 1.3.1

number of tickets sold to a concert

S 1.3.1

quantitative discrete, 150

Q 1.3.2

percent of body fat

Q 1.3.3

favorite baseball team

S 1.3.3

qualitative, Oakland A's

Q 1.3.4

time in line to buy groceries

Q 1.3.5

number of students enrolled at Evergreen Valley College

S 1.3.5

quantitative discrete, 11,234 students

Q 1.3.6

most-watched television show

Q 1.3.7

brand of toothpaste

S 1.3.7

qualitative, Crest

Q 1.3.8

distance to the closest movie theater

Q 1.3.9

age of executives in Fortune 500 companies

S 1.3.9

quantitative continuous, 47.3 years

Q 1.3.10

number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park

was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Q 1.3.11

"Number of times per week" is what type of data?

1. qualitative
2. quantitative discrete
3. quantitative continuous

S 1.3.11

b

Q 1.3.12

"Duration (amount of time)" is what type of data?

1. qualitative
2. quantitative discrete
3. quantitative continuous

Q 1.3.13

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

S 1.3.13

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Q 1.3.14

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Q 1.3.15

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

S 1.3.15

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Q 1.3.16

List some practical difficulties involved in getting accurate results from a telephone survey.

Q 1.3.17

List some practical difficulties involved in getting accurate results from a mailed survey.

S 1.3.17

Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

Q 1.3.18

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

Q 1.3.19

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

S 1.3.19

b

Q 1.3.20

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

Q 1.3.21

Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

S 1.3.21

a. convenience
b. cluster
c. stratified
d. systematic
e. simple random

Q 1.3.22

A “random survey” was conducted of 3,274 people of the “microprocessor generation” (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why? Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called “America’s Smithsonian.”
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Q 1.3.23

The Gallup-Healthways Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

S 1.3.23

a. qualitative
b. quantitative discrete
c. quantitative discrete
d. qualitative

Q 1.3.24

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

Q 1.3.25

Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in [link] could explain this connection?

S 1.3.26

Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

Q 1.3.27

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

"Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[1]

As of April 25, 11 people responded to this question. Each participant answered "NO!"

Which of the potential problems with samples discussed in this module could explain this connection?

Q 1.3.28

A scholarly article about response rates begins with the following quote:

"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[2]

The Pew Research Center for People and the Press admits:

"The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more."[3]

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

S 1.3.28

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

1.4: Frequency, Frequency Tables, and Levels of Measurement

Q 1.4.1

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

Part-time Student Course Loads

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

a. Fill in the blanks in Table.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Q 1.4.2

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in Table.

Flossing Frequency for Adults with Gum Disease

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

a. Fill in the blanks in Table.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

S 1.4.2

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

Q 1.4.3

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 45; 10 .

Table was produced.

Frequency of Immigrant Survey Responses

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	219219	0.1053
2	3	319319	0.2632
4	1	119119	0.3158
5	3	319319	0.4737
7	2	219219	0.5789
10	2	219219	0.6842
12	2	219219	0.7895
15	1	119119	0.8421
20	1	119119	1.0000

a. Fix the errors in Table. Also, explain how someone might have arrived at the incorrect number(s).

b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Q 1.4.4

How much time does it take to travel to work? Table shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

S 1.4.4

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state’s travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

Q 1.4.5

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. Table shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?

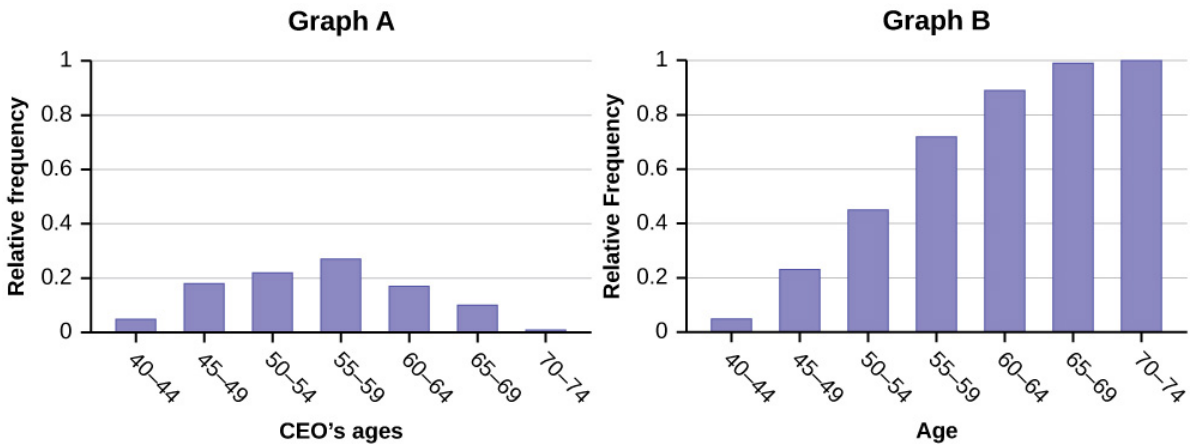


Figure 1.4.1. (a) Figure 1.4.1. (b)

Use the following information to answer the next two exercises: Table contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Frequency of Hurricane Direct Hits

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Q 1.4.6

What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

S 1.4.6

b

Q 1.4.7

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

1.5: Experimental Design and Ethics

Q 1.5.1

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments

were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

Use key terms from this module to describe the design of this experiment.

S 1.5.1

Explanatory variable: amount of sleep

Response variable: performance measured in assigned tasks

Treatments: normal sleep and 27 hours of total sleep deprivation

Experimental Units: 19 professional drivers

Lurking variables: none – all drivers participated in both treatments

Random assignment: treatments were assigned in random order; this eliminated the effect of any "learning" that may take place during the first experimental session

Control/Placebo: completing the experimental session under normal sleep conditions

Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time

Q 1.5.2

An advertisement for Acme Investments displays the two graphs in Figures to show the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

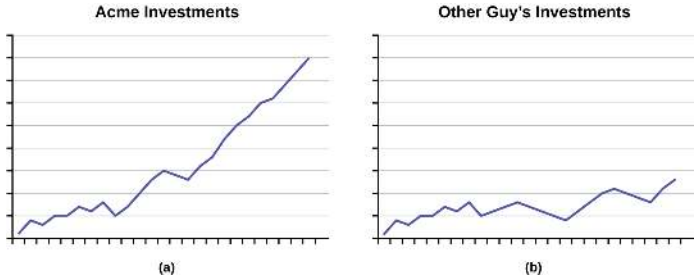


Figure 1.5.1. (a) Figure 1.5.1. (b)

As the graphs show, Acme consistently outperforms the Other Guys!

Q 1.5.3

The graph in Figure shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?

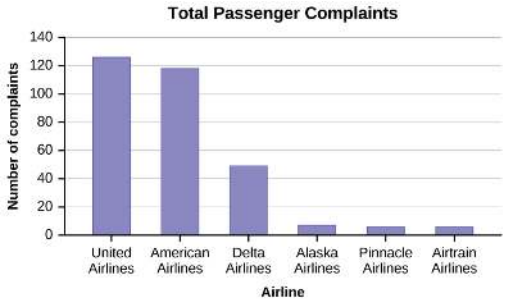


Figure 1.5.2.

S 1.5.3

You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

1.6: Data Collection Experiment

1.7: Sampling Experiment

This page titled 8.E: Sampling and Data (Exercises) 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.E: Sampling and Data (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

9: Descriptive Statistics

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

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9: Descriptive Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Prelude to Descriptive Statistics

Skills to Develop

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

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

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

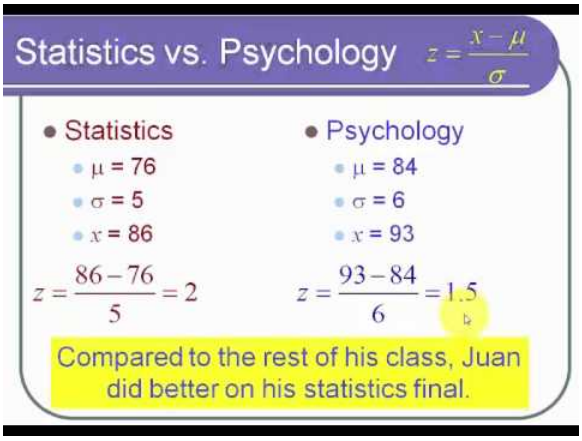
Figure 2.1.1: When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

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

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

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at 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.

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



Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.1: Prelude to Descriptive Statistics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

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 9.2.1

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 9.2.2

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

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

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 9.2.3

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 9.2.4

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

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
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.

Example 9.2.5: 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 9.2.1 and 9.2.2 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 9.2.1: *Presidential Ages at Inauguration*

President	Age*at Inauguration*	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51	Trump	70

9.2.2 *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
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		

President	Age	President	Age	President	Age
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 9.2.6

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
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006

Losses	Wins	Year	Losses	Wins	Year
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

9.2.1 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

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in Example, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example 9.2.7

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in Table and in Figure.

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

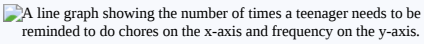


Figure 9.2.1: A line graph showing the number of times a teenager needs to be reminded to do chores on the x-axis and frequency on the y-axis.

Exercise 9.2.8

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in Table. Construct a line graph.

Number of times in shop	Frequency
0	7

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

Answer

Figure 9.2.2: A line graph showing the number of times a car is in the shop on the x-axis and frequency on the y-axis.

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in Example 9.2.9 has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 9.2.9

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

Figure 9.2.3: This is a bar graph that matches the supplied data. The x-axis shows age groups and the y-axis show the percentages of Facebook users

Exercise 9.2.10

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%
Retirees	131,662	38%

Answer

Figure 9.2.4: This is a bar graph that matches the supplied data. The x-axis shows age groups, and the y-axis shows the percentages of Park City's population.

Example 9.2.11

The columns in 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

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

Figure 9.2.5: This is a bar graph that matches the supplied data. The x-axis shows race and ethnicity, and the y-axis shows the percentages of AP examinees.

Exercise 9.2.12

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Answer

Figure 9.2.6: This is a bar graph that matches the supplied data. The x-axis shows Park City voting districts, and the y-axis shows the percentages of the registered voter population.

Summary

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

References

1. Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://www.kenburbary.com/2011/03/fa...-statistics-2/ (accessed August 21, 2013).
2. “9th Annual AP Report to the Nation.” CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goa...omoting-equity (accessed September 13, 2013).

3. “Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.3: Histograms, Frequency Polygons, and Time Series Graphs

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

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

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

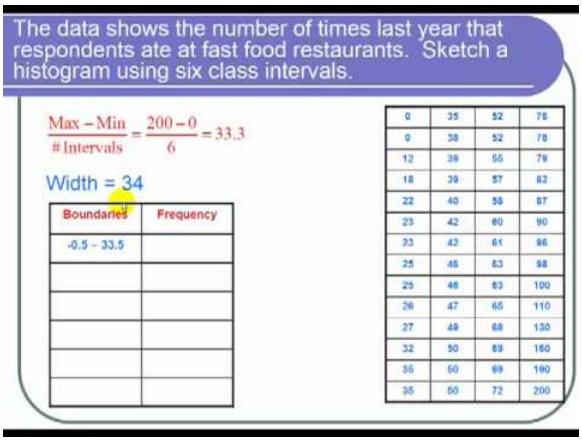
- f is frequency
- n is total number of data values (or the sum of the individual frequencies), and
- RF is relative frequency,

then:

$$RF = \frac{f}{n} \tag{9.3.1}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and RF = fn = 340 = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

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



Example 2.3.1

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

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 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; 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.

$$\frac{74.05 - 59.95}{8} = 1.76 \tag{9.3.2}$$

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

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95

- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

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

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

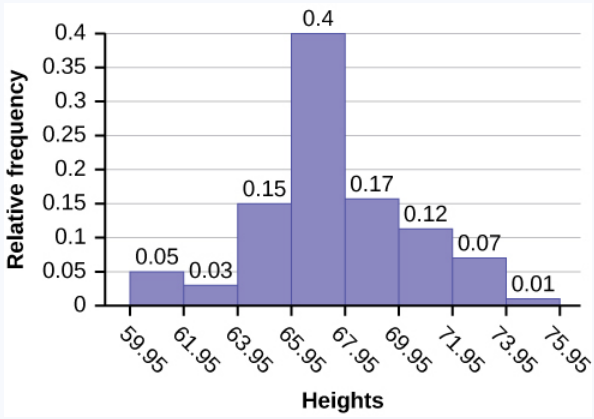


Figure 2.3.1: Histogram of something

Exercise 2.3.1

continuous data since shoe size is measured. 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; 11.5; 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 suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example 2.3.2

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; 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.

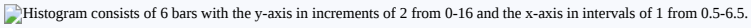


Figure 2.3.2.

Note

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

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2^{nd} Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2^{nd} Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3^{rd}picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2^{nd} 1). Arrow down to Freq. Enter L2 (2^{nd} 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Exercise 2.3.2

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
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 2.3.3

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 2.3.3.

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

Exercise 2.3.3

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

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

Use 10–19 as the first interval.

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

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons. To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 2.3.4

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Figure 2.3.4.

The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Exercise 2.3.4

Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in Table.

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Answer

The first label on the *x*-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 44 represents the next interval, or the first "real" interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

.

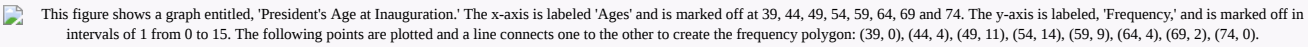


Figure 2.3.5.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 2.3.5

We will construct an overlay frequency polygon comparing the scores from Example with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
Lower Bound Distribution for Calculus Final Grades	Upper Bound	Frequency	Cumulative Frequency
Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

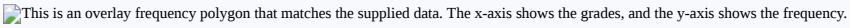

This is an overlay frequency polygon that matches the supplied data. The x-axis shows the grades, and the y-axis shows the frequency.

Figure 2.3.6.

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.

Constructing a Time Series Graph

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

Example 2.3.6

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

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

Answer

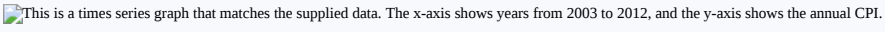

This is a times series graph that matches the supplied data. The x-axis shows years from 2003 to 2012, and the y-axis shows the annual CPI.

Figure 2.3.7.

Exercise 2.3.5

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.

CO2 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

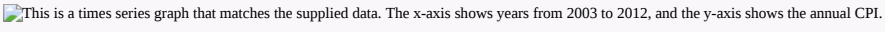

This is a times series graph that matches the supplied data. The x-axis shows years from 2003 to 2012, and the y-axis shows the annual CPI.

Figure 2.3.8.

Uses of a Time Series Graph

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

Chapter Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.Glossary

References

1. Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker
2. "Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/a...-us-presidents (accessed April 3, 2013).
3. "Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).
4. "Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).
5. "Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).
6. "CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).
7. "Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/stati...me-series.html (accessed April 3, 2013).
8. "Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).
9. Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.
10. "Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Frequency

the number of times a value of the data occurs

Histogram

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

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.3: Histograms, Frequency Polygons, and Time Series Graphs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.4: Measures of the Location of the Data

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

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

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

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

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

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

Ordered from smallest to largest:

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

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

$$\frac{6.8+7.2}{2}=7 \tag{9.4.1}$$

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

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

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

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

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

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

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

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

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

$$IQR = Q_3 - Q_1 \tag{2.4.1}$$

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

Definition: Outliers

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

Example 2.4.1

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

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

Answer

Order the data from smallest to largest.

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

$$M = 488,800$$

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

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

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

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

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

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

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

Exercise 2.4.1

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

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

Answer

Order the data from smallest to largest.

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

Median = $54,000

$$Q_1 = \$40,500$$

$$Q_3 = \$69,000$$

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

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

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

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

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

Example 2.4.2

For the two data sets in the test scores example, find the following:

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

Answer

The five number summary for the day and night classes is

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

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

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

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

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

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

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

Night class outliers are calculated as:

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

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

Exercise 2.4.2

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

Test Scores for Class *A*

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

Test Scores for Class *B*

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

Answer

Class *A*

Order the data from smallest to largest.

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

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

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

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

$IQR = 90.5 - 72.5 = 18$

Class *B*

Order the data from smallest to largest.

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

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

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

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

$IQR = 93.5 - 72.5 = 21$

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

Example 2.4.3

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

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

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

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

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

Exercise 2.4.3

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

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

Answer

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

The 65th percentile is 3.5.

Example 2.4.4

Using Table:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution

Using the data from the frequency table, we have:

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

Exercise 2.4.4

Refer to the Table. Find the third quartile. What is another name for the third quartile?

Answer

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

COLLABORATIVE STATISTICS

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

a. How many students were surveyed?
b. What kind of sampling did you do?
c. Construct two different histograms. For each, starting value = _____ ending value = ____.
d. Find the median, first quartile, and third quartile.
e. Construct a table of the data to find the following:
   i. the 10th percentile
   ii. the 70th percentile
   iii. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

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

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

Order the data from smallest to largest.

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

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

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

Example 2.4.5

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 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 2.4.5

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

Calculate the 20th percentile and the 55th percentile.

Answer

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

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

Note 2.4.2

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

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

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

Example 2.4.6

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 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. There is one value of 58.

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

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

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

Exercise 2.4.6

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

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

Find the percentiles for 47 and 31.

Answer

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

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

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

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

Interpreting Percentiles, Quartiles, and Median

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

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

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

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

GUIDELINE

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

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

Example 2.4.7

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.

Answer

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

Exercise 2.4.7

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

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.

Example 2.4.8

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

Answer

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

Exercise 2.4.8

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

Answer

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

Example 2.4.9

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

Answer

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

Exercise 2.4.9

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

Answer

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

Example 2.4.10

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

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

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

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

Determine the following five values.

- Min = 0
- Q_1 = 20

- Med = 40
- Q_3 = 60
- Max = 300

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

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

$$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120 \tag{9.4.2}$$

.

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

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

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

References

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

Chapter Review

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

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

Formula Review

$$i = \frac{k}{100}(n+1) \tag{9.4.3}$$

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

k = the kth percentile,

n = total number of data.

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

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

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

$n =$ total number of data

Glossary

Interquartile Range

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

Outlier

an observation that does not fit the rest of the data

Percentile

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

Quartiles

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

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.4: Measures of the Location of the Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.4E: Measures of the Location of the Data (Exercises)

Exercise 2.4.10

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

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

a. Find the 40th percentile.
b. Find the 78th percentile.

Answer

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

Exercise 2.4.11

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

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

a. Find the percentile of 37.
b. Find the percentile of 72.

Exercise 2.4.12

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Answer

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercise 2.4.13

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 2.4.14

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Answer

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise 2.4.15

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise 2.4.16

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

Answer

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise 2.4.17

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise 2.4.18

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

Answer

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise 2.4.19

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise 2.4.20

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Answer

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Use Exercise to calculate the following values:

Exercise 2.4.21

First quartile = _____

Exercise 2.4.22

Second quartile = median = 50th percentile = _____

Answer

4

Exercise 2.4.23

Third quartile = _____

Exercise 2.4.24

Interquartile range (*IQR*) = _____ – _____ = _____

Answer

$6 - 4 = 2$

Exercise 2.4.25

10^{th} percentile = _____

Exercise 2.4.26

70^{th} percentile = _____

Answer

6

9.4E: Measures of the Location of the Data (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

9.5: Box Plots

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

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

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

Consider, again, this dataset.

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

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

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

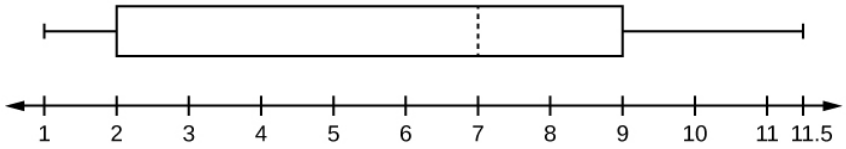
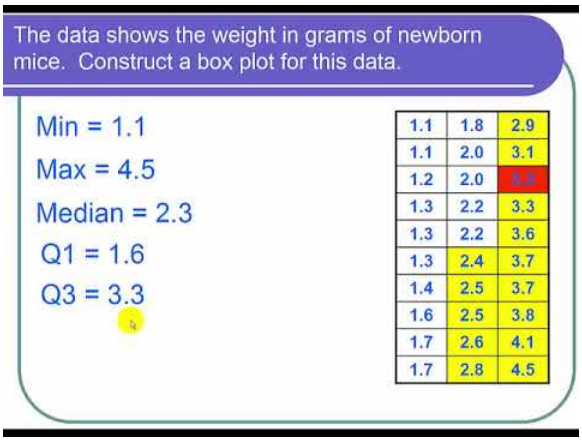


Figure 9.5.1

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

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



Play the Draw the box plot game and see if you can get a perfect score of 70.

This is a game that will guide you through the process of drawing a box plot.

Score 0

Min:	8	σ:	5	Q1:	16	Mean:	24
Median:	27	Q3:	32	Max:	45		

Click on each of the values that are part of the five point summary for the box plot.

Example 9.5.1

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

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

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

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

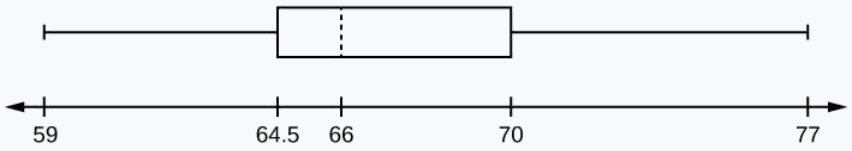


Figure 9.5.2

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

Calculator

To find the minimum, maximum, and quartiles:

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

Put the data values into the list L1.

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

Press ENTER.

Use the down and up arrow keys to scroll.

Smallest value = 59.

Largest value = 77.

Q_1: First quartile = 64.5.

Q_2: Second quartile or median = 66.

Q_3: Third quartile = 70.

To construct the box plot:

Press 4: Plotsoff. Press ENTER.

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

Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.

Press Zoom. Press 9: ZoomStat.

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

Exercise 9.5.1

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234; 240; 255; 270; 275; 290; 301; 303; 315; 317; 318; 326; 333; 343; 349; 360; 369; 377; 388; 391; 392; 398; 400; 402; 405; 408; 422; 429; 450; 475; 512

Answer

Figure 9.5.3

$IQR = 158$

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

Figure 9.5.4

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

Example 9.5.2

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

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

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

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

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

Answer

a.
- Min = 32
- Q_1 = 56
- M = 74.5

- $Q_3 = 82.5$
- Max = 99

b.
- Min = 25.5
- $Q_1 = 78$
- $M = 81$
- $Q_3 = 89$
- Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

d.

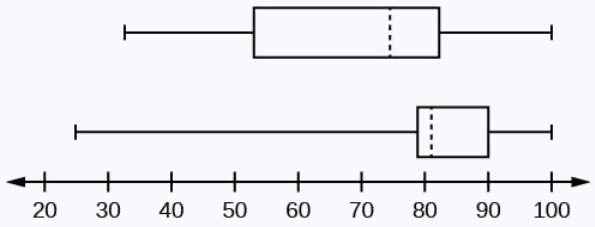


***Figure* 9.5.5**

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

Exercise 9.5.2

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

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

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

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

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

Answer

Figure 9.5.6

IQR for the boys = 4

IQR for the girls = 5

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

Example 9.5.3

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

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

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

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

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

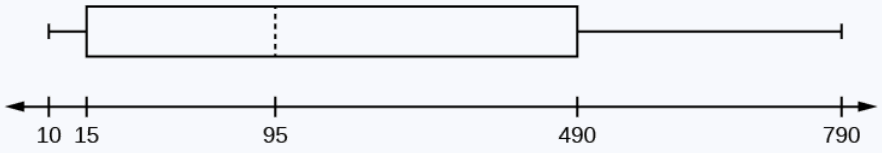


Figure 9.5.7

Exercise 9.5.3

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

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

Answer

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

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

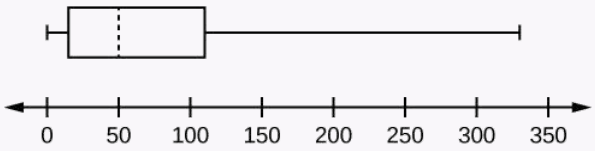


Figure 9.5.8

References

1. Data from *West Magazine*.

Chapter Review

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

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

Exercise 2.5.4

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

Exercise 2.5.5

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

Answer

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

Bringing It Together

Exercise 2.5.6

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

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

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

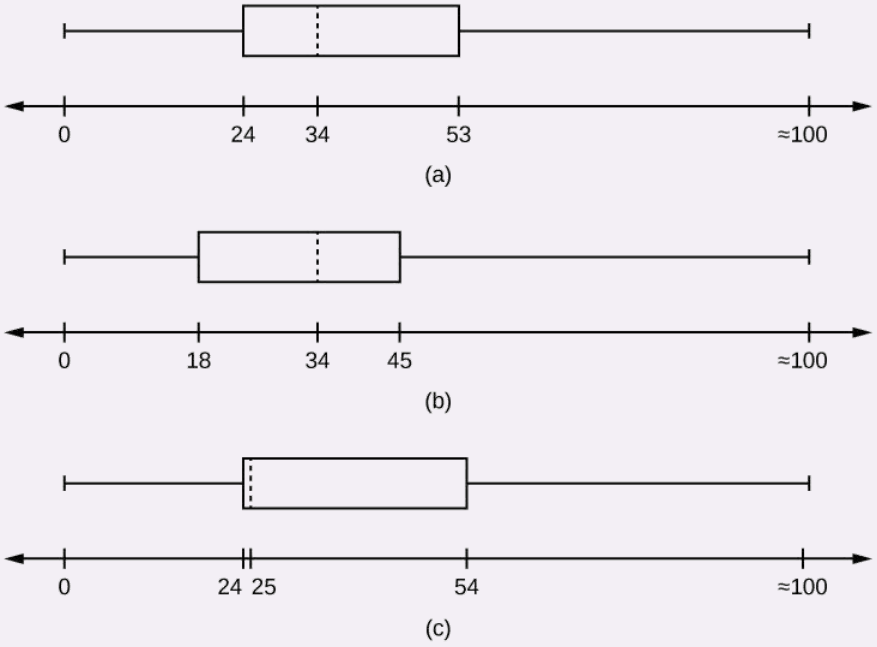


Figure 2.5.9.

Answer

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

Glossary

Box plot

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

First Quartile

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

Frequency Polygon

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

Interval

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

Paired Data Set

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

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

Skewed

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

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.5: Box Plots is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.6: Measures of the Center of the Data

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

> The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

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

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

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

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

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7 \tag{9.6.1}$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7 \tag{9.6.2}$$

In the second example, the frequencies are

$$3(1)+2(2)+1(3)+5(4). \tag{9.6.3}$$

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

$$\frac{n+1}{2} \tag{9.6.4}$$

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

$$\frac{n+1}{2} = \frac{97+1}{2} = 49. \tag{9.6.5}$$

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

$$\frac{n+1}{2} = \frac{100+1}{2} = 50.5. \tag{9.6.6}$$

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 9.6.1

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:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\ldots+35+37+40+(44)(2)+47]}{40} = 23.6 \tag{9.6.7}$$

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

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5 \tag{9.6.8}$$

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

$$M = \frac{24+24}{2} = 24 \tag{9.6.9}$$

Calculator

To find the mean and the median:

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

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

Put the data values into list L1.

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

Press the down and up arrow keys to scroll.

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

Online Mean and Median Calculator

You can find the mean, $\bar{x}$, and median, M, quickly by entering the numbers separated by commas and clicking Calculate

Calculate

$\bar{x}$ =

M =

Exercise 9.6.1

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

$$\frac{544}{39} = 13.95 \tag{9.6.10}$$

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

Example 9.6.2

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

$$\bar{x} = \frac{5,000,000 + 49(30,000)}{50} = 129,400 \quad (9.6.11)$$

$M = 30,000$

(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 9.6.2

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

Answer

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

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

Example 9.6.3

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 9.6.3

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.

Example 9.6.4

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

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

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

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

Exercise 9.6.4

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

Answer

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

The Law of Large Numbers and the Mean

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

Sampling Distributions and Statistic of a Sampling Distribution

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

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

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

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

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean:

$$mean = \frac{\text{data sum}}{\text{number of data values}}. \tag{9.6.12}$$

We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is

$$\frac{\text{lower boundary}+\text{upper boundary}}{2}. \tag{9.6.13}$$

We can now modify the mean definition to be

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{9.6.14}$$

where f is the frequency of the interval and m is the midpoint of the interval.

Example 9.6.5

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4

Grade Interval	Number of Students
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint.
  $\sum fm53.25(1)+59.5(0)+65.5(4)+71.5(4)+77.5(2)+83.5(3)+89.5(4)+95.5(1)=1460.25$
- $\mu=\dfrac{\sum fm}{\sum f}=\dfrac{1460.25}{19}=76.86$

Exercise 9.6.5

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Answer

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers

The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.

$$Mean=(1.75)(3)+(5.5)(7)+(9.5)(12)+(13.5)(7)+(17.5)(9)=409.75 \quad (9.6.15)$$

Online Mean, median, and Mode Calculator from a frequency table

Enter the lower bounds, the upper bounds, and the frequencies for each of the intervals of the frequency table and then hit Calculate. Leave the bottom rows that do not have any intervals blank.

Lower Bounds	Upper Bounds	Frequencies

Calculate	Mean =	Median =	Mode =

References

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

Chapter Review

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

Formula Review

$$\mu = \frac{\sum fm}{\sum f} \tag{9.6.16}$$

where f = interval frequencies and m = interval midpoints.

Exercise 2.6.6

Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

Exercise 2.6.7

Calculate the mean.

Answer

Mean: $16+17+19+20+20+21+23+24+25+25+25+26+26+27+27+27+28+29+30+32+33+33+34+35+37+39+40=738$;

$\frac{738}{27}=27.33$

Exercise 2.6.8

Identify the median.

Exercise 2.6.9

Identify the mode.

Answer

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise 2.6.10

sample mean = $\bar{x}$ = _____

Exercise 2.6.11

median = _____

Answer

4

Bringing It Together

Exercise 2.6.12

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?
<figure >

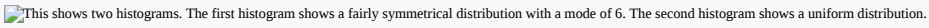


Figure 2.6.1

</figure>
d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?
<figure >

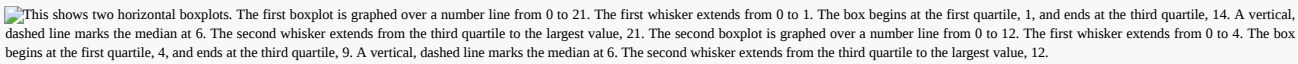


Figure 2.6.2

</figure>

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise 2.6.13

What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Answer

a

Exercise 2.6.14

What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise 2.6.15

Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Answer

b

Glossary

Frequency Table

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

Mean

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

Median

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

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.6: Measures of the Center of the Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.7: Skewness and the Mean, Median, and Mode

Consider the following data set.

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

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

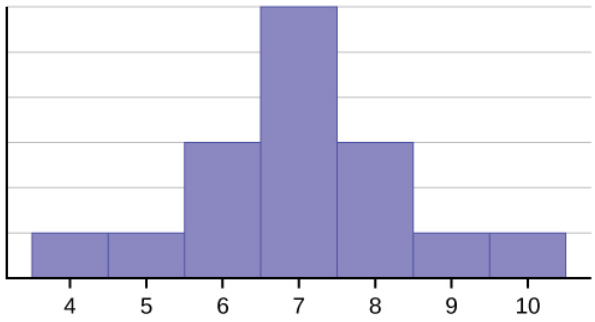


Figure 9.7.1

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

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

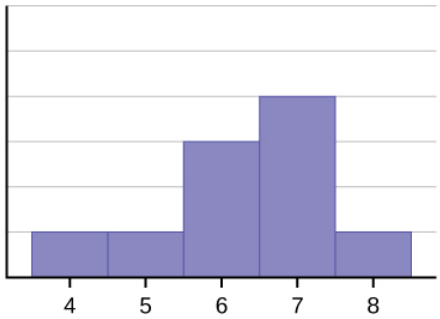


Figure 9.7.2

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

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

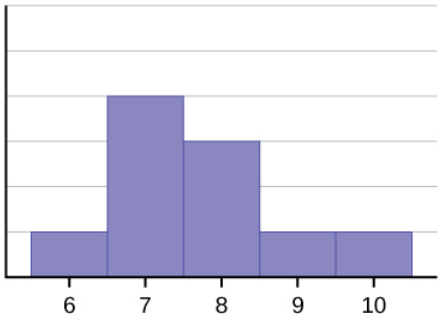


Figure 9.7.3

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

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

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

Example 9.7.1

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

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

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

Solution

a.

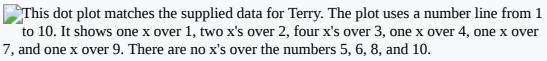


Figure 9.7.4: This dot plot matches the supplied data for Terry. The plot uses a number line from 1 to 10. It shows one x over 1, two x's over 2, four x's over 3, one x over 4, one x over 7, and one x over 9. There are no x's over the numbers 5, 6, 8, and 10.

Figure 9.7.5: Copy and Paste Caption here. (Copyright; author via source)

Figure 9.7.6: Copy and Paste Caption here. (Copyright; author via source)

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

Exercise 9.7.1

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

a.

Figure 9.7.7: This dot plot matches the supplied data. The plot uses a number line from 0 to 14. It shows two x's over 0, four x's over 1, three x's over 2, one x over 3, two x's over the number 4, 5, 6, and 9, and 1 x each over 10 and 14. There are no x's over the numbers 7, 8, 11, 12, and 13.

b.

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

The Ages Former U.S Presidents Died	
9	0 0 3 3

Key: 8|0 means 80.

c.

Figure 9.7.8: This is a histogram titled Hours Spent Playing Video Games on Weekends. The x-axis shows the number of hours spent playing video games with bars showing values at intervals of 5. The y-axis shows the number of students. The first bar for 0 - 4.99 hours has a height of 2. The second bar from 5 - 9.99 has a height of 3. The third bar from 10 - 14.99 has a height of 4. The fourth bar from 15 - 19.99 has a height of 7. The fifth bar from 20 - 24.99 has a height of 9.

Chapter Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions:

- A **right (or positive) skewed** distribution has a shape like Figure 9.7.3.
- A **left (or negative) skewed** distribution has a shape like Figure 9.7.2.
- A **symmetrical** distribution looks like Figure 9.7.1.

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

Exercise 2.7.2

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

Answer

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

Exercise 2.7.3

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

Exercise 2.7.4

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

Answer

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

Exercise 2.7.5

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

Exercise 2.7.6

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

Answer

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

Exercise 2.7.7

What word describes a distribution that has two modes?

Exercise 2.7.8

Describe the shape of this distribution.

Figure 9.7.9: This is a historgram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right.

Answer

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

Exercise 2.7.9

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

Figure 9.7.10: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Exercise 2.7.10

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

Figure 9.7.11: This is a histogram which consists of 5 adjacent bars with the x-axis split into intervals of 1 from 3 to 7. The bar heights peak at the first bar and taper lower to the right. The bar heights from left to right are: 8, 4, 2, 2, 1.

Answer

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

Exercise 2.7.11

Describe the shape of this distribution.

Figure 9.7.12

Exercise 2.7.12

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

Figure 9.7.13

Answer

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

Exercise 2.7.13

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

Figure 9.7.14

Exercise 2.7.14

Describe the shape of this distribution.

Figure 9.7.15

Answer

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

Exercise 2.7.15

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

Figure 9.7.16: Copy and Paste Caption here. (Copyright; author via source)

Exercise 2.7.16

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

Figure 9.7.17

Answer

The mean and the median are both six.

Exercise 2.7.17

The mean and median for the data are the same.

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

Is the data perfectly symmetrical? Why or why not?

Exercise 2.7.18

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

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

Answer

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

Exercise 2.7.19

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

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

Exercise 2.7.20

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

Answer

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

Exercise 2.7.21

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

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.7: Skewness and the Mean, Median, and Mode is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.8: Measures of the Spread of the Data

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

The standard deviation

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

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

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

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

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

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

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

Rosa waits for seven minutes:

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

Binh waits for one minute.

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

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

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

Figure 9.8.1

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: $1 = 5 + (-2)(2)$.

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

- sample:

$$x = \bar{x} + (\#\text{ofSTDEV})(s) \tag{9.8.1}$$

- Population:

$$x = \mu + (\#\text{ofSTDEV})(s) \tag{9.8.2}$$

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

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

Calculating the Standard Deviation

If x is a number, then the difference "x – 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

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{9.8.3}$$

or

$$s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n - 1}} \tag{9.8.4}$$

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

Formulas for the Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \tag{9.8.5}$$

or

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \tag{9.8.6}$$

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

In Equations 9.8.4 and 9.8.6, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three.

Standard Deviation Calculator

Type in the values from the data set separated by commas, for example, 2,4,5,8,11,2, and click Calculate.

- ◯ Population Standard Deviation
- ◯ Sample Standard Deviation

Standard Deviation

Calculate

Play the Guess the Standard Deviation Game.

The goal is to guess the standard deviation rounded to the nearest whole number. Try to do this visually without doing any calculations. You score a point if you get it correct, don't lose a point if you are within 1 of the standard deviation, and lose a point if you are off by 2 or more. Enjoy the game!

σ= 1 | σ= 2 | σ= 3 | σ= 4

Next | Click Next to start the Game | Score: 0

Sampling Variability of a Statistic

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

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

In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you

should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example 9.8.1

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

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

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525 \quad (9.8.7)$$

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):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125 \quad (9.8.8)$$

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

$$s = \sqrt{0.5125} = 0.715891 \quad (9.8.9)$$

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

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

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)σ
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution

a.

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

b. $(\bar{x}+1s) = 10.53 + (1)(0.72) = 11.25$

c. $(\bar{x}-2s) = 10.53 - (2)(0.72) = 9.09$

d.
- $(\bar{x}-1.5s) = 10.53 - (1.5)(0.72) = 9.45$
- $(\bar{x}+1.5s) = 10.53 + (1.5)(0.72) = 11.61$

Exercise 2.8.1

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

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

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

Answer

$\mu = 30.68$

$s = 6.09$

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

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For Example 9.8.1, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

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

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

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

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

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

Example 9.8.2

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

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

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

Answer

a. See Table
b. i. The sample mean = 73.5
   ii. The sample standard deviation = 17.9
   iii. The median = 73
   iv. The first quartile = 61
   v. The third quartile = 90
   vi. *IQR* = 90 – 61 = 29
c. The x-axis goes from 32.5 to 100.5; y-axis goes from -2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is $(100.5 - 32.5)$ divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, $32.5 + 13.6 = 46.1$, $46.1 + 13.6 = 59.7$, $59.7 + 13.6 = 73.3$, $73.3 + 13.6 = 86.9$, $86.9 + 13.6 = 100.5 =$ the ending value; No data values fall on an interval boundary.

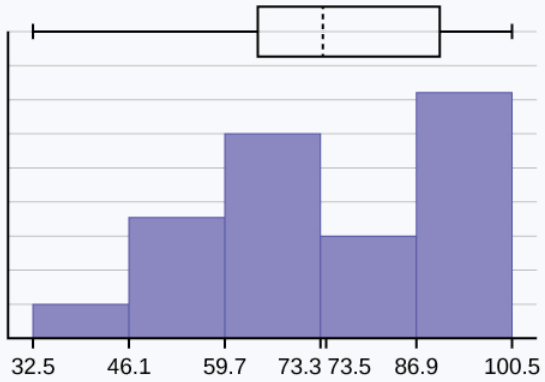


Figure 2.9.2.

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

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Exercise 9.8.2

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

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

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

Answer

$\mu = 9.3$ and $s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$\text{Mean of Frequency Table} = \frac{\sum fm}{\sum f} \tag{9.8.10}$$

where f interval frequencies and $m =$ interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how “unusual” individual data is compared to the mean.

Example 9.8.3

Find the standard deviation for the data in Table 9.8.3.

Table 9.8.3

Class	Frequency, f	Midpoint, m	m^2	$\bar{x}$	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5

Class	Frequency, f	Midpoint, m	m^2	$\bar{x}$	fm^2	Standard Deviation
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, $\bar{x}$ = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{\sum f(m-\bar{x})^2}{n-1}}$ where s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Find the standard deviation for the data from the previous example

Class	0-2	3-5	6-8	9–11	12–14	15–17
Frequency, f	1	6	10	7	0	2

First, press the **STAT** key and select **1:Edit**

Figure 2.8.3

Input the midpoint values into **L1** and the frequencies into **L2**

Figure 2.8.4

Select **STAT**, **CALC**, and **1: 1-Var Stats**

Figure 2.8.5

Select **2nd** then **1** then , **2nd** then **2 Enter**

Figure 2.8.6

You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Mean and Standard Deviation for grouped frequency Tables Calculator

Type in the midpoints and frequencies below. Put the midpoints in increasing order and do not include any values with zero frequency.

Midpoints

Frequencies

- ◯ Population Standard Deviation
- ◯ Sample Standard Deviation

Mean

Standard Deviation

Five Point Summary

Sample Size

Calculate

Comparing Values from Different Data Sets

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

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

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

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

Z-Score Calculator

x	μ	σ

z-score

Enter any three of them, leave the fourth blank, and click Calculate.

Calculate

Example 9.8.4

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

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

Answer

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

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

For John,

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

For Ali,

$$z = \#\text{ofSTDEVs} = \left(\frac{77-80}{10}\right) = -0.3$$

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

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

Exercise 9.8.4

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

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

Answer

For Angie:

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

For Beth:

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

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

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

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

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

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.

- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

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

Chapter Review

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

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

Formula Review

$$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} \tag{9.8.11}$$

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

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

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

Exercise 2.8.4

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

Answer

s = 34.5

Exercise 2.8.5

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

Exercise 2.8.6

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

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

Answer

For Fredo:

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

For Karl:

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

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

Exercise 2.8.7

Use Table to find the value that is three standard deviations:

- above the mean
- below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise 2.8.5

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3

Grade	Frequency
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Answer

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Bringing It Together

Exercise 2.8.7

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

a. Find the sample mean $\bar{x}$.
b. Find the approximate sample standard deviation, s.

Answer

a. 1.48
b. 1.12

Exercise 2.8.8

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let $X =$ the number of pairs of sneakers owned. The results are as follows:

X	**Frequency**
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40^{th} percentile.
k. Find the 90^{th} percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise 2.8.9

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _____ to _____.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:
    i. the population mean, μ.

ii. the population standard deviation, σ.
iii. the weight that is two standard deviations below the mean.
iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Answer

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5
e. A box plot with a whisker between 174 and 205.5, a solid line at 205.5, a dashed line at 241, a solid line at 272.5, and a whisker between 272.5 and 302.
f. 205.5, 272.5
g. sample
h. population
i. i. 236.34
ii. 37.50
iii. 161.34
iv. 0.84 std. dev. below the mean
j. Young

Exercise 2.8.10

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

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

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise 2.8.11

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

<figure >

This shows three graphs. The first is a histogram with a mode of 3 and fairly symmetrical distribution between 1 (minimum value) and 5 (maximum value). The second graph is a histogram with peaks at 1 (minimum value) and 5 (maximum value) with 3 having the lowest frequency. The third graph is a box plot. The first whisker extends from 0 to 1. The box begins at the firs quartile, 1, and ends at the third quartile,6. A vertical, dashed line marks the median at 3. The second whisker extends from 6 on.

Figure 2.8.6.

</figure>

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

Answer

a. True
b. True
c. True
d. False

Exercise 2.8.12

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _____ days to _____ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise 2.8.13

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Answer

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. $X =$ the number of days per week that 100 clients use a particular exercise facility.

x	**Frequency**
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Exercise 2.8.14

The 80^{th} percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise 2.8.15

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Answer

a

Exercise 2.8.16

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the Table.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR*to identify outliers, if any, and clearly state your conclusion.

b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 9.8: Measures of the Spread of the Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.9: Descriptive Statistics (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will construct a histogram and a box plot.
- The student will calculate univariate statistics.
- The student will examine the graphs to interpret what the data implies.

Collect the Data

Record the number of pairs of shoes you own.

1. Randomly survey 30 classmates about the number of pairs of shoes they own. Record their values.

Survey Results

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil and scale the axes.

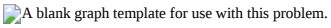

A blank graph template for use with this problem.

Figure 9.9.1

3. Calculate the following values.
   a. $\bar{x}$ = _____
   b. s = _____
4. Are the data discrete or continuous? How do you know?
5. In complete sentences, describe the shape of the histogram.
6. Are there any potential outliers? List the value(s) that could be outliers. Use a formula to check the end values to determine if they are potential outliers.

Analyze the Data

1. Determine the following values.
   a. Min = _____
   b. M = _____
   c. Max = _____
   d. Q_1 = _____
   e. Q_3 = _____
   f. IQR = _____
2. Construct a box plot of data
3. What does the shape of the box plot imply about the concentration of data? Use complete sentences.
4. Using the box plot, how can you determine if there are potential outliers?
5. How does the standard deviation help you to determine concentration of the data and whether or not there are potential outliers?

6. What does the *IQR* represent in this problem?
7. Show your work to find the value that is 1.5 standard deviations:
    a. above the mean.
    b. below the mean.

This page titled 9.9: Descriptive Statistics (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

9.E: Descriptive Statistics (Exercises)

2.2: Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Q 2.2.1

Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Q 2.2.2

The table below contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

S 2.2.2

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.
   - Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
   - Press MATH
   - Arrow over to PRB
   - Press 5:randInt(
   - Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

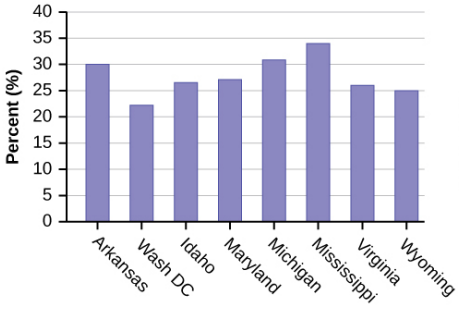


Figure 9.E.1: (a)

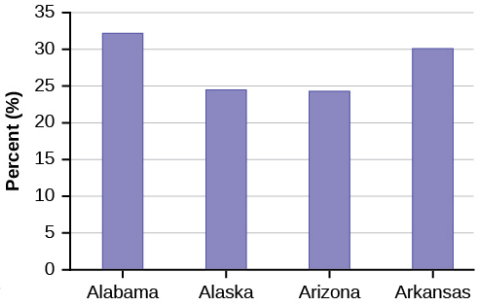


Figure 9.E.1: (b)

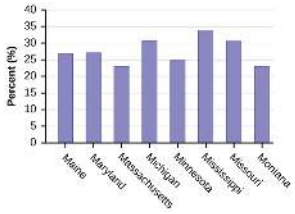


Figure 9.E.1: (c)

For each of the following data sets, create a stem plot and identify any outliers.

Exercise 2.2.7

The miles per gallon rating for 30 cars are shown below (lowest to highest).

19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Answer

Stem	Leaf
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Answer

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

The data are daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

For the next three exercises, use the data to construct a line graph.

Exercise 2.2.8

In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in the Table below.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Answer

Figure 9.E. 2:

Exercise 2.2.9

In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in Table.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

Exercise 2.2.10

Several children were asked how many TV shows they watch each day. The results of the survey are shown in the Table below.

Number of TV Shows	Frequency

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Answer

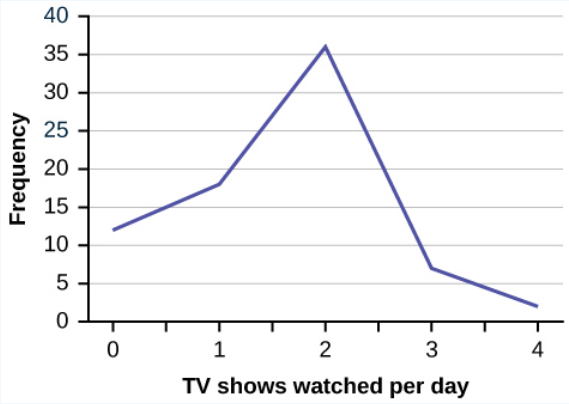


Figure 9.*E*. 3:

Exercise 2.2.11

The students in Ms. Ramirez's math class have birthdays in each of the four seasons. Table shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Using the data from Mrs. Ramirez's math class supplied in the table above, construct a bar graph showing the percentages.

Answer

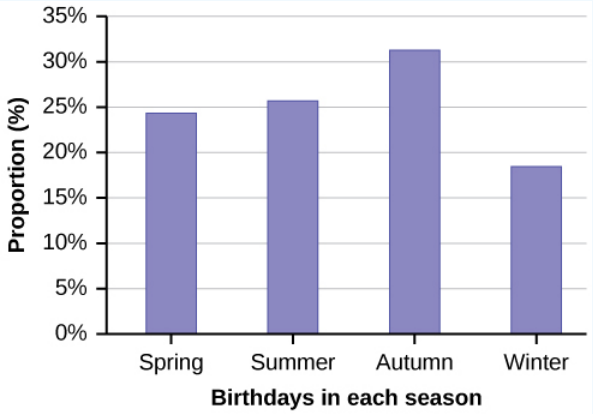


Figure 9.E. 4:

Exercise 2.2.12

David County has six high schools. Each school sent students to participate in a county-wide science competition. Table shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Use the data from the David County science competition supplied in Exercise. Construct a bar graph that shows the county-wide population percentage of students at each school.

Answer

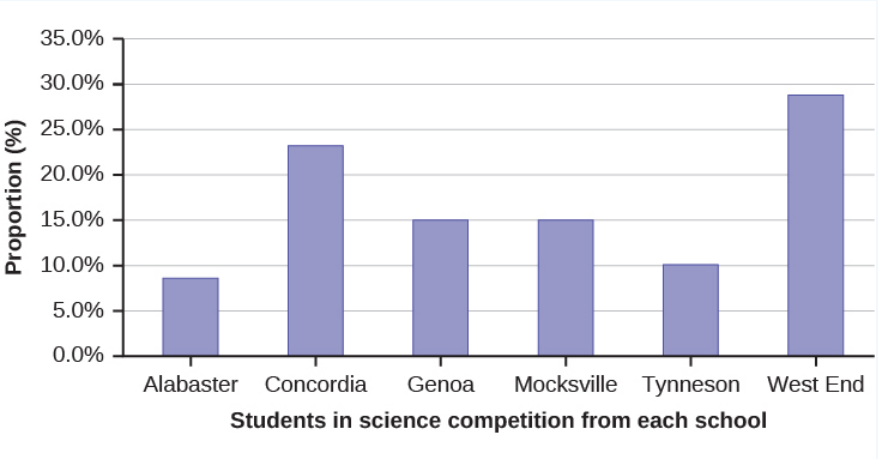


Figure 9.E.5:

2.3: Histograms, Frequency, Polygons, and Time Series Graphs

Q 2.3.1

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	

# of books	Freq.	Rel. Freq.
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher C

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Q 2.3.2

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Singles

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Couples

Amount($)	Frequency	Rel. Frequency
100–150	5	

Amount($)	Frequency	Rel. Frequency
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

S 2.3.2

Singles

Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Couples

Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07

Amount($)	Frequency	Relative Frequency
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

a. See the tables above

b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

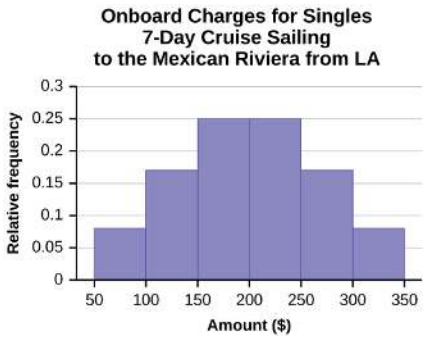


Figure 9.E.6:

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

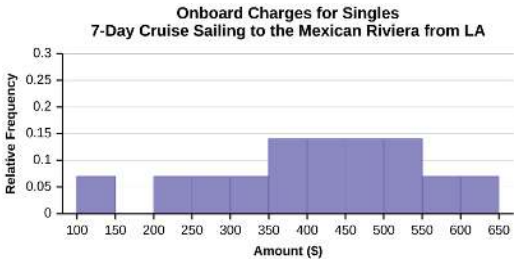


Figure 9.E.7:

d. Compare the two graphs:

i. Answers may vary. Possible answers include:

- Both graphs have a single peak.
- Both graphs use class intervals with width equal to $50.

ii. Answers may vary. Possible answers include:

- The couples graph has a class interval with no values.
- It takes almost twice as many class intervals to display the data for couples.

iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the Singles with the new graph for the Couples:
   i.
      - Both graphs have a single peak.
      - Both graphs display 6 class intervals.
      - Both graphs show the same general pattern.
   ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Q 2.3.3

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.

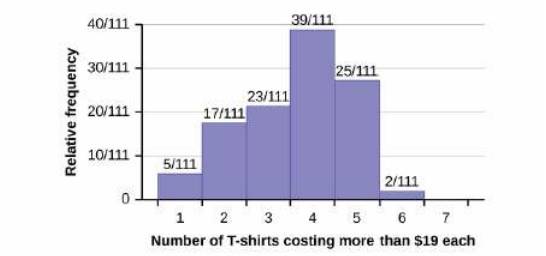


Figure 9.*E*. 8:

Q 2.3.4

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

S 2.3.4

c

Q 2.3.5

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

Q 2.3.6

Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the x-axis with the states.

S 2.3.7

Answers will vary.

Exercise 2.3.6

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

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Exercise 2.3.7

What does the frequency column in the Table above sum to? Why?

Answer

65

Exercise 2.3.8

What does the relative frequency column in in the Table above sum to? Why?

Exercise 2.3.9

What is the difference between relative frequency and frequency for each data value in in the Table above ?

Answer

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

Exercise 2.3.10

What is the difference between cumulative relative frequency and relative frequency for each data value?

Exercise 2.3.11

To construct the histogram for the data in in the Table above , determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Figure 9.E.9:

Answer

Answers will vary. One possible histogram is shown:

Figure 9.E.10:

Exercise 2.3.12

Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1

Pulse Rates for Women	Frequency
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Exercise 2.3.13

Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Answer

Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.

Figure 9.*E*.11:

Exercise 2.3.14

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Exercise 2.3.15

Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Sex/Year	1871	1870	1872	1871	1872	1827	1874	1875
Female	56,099	56,431	57,472	56,099	57,472	58,233	60,109	60,146
Male	60,029	58,959	61,293	60,029	61,293	61,467	63,602	63,432
Total	116,128	115,390	118,765	116,128	118,765	119,700	123,711	123,578

Answer

Figure 9.E.12:

Exercise 2.3.16

https://stats.libretexts.org/@go/page/24877

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

2.4: Measures of the Location of the Data

Q 2.4.1

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Q 2.4.2

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in the Table below. Also, include left endpoint, but not the right endpoint.

Salary ($)	**Relative Frequency**
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
    i. Should all bars have the same width, based on the data? Why or why not?
    ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40^{th} and 80^{th} percentiles
e. Construct a bar graph of the data

S 2.4.2

a. $1-(0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01)=0.06$
b. $0.19+0.26+0.18=0.63$
c. Check student's solution.
d. 40th percentile will fall between 30,000 and 40,000

80th percentile will fall between 50,000 and 75,000

e. Check student's solution.

Q 2.4.3

Given the following box plot:

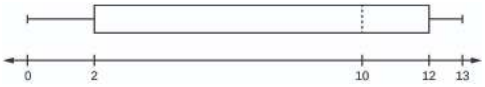


Figure 9.E.13:

a. which quarter has the smallest spread of data? What is that spread?
b. which quarter has the largest spread of data? What is that spread?
c. find the interquartile range (*IQR*).
d. are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
e. which interval has the fewest data in it? How do you know this?
   i. 0–2
   ii. 2–4
   iii. 10–12
   iv. 12–13
   v. need more information

Q 2.4.4

The following box plot shows the U.S. population for 1990, the latest available year.

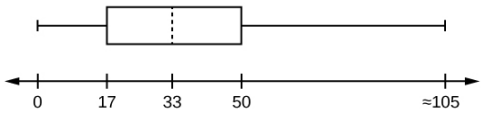


Figure 9.E.14:

a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

S 2.4.4

a. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
b. 62.4%

2.5: Box Plots

Q 2.5.1

In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.

https://stats.libretexts.org/@go/page/24877

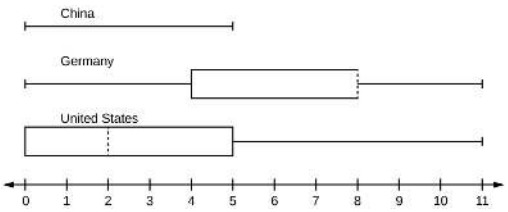


Figure 9.E.15:

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

Q 2.5.2

Given the following box plot, answer the questions.

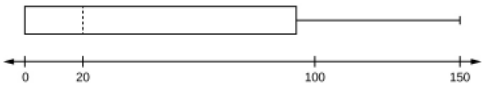


Figure 9.E.16:

a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

S 2.5.2

a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

Q 2.5.3

Given the following box plots, answer the questions.

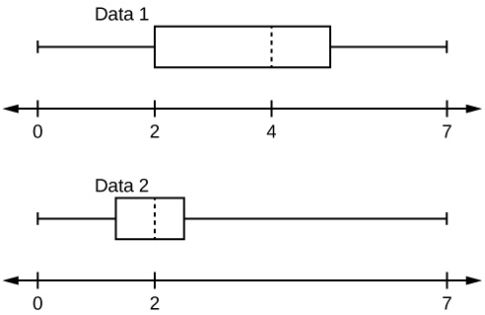


Figure 9.E.17:

a. In complete sentences, explain why each statement is false.
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.
b. For which group, Data 1 or Data 2, is the value of “7” more likely to be an outlier? Explain why in complete sentences.

Q 2.5.4

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.

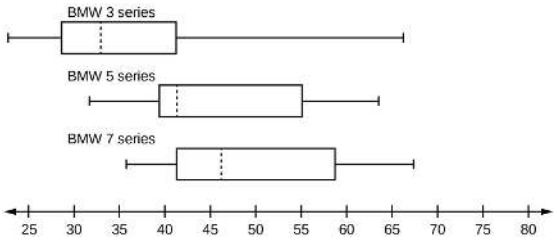


Figure 9.E.18:

a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 series. Estimate the interquartile range (IQR).
g. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?
    i. 31–35
    ii. 38–41
    iii. 41–64

S 2.5.4

a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.
b. The BMW 3 series is most likely to have an outlier. It has the longest whisker.
c. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
f. $IQR \sim 17$ years
g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

Q 2.5.5

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Construct a box plot of the data.

2.6: Measures of the Center of the Data

Q 2.6.1

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

Q 2.6.2

The table below gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

S 2.6.2

The mean percentage, $\bar{x} = \frac{1328.65}{50} = 26.75$

2.7: Skewness and the Mean, Median, and Mode

Q 2.7.1

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.8: Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- $\mu = 1000$ FTES
- median = 1,014 FTES
- $\sigma = 474$ FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- $n = 29$ years

Q 2.8.1

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

S 2.8.1

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Q 2.8.2

75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

Q 2.8.3

The population standard deviation = _____

S 2.8.3

474 FTES

Q 2.8.4

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Q 2.8.5

What is the *IQR*? What does the *IQR* represent?

S 2.8.5

919

Q 2.8.6

How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Q 2.8.7

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the*IQR*. Round to one decimal place.

S 2.8.7

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935

- *IQR* = 245

Q 2.8.8

Construct a box plot for the FTES for 2005–2006 through 2010–2011 and a box plot for the FTES for 1976–1977 through 2004–2005.

Q 2.8.9

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

S 2.8.10

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Q 2.8.11

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Q 2.8.12

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

S 2.8.12

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Q 2.8.13

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Q 2.8.14

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the table belo2

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13

Percent of Population Obese	Number of Countries
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

S 2.8.14

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that $23.32 + 12.95 = 36.27$is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not have an unusually high percentage of obese people.

Q 2.8.15

The Table below gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

This page titled 9.E: Descriptive Statistics (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

CHAPTER OVERVIEW

10: Probability Topics

Probability theory is concerned with probability, the analysis of random phenomena. The central objects of probability theory are random variables, stochastic processes, and events: mathematical abstractions of non-deterministic events or measured quantities that may either be single occurrences or evolve over time in an apparently random fashion.

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 10: Probability Topics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.1: Introduction

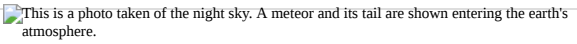


Figure 3.1.1. Meteor showers are rare, but the probability of them occurring can be calculated. (credit: Navicore/flickr)

CHAPTER OBJECTIVES

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

- Understand and use the terminology of probability.
- Determine whether two events are mutually exclusive and whether two events are independent.
- Calculate probabilities using the Addition Rules and Multiplication Rules.
- Construct and interpret Contingency Tables.
- Construct and interpret Venn Diagrams.
- Construct and interpret Tree Diagrams.

It is often necessary to "guess" about the outcome of an event in order to make a decision. Politicians study polls to guess their likelihood of winning an election. Teachers choose a particular course of study based on what they think students can comprehend. Doctors choose the treatments needed for various diseases based on their assessment of likely results. You may have visited a casino where people play games chosen because of the belief that the likelihood of winning is good. You may have chosen your course of study based on the probable availability of jobs.

You have, more than likely, used probability. In fact, you probably have an intuitive sense of probability. Probability deals with the chance of an event occurring. Whenever you weigh the odds of whether or not to do your homework or to study for an exam, you are using probability. In this chapter, you will learn how to solve probability problems using a systematic approach.

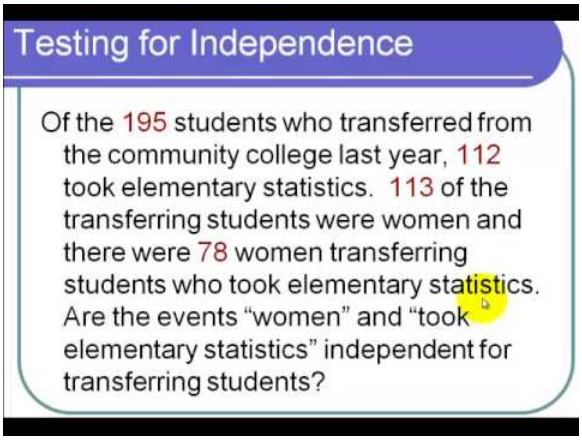
COLLABORATIVE EXERCISE

Your instructor will survey your class. Count the number of students in the class today.

- Raise your hand if you have any change in your pocket or purse. Record the number of raised hands.
- Raise your hand if you rode a bus within the past month. Record the number of raised hands.
- Raise your hand if you answered "yes" to BOTH of the first two questions. Record the number of raised hands.

Use the class data as estimates of the following probabilities. $P(\text{change})$ means the probability that a randomly chosen person in your class has change in his/her pocket or purse. $P(\text{bus})$ means the probability that a randomly chosen person in your class rode a bus within the last month and so on. Discuss your answers.

- Find $P(\text{change})$.
- Find $P(\text{bus})$.
- Find $P(\text{change AND bus})$. Find the probability that a randomly chosen student in your class has change in his/her pocket or purse and rode a bus within the last month.
- Find $P(\text{change}|\text{bus})$. Find the probability that a randomly chosen student has change given that he or she rode a bus within the last month. Count all the students that rode a bus. From the group of students who rode a bus, count those who have change. The probability is equal to those who have change and rode a bus divided by those who rode a bus.



This page titled 10.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.2: Terminology

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter *S* is used to denote the sample space. For example, if you flip one fair coin, $S = \{\text{H, T}\}$ where $\text{H} =$ heads and $\text{T} =$ tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(\text{A})$.

Definition: probability

The *probability* of any outcome is the long-term relative frequency of that outcome. Probabilities are between zero and one, inclusive (that is, zero and one and all numbers between these values).

- $P(\text{A}) = 0$ means the event A can never happen.
- $P(\text{A}) = 1$ means the event A always happens.
- $P(\text{A}) = 0.5$ means the event A is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{\text{HH, TH, HT, TT}\}$ where $\text{T} =$ tails and $\text{H} =$ heads. The sample space has four outcomes. $\text{A} =$ getting one head. There are two outcomes that meet this condition $\{\text{HT, TH}\}$, so $P(\text{A}) = \frac{2}{4} = 0.5$.

Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event $\text{E} =$ rolling a number that is at least five. There are two outcomes {5, 6}. $P(\text{E}) = \frac{2}{6}$. If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, $\frac{2}{6}$ of the rolls would result in an outcome of "at least five". You would not expect exactly $\frac{2}{6}$. The long-term relative frequency of obtaining this result would approach the theoretical probability of $\frac{2}{6}$ as the number of repetitions grows larger and larger.

Definition: law of large numbers

This important characteristic of probability experiments is known as the law of large numbers which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as

the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

The "OR" Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. For example, let $A = \{1, 2, 3, 4, 5\}$ and $B = \{4, 5, 6, 7, 8\}$. A OR B $= \{1, 2, 3, 4, 5, 6, 7, 8\}$. Notice that 4 and 5 are NOT listed twice.

The "AND" Event

An outcome is in the event A AND B if the outcome is in both A and B at the same time. For example, let A and B be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then A AND B $= 4, 5$.

The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that $P(A) + P(A') = 1$. For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = 1, 2, 3, 4$. Then, $A' = 5, 6$. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, and $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

The conditional probability of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space**. We calculate the probability of A from the reduced sample space B. The formula to calculate $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let A = face is 2 or 3 and B = face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes 2 or 3 in the sample space $B = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{the number of outcomes that are 2 or 3 and even in S})}{6}}{\frac{(\text{the number of outcomes that are even in S})}{6}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3} \quad (10.2.1)$$

Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

Example 10.2.1

The sample space S is the whole numbers starting at one and less than 20.

a. $S =$ _____

Let event $A =$ the even numbers and event $B =$ numbers greater than 13.

b. $A =$ _____, $B =$ _____
c. $P(A) =$ _____, $P(B) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(A \text{ AND } B) =$ _____, $P(A \text{ OR } B) =$ _____
f. $A' =$ _____, $P(A') =$ _____
g. $P(A) + P(A') =$ _____
h. $P(A|B) =$ _____, $P(B|A) =$ _____; are the probabilities equal?

Answer

a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}, B = \{14, 15, 16, 17, 18, 19\}$
c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
d. A AND B $= \{14, 16, 18\}$, A OR B $= \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
e. $P(A \text{ AND } B) = \frac{3}{19}$, $P(A \text{ OR } B) = \frac{12}{19}$

f. $\mathrm{A}' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$, $P(\mathrm{A}') = \frac{10}{19}$
g. $P(\mathrm{A}) + P(\mathrm{A}') = 1$ $\left(\left(\frac{9}{19} + \frac{10}{19} = 1\right)\right)$
h. $P(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(B)}} = \frac{3}{6}$, $P(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P(A\ AND\ B)}}{\mathrm{P(A)}} = \frac{3}{9}$, No

Exercise 10.2.1

The sample space *S* is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

a. $S =$ _____
Let event $A =$ the sum is even and event $B =$ the first number is prime.
b. $A =$ _____, $B =$ _____
c. $P(\mathrm{A}) =$ _____, $P(\mathrm{B}) =$ _____
d. A AND B = _____, A OR B = _____
e. $P(\mathrm{A\ AND\ B}) =$ _____, $P(\mathrm{A\ OR\ B}) =$ _____
f. $\mathrm{B}' =$ _____, $P(\mathrm{B}') =$ _____
g. $P(\mathrm{A}) + P(\mathrm{A}') =$ _____
h. $P(\mathrm{A}|\mathrm{B}) =$ _____, $P(\mathrm{B}|\mathrm{A}) =$ _____; are the probabilities equal?

Answer

a. $\mathrm{S} = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
b. $\mathrm{A} = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$
$\mathrm{B} = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
c. $P(\mathrm{A}) = \frac{1}{2}$, $P(\mathrm{B}) = \frac{2}{3}$
d. A AND B = $\{(2,2), (2,4), (3,1), (3,3)\}$
A OR B = $\{(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
e. $P(\mathrm{A\ AND\ B}) = \frac{1}{3}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
f. $\mathrm{B}' = \{(1,1), (1,2), (1,3), (1,4)\}$, $P(\mathrm{B}') = \frac{1}{3}$
g. $P(\mathrm{B}) + P(\mathrm{B}') = 1$
h. $P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{1}{2}$, $P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{A\ AND\ B})}{P(\mathrm{B})} = \frac{2}{3}$, No.

Example 10.2.2*A*

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

a. Event T = the outcome is two.
b. Event A = the outcome is an even number.
c. Event B = the outcome is less than four.
d. The complement of A.
e. A GIVEN B
f. B GIVEN A
g. A AND B
h. A OR B
i. A OR B′
j. Event N = the outcome is a prime number.
k. Event I = the outcome is seven.

Solution

a. $\mathrm{T} = \{2\}$, $P(\mathrm{T}) = \frac{1}{6}$
b. $A = \{2, 4, 6\}$, $P(\mathrm{A}) = \frac{1}{2}$
c. $\mathrm{B} = \{1, 2, 3\}$, $P(\mathrm{B}) = \frac{1}{2}$
d. $\mathrm{A}' = \{1, 3, 5\}$, $P(\mathrm{A}') = \frac{1}{2}$
e. $\mathrm{A}|\mathrm{B} = \{2\}$, $P(\mathrm{A}|\mathrm{B}) = \frac{1}{3}$

f. $\mathrm{B}|\mathrm{A} = \{2\}$, $P(\mathrm{B}|\mathrm{A}) = \frac{1}{3}$
g. $\mathrm{A\ AND\ B} = 2$, $P(\mathrm{A\ AND\ B}) = \frac{1}{6}$
h. $\mathrm{A\ OR\ B} = \{1, 2, 3, 4, 6\}$, $P(\mathrm{A\ OR\ B}) = \frac{5}{6}$
i. $\mathrm{A\ OR\ B'} = \{2, 4, 5, 6\}$, $P(\mathrm{A\ OR\ B'}) = \frac{2}{3}$
j. $\mathrm{N} = \{2, 3, 5\}$, $P(\mathrm{N}) = \frac{1}{2}$
k. A six-sided die does not have seven dots. $P(7) = 0$.

Example 10.2.2B

Table describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events $M =$ the subject is male, $F =$ the subject is female, $R =$ the subject is right-handed, $L =$ the subject is left-handed. Compute the following probabilities:

a. $P(\mathrm{M})$
b. $P(\mathrm{F})$
c. $P(\mathrm{R})$
d. $P(\mathrm{L})$
e. $P(\mathrm{M\ AND\ R})$
f. $P(\mathrm{F\ AND\ L})$
g. $P(\mathrm{M\ OR\ F})$
h. $P(\mathrm{M\ OR\ R})$
i. $P(\mathrm{F\ OR\ L})$
j. $P(\mathrm{M'})$
k. $P(\mathrm{R}|\mathrm{M})$
l. $P(\mathrm{F}|\mathrm{L})$
m. $P(\mathrm{L}|\mathrm{F})$

Answer

a. $P(\mathrm{M}) = 0.52$
b. $P(\mathrm{F}) = 0.48$
c. $P(\mathrm{R}) = 0.87$
d. $P(\mathrm{L}) = 0.13$
e. $P(\mathrm{M\ AND\ R}) = 0.43$
f. $P(\mathrm{F\ AND\ L}) = 0.04$
g. $P(\mathrm{M\ OR\ F}) = 1$
h. $P(\mathrm{M\ OR\ R}) = 0.96$
i. $P(\mathrm{F\ OR\ L}) = 0.57$
j. $P(\mathrm{M'}) = 0.48$
k. $P(\mathrm{R}|\mathrm{M}) = 0.8269$ (rounded to four decimal places)
l. $P(\mathrm{F}|\mathrm{L}) = 0.3077$ (rounded to four decimal places)
m. $P(\mathrm{L}|\mathrm{F}) = 0.0833$

References

1. "Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(\text{S}) = 1$ where S is the sample space

$0 \leq P(\text{A}) \leq 1$

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

Glossary

Conditional Probability

the likelihood that an event will occur given that another event has already occurred

Equally Likely

Each outcome of an experiment has the same probability.

Event

a subset of the set of all outcomes of an experiment; the set of all outcomes of an experiment is called a **sample space** and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment

a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

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

Sample Space

the set of all possible outcomes of an experiment

The AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Complement Event

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

The Conditional Probability of *A* GIVEN *B*

$P(\text{A}|\text{B})$ is the probability that event A will occur given that the event B has already occurred.

The Or Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

Exercise 3.2.2

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

a. The probability that a student does not have long hair.
b. The probability that a student is male or has short hair.
c. The probability that a student is a female and has long hair.
d. The probability that a student is male, given that the student has long hair.
e. The probability that a student has long hair, given that the student is male.
f. Of all the female students, the probability that a student has short hair.
g. Of all students with long hair, the probability that a student is female.
h. The probability that a student is female or has long hair.
i. The probability that a randomly selected student is a male student with short hair.
j. The probability that a student is female.

Answer

a. $P(\text{L}') = P(\text{S})$
b. $P(\text{M OR S})$
c. $P(\text{F AND L})$
d. $P(\text{M}|\text{L})$
e. $P(\text{L}|\text{M})$
f. $P(\text{S}|\text{F})$
g. $P(\text{F}|\text{L})$
h. $P(\text{F OR L})$
i. $P(\text{M AND S})$
j. $P(\text{F})$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.

Let $H =$ the event of getting a hat.

Let $N =$ the event of getting a noisemaker.

Let $F =$ the event of getting a finger trap.

Let $C =$ the event of getting a bag of confetti.

Exercise 3.2.3

Find $P(\text{H})$.

Exercise 3.2.4

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{15}{42} = \frac{5}{14} = 0.36$

Exercise 3.2.5

Find $P(\text{F})$.

Exercise 3.2.6

Find $P(\text{C})$.

Answer

$P(\text{C}) = \frac{5}{42} = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let $B =$ the event of getting a blue jelly bean

Let $G =$ the event of getting a green jelly bean.

Let $O =$ the event of getting an orange jelly bean.

Let $P =$ the event of getting a purple jelly bean.

Let $R =$ the event of getting a red jelly bean.

Let $Y =$ the event of getting a yellow jelly bean.

Exercise 3.2.7

Find $P(\text{B})$.

Exercise 3.2.8

Find $P(\text{G})$.

Answer

$P(\text{G}) = \frac{20}{150} = \frac{2}{15} = 0.13$

Exercise 3.2.9

Find $P(\text{P})$.

Exercise 3.2.10

Find $P(\text{R})$.

Answer

$P(\text{R}) = \frac{22}{150} = \frac{11}{75} = 0.15$

Exercise 3.2.11

Find $P(\text{Y})$.

Exercise 3.2.12

Find $P(\text{O})$.

Answer

$P(textO) = \frac{150-22-38-20-28-26}{150} = \frac{16}{150} = \frac{8}{75} = 0.11$

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).

Let $\text{A} =$ the event that a country is in Asia.

Let $\text{E} =$ the event that a country is in Europe.

Let F = the event that a country is in Africa.

Let N = the event that a country is in North America.

Let O = the event that a country is in Oceania.

Let S = the event that a country is in South America.

Exercise 3.2.13

Find $P(\text{A})$.

Exercise 3.2.14

Find $P(\text{E})$.

Answer

$P(\text{E}) = \frac{47}{194} = 0.24$

Exercise 3.2.15

Find $P(\text{F})$.

Exercise 3.2.16

Find $P(\text{N})$.

Answer

$P(\text{N}) = \frac{23}{194} = 0.12$

Exercise 3.2.17

Find $P(\text{O})$.

Exercise 3.2.18

Find $P(\text{S})$.

Answer

$P(\text{S}) = \frac{12}{194} = \frac{6}{97} = 0.06$

Exercise 3.2.19

What is the probability of drawing a red card in a standard deck of 52 cards?

Exercise 3.2.20

What is the probability of drawing a club in a standard deck of 52 cards?

Answer

$\frac{13}{52} = \frac{1}{4} = 0.25$

Exercise 3.2.21

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

Exercise 3.2.22

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

Answer

$\frac{3}{6} = \frac{1}{2} = 0.5$

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.

Figure 3.2.1.

Let B = the event of landing on blue.

Let R = the event of landing on red.

Let G = the event of landing on green.

Let Y = the event of landing on yellow.

Exercise 3.2.23

If you land on Y, you get the biggest prize. Find $P(\mathrm{Y})$.

Exercise 3.2.24

If you land on red, you don't get a prize. What is $P(\mathrm{R})$?

Answer

$\mathrm{P}(R) = \frac{4}{8} = 0.5$

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.

Let I = the event that a player in an infielder.

Let O = the event that a player is an outfielder.

Let H = the event that a player is a great hitter.

Let N = the event that a player is not a great hitter.

Exercise 3.2.25

Write the symbols for the probability that a player is not an outfielder.

Exercise 3.2.26

Write the symbols for the probability that a player is an outfielder or is a great hitter.

Answer

$P(\mathrm{O\ OR\ H})$

Exercise 3.2.27

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Exercise 3.2.28

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

Answer

$P(\mathrm{H}|\mathrm{I})$

Exercise 3.2.29

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Exercise 3.2.30

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

Answer

$P(\mathrm{N}|\mathrm{O})$

Exercise 3.2.31

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Exercise 3.2.32

Write the symbols for the probability that a player is an infielder or is not a great hitter.

Answer

$P(\text{I OR N})$

Exercise 3.2.33

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Exercise 3.2.34

Write the symbols for the probability that a player is an infielder.

Answer

$P(\text{I})$

Exercise 3.2.35

What is the word for the set of all possible outcomes?

Exercise 3.2.36

What is conditional probability?

Answer

The likelihood that an event will occur given that another event has already occurred.

Exercise 3.2.37

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book

Let $\text{F} =$ event that book is fiction

Let $\text{N} =$ event that book is nonfiction

What is the sample space?

Exercise 3.2.38

What is the sum of the probabilities of an event and its complement?

Answer

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let $\text{E} =$ the event that it lands on an even number. Let $\text{M} =$ the event that it lands on a multiple of three.

Exercise 3.2.39

What does $P(\text{E}|\text{M})$ mean in words?

Exercise 3.2.40

What does $P(\text{E OR M})$ mean in words?

Answer

the probability of landing on an even number or a multiple of three

This page titled 10.2: Terminology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.3: Independent and Mutually Exclusive Events

Independent and *mutually exclusive* do not mean the same thing.

Independent Events

Two events are independent if the following are true:

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

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show only one of the above conditions. If two events are NOT independent, then we say that they are dependent.

Sampling a population

Sampling may be done with replacement or without replacement (Figure $10.3.1$):

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be *independent*, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be *dependent* or *not independent*.

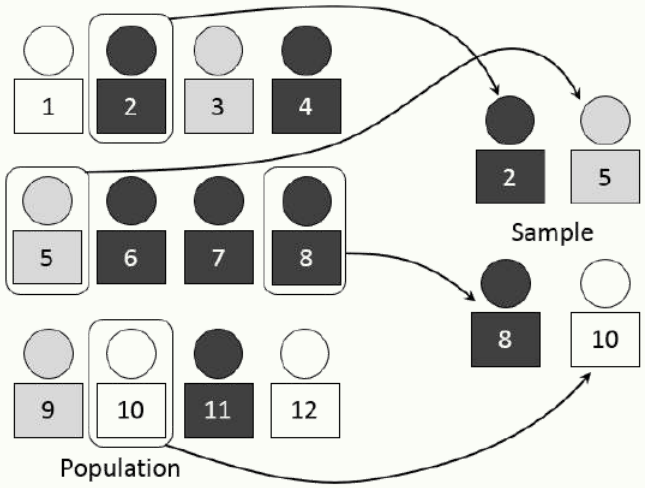
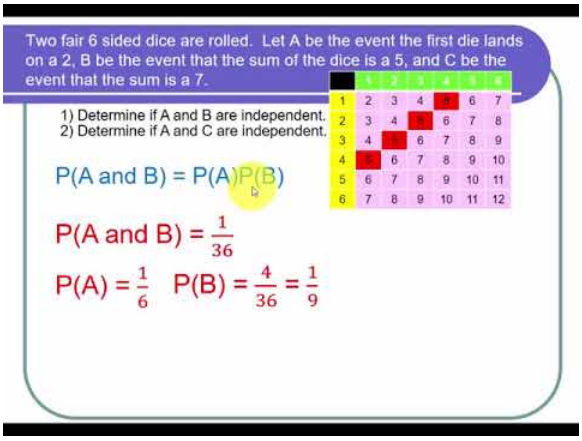


Figure 10.3.1: *A visual representation of the sampling process. If the sample items are replaced after each sampling event, then this is "sampling with replacement" if not, then it is "sampling without replacement". Image used with permission (CC BY-SA 4.0; Dan Kernler).*

If it is not known whether A and B are independent or dependent, assume they are dependent until you can show otherwise.



Example 10.3.1: Sampling with and without replacement

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the Q of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the Q of spades again. Your picks are {Q of spades, ten of clubs, Q of spades}. You have picked the Q of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the K of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the J of spades. Your picks are {K of hearts, three of diamonds, J of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

Exercise 10.3.1

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), K (king) of that suit. Three cards are picked at random.

a. Suppose you know that the picked cards are Q of spades, K of hearts and Q of spades. Can you decide if the sampling was with or without replacement?
b. Suppose you know that the picked cards are Q of spades, K of hearts, and J of spades. Can you decide if the sampling was with or without replacement?

Answer

a. With replacement
b. No

Example 10.3.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs.

a. Suppose you pick four cards, but do not put any cards back into the deck. Your cards are QS, 1D, 1C, QD.
b. Suppose you pick four cards and put each card back before you pick the next card. Your cards are KH, 7D, 6D, KH.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

Solution

a. Without replacement
b. With replacement

Exercise 10.3.2

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (jack), Q (queen), and K (king) of that suit. S = spades, H = Hearts, D = Diamonds, C = Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

a. QS, 1D, 1C, QD
b. KH, 7D, 6D, KH
c. QS, 7D, 6D, KS

Answer

without replacement: a. Possible; b. Impossible, c. Possible

with replacement: a. Possible; c. Possible, c. Possible

Mutually Exclusive Events

A and B are mutually exclusive events if they **cannot** occur at the same time. This means that A and B do not share any outcomes and $P(\text{A AND B}) = 0$.

For example, suppose the sample space

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}. \tag{10.3.1}$$

Let $\text{A} = \{1, 2, 3, 4, 5\}$, $\text{B} = \{4, 5, 6, 7, 8\}$ and $\text{C} = \{7, 9\}$. A AND B $= \{4, 5\}$.

$$P(\text{A AND B}) = \frac{2}{10} \tag{10.3.2}$$

and is not equal to zero. Therefore, A and B are not mutually exclusive. A and C do not have any numbers in common so $P(\text{A AND C}) = 0$. Therefore, A and C are mutually exclusive.

If it is not known whether A and B are mutually exclusive, **assume they are not until you can show otherwise**. The following examples illustrate these definitions and terms.

Example 10.3.3

Flip two fair coins.

The sample space is $\{HH, HT, TH, TT\}$ where $T =$ tails and $H =$ heads. The outcomes are HH, HT, TH, and TT. The outcomes HT and TH are different. The HT means that the first coin showed heads and the second coin showed tails. The TH means that the first coin showed tails and the second coin showed heads.

- Let A = the event of getting **at most one tail**. (At most one tail means zero or one tail.) Then A can be written as $\{HH, HT, TH\}$. The outcome HH shows zero tails. HT and TH each show one tail.
- Let B = the event of getting all tails. B can be written as $\{TT\}$. B is the **complement** of A, so B = A′. Also, $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$.
- The probabilities for A and for B are $P(\text{A}) = \frac{3}{4}$ and $P(\text{B}) = \frac{1}{4}$.
- Let C = the event of getting all heads. C = $\{HH\}$. Since B = $\{TT\}$, $P(\text{B AND C}) = 0$. B and Care mutually exclusive. B and C have no members in common because you cannot have all tails and all heads at the same time.)
- Let D = event of getting **more than one** tail. D = $\{TT\}$. $P(\text{D}) = \frac{1}{4}$
- Let E = event of getting a head on the first roll. (This implies you can get either a head or tail on the second roll.) E = $\{HT, HH\}$. $P(\text{E}) = \frac{2}{4}$
- Find the probability of getting **at least one** (one or two) tail in two flips. Let F = event of getting at least one tail in two flips. F = $\{HT, TH, TT\}$. $P(\text{F}) = \frac{3}{4}$

Exercise 10.3.3

Draw two cards from a standard 52-card deck with replacement. Find the probability of getting at least one black card.

Answer

The sample space of drawing two cards with replacement from a standard 52-card deck with respect to color is $\{BB, BR, RB, RR\}$.

Event $A =$ Getting at least one black card = $\{BB, BR, RB\}$

$P(\text{A}) = \frac{3}{4} = 0.75$

Example 10.3.4

Flip two fair coins. Find the probabilities of the events.

a. Let F = the event of getting at most one tail (zero or one tail).
b. Let G = the event of getting two faces that are the same.
c. Let H = the event of getting a head on the first flip followed by a head or tail on the second flip.
d. Are F and G mutually exclusive?
e. Let J = the event of getting all tails. Are J and H mutually exclusive?

Solution

Look at the sample space in Example 10.3.3

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(\text{F}) = \frac{3}{4}$
2. Two faces are the same if HH or TT show up. $P(\text{G}) = \frac{2}{4}$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(\text{H}) = \frac{2}{4}$
4. F and G share HH so $P(\text{F AND G})$ is not equal to zero (0). F and G are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (TT). H's outcomes are HH and HT.

J and H have nothing in common so $P(\text{J AND H}) = 0$. J and H are mutually exclusive.

Exercise 10.3.4

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

a. Let F = the event of getting the white ball twice.
b. Let G = the event of getting two balls of different colors.
c. Let H = the event of getting white on the first pick.
d. Are F and G mutually exclusive?
e. Are G and H mutually exclusive?

Answer

a. $P(\text{F}) = \frac{1}{4}$
b. $P(\text{G}) = \frac{1}{2}$
c. $P(\text{H}) = \frac{1}{2}$
d. Yes
e. No

Example 10.3.5

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A′. The complement of A, A′, is B because A and B together make up the sample space. $P(\text{A}) + P(\text{B}) = P(\text{A}) + P(\text{A}') = 1$. Also, $P(\text{A}) = \frac{3}{6}$ and $P(\text{B}) = \frac{3}{6}$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(\text{C AND D}) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Are C and E mutually exclusive events? (Answer yes or no.) Why or why not?

Answer

No. C = {3, 5} and E = {1, 2, 3, 4}. $P(\text{C AND E}) = \frac{1}{6}$. To be mutually exclusive, $P(\text{C AND E})$ must be zero.

- Find $P(\text{C}|\text{A})$. This is a conditional probability. Recall that the event C is {3, 5} and event A is {1, 3, 5}. To find $P(\text{C}|\text{A})$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space {1, 2, 3, 4, 5, 6} to {1, 3, 5}. So, $P(\text{C}|\text{A}) = \frac{2}{3}$.

Exercise 10.3.5

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(\text{A}) = 0.4$ and $P(\text{B}) = 0.2$. $P(\text{A AND B}) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

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

Answer

$$P(\text{A}|\text{B}) = \frac{\text{P(A AND B)}}{P(\text{B})} = \frac{0.08}{0.2} = 0.4 = P(\text{A}) \quad (10.3.3)$$

The events are independent because $P(\text{A}|\text{B}) = P(\text{A})$.

Example 10.3.6

Let event G = taking a math class. Let event H = taking a science class. Then, G AND H = taking a math class and a science class. Suppose $P(\text{G}) = 0.6$, $P(\text{H}) = 0.5$, and $P(\text{G AND H}) = 0.3$. Are G and H independent?

If G and H are independent, then you must show **ONE** of the following:

- $P(\text{G}|\text{H}) = P(\text{G})$
- $P(\text{H}|\text{G}) = P(\text{H})$
- $P(\text{G AND H}) = P(\text{G})P(\text{H})$

The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. a. Show that $P(\text{G}|\text{H}) = P(\text{G})$.
b. b. Show $P(\text{G AND H}) = P(\text{G})P(\text{H})$.

Solution

a. $P(\text{G}|\text{H}) = \dfrac{P(\text{G AND H})}{P(\text{H})} = \dfrac{0.3}{0.5} = 0.6 = P(\text{G})$
b. $P(\text{G})P(\text{H}) = (0.6)(0.5) = 0.3 = P(\text{G AND H})$

Since G*and*H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(\text{H}|\text{G}) = P(\text{H})$ to show that G and H are independent events.

Exercise 10.3.6

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is S = $\{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(\text{G AND O})$?

Answer

Event G and O = $\{G1, G3\}$

$P(\text{G and O}) = \dfrac{2}{10} = 0.2$

Example 10.3.7

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(\text{C}) = 0.75$, $P(\text{D}) = 0.3$, $P(\text{C}|\text{D}) = 0.75$ and $P(\text{C AND D}) = 0.225$.

Justify your answers to the following questions numerically.

a. Are C and D independent?
b. Are C and D mutually exclusive?
c. What is $P(\text{D}|\text{C})$?

Solution

a. Yes, because $P(\text{C}|\text{D}) = P(\text{C})$.
b. No, because $P(\text{C AND D})$ is not equal to zero.
c. $P(\text{D}|\text{C}) = \dfrac{P(\text{C AND D})}{P(\text{C})} = \dfrac{0.225}{0.75} = 0.3$

Exercise 10.3.7

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{B AND D}) = 0.20$.

a. Find $P(\text{B}|\text{D})$.
b. Find $P(\text{D}|\text{B})$.
c. Are B and D independent?
d. Are B and D mutually exclusive?

Answer

a. $P(\text{B}|\text{D}) = 0.6667$
b. $P(\text{D}|\text{B}) = 0.5$
c. No
d. No

Example 10.3.8

In a box there are three red cards and five blue cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.

Let

- R = red card is drawn,
- B = blue card is drawn,
- E = even-numbered card is drawn.

The sample space $S = R1, R2, R3, B1, B2, B3, B4, B5$.

S has eight outcomes.

- $P(\text{R}) = \frac{3}{8}$. $P(\text{B}) = \frac{5}{8}$. $P(\text{R AND B}) = 0$. (You cannot draw one card that is both red and blue.)
- $P(\text{E}) = \frac{3}{8}$. (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(\text{E}|\text{B}) = \frac{2}{5}$. (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(\text{B}|\text{E}) = \frac{2}{3}$. (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(\text{R AND B}) = 0$.
- Let G = card with a number greater than 3. $\text{G} = \{B4, B5\}$. $P(\text{G}) = \frac{2}{8}$. Let H = blue card numbered between one and four, inclusive. $\text{H} = \{B1, B2, B3, B4\}$. $P(\text{G}|\text{H}) = frac14$. (The only card in H that has a number greater than three is B4.) Since $\frac{2}{8} = \frac{1}{4}$, $P(\text{G}) = P(\text{G}|\text{H})$, which means that G and H are independent.

Exercise 10.3.8

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.

Let B be the event that a fan is wearing blue.

Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

Answer

- $P(\text{B}|\text{A}) = 0.67$
- $P(\text{B}) = 0.25$

So $P(\text{B})$ does not equal $P(\text{B}|\text{A})$ which means that $\text{B} and \text{A}$ are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND A}) = 0.20$, not 0.

Example 10.3.9

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let F be the event that a student is female. Let L be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(\text{F}) = 0.60$; $P(\text{L}) = 0.50$
- $P(\text{F AND L}) = 0.45$
- $P(\text{L}|\text{F}) = 0.75$

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(\text{F}|\text{L})$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(\text{F AND L}) = P(\text{F})P(\text{L})$. We are given that $P(\text{F AND L}) = 0.45$, but $P(\text{F})P(\text{L}) = (0.60)(0.50) = 0.30$ The events of being female and having long hair are not independent because $P(\text{F AND L})$ does not equal $P(\text{F})P(\text{L})$.

Solution 2

Check whether $P(\text{L}|\text{F})$ equals $P(\text{L})$. We are given that $P(\text{L}|\text{F}) = 0.75$, but $P(\text{L}) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Exercise 10.3.9

Mark is deciding which route to take to work. His choices are I = the Interstate and F = Fifth Street.

- $P(\text{I}) = 0.44$ and $P(\text{F}) = 0.55$
- $P(\text{I AND F}) = 0$ because Mark will take only one route to work.

What is the probability of $P(\text{I OR F})$?

Answer

Because $P(\text{I AND F}) = 0$,

$P(\text{I OR F}) = P(\text{I}) + P(\text{F}) - P(\text{I AND F}) = 0.44 + 0.56 - 0 = 1$

Example 10.3.10

a. Toss one fair coin (the coin has two sides, H and T). The outcomes are _____. Count the outcomes. There are ____ outcomes.
b. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are _____. Count the outcomes. There are ___ outcomes.
c. Multiply the two numbers of outcomes. The answer is _____.
d. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are $H1$ and $T6$.)
e. Event A = heads (H) on the coin followed by an even number (2, 4, 6) on the die.
A = {_____}. Find $P(\text{A})$.
f. Event B = heads on the coin followed by a three on the die. B = {_____}. Find $P(\text{B})$.

g. Are A and B mutually exclusive? (Hint: What is $P(\text{A AND B})$? If $P(\text{A AND B}) = 0$, then A and B are mutually exclusive.)
h. Are A and B independent? (Hint: Is $P(\text{A AND B}) = P(\text{A})P(\text{B})$? If $P(\text{A AND B}) = P(\text{A})P(\text{B})$, then A and B are independent. If not, then they are dependent).

Solution

a. H and T; 2
b. 1, 2, 3, 4, 5, 6; 6
c. 2(6) = 12
d. $T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6$
e. $\text{A} = \{H2, H4, H6\}$; $P(\text{A}) = \frac{3}{12}$
f. $\text{B} = \{H3\}$; $P(\text{B}) = \frac{1}{12}$
g. Yes, because $P(\text{A AND B}) = 0$
h. $P(\text{A AND B}) = 0$. $P(\text{A})P(\text{B}) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(\text{A AND B})$ does not equal $P(\text{A})P(\text{B})$, so A and B are dependent.

Exercise 10.3.10

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let $textT$ be the event of getting the white ball twice, F the event of picking the white ball first, S the event of picking the white ball in the second drawing.

a. Compute $P(\text{T})$.
b. Compute $P(\text{T}|\text{F})$.
c. Are $textT$ and F independent?.
d. Are F and S mutually exclusive?
e. Are F and S independent?

Answer

a. $P(\text{T}) = \frac{1}{4}$
b. $P(\text{T}|\text{F}) = \frac{1}{2}$
c. No
d. No
e. Yes

References

1. Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/te...workplace.aspx (accessed May 2, 2013).
2. Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

- If A and B are independent, $P(\text{A AND B}) = P(\text{A})P(\text{B})$, $P(\text{A}|\text{B}) = P(\text{A})$ and $P(\text{B}|\text{A}) = P(\text{B})$.
- If A and B are mutually exclusive, $P(\text{A OR B}) = P(textA) + P(\text{B})andP(\text{A AND B}) = 0$.

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

Exercise 3.3.11

E and F are mutually exclusive events. $P(\text{E}) = 0.4$; $P(\text{F}) = 0.5$. Find $P(\text{E}|\text{F})$.

Exercise 3.3.12

J and K are independent events. $P(\text{J}|\text{K}) = 0.3$. Find $P(\text{J})$.

Answer

$P(\text{J}) = 0.3$

Exercise 3.3.13

U and V are mutually exclusive events. $P(\text{U}) = 0.26$; $P(\text{V}) = 0.37$. Find:

a. $P(\text{U AND V}) =$
b. $P(\text{U}|\text{V}) =$
c. $P(\text{U OR V}) =$

Exercise 3.3.14

Q and R are independent events. $P(\text{Q}) = 0.4$ and $P(\text{Q AND R}) = 0.1$. Find $P(\text{R})$.

Answer

$P(\text{Q AND R}) = P(\text{Q})P(\text{R})$

$0.1 = (0.4)P(\text{R})$

$P(\text{R}) = 0.25$

Bringing It Together

Exercise 3.3.16

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into Table.

Shirt#	≤ 210	211–250	251–290	290≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt}\#1-33| \leq 210 \text{ pounds})$?

Exercise 3.3.17

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write "not enough information" for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

a. $P(\text{C}) =$ _____
b. $P(\text{P}|\text{C}) =$ _____
c. $P(\text{P}|\text{C'}) =$ _____
d. If a test comes up positive, based upon numerical values, can you assume that man has cancer? Justify numerically and explain why or why not.

Answer

a. $P(\text{C}) = 0.4567$
b. not enough information
c. not enough information
d. No, because over half (0.51) of men have at least one false positive text

Exercise 3.3.18

Given events G and H : $P(\text{G}) = 0.43$; $P(\text{H}) = 0.26$; $P(\text{H AND G}) = 0.14$

a. Find $P(\text{H OR G})$.
b. Find the probability of the complement of event (H AND G).
c. Find the probability of the complement of event (H OR G).

Exercise 3.3.19

Given events J and K : $P(\text{J}) = 0.18$; $P(\text{K}) = 0.37$; $P(\text{J OR K}) = 0.45$

a. Find $P(\text{J AND K})$.
b. Find the probability of the complement of event (J AND K).
c. Find the probability of the complement of event (J AND K).

Answer

a. $P(\text{J OR K}) = P(\text{J}) + P(\text{K}) - P(\text{J AND K}); 0.45 = 0.18 + 0.37 - P(\text{J AND K})$; solve to find $P(\text{J AND K}) = 0.10$
b. $P(\text{NOT (J AND K)}) = 1 - P(\text{J AND K}) = 1 - 0.10 = 0.90$
c. $P(\text{NOT (J OR K)}) = 1 - P(\text{J OR K}) = 1 - 0.45 = 0.55$

Glossary

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

The Conditional Probability of One Event Given Another Event

P(*A*|*B*) is the probability that event *A* will occur given that the event *B* has already occurred.

The OR of Two Events

An outcome is in the event *A* OR *B* if the outcome is in *A*, is in *B*, or is in both *A* and *B*.

This page titled 10.3: Independent and Mutually Exclusive Events is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.4: Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If A and B are two events defined on a sample space, then:

$$P(A \text{ AND } B) = P(B)P(A|B) \tag{10.4.1}$$

This rule may also be written as:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} \tag{10.4.2}$$

(The probability of A given B equals the probability of A and B divided by the probability of B.)

If A and B are *independent*, then

$$P(A|B) = P(A). \tag{10.4.3}$$

and Equation 10.4.1 becomes

$$P(A \text{ AND } B) = P(A)P(B). \tag{10.4.4}$$

The Addition Rule

If A and B are defined on a sample space, then:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B) \tag{10.4.5}$$

If A and B are **mutually exclusive**, then

$$P(A \text{ AND } B) = 0. \tag{10.4.6}$$

and Equation 10.4.5 becomes

$$P(A \text{ OR } B) = P(A) + P(B). \tag{10.4.7}$$

Example 10.4.1

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses A is $P(\text{A}) = 0.6$ and the probability that he chooses B is $P(\text{B}) = 0.35$.
- $P(\text{A AND B}) = 0$ because Klaus can only afford to take one vacation
- Therefore, the probability that he chooses either New Zealand or Alaska is $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) = 0.6 + 0.35 = 0.95$. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(\text{A}) = 0.65$. B = the event Carlos is successful on his second attempt. $P(\text{B}) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal **GIVEN** that he made the first goal is 0.90.

a. What is the probability that he makes both goals?
b. What is the probability that Carlos makes either the first goal or the second goal?
c. Are A and B independent?
d. Are A and B mutually exclusive?

Solutions

a. The problem is asking you to find $P(\text{A AND B}) = P(\text{B AND A})$. Since $P(\text{B}|\text{A}) = 0.90 : P(\text{B AND A}) = P(\text{B}|\text{A})P(\text{A}) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. The problem is asking you to find $P(\text{A OR B})$.

$$P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B}) = 0.65 + 0.65 - 0.585 = 0.715 \quad (10.4.8)$$

Carlos makes either the first goal or the second goal with probability 0.715.

c. No, they are not, because $P(\text{B AND A}) = 0.585$.

$$P(\text{B})P(\text{A}) = (0.65)(0.65) = 0.423 \quad (10.4.9)$$

$$0.423 \neq 0.585 = P(\text{B AND A}) \quad (10.4.10)$$

So, $P(\text{B AND A})$ is **not** equal to $P(\text{B})P(\text{A})$.

d. No, they are not because $P(\text{A and B}) = 0.585$.

To be mutually exclusive, $P(\text{A AND B})$ must equal zero.

Exercise 10.4.1

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(\text{C}) = 0.75$. D = the event Helen makes the second shot. $P(\text{D}) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

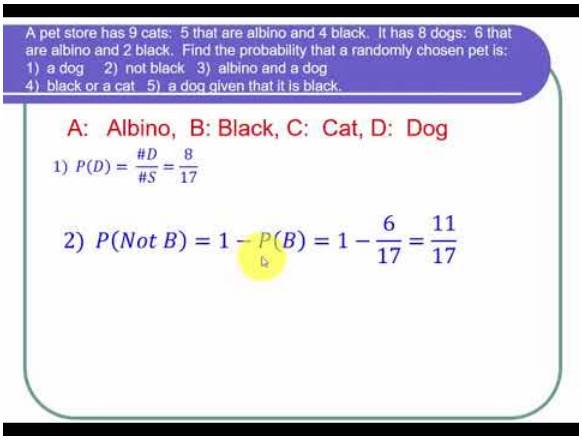
Answer

$$P(\text{D}|\text{C}) = 0.85 \quad (10.4.11)$$

$$P(\text{C AND D}) = P(\text{D AND C}) \quad (10.4.12)$$

$$P(\text{D AND C}) = P(\text{D}|\text{C})P(\text{C}) = (0.85)(0.75) = 0.6375 \quad (10.4.13)$$

Helen makes the first and second free throws with probability 0.6375.



Example 10.4.2

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?
b. What is the probability that the member practices four times a week?
c. What is the probability that the member is an advanced swimmer and practices four times a week?
d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?
e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

Answer

a. $\frac{28}{150}$
b. $\frac{80}{150}$
c. $\frac{40}{150}$
d. $P(\text{advanced AND intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.
e. No, these are not independent events.

$$P(\text{novice AND practices four times per week}) = 0.0667 \tag{10.4.14}$$

$$P(\text{novice})P(\text{practices four times per week}) = 0.0996 \tag{10.4.15}$$

$$0.0667 \neq 0.0996 \tag{10.4.16}$$

Exercise 10.4.2

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

Answer

$$P = \frac{200 - 140 - 40}{200} = \frac{20}{200} = 0.1 \tag{10.4.17}$$

Example 10.4.3

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class GIVEN that she enrolls in speech class is 0.25.

Let: M = math class, S = speech class, M|S = math given speech

a. What is the probability that Felicity enrolls in math and speech?
   Find $P(\text{M AND S}) = P(\text{M}|\text{S})P(\text{S})$.
b. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(\text{M OR S}) = P(\text{M}) + P(\text{S}) - P(\text{M AND S})$.
c. Are M and S independent? Is $P(\text{M}|\text{S}) = P(\text{M})$?
d. Are M and S mutually exclusive? Is $P(\text{M AND S}) = 0$?

Answer

a. 0.1625, b. 0.6875, c. No, d. No

Exercise 10.4.3

A student goes to the library. Let events B = the student checks out a book and D = the student check out a DVD. Suppose that $P(\text{B}) = 0.40$, $P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B AND D})$.
b. Find $P(\text{B OR D})$.

Answer

a. $P(\text{B AND D}) = P(\text{D}|\text{B})P(\text{B}) = (0.5)(0.4) = 0.20$.
b. $P(\text{B OR D}) = P(\text{B}) + P(\text{D}) - P(\text{B AND D}) = 0.40 + 0.30 - 0.20 = 0.50$

Example 10.4.4

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let B = woman develops breast cancer and let N = tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
b. Given that the woman has breast cancer, what is the probability that she tests negative?
c. What is the probability that the woman has breast cancer AND tests negative?
d. What is the probability that the woman has breast cancer or tests negative?
e. Are having breast cancer and testing negative independent events?
f. Are having breast cancer and testing negative mutually exclusive?

Answers

a. $P(\text{B}) = 0.143$; $P(\text{N}) = 0.85$
b. $P(\text{N}|\text{B}) = 0.02$
c. $P(\text{B AND N}) = P(\text{B})P(\text{N}|\text{B}) = (0.143)(0.02) = 0.0029$
d. $P(\text{B OR N}) = P(\text{B}) + P(\text{N}) - P(\text{B AND N}) = 0.143 + 0.85 - 0.0029 = 0.9901$
e. No. $P(\text{N}) = 0.85$; $P(\text{N}|\text{B}) = 0.02$. So, $P(\text{N}|\text{B})$ does not equal $P(\text{N})$.
f. No. $P(\text{B AND N}) = 0.0029$. For B and N to be mutually exclusive, $P(\text{B AND N})$ must be zero

Exercise 10.4.4

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Answer

Let A = student is a senior going to college.

Let B = student plays sports.

$P(\text{B}) = \frac{140}{200}$

$P(\text{B}|\text{A}) = \frac{50}{140}$

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

$P(\text{A AND B}) = (\frac{140}{200})(\frac{50}{140}) = \frac{1}{4}$

Example 10.4.5

Refer to the information in Example 10.4.4 P = tests positive.

a. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(\text{P}|\text{B}) = 1 - P(\text{N}|\text{B})$.
b. What is the probability that a woman develops breast cancer and tests positive. Find $P(\text{B AND P}) = P(\text{P}|\text{B})P(\text{B})$.
c. What is the probability that a woman does not develop breast cancer. Find $P(\text{B}') = 1 - P(\text{B})$.
d. What is the probability that a woman tests positive for breast cancer. Find $P(\text{P}) = 1 - P(\text{N})$.

Answer

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Exercise 10.4.5

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(\text{B}) = 0.40, P(\text{D}) = 0.30$ and $P(\text{D}|\text{B}) = 0.5$.

a. Find $P(\text{B}')$.
b. Find $P(\text{D AND B})$.
c. Find $P(\text{B}|\text{D})$.
d. Find $P(\text{D AND B}')$.
e. Find $P(\text{D}|\text{B}')$.

Answer

a. $P(\text{B}') = 0.60$
b. $P(\text{D AND B}) = P(\text{D}|\text{B})P(\text{B}) = 0.20$
c. $P(\text{B}|\text{D}) = \frac{P(\text{B AND D})}{P(\text{D})} = \frac{(0.20)}{(0.30)} = 0.66$
d. $P(\text{D AND B}') = P(\text{D}) - P(\text{D AND B}) = 0.30 - 0.20 = 0.10$
e. $P(\text{D}|\text{B}') = P(\text{D AND B}')P(\text{B}') = (P(\text{D}) - P(\text{D AND B}))(0.60) = (0.10)(0.60) = 0.06$

References

1. DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://www.field.com/fieldpollonline...rs/Rls2443.pdf (accessed May 2, 2013).
2. Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011..._suggests.html (accessed May 2, 2013).
3. “Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).
4. “Roulette.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).
5. Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/l...acs/ACS-12.pdf (accessed May 2, 2013).
6. Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).
7. Data from U.S. Census Bureau.
8. Data from the Wall Street Journal.

9. Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).
10. Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Chapter Review

The multiplication rule and the addition rule are used for computing the probability of A and B, as well as the probability of A or B for two given events A, B defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events A and B are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(\text{A AND B}) = P(\text{A}|\text{B})P(\text{B})$

The addition rule: $P(\text{A OR B}) = P(\text{A}) + P(\text{B}) - P(\text{A AND B})$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

Exercise 3.4.5

Find $P(\text{C})$.

Exercise 3.4.6

Find $P(\text{L})$.

Answer

0.376

Exercise 3.4.7

Find $P(\text{C}|\text{L})$.

Exercise 3.4.8

In words, what is $\text{C}|\text{L}$?

Answer

$\text{C}|\text{L}$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Exercise 3.4.9

Find $P(\text{L AND C})$

Exercise 3.4.10

In words, what is L AND C?

Answer

L AND C is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Exercise 3.4.11

Are L and C independent events? Show why or why not.

Exercise 3.4.12

Find $P(\text{L OR C})$.

Answer

0.6492

Exercise 3.4.13

In words, what is L OR C?

Exercise 3.4.14

Are L and C mutually exclusive events? Show why or why not.

Answer

No, because $P(\text{L AND C})$ does not equal 0.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(\text{A}|\text{B}) = P(\text{A})$
2. $P(\text{B}|\text{A}) = P(\text{B})$
3. $P(\text{A AND B}) = P(\text{A})P(\text{B})$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(\text{A AND B}) = 0$.

This page titled 10.4: Two Basic Rules of Probability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.5: Contingency Tables

A *contingency table* provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Example 10.5.1

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	No speeding violation in the last year	Total
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.

Calculate the following probabilities using the table.

a. Find $P(\text{Person is a car phone user})$.
b. Find $P(\text{person had no violation in the last year})$.
c. Find $P(\text{Person had no violation in the last year AND was a car phone user})$.
d. Find $P(\text{Person is a car phone user OR person had no violation in the last year})$.
e. Find $P(\text{Person is a car phone user GIVEN person had a violation in the last year})$.
f. Find $P(\text{Person had no violation last year GIVEN person was not a car phone user})$

Answer

a. $\dfrac{\text{number of car phone users}}{\text{total number in study}} = \dfrac{305}{755}$
b. $\dfrac{\text{number that had no violation}}{\text{total number in study}} = \dfrac{685}{755}$
c. $\dfrac{280}{755}$
d. $\left(\dfrac{305}{755} + \dfrac{685}{755}\right) - \dfrac{280}{755} = \dfrac{710}{755}$
e. $\dfrac{25}{70}$ (The sample space is reduced to the number of persons who had a violation.)
f. $\dfrac{405}{450}$ (The sample space is reduced to the number of persons who were not car phone users.)

Exercise 10.5.1

Table shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	No injury in last year	Total
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

a. What is $P(\text{athlete stretches before exercising})$?
b. What is $P(\text{athlete stretches before exercising} | \text{no injury in the last year})$?

Answer

a. $P(\text{athlete stretches before exercising}) = \frac{350}{800} = 0.4375$

b. $P(\text{athlete stretches before exercising}|\text{no injury in the last year}) = \frac{295}{514} = 0.5739$

Example 10.5.2

Table shows a random sample of 100 hikers and the areas of hiking they prefer.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

a. Complete the table.

b. Are the events "being female" and "preferring the coastline" independent events? Let *F* = being female and let *C* = preferring the coastline.

1. Find *P*(*F* AND *C*).
2. Find *P*(*F*)*P*(*C*)
3. Are these two numbers the same? If they are, then *F* and *C* are independent. If they are not, then *F* and *C* are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let $\mathrm{M} =$ being male, and let $\mathrm{L} =$ prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: P(___|___) = ___.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let $\mathrm{F} =$ being female, and let $\mathrm{P} =$ prefers mountain peaks.

1. Find $P(\mathrm{F})$.
2. Find $P(\mathrm{P})$.
3. Find $P(\mathrm{F\ AND\ P})$.
4. Find $P(\mathrm{F\ OR\ P})$.

Answers

a.

Hiking Area Preference

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	**11**	45
Male	**16**	**25**	14	55
Total	**34**	41	**25**	**100**

b.

$$P(\mathrm{F\ AND\ C}) = \frac{18}{100} = 0.18$$

$$P(\mathrm{F})P(\mathrm{C}) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$$

$P(\mathrm{F\ AND\ C}) \neq P(\mathrm{F})P(\mathrm{C})$, so the events F and C are not independent.

c.

1. The word 'given' tells you that this is a conditional.
2. $P(\text{M}|\text{L}) = \frac{25}{41}$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d.

a. Find $P(\text{F})$.
b. Find $P(\text{P})$.
c. Find $P(\text{F AND P})$.
d. Find $P(\text{F OR P})$.

d.

1. $P(\text{F}) = \frac{45}{100}$
2. $P(\text{P}) = \frac{25}{100}$
3. $P(\text{F AND P}) = \frac{11}{100}$
4. $P(\text{F OR P}) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Exercise 10.5.2

Table shows a random sample of 200 cyclists and the routes they prefer. Let $\text{M} =$ males and $\text{H} =$ hilly path.

Gender	Lake Path	Hilly Path	Wooded Path	Total
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

a. Out of the males, what is the probability that the cyclist prefers a hilly path?
b. Are the events “being male” and “preferring the hilly path” independent events?

Answer

a. $P(H|M) = \frac{52}{90} = 0.5778$
b. For M and H to be independent, show $P(H|M) = P(H)$
$P(H|M) = 0.5778$, $P(H) = \frac{90}{200} = 0.45$
$P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Example 10.5.3

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is $\frac{1}{5}$ and the probability he is not caught is $\frac{4}{5}$. If he goes out the second door, the probability he gets caught by Alissa is $\frac{1}{4}$ and the probability he is not caught is $\frac{3}{4}$. The probability that Alissa catches Muddy coming out of the third door is $\frac{1}{2}$ and the probability she does not catch Muddy is $\frac{1}{2}$. It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is $\frac{1}{3}$.

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	____
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	____
Total	____	____	____	1

- The first entry $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Caught})$
- The entry $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ is $P(\text{Door One AND Not Caught})$

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.
b. What is the probability that Alissa does not catch Muddy?
c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

Solution

Door Choice

Caught or Not	Door One	Door Two	Door Three	Total
Caught	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Not Caught	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Total	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. $\frac{41}{60}$
c. $\frac{9}{19}$

Example 10.5.4

Table contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

TOTAL each column and each row. Total data = 4,520.7

a. Find $P(2009 \text{ AND Robbery})$.
b. Find $P(2010 \text{ AND Burglary})$.
c. Find $P(2010 \text{ OR Burglary})$.
d. Find $P(2011|\text{Rape})$.

e. Find $P(\text{Vehicle}|2008)$

Answer

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Exercise 10.5.3

Table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

a. Find the total for each row and column
b. Find the probability that a randomly chosen individual from this group is Tall.
c. Find the probability that a randomly chosen individual from this group is Obese and Tall.
d. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
e. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
f. Find the probability a randomly chosen individual from this group is Tall and Underweight.
g. Are the events Obese and Tall independent?

Answer

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

a. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
b. $P(\text{Tall}) = \frac{50}{205} = 0.244$
c. $P(\text{Obese AND Tall}) = \frac{18}{205} = 0.088$
d. $P(\text{Tall}|\text{Obese}) = \frac{18}{60} = 0.3$
e. $P(\text{Obese}|\text{Tall}) = \frac{18}{50} = 0.36$
f. $P(\text{Tall AND Underweight}) = \frac{12}{205} = 0.0585$
g. No. $P(\text{Tall})$ does not equal $P(\text{Tall}|\text{Obese})$.

References

1. "Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-a...od/blood-types (accessed May 3, 2013).
2. Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.
3. Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).
4. Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of

Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).
5. “Human Blood Types.” Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).
6. Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_st...ive-blood.html (accessed May 2, 2013).
7. “United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Chapter Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilities that have multiple dependent variables.

Use the following information to answer the next four exercises. Table shows a random sample of musicians and how they learned to play their instruments.

Gender	Self-taught	Studied in School	Private Instruction	Total
Female	12	38	22	72
Male	19	24	15	58
Total	31	62	37	130

Exercise 3.5.4

Find *P*(musician is a female).

Exercise 3.5.5

Find $P(\text{musician is a male AND had private instruction})$.

Answer

$$P(\text{musician is a male AND had private instruction}) = \frac{15}{130} = \frac{3}{26} = 0.12$$

Exercise 3.5.6

Find *P*(musician is a female OR is self taught).

Exercise 3.5.7

Are the events “being a female musician” and “learning music in school” mutually exclusive events?

Answer

$$P(\text{being a female musician AND learning music in school}) = \frac{38}{130} = \frac{19}{65} = 0.29$$

$$P(\text{being a female musician})\text{P}(\text{learning music in school}) = \left(\frac{72}{130}\right)\left(\frac{62}{130}\right) = \frac{4,464}{16,900} = \frac{1,116}{4,225} = 0.26$$

No, they are not independent because $P(\text{being a female musician AND learning music in school})$ is not equal to $P(\text{being a female musician})P(\text{learning music in school})$.

Bringing it Together

Use the following information to answer the next seven exercises. An article in the *New England Journal of Medicine*, reported about a study of smokers in California and Hawaii. In one part of the report, the self-reported ethnicity and smoking levels per day were given. Of the people smoking at most ten cigarettes per day, there were 9,886 African Americans, 2,745 Native Hawaiians, 12,831 Latinos, 8,378 Japanese Americans, and 7,650 Whites. Of the people smoking 11 to 20 cigarettes per day, there were 6,514 African Americans, 3,062 Native Hawaiians, 4,932 Latinos, 10,680 Japanese Americans, and 9,877 Whites. Of the people smoking 21 to 30 cigarettes per day, there were 1,671 African Americans, 1,419 Native Hawaiians, 1,406 Latinos, 4,715 Japanese

Americans, and 6,062 Whites. Of the people smoking at least 31 cigarettes per day, there were 759 African Americans, 788 Native Hawaiians, 800 Latinos, 2,305 Japanese Americans, and 3,970 Whites.

Exercise 3.5.8

Complete the table using the data provided. Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Smoking Levels by Ethnicity

Smoking Level	African American	Native Hawaiian	Latino	Japanese Americans	White	TOTALS
1–10						
11–20						
21–30						
31+						
TOTALS						

Exercise 3.5.9

Suppose that one person from the study is randomly selected. Find the probability that person smoked 11 to 20 cigarettes per day.

Answer

$\frac{35,065}{100,450}$

Exercise 3.5.10

Find the probability that the person was Latino.

Exercise 3.5.11

In words, explain what it means to pick one person from the study who is "Japanese American **AND** smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American AND smokes 21 to 30 cigarettes per day means that the person has to meet both criteria: both Japanese American and smokes 21 to 30 cigarettes. The sample space should include everyone in the study. The probability is $\frac{4,715}{100,450}$.

Exercise 3.5.12

In words, explain what it means to pick one person from the study who is "Japanese American **OR** smokes 21 to 30 cigarettes per day." Also, find the probability.

Exercise 3.5.13

In words, explain what it means to pick one person from the study who is "Japanese American **GIVEN** that person smokes 21 to 30 cigarettes per day." Also, find the probability.

Answer

To pick one person from the study who is Japanese American given that person smokes 21-30 cigarettes per day, means that the person must fulfill both criteria and the sample space is reduced to those who smoke 21-30 cigarettes per day. The probability is $\frac{4,715}{15,273}$.

Exercise 3.5.14

Prove that smoking level/day and ethnicity are dependent events.

Glossary

contingency table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

This page titled 10.5: Contingency Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.6: Tree and Venn Diagrams

Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams and Venn diagrams are two tools that can be used to visualize and solve conditional probabilities.

Tree Diagrams

A *tree diagram* is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

Example 10.6.1: Probabilities from Sampling with replacement

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, with replacement (remember that "with replacement" means that you put the first ball back in the urn before you select the second ball). The tree diagram using frequencies that show all the possible outcomes follows.

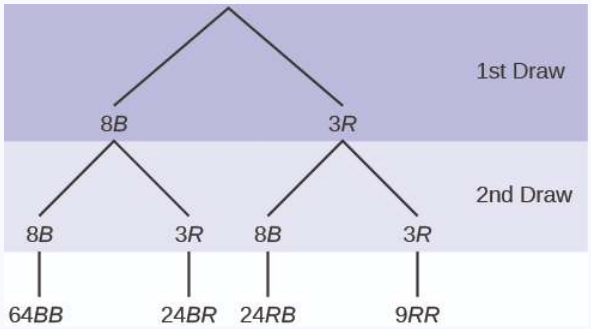


Figure 10.6.1: Total = 64 + 24 + 24 + 9 = 121

The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as $R1$, $R2$, and $R3$ and each blue ball as $B1$, $B2$, $B3$, $B4$, $B5$, $B6$, $B7$, and $B8$. Then the nine RR outcomes can be written as:

$R1R1$ $R1R2$ $R1R3$ $R2R1$ $R2R2$ $R2R3$ $R3R1$ $R3R2$ $R3R3$

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the sample space.

Exercise 10.6.1

In a standard deck, there are 52 cards. 12 cards are face cards (event F) and 40 cards are not face cards (event N). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate $P(\text{FF})$.

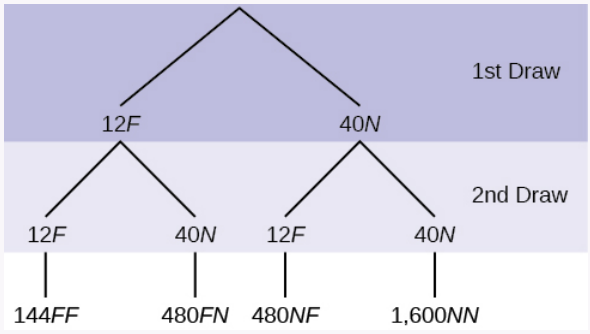


Figure 10.6.2:

Answer

Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

$$P(\text{FF}) = \frac{144}{144+480+480+1,600} = \frac{144}{2,704} = \frac{9}{169} \quad (10.6.1)$$

1. a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...
2. b. Using the tree diagram, calculate *P*(*RR*).
3. c. Using the tree diagram, calculate $P(\text{RB OR BR})$.
4. d. Using the tree diagram, calculate $P(\text{R on 1st draw AND B on 2nd draw})$.
5. e. Using the tree diagram, calculate *P*(*R* on 2nd draw GIVEN *B* on 1st draw).
6. Using the tree diagram, calculate $P(\text{BB})$.
7. g. Using the tree diagram, calculate $P(\text{B on the 2nd draw given R on the first draw})$.

Solution

a. *B*1*R*1; *B*1*R*2; *B*1*R*3; *B*2*R*1; *B*2*R*2; *B*2*R*3; *B*3*R*1; *B*3*R*2; *B*3*R*3; *B*4*R*1; *B*4*R*2; *B*4*R*3; *B*5*R*1; *B*5*R*2; *B*5*R*3; *B*6*R*1; *B*6*R*2; *B*6*R*3 *B*7*R*1; *B*7*R*2; *B*7*R*3; *B*8*R*1; *B*8*R*2; *B*8*R*3

b. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. $P(\text{R on 1st draw AND B on 2nd draw}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. *P*(*R* on 2nd draw GIVEN *B* on 1st draw) = *P*(*R* on 2nd|*B* on 1st) = $\frac{24}{88} = \frac{3}{11}$ This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are 24 + 64 = 88 possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $\frac{24}{88} = \frac{3}{11}$

f. $P(\text{BB}) = \frac{64}{121}$

g. $P(\text{B on 2nd draw}|\text{R on 1st draw}) = \frac{8}{11}$. There are 9 + 24 outcomes that have R on the first draw (9 *RR* and 24 *RB*). The sample space is then 9 + 24 = 33. 24 of the 33 outcomes have B on the second draw. The probability is then $\frac{24}{33}$.

Example 10.6.2: Probabilities from Sampling without replacement

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. (remember that "without replacement" means that you do not put the first ball back before you select the second marble). Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.

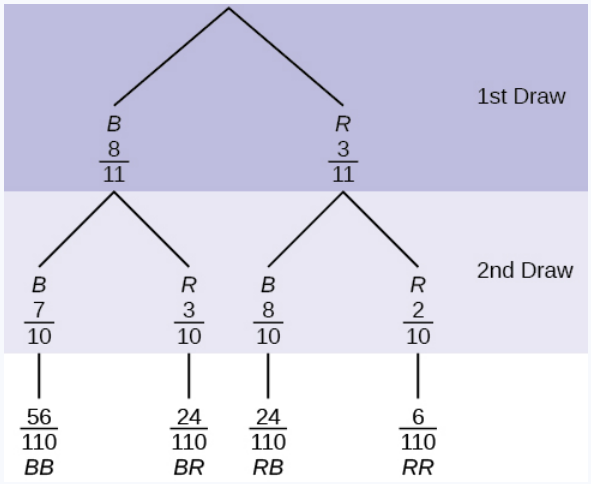


Figure 10.6.3: Total $= \frac{56+24+24+6}{110} = \frac{110}{110} = 1$

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it. You draw ***without replacement****, so that on the second draw there are ten marbles left in the urn.*

Calculate the following probabilities using the tree diagram.

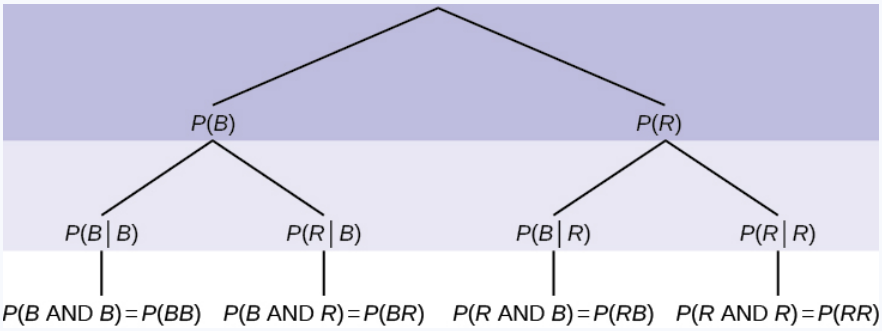
a. $P(\text{RR}) =$ _____

b. Fill in the blanks: $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\_\_\_)(\_\_\_) = \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) =$
d. Fill in the blanks: $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \ (\_\_\_)(\_\_\_) = \frac{24}{100}$
e. Find $P(\text{BB})$.
f. Find $P(\text{B on 2nd}|\text{R on 1st})$.

Answers

a. $P(\text{RR}) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$
b. $P(\text{RB OR BR}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$
c. $P(\text{R on 2nd}|\text{B on 1st}) = \frac{3}{10}$
d. $P(\text{R on 1st AND B on 2nd}) = P(\text{RB}) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{100}$
e. $P(\text{BB}) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$
f. Using the tree diagram, $P(\text{B on 2nd}|\text{R on 1st}) = P(\text{R}|\text{B}) = \frac{8}{10}$.

If we are using probabilities, we can label the tree in the following general way.



- $P(\text{R}|\text{R})$ here means $P(\text{R on 2nd}|\text{R on 1st})$
- $P(\text{B}|\text{R})$ here means $P(\text{B on 2nd}|\text{R on 1st})$
- $P(\text{R}|\text{B})$ here means $P(\text{R on 2nd}|\text{B on 1st})$
- $P(\text{B}|\text{B})$ here means $P(\text{B on 2nd}|\text{B on 1st})$

Exercise 10.6.2

In a standard deck, there are 52 cards. Twelve cards are face cards (F) and 40 cards are not face cards (N). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.

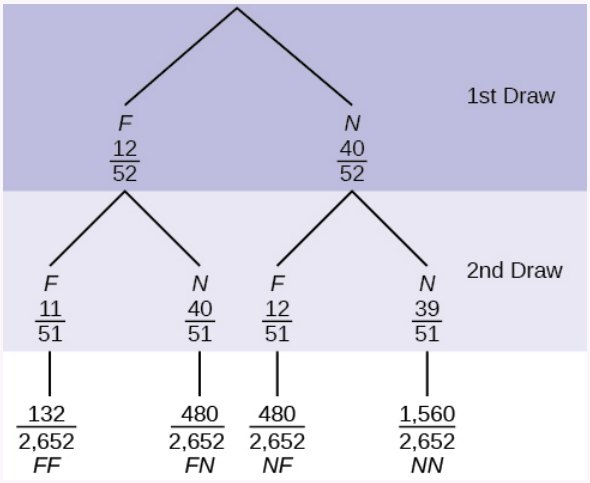


Figure 10.6.4:

a. Find $P(\text{FN OR NF})$.
b. Find $P(\text{N}|\text{F})$.

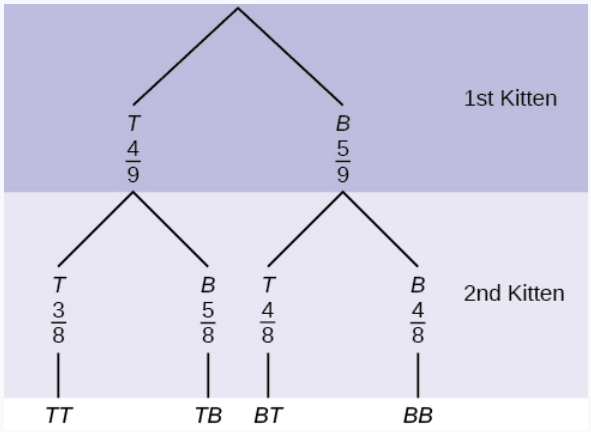
c. Find $P(\text{at most one face card})$.
Hint: "At most one face card" means zero or one face card.
d. Find $P(\text{at least on face card})$.
Hint: "At least one face card" means one or two face cards.

Answer

a. $P(\text{FN OR NF}) = \frac{480}{2,652} + \frac{480}{2,652} = \frac{960}{2,652} = \frac{80}{221}$
b. $P(\text{N}|\text{F}) = \frac{40}{51}$
c. $P(\text{at most one face card}) = \frac{(480+480+1,560)}{2,652} = \frac{2,520}{2,652}$
d. $P(\text{at least one face card}) = \frac{(132+480+480)}{2,652} = \frac{1,092}{2,652}$

Example 10.6.3

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



a. What is the probability that both kittens are tabby?
a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$
b. What is the probability that one kitten of each coloring is selected?
a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$
c. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
d. What is the probability of choosing two kittens of the same color?

Answer

a. c, b. d, c. $\frac{4}{8}$, d. $\frac{32}{72}$

Exercise 10.6.3

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

Answer

$\left(\frac{4}{7}\right)\left(\frac{3}{6}\right) + \left(\frac{3}{7}\right)\left(\frac{4}{6}\right)$

Venn Diagram

A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space S together with circles or ovals. The circles or ovals represent events.

Example 10.6.4

Suppose an experiment has the outcomes 1, 2, 3, ... , 12 where each outcome has an equal chance of occurring. Let event $A = \{1, 2, 3, 4, 5, 6\}$ and event $B = \{6, 7, 8, 9\}$. Then A AND $B = \{6\}$ and A OR $B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. The Venn diagram is as follows:

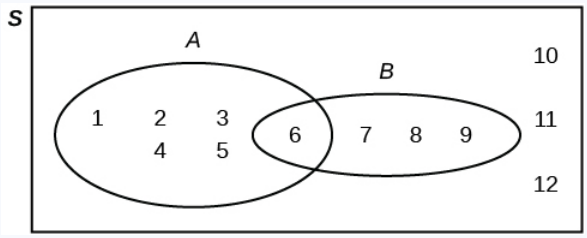


Figure 10.6.5:

Exercise 10.6.4

Suppose an experiment has outcomes black, white, red, orange, yellow, green, blue, and purple, where each outcome has an equal chance of occurring. Let event $C = \{\text{green, blue, purple}\}$ and event $P = \{\text{red, yellow, blue}\}$. Then C AND $P = \{\text{blue}\}$ and C OR $P = \{\text{green, blue, purple, red, yellow}\}$. Draw a Venn diagram representing this situation.

Answer

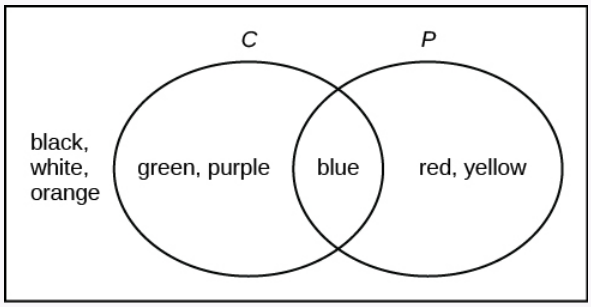


Figure 10.6.6:

Example 10.6.5

Flip two fair coins. Let $A =$ tails on the first coin. Let $B =$ tails on the second coin. Then $A = \{TT, TH\}$ and $B = \{TT, HT\}$. Therefore, A AND $B = \{TT\}$. A OR $B = \{TH, TT, HT\}$.

The sample space when you flip two fair coins is $X = \{HH, HT, TH, TT\}$. The outcome HH is in NEITHER A NOR B. The Venn diagram is as follows:

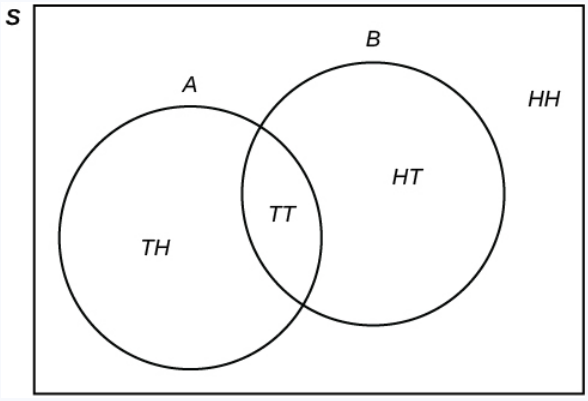
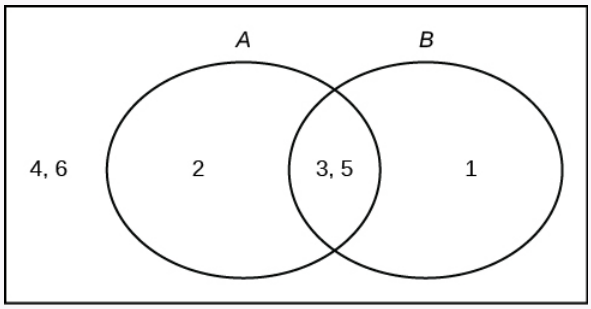


Figure 10.6.7:

Exercise 10.6.5

Roll a fair, six-sided die. Let $A =$ a prime number of dots is rolled. Let $B =$ an odd number of dots is rolled. Then $A = \{2, 3, 5\}$ and $B = \{1, 3, 5\}$. Therefore, A AND $B = \{3, 5\}$. A OR $B = \{1, 2, 3, 5\}$. The sample space for rolling a fair die is $S = \{1, 2, 3, 4, 5, 6\}$. Draw a Venn diagram representing this situation.

Answer



Example 10.6.6: Probability and Venn Diagrams

Forty percent of the students at a local college belong to a club and 50% work part time. Five percent of the students work part time and belong to a club. Draw a Venn diagram showing the relationships. Let $\mathrm{C} =$ student belongs to a club and $\mathrm{PT} =$ student works part time.

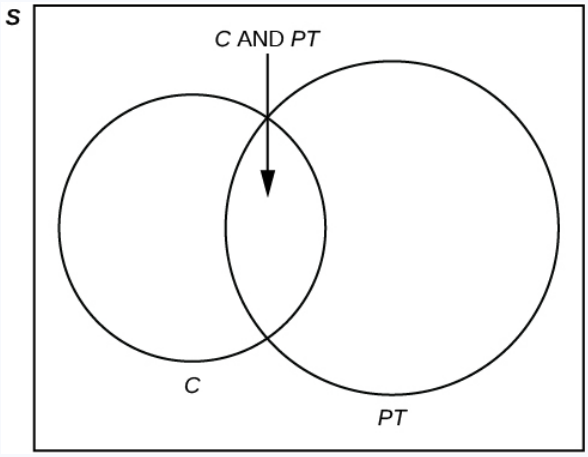


Figure 10.6.8:

If a student is selected at random, find

- the probability that the student belongs to a club. $P(\mathrm{C}) = 0.40$
- the probability that the student works part time. $P(\mathrm{PT}) = 0.50$
- the probability that the student belongs to a club AND works part time. $P(\mathrm{C\ AND\ PT}) = 0.05$
- the probability that the student belongs to a club **given** that the student works part time.
  $P(\mathrm{C}|\mathrm{PT}) = \frac{P(\mathrm{C\ AND\ PT})}{P(\mathrm{PT})} = \frac{0.05}{0.50} = 0.1$
- the probability that the student belongs to a club **OR** works part time.
  $P(\mathrm{C\ OR\ PT}) = P(\mathrm{C}) + P(\mathrm{PT}) - P(\mathrm{C\ AND\ PT}) = 0.40 + 0.50 - 0.05 = 0.85$

Exercise 10.6.6

Fifty percent of the workers at a factory work a second job, 25% have a spouse who also works, 5% work a second job and have a spouse who also works. Draw a Venn diagram showing the relationships. Let $\mathrm{W} =$ works a second job and $\mathrm{S} =$ spouse also works.

Answer

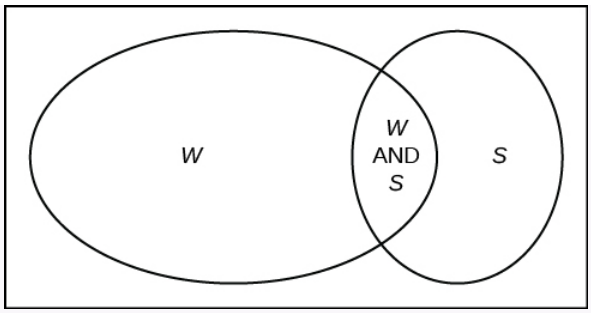


Figure 10.6.9:

Example 10.6.7

A person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Four percent of African Americans have type O blood and a negative RH factor, 5−10% of African Americans have the Rh- factor, and 51% have type O blood.

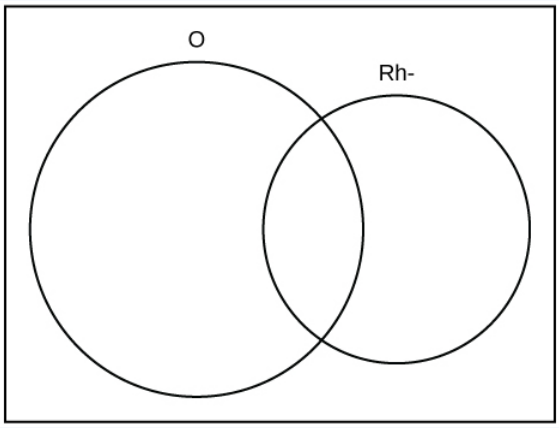


Figure 10.6.10:

The “O” circle represents the African Americans with type O blood. The “Rh-“ oval represents the African Americans with the Rh- factor.

We will take the average of 5% and 10% and use 7.5% as the percent of African Americans who have the Rh- factor. Let $\mathrm{O} =$ African American with Type O blood and $\mathrm{R} =$ African American with Rh- factor.

a. $P(\mathrm{O}) =$ _____
b. $P(\mathrm{R}) =$ _____
c. $P(\mathrm{O\ AND\ R}) =$ _____
d. $P(\mathrm{O\ OR\ R}) =$ _____
e. In the Venn Diagram, describe the overlapping area using a complete sentence.
f. In the Venn Diagram, describe the area in the rectangle but outside both the circle and the oval using a complete sentence.

Answer

a. 0.51; b. 0.075; c. 0.04; d. 0.545; e. The area represents the African Americans that have type O blood and the Rh- factor. f. The area represents the African Americans that have neither type O blood nor the Rh- factor.

Exercise 10.6.7

In a bookstore, the probability that the customer buys a novel is 0.6, and the probability that the customer buys a non-fiction book is 0.4. Suppose that the probability that the customer buys both is 0.2.

a. Draw a Venn diagram representing the situation.
b. Find the probability that the customer buys either a novel or anon-fiction book.
c. In the Venn diagram, describe the overlapping area using a complete sentence.
d. Suppose that some customers buy only compact disks. Draw an oval in your Venn diagram representing this event.

Answer

a. and d. In the following Venn diagram below, the blue oval represent customers buying a novel, the red oval represents customer buying non-fiction, and the yellow oval customer who buy compact disks.

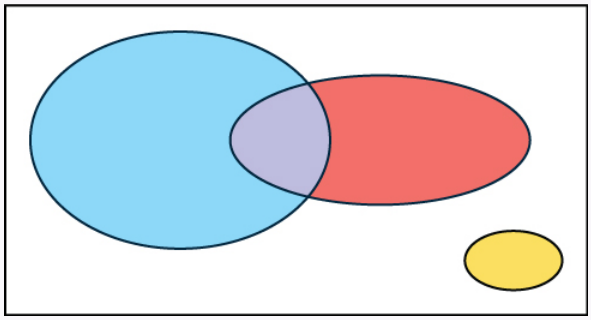


Figure 10.6.11:

b. $P(\text{novel or non-fiction}) = P(\text{Blue OR Red}) = P(\text{Blue}) + P(\text{Red}) - P(\text{Blue AND Red}) = 0.6 + 0.4 - 0.2 = 0.8$.

c. The overlapping area of the blue oval and red oval represents the customers buying both a novel and a nonfiction book.

References

1. Data from Clara County Public H.D.
2. Data from the American Cancer Society.
3. Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).
4. Data from the Federal Highway Administration, part of the United States Department of Transportation.
5. Data from the United States Census Bureau, part of the United States Department of Commerce.
6. Data from USA Today.
7. "Environment." The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).
8. "Search for Datasets." Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/dat..._datasets.html (accessed May 2, 2013).

Chapter Review

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize. A Venn diagram is a picture that represents the outcomes of an experiment. It generally consists of a box that represents the sample space *S* together with circles or ovals. The circles or ovals represent events. A Venn diagram is especially helpful for visualizing the OR event, the AND event, and the complement of an event and for understanding conditional probabilities.

Glossary

Tree Diagram

the useful visual representation of a sample space and events in the form of a "tree" with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Venn Diagram

the visual representation of a sample space and events in the form of circles or ovals showing their intersections

This page titled 10.6: Tree and Venn Diagrams is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.7: Probability Topics (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will use theoretical and empirical methods to estimate probabilities.
- The student will appraise the differences between the two estimates.
- The student will demonstrate an understanding of long-term relative frequencies.

Do the Experiment

Count out 40 mixed-color M&Ms® which is approximately one small bag's worth. Record the number of each color in Table. Use the information from this table to complete Table. Next, put the M&Ms in a cup. The experiment is to pick two M&Ms, one at a time. Do **not** look at them as you pick them. The first time through, replace the first M&M before picking the second one. Record the results in the "With Replacement" column of Table. Do this 24 times. The second time through, after picking the first M&M, do **not** replace it before picking the second one. Then, pick the second one. Record the results in the "Without Replacement" column section of Table. After you record the pick, put **both** M&Ms back. Do this a total of 24 times, also. Use the data from Table to calculate the empirical probability questions. Leave your answers in unreduced fractional form. Do **not** multiply out any fractions.

Population

Color	Quantity
Yellow (Y)	
Green (G)	
Blue (BL)	
Brown (B)	
Orange (O)	
Red (R)	

Theoretical Probabilities

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

G_2 = green on second pick; R_1 = red on first pick; B_1 = brown on first pick; B_2 = brown on second pick; doubles = both picks are the same colour.

Empirical Results

With Replacement	Without Replacement
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)
(__ , __) (__ , __)	(__ , __) (__ , __)

Empirical Probabilities

	With Replacement	Without Replacement
P(2 reds)		
$P(R_1B_2$ OR $B_1R_2)$		
$P(R_1$ AND $G_2)$		
$P(G_2\|R_1)$		
P(no yellows)		
P(doubles)		
P(no doubles)		

Discussion Questions

a. Why are the “With Replacement” and “Without Replacement” probabilities different?
b. Convert P(no yellows) to decimal format for both Theoretical “With Replacement” and for Empirical “With Replacement”. Round to four decimal places.
    1. Theoretical “With Replacement”: P(no yellows) = _____
    2. Empirical “With Replacement”: P(no yellows) = _____
    3. Are the decimal values “close”? Did you expect them to be closer together or farther apart? Why?
c. If you increased the number of times you picked two M&Ms to 240 times, why would empirical probability values change?
d. Would this change (see part 3) cause the empirical probabilities and theoretical probabilities to be closer together or farther apart? How do you know?
e. Explain the differences in what $P(G_1$ AND $R_2)$ and $P(R_1|G_2)$ represent. Hint: Think about the sample space for each probability.

This page titled 10.7: Probability Topics (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

10.E: Probability Topics (Exericses)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

3.1: Introduction

3.2: Terminology

Q 3.2.1

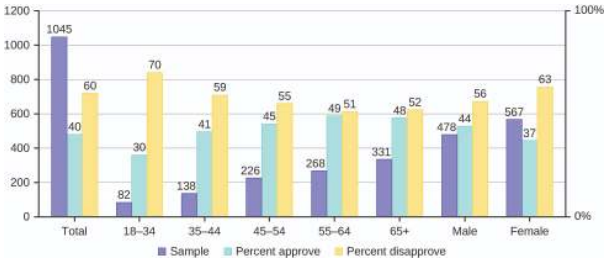


Figure 3.2.3.2.11.

The graph in Figure 3.2.1 displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

a. Define three events in the graph.
b. Describe in words what the entry 40 means.
c. Describe in words the complement of the entry in question 2.
d. Describe in words what the entry 30 means.
e. Out of the males and females, what percent are males?
f. Out of the females, what percent disapprove of Mayor Ford?
g. Out of all the age groups, what percent approve of Mayor Ford?
h. Find $P(\text{Approve}|\text{Male})$.
i. Out of the age groups, what percent are more than 44 years old?
j. Find $P(\text{Approve}|\text{Age} < 35)$.

Q 3.2.2

Explain what is wrong with the following statements. Use complete sentences.

a. If there is a 60% chance of rain on Saturday and a 70% chance of rain on Sunday, then there is a 130% chance of rain over the weekend.
b. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

S 3.2.2

a. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
b. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

3.3: Independent and Mutually Exclusive Events

Use the following information to answer the next 12 exercises. The graph shown is based on more than 170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.

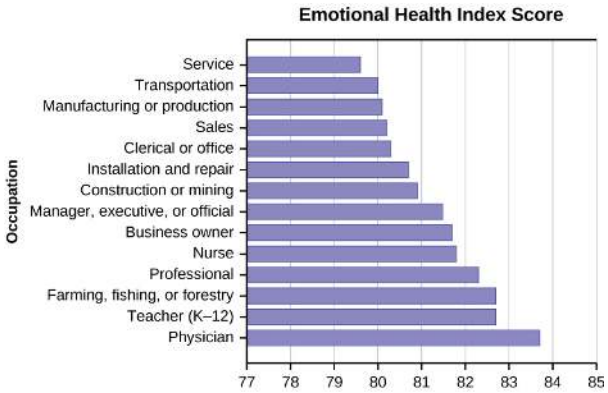


Figure 3.3.1.

Q 3.3.1

Find the probability that an Emotional Health Index Score is 82.7.

Q 3.3.2

Find the probability that an Emotional Health Index Score is 81.0.

S 3.3.2

0

Q 3.3.3

Find the probability that an Emotional Health Index Score is more than 81?

Q 3.3.4

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

S 3.3.4

0.3571

Q 3.3.5

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

Q 3.3.6

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

S 3.3.6

0.2142

Q 3.3.7

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

Q 3.3.8

What occupation has the highest emotional index score?

S 3.3.8

Physician (83.7)

Q 3.3.9

What occupation has the lowest emotional index score?

Q 3.3.10

What is the range of the data?

S 3.3.10

83.7 − 79.6 = 4.1

Q 3.3.11

Compute the average EHIS.

Q 3.3.12

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

S 3.3.12

$P(\text{Occupation} < 81.3) = 0.5$

3.4: Two Basic Rules of Probability

Q 3.4.1

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- $\mathrm{C} =$ California registered voters who support same-sex marriage.
- $\mathrm{B} =$ California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- $\mathrm{A} =$ California registered voters who are 18 to 39 years old.

a. Find $P(\mathrm{C})$.
b. Find $P(\mathrm{B})$.
c. Find $P(\mathrm{C}|\mathrm{A})$.
d. Find $P(\mathrm{B}|\mathrm{C})$.
e. In words, what is $\mathrm{C}|\mathrm{A}$?
f. In words, what is $\mathrm{B}|\mathrm{C}$?
g. Find $P(\mathrm{C\ AND\ B})$.
h. In words, what is $\mathrm{C\ AND\ B}$?
i. Find $P(\mathrm{C\ OR\ B})$.
j. Are C and B mutually exclusive events? Show why or why not.

Q 3.4.2

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

a. What is the sample size for this study?
b. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?
c. How many people polled responded that they approved of Mayor Ford in late 2011?
d. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?

e. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

S 3.4.2

a. The Forum Research surveyed 1,046 Torontonians.
b. 58%
c. 42% of 1,046 = 439 (rounding to the nearest integer)
d. 0.57
e. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.

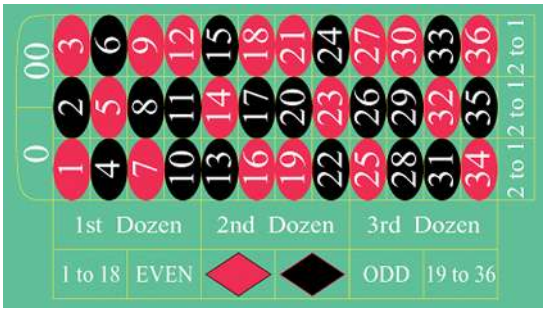


Figure 3.4.1

Q 3.4.3

a. List the sample space of the 38 possible outcomes in roulette.
b. You bet on red. Find $P(\text{red})$.
c. You bet on -1st 12- (1st Dozen). Find $P(\text{-1st 12-})$.
d. You bet on an even number. Find $P(\text{even number})$.
e. Is getting an odd number the complement of getting an even number? Why?
f. Find two mutually exclusive events.
g. Are the events Even and 1st Dozen independent?

Q 3.4.4

Compute the probability of winning the following types of bets:

a. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6
b. Betting on three numbers in a line, as in 1-2-3
c. Betting on one number
d. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
e. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
f. Betting on 0-00-1-2-3
g. Betting on 0-1-2; or 0-00-2; or 00-2-3

S 3.4.4

1. $P(\text{Betting on two line that touch each other on the table}) = \frac{6}{38}$
2. $P(\text{Betting on three numbers in a line}) = \frac{3}{38}$
3. $P(\text{Betting on one number}) = \frac{1}{38}$
4. $P(\text{Betting on four number that touch each other to form a square}) = \frac{4}{38}$
5. $P(\text{Betting on two number that touch each other on the table}) = \frac{2}{38}$
6. $P(\text{Betting on 0-00-1-2-3}) = \frac{5}{38}$
7. $P(\text{Betting on 0-1-2; or 0-00-2; or 00-2-3}) = \frac{3}{38}$

Q 3.4.5

Compute the probability of winning the following types of bets:

a. Betting on a color
b. Betting on one of the dozen groups
c. Betting on the range of numbers from 1 to 18
d. Betting on the range of numbers 19–36
e. Betting on one of the columns
f. Betting on an even or odd number (excluding zero)

Q 3.4.6

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- G = card drawn is green
- E = card drawn is even-numbered

a. List the sample space.
b. $P(\text{G}) =$ _____
c. $P(\text{G}|\text{E}) =$ _____
d. $P(\text{G AND E}) =$ _____
e. $P(\text{G OR E}) =$ _____
f. Are G and E mutually exclusive? Justify your answer numerically.

S 3.4.6

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. $\frac{5}{8}$
3. $\frac{2}{3}$
4. $\frac{2}{8}$
5. $\frac{}{6}8$
6. No, because $P(\text{G AND E})$ does not equal 0.

Q 3.4.7

Roll two fair dice. Each die has six faces.

a. List the sample space.
b. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(\text{A})$.
c. Let *B* be the event that the sum of the two rolls is at most seven. Find $P(\text{B})$.
d. In words, explain what "$P(\text{A}|\text{B})$" represents. Find $P(\text{A}|\text{B})$.
e. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
f. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.8

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

a. List the sample space.
b. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(\text{A})$.
c. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
d. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

S 3.4.9

The coin toss is independent of the card picked first.

a. $\{(G, H)(G, T)(B, H)(B, T)(R, H)(R, T)\}$
b. $P(\text{A}) = P(\text{blue})P(\text{head}) = \left(\frac{3}{10}\right)\left(\frac{1}{2}\right) = \frac{3}{20}$
c. Yes, A and B are mutually exclusive because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(\text{A AND B}) = 0$
d. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(\text{A AND C}) = P(\text{A}) = 320$

Q 3.4.10

An experiment consists of first rolling a die and then tossing a coin.

a. List the sample space.
b. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(\text{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

Q 3.4.11

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

a. List the sample space.
b. Let A be the event that there are at least two tails. Find $P(\text{A})$.
c. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

S 3.4.12

a. $S = (\text{HHH}), (\text{HHT}), (\text{HTH}), (\text{HTT}), (\text{THH}), (\text{THT}), (\text{TTH}), (\text{TTT})$
b. $\frac{4}{8}$
c. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(\text{A AND B}) = 0$.

Q 3.4.13

Consider the following scenario:

Let $P(\text{C}) = 0.4$.

Let $P(\text{D}) = 0.5$.

Let $P(\text{C}|\text{D}) = 0.6$.

a. Find $P(CANDD)$.
b. Are C and D mutually exclusive? Why or why not?
c. Are C and D independent events? Why or why not?
d. Find $P(\text{C OR D})$.
e. Find $P(\text{D}|\text{C})$.

Q 3.4.14

Y and Z are independent events.

a. Rewrite the basic Addition Rule $P(\text{Y OR Z}) = P(\text{Y}) + P(\text{Z}) - P(\text{Y AND Z})$ using the information that Y and Z are independent events.
b. Use the rewritten rule to find $P(\text{Z})$ if $P(\text{Y OR Z}) = 0.71$ and $P(\text{Y}) = 0.42$.

S 3.4.14

a. If Y and Z are independent, then $P(\text{Y AND Z}) = P(\text{Y})P(\text{Z})$, so $P(\text{Y OR Z}) = P(\text{Y}) + P(\text{Z}) - P(\text{Y})P(\text{Z})$.
b. 0.5

Q 3.4.15

G and H are mutually exclusive events. $P(\text{G}) = 0.5 P(\text{H}) = 0.3$

a. Explain why the following statement MUST be false: $P(\text{H}|\text{G}) = 0.4$.
b. Find $P(\text{H OR G})$.
c. Are G and H independent or dependent events? Explain in a complete sentence.

Q 3.4.16

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: E = speaks English at home; E′ = speaks another language at home; S = speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	Answers
a. $P(\text{E}') =$	i. 0.8043
b. $P(\text{E}) =$	ii. 0.623
c. $P(\text{S and E}') =$	iii. 0.1957
d. $P(\text{S}|\text{E}') =$	iv. 0.1219

S 3.4.16

a. iii
b. i
c. iv
d. ii

Q 3.4.17

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let $G =$ won green card.

a. What was Renate's chance of winning a Green Card? Write your answer as a probability statement.
b. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let F = was a finalist.
c. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
d. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Q 3.4.18

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: R = money returned; E = economics classes; O = other classes

a. Write a probability statement for the overall percent of money returned.
b. Write a probability statement for the percent of money returned out of the economics classes.
c. Write a probability statement for the percent of money returned out of the other classes.
d. Is money being returned independent of the class? Justify your answer numerically and explain it.
e. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

S 3.4.18

a. $P(\mathrm{R}) = 0.44$
b. $P(\mathrm{R}|\mathrm{E}) = 0.56$
c. $P(\mathrm{R}|\mathrm{O}) = 0.31$
d. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(\mathrm{R}|\mathrm{E}) \neq P(\mathrm{R})$.
e. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a higher rate than the money place in all classes collectively; $P(\mathrm{R}|\mathrm{E}) > P(\mathrm{R})$.

Q 3.4.19

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home Run	Total Hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

a. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$
b. No, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit is a double})$
c. No, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) \neq P(\text{hit by Hank Aaron})$
d. Yes, because $P(\text{hit is by Hank Aaron}|\text{hit is a double}) = P(\text{hit is a double})$

Q 3.4.29

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any blood type. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

a. Find the probability that a person has both type O blood and the Rh- factor.
b. Find the probability that a person does NOT have both type O blood and the Rh- factor.

S 3.4.30

a. $P(\text{type O OR Rh-}) = P(\text{type O}) + P(\text{Rh-}) - P(\text{type O AND Rh-})$

$0.52 = 0.43 + 0.15 - P(\text{type O AND Rh-})$; solve to find $P(\text{type O AND Rh-}) = 0.06$

6% of people have type O, Rh- blood

b. $P(\text{NOT}(\text{type O AND Rh-})) = 1 - P(\text{type O AND Rh-}) = 1 - 0.06 = 0.94$

94% of people do not have type O, Rh- blood

Q 3.4.31

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let *F* be the event that a course has a final exam. Let *R* be the event that a course requires a research paper.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

Q 3.4.32

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

S 3.4.32

a. Let $C =$ be the event that the cookie contains chocolate. Let $N =$ the event that the cookie contains nuts.
b. $P(\text{C OR N}) = P(\text{C}) + P(\text{N}) - P(\text{C AND N}) = 0.36 + 0.12 - 0.08 = 0.40$
c. $P(\text{NEITHER chocolate NOR nuts}) = 1 - P(\text{C OR N}) = 1 - 0.40 = 0.60$

Q 3.4.33

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let $\text{D} =$ event that a student takes a distance learning class and $\text{E} =$ event that a student is a part time student

a. Find $P(\text{D AND E})$.
b. Find $P(\text{E}|\text{D})$.
c. Find $P(\text{D OR E})$.
d. Using an appropriate test, show whether D and E are independent.
e. Using an appropriate test, show whether D and E are mutually exclusive.

3.5: Contingency Tables

Use the information in the Table to answer the next eight exercises. The table shows the political party affiliation of each of 67 members of the US Senate in June 2012, and when they are up for reelection.

Up for reelection:	Democratic Party	Republican Party	Other	Total
November 2014	20	13	0	
November 2016	10	24	0	
Total				

Q 3.5.1

What is the probability that a randomly selected senator has an “Other” affiliation?

S 3.5.1

0

Q 3.5.2

What is the probability that a randomly selected senator is up for reelection in November 2016?

Q 3.5.3

What is the probability that a randomly selected senator is a Democrat and up for reelection in November 2016?

S 3.5.3

$\frac{10}{67}$

Q 3.5.4

What is the probability that a randomly selected senator is a Republican or is up for reelection in November 2014?

Q 3.5.5

Suppose that a member of the US Senate is randomly selected. Given that the randomly selected senator is up for reelection in November 2016, what is the probability that this senator is a Democrat?

S 3.5.5

$\frac{10}{34}$

Q 3.5.6

Suppose that a member of the US Senate is randomly selected. What is the probability that the senator is up for reelection in November 2014, knowing that this senator is a Republican?

Q 3.5.7

The events “Republican” and “Up for reelection in 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

S 3.5.7

d

Q 3.5.8

The events “Other” and “Up for reelection in November 2016” are _____

a. mutually exclusive.
b. independent.
c. both mutually exclusive and independent.
d. neither mutually exclusive nor independent.

Q 3.5.9

This table gives the number of participants in the recent National Health Interview Survey who had been treated for cancer in the previous 12 months. The results are sorted by age, race (black or white), and sex. We are interested in possible relationships between age, race, and sex.

Race and Sex	15-24	25-40	41-65	over 65	TOTALS
white, male	1,165	2,036	3,703		8,395
white, female	1,076	2,242	4,060		9,129
black, male	142	194	384		824
black, female	131	290	486		1,061
all others					
TOTALS	2,792	5,279	9,354		21,081

Do not include "all others" for parts f and g.

a. Fill in the column for cancer treatment for individuals over age 65.
b. Fill in the row for all other races.
c. Find the probability that a randomly selected individual was a white male.
d. Find the probability that a randomly selected individual was a black female.
e. Find the probability that a randomly selected individual was black
f. Find the probability that a randomly selected individual was a black or white male.
g. Out of the individuals over age 65, find the probability that a randomly selected individual was a black or white male.

S 3.5.9

a.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others				100	
TOTALS	310	4,650	18,780	6,020	29,760

b.

Race and Sex	1–14	15–24	25–64	over 64	TOTALS
white, male	210	3,360	13,610	4,870	22,050
white, female	80	580	3,380	890	4,930
black, male	10	460	1,060	140	1,670
black, female	0	40	270	20	330
all others	10	210	460	100	780
TOTALS	310	4,650	18,780	6,020	29,760

c. $\frac{22,050}{29,760}$
d. $\frac{330}{29,760}$
e. $\frac{2,000}{29,760}$
f. $\frac{23,720}{29,760}$
g. $\frac{5,010}{6,020}$

Use the following information to answer the next two exercises. The table of data obtained from*www.baseball-almanac.com* shows hit information for four well known baseball players. Suppose that one hit from the table is randomly selected.

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
TOTAL	8,471	1,577	583	1,720	12,351

Q 3.5.10

Find $P(\text{hit was made by Babe Ruth})$.

a. $\frac{1518}{2873}$
b. $\frac{2873}{12351}$
c. $\frac{583}{12351}$
d. $\frac{4189}{12351}$

Q 3.5.11

Find $P(\text{hit was made by Ty Cobb}|\text{The hit was a Home Run})$.

a. $\frac{4189}{12351}$
b. $\frac{114}{1720}$
c. $\frac{1720}{4189}$
d. $\frac{114}{12351}$

S 3.5.11

b

Q 3.5.12

Table identifies a group of children by one of four hair colors, and by type of hair.

Hair Type	Brown	Blond	Black	Red	Totals
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

a. Complete the table.
b. What is the probability that a randomly selected child will have wavy hair?
c. What is the probability that a randomly selected child will have either brown or blond hair?
d. What is the probability that a randomly selected child will have wavy brown hair?
e. What is the probability that a randomly selected child will have red hair, given that he or she has straight hair?
f. If B is the event of a child having brown hair, find the probability of the complement of B.
g. In words, what does the complement of B represent?

Q 3.5.13

In a previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data were compiled into the following table.

Shirt#	≤ 210	211–250	251–290	> 290
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

a. Find the probability that his shirt number is from 1 to 33.
b. Find the probability that he weighs at most 210 pounds.
c. Find the probability that his shirt number is from 1 to 33 AND he weighs at most 210 pounds.
d. Find the probability that his shirt number is from 1 to 33 OR he weighs at most 210 pounds.
e. Find the probability that his shirt number is from 1 to 33 GIVEN that he weighs at most 210 pounds.

S 3.5.13

a. $\frac{26}{106}$
b. $\frac{33}{106}$
c. $\frac{21}{106}$
d. $\left(\frac{26}{106}\right)+\left(\frac{33}{106}\right)-\left(\frac{21}{106}\right)=\left(\frac{38}{106}\right)$
e. $\frac{21}{33}$

3.6: Tree and Venn Diagrams

Exercise 3.6.8

The probability that a man develops some form of cancer in his lifetime is 0.4567. The probability that a man has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Let: $\mathrm{C} =$ a man develops cancer in his lifetime; $\mathrm{P} =$ man has at least one false positive. Construct a tree diagram of the situation.

Answer

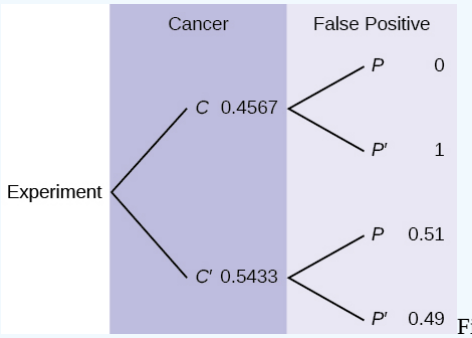


Figure 3.6.12

Bring It Together

Use the following information to answer the next two exercises. Suppose that you have eight cards. Five are green and three are yellow. The cards are well shuffled.

Exercise 3.6.9

Suppose that you randomly draw two cards, one at a time, **with replacement**.

Let $\mathrm{G}_1 =$ first card is green
Let $\mathrm{G}_2 =$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(\mathrm{G}_1 \mathrm{AND} \mathrm{G}_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(\mathrm{G}_2|\mathrm{G}_1)$.
e. Are G_1 and G_2 independent events? Explain why or why not.

Answer

a.

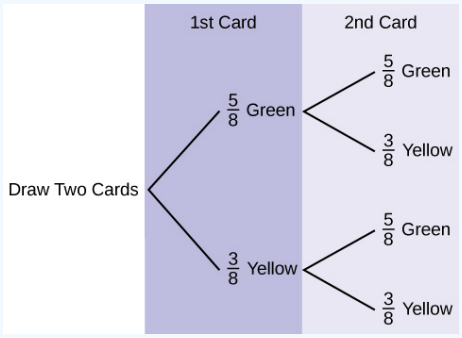


Figure 3.6.14

b. $P(\mathrm{GG}) = \left(\frac{5}{8}\right)\left(\frac{5}{8}\right) = \frac{25}{64}$
c. $P(\text{at least one green}) = P(\mathrm{GG}) + P(\mathrm{GY}) + P(\mathrm{YG}) = \frac{25}{64} + \frac{15}{64} + \frac{15}{64} = \frac{55}{64}$
d. $P(\mathrm{G}|\mathrm{G}) = \frac{5}{8}$
e. Yes, they are independent because the first card is placed back in the bag before the second card is drawn; the composition of cards in the bag remains the same from draw one to draw two.

Exercise 3.6.10

Suppose that you randomly draw two cards, one at a time, **without replacement**.

$G_1 =$ first card is green
$G_2 =$ second card is green

a. Draw a tree diagram of the situation.
b. Find $P(G_1 \text{AND } G_2)$.
c. Find $P(\text{at least one green})$.
d. Find $P(G_2|G_1)$.
e. Are G_2 and G_1 independent events? Explain why or why not.

Use the following information to answer the next two exercises. The percent of licensed U.S. drivers (from a recent year) that are female is 48.60. Of the females, 5.03% are age 19 and under; 81.36% are age 20–64; 13.61% are age 65 or over. Of the licensed U.S. male drivers, 5.04% are age 19 and under; 81.43% are age 20–64; 13.53% are age 65 or over.

Exercise 3.6.11

Complete the following.

a. Construct a table or a tree diagram of the situation.
b. Find $P(\text{driver is female})$.
c. Find $P(\text{driver is age 65 or over}|\text{driver is female})$.
d. Find $P(\text{driver is age 65 or over AND female})$.
e. In words, explain the difference between the probabilities in part c and part d.
f. Find $P(\text{driver is age 65 or over})$.
g. Are being age 65 or over and being female mutually exclusive events? How do you know?

Answer

a.

	<20	20–64	>64	Totals
Female	0.0244	0.3954	0.0661	0.486
Male	0.0259	0.4186	0.0695	0.514
Totals	0.0503	0.8140	0.1356	1

b. $P(\text{F}) = 0.486$
c. $P(>64|\text{F}) = 0.1361$
d. $P(>64 \text{ and } \text{F}) = P(\text{F})P(>64|\text{F}) = (0.486)(0.1361) = 0.0661$
e. $P(>64|\text{F})$ is the percentage of female drivers who are 65 or older and $P(>64 \text{ and } \text{F})$ is the percentage of drivers who are female and 65 or older.
f. $P(>64) = P(>64 \text{ and } \text{F}) + P(>64 \text{ and } \text{M}) = 0.1356$
g. No, being female and 65 or older are not mutually exclusive because they can occur at the same time $P(>64 \text{ and } \text{F}) = 0.0661$.

Exercise 3.6.12

Suppose that 10,000 U.S. licensed drivers are randomly selected.

a. How many would you expect to be male?
b. Using the table or tree diagram, construct a contingency table of gender versus age group.
c. Using the contingency table, find the probability that out of the age 20–64 group, a randomly selected driver is female.

Exercise 3.6.13

Approximately 86.5% of Americans commute to work by car, truck, or van. Out of that group, 84.6% drive alone and 15.4% drive in a carpool. Approximately 3.9% walk to work and approximately 5.3% take public transportation.

a. Construct a table or a tree diagram of the situation. Include a branch for all other modes of transportation to work.
b. Assuming that the walkers walk alone, what percent of all commuters travel alone to work?
c. Suppose that 1,000 workers are randomly selected. How many would you expect to travel alone to work?
d. Suppose that 1,000 workers are randomly selected. How many would you expect to drive in a carpool?

Answer

a.

	Car, Truck or Van	Walk	Public Transportation	Other	Totals
Alone	0.7318				
Not Alone	0.1332				
Totals	0.8650	0.0390	0.0530	0.0430	1

b. If we assume that all walkers are alone and that none from the other two groups travel alone (which is a big assumption) we have: $P(\text{Alone}) = 0.7318 + 0.0390 = 0.7708$
c. Make the same assumptions as in (b) we have: (0.7708)(1,000) = 771
d. (0.1332)(1,000) = 133

Exercise 3.6.14

When the Euro coin was introduced in 2002, two math professors had their statistics students test whether the Belgian one Euro coin was a fair coin. They spun the coin rather than tossing it and found that out of 250 spins, 140 showed a head (event H) while 110 showed a tail (event T). On that basis, they claimed that it is not a fair coin.

a. Based on the given data, find $P(\text{H})$ and $P(\text{T})$.
b. Use a tree to find the probabilities of each possible outcome for the experiment of tossing the coin twice.
c. Use the tree to find the probability of obtaining exactly one head in two tosses of the coin.
d. Use the tree to find the probability of obtaining at least one head.

Exercise 3.6.15

Use the following information to answer the next two exercises. The following are real data from Santa Clara County, CA. As of a certain time, there had been a total of 3,059 documented cases of AIDS in the county. They were grouped into the following categories:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	____
Male	2,146	463	60	135	____
Totals	____	____	____	____	____

Suppose a person with AIDS in Santa Clara County is randomly selected.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person has a risk factor heterosexual contact})$.
c. Find $P(\text{Person is female OR has a risk factor of IV drug user})$.
d. Find $P(\text{Person is female AND has a risk factor of homosexual/bisexual})$.
e. Find $P(\text{Person is male AND has a risk factor of IV drug user})$.

f. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$.
g. Construct a Venn diagram. Make one group females and the other group heterosexual contact.

Answer

The completed contingency table is as follows:

* includes homosexual/bisexual IV drug users

	Homosexual/Bisexual	IV Drug User*	Heterosexual Contact	Other	Totals
Female	0	70	136	49	**255**
Male	2,146	463	60	135	**2,804**
Totals	**2,146**	**533**	**196**	**184**	**3,059**

a. $\frac{255}{2059}$
b. $\frac{196}{3059}$
c. $\frac{718}{3059}$
d. 0
e. $\frac{463}{3059}$
f. $\frac{136}{196}$
g.

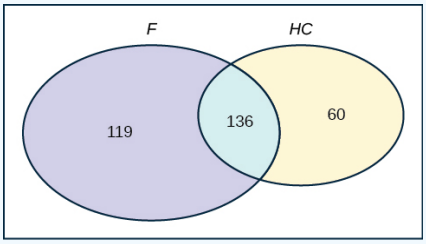


Figure 3.6.15

Exercise 3.6.16

Answer these questions using probability rules. Do NOT use the contingency table. Three thousand fifty-nine cases of AIDS had been reported in Santa Clara County, CA, through a certain date. Those cases will be our population. Of those cases, 6.4% obtained the disease through heterosexual contact and 7.4% are female. Out of the females with the disease, 53.3% got the disease from heterosexual contact.

a. Find $P(\text{Person is female})$.
b. Find $P(\text{Person obtained the disease through heterosexual contact})$.
c. Find $P(\text{Person is female GIVEN person got the disease from heterosexual contact})$
d. Construct a Venn diagram representing this situation. Make one group females and the other group heterosexual contact. Fill in all values as probabilities.

Use the following information to answer the next two exercises. This tree diagram shows the tossing of an unfair coin followed by drawing one bead from a cup containing three red (R), four yellow (Y) and five blue (B) beads. For the coin, $P(\text{H}) = \frac{2}{3}$ and $P(\text{T}) = \frac{1}{3}$ where H is heads and T is tails.

https://stats.libretexts.org/@go/page/24886

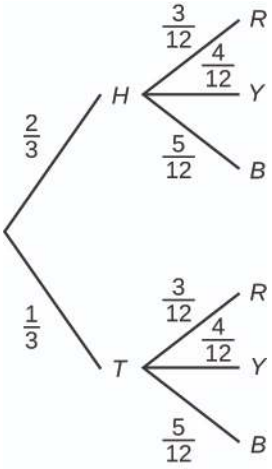


Figure 3.6.1.

Q 3.6.1

Find $P(\text{tossing a Head on the coin AND a Red bead})$

a. $\frac{2}{3}$
b. $\frac{5}{15}$
c. $\frac{6}{36}$
d. $\frac{5}{36}$

Q 3.6.2

Find $P(\text{Blue bead})$.

a. $\frac{15}{36}$
b. $\frac{10}{36}$
c. $\frac{10}{12}$
d. $\frac{6}{36}$

S 3.6.2

a

Q 3.6.3

A box of cookies contains three chocolate and seven butter cookies. Miguel randomly selects a cookie and eats it. Then he randomly selects another cookie and eats it. (How many cookies did he take?)

a. Draw the tree that represents the possibilities for the cookie selections. Write the probabilities along each branch of the tree.
b. Are the probabilities for the flavor of the SECOND cookie that Miguel selects independent of his first selection? Explain.
c. For each complete path through the tree, write the event it represents and find the probabilities.
d. Let S be the event that both cookies selected were the same flavor. Find $P(\text{S})$.
e. Let T be the event that the cookies selected were different flavors. Find $P(\text{T})$ by two different methods: by using the complement rule and by using the branches of the tree. Your answers should be the same with both methods.
f. Let U be the event that the second cookie selected is a butter cookie. Find $P(\text{U})$.

3.7: Probability Topics

This page titled 10.E: Probability Topics (Exericses) 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.E: Probability Topics (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

CHAPTER OVERVIEW

11: The Normal Distribution

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ).

Contributors and Attributions

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11: The Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.1: Prelude to The Normal Distribution

Skills to Develop

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

- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world.

Figure 11.1.1: If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them. The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\left(-\frac{1}{2}\right) \cdot \left(\frac{x-\mu}{\sigma}\right)^2} \tag{11.1.1}$$

The probability density function is a rather complicated function. **Do not memorize it**. It is not necessary.

The cumulative distribution function is $P(X < x)$. It is calculated either by a calculator or a computer, or it is looked up in a table. Technology has made the tables virtually obsolete. For that reason, as well as the fact that there are various table formats, we are not including table instructions.

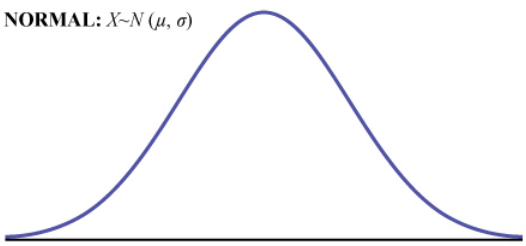


Figure 11.1.2: The standard normal distribution

The curve is symmetrical about a vertical line drawn through the mean, μ. In theory, the mean is the same as the median, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard

deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

COLLABORATIVE CLASSROOM ACTIVITY

Your instructor will record the heights of both men and women in your class, separately. Draw histograms of your data. Then draw a smooth curve through each histogram. Is each curve somewhat bell-shaped? Do you think that if you had recorded 200 data values for men and 200 for women that the curves would look bell-shaped? Calculate the mean for each data set. Write the means on the *x*-axis of the appropriate graph below the peak. Shade the approximate area that represents the probability that one randomly chosen male is taller than 72 inches. Shade the approximate area that represents the probability that one randomly chosen female is shorter than 60 inches. If the total area under each curve is one, does either probability appear to be more than 0.5?

Formula Review

- $X \sim N(\mu, \sigma)$
- $\mu =$ the mean $\sigma =$ the standard deviation

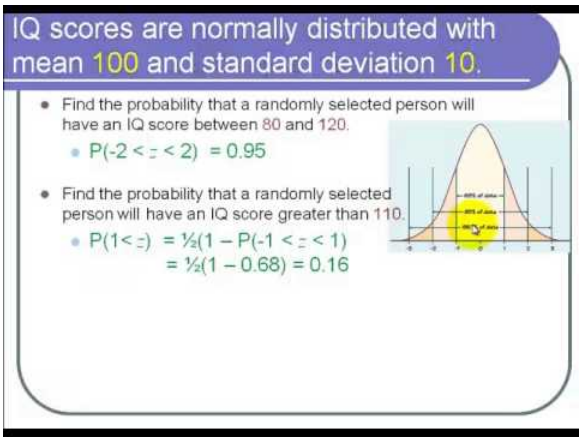
Glossary

Normal Distribution

a continuous random variable (RV) with pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x \cdot \mu)^2}{2\sigma^2}} \tag{11.1.2}$$

, where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called the **standard normal distribution**.



Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.1: Prelude to The Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.2: The Standard Normal Distribution

Z-Scores

The standard normal distribution is a normal distribution of standardized values called *z-scores*. A z-score is measured in units of the standard deviation.

Definition: Z-Score

If X is a normally distributed random variable and $X \sim N(\mu, \sigma)$, then the *z*-score is:

$$z = \frac{x - \mu}{\sigma} \tag{11.2.1}$$

The *z*-score tells you how many standard deviations the value x is above (to the right of) or below (to the left of) the mean, μ. Values of x that are larger than the mean have positive z-scores, and values of x that are smaller than the mean have negative z-scores. If x equals the mean, then x has a z-score of zero. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$\begin{aligned} x &= \mu + (z)(\sigma) \\ &= 5 + (3)(2) = 11 \end{aligned}$$

The *z*-score is three.

Since the mean for the standard normal distribution is zero and the standard deviation is one, then the transformation in Equation 11.2.1 produces the distribution $Z \sim N(0, 1)$. The value x comes from a normal distribution with mean μ and standard deviation σ.

> *A z-score is measured in units of the standard deviation.*

Example 11.2.1

Suppose $X \sim N(5, 6)$. This says that x is a normally distributed random variable with mean $\mu = 5$ and standard deviation $\sigma = 6$. Suppose $x = 17$. Then (via Equation 11.2.1):

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

This means that $x = 17$ is **two** standard deviations (2σ) above or to the right of the mean $\mu = 5$. The standard deviation is $\sigma = 6$.

Notice that: $5 + (2)(6) = 17$ (The pattern is $\mu + z\sigma = x$)

Now suppose $x = 1$. Then:

$$z = \frac{x = \mu}{\sigma} = \frac{1 - 5}{6} = -0.67$$

(rounded to two decimal places)

This means that $x = 1$ is 0.67 standard deviations (-0.67σ) below or to the left of the mean $\mu = 5$. Notice that: $5 + (-0.67)(6)$ is approximately equal to one (This has the pattern $\mu + (-0.67)\sigma = 1$)

Summarizing, when z is positive, x is above or to the right of μ and when z is negative, x is to the left of or below μ. Or, when z is positive, x is greater than μ, and when z is negative x is less than μ.

Exercise 11.2.1

What is the z-score of x, when $x = 1$ and $X \sim N(12, 3)$?

Answer

$$z = \frac{1 - 12}{3} \approx -3.67$$

Example 11.2.2

Some doctors believe that a person can lose five pounds, on the average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let $X =$ the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$. Fill in the blanks.

a. Suppose a person **lost** ten pounds in a month. The z-score when $x = 10$ pounds is $x = 2.5$ (verify). This z-score tells you that $x = 10$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _____. This z-score tells you that $x = -3$ is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. This z-score tells you that $x = 10$ is 2.5 standard deviations to the right of the mean five.

b. Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

$$z = \frac{y - \mu}{\sigma} = \frac{4 - 2}{1} = 2$$

where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of their own) standard deviations to the right of their respective means.

The z-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Exercise 11.2.2

Fill in the blanks.

Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16, 4)$. Suppose Jerome scores ten points in a game. The z–score when $x = 10$ is -1.5. This score tells you that $x = 10$ is _____ standard deviations to the _____(right or left) of the mean_____(What is the mean?).

Answer

1.5, left, 16

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the *Empirical Rule* says the following:

- About 68% of the x values lie between –1σ and +1σ of the mean μ (within one standard deviation of the mean).
- About 95% of the x values lie between –2σ and +2σ of the mean μ (within two standard deviations of the mean).
- About 99.7% of the x values lie between –3σ and +3σ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for +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.

The empirical rule is also known as the 68-95-99.7 rule.

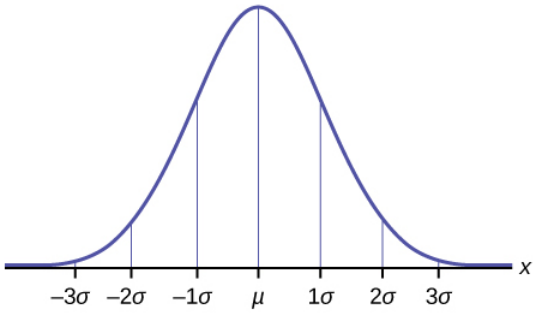


Figure 11.2.1

Example 11.2.3

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when $x = 168$ cm is $z =$ _____. This z-score tells you that $x = 168$ is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = 1.27$. What is the male's height? The z-score ($z = 1.27$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answers

a. –0.32, 0.32, left, 170
b. 177.98, 1.27, right

Exercise 11.2.3

Use the information in Example $11.2.3$ to answer the following questions.

a. Suppose a 15 to 18-year-old male from Chile was 176 cm tall from 2009 to 2010. The z-score when $x = 176$ cm is $z =$ _____. This z-score tells you that $x = 176$ cm is _____ standard deviations to the _____ (right or left) of the mean _____ (What is the mean?).
b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a z-score of $z = -2$. What is the male's height? The z-score ($z = -2$) tells you that the male's height is _____ standard deviations to the _____ (right or left) of the mean.

Answer a

Solve the equation $z = \dfrac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$z = \dfrac{176 - 170}{6.28}$, This z-score tells you that $x = 176$ cm is 0.96 standard deviations to the right of the mean 170 cm.

Answer b

Solve the equation $z = \dfrac{x - \mu}{\sigma}$ for z. $x = \mu + (z)(\sigma)$

$X = 157.44$ cm, The z-score($z = -2$) tells you that the male's height is two standard deviations to the left of the mean.

Example 11.2.4

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males from 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let $X =$ the height of a 15 to 18-year-old male from Chile in 2009 to

2010. Then $X \sim N(170, 6.28)$.

Find the z-scores for $x = 160.58$ cm and $y = 162.85$ cm. Interpret each z-score. What can you say about $x = 160.58$ cm and $y = 162.85$ cm?

Answer

- The z-score (Equation 11.2.1) for $x = 160.58$ is $z = -1.5$.
- The z-score for $y = 162.85$ is $z = -1.5$.

Both $x = 160.58$ and $y = 162.85$ deviate the same number of standard deviations from their respective means and in the same direction.

Exercise 11.2.4

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let $X =$ a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the z-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each z-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

Answer

The z-score (Equation 11.2.1) for $x_1 = 325$ is $z_1 = -1.15$.

The z-score (Equation 11.2.1) for $x_2 = 366.21$ is $z_2 = -1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Example 11.2.5

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between $-1\sigma = (-1)(6) = -6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 - 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie between $-2\sigma = (-2)(6) = -12$ and $2\sigma = (2)(6) = 12$. The values $50 - 12 = 38$ and $50 + 12 = 62$ are within two standard deviations of the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between $-3\sigma = (-3)(6) = -18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 - 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Exercise 11.2.5

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

Answer

between 20 and 30.

Example 11.2.6

From 1984 to 1985, the mean height of 15 to 18-year-old males from Chile was 172.36 cm, and the standard deviation was 6.34 cm. Let $Y =$ the height of 15 to 18-year-old males in 1984 to 1985. Then $Y \sim N(172.36, 6.34)$.

a. About 68% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____ respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer

a. About 68% of the values lie between 166.02 and 178.7. The z-scores are –1 and 1.
b. About 95% of the values lie between 159.68 and 185.04. The z-scores are –2 and 2.
c. About 99.7% of the values lie between 153.34 and 191.38. The z-scores are –3 and 3.

Exercise 11.2.6

The scores on a college entrance exam have an approximate normal 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 _____. The z-scores are _____, respectively.
b. About 95% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.
c. About 99.7% of the y values lie between what two values? These values are _____. The z-scores are _____, respectively.

Answer a

About 68% of the values lie between the values 41 and 63. The z-scores are –1 and 1, respectively.

Answer b

About 95% of the values lie between the values 30 and 74. The z-scores are –2 and 2, respectively.

Answer c

About 99.7% of the values lie between the values 19 and 85. The z-scores are –3 and 3, respectively.

Summary

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If y is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Formula Review

$Z \sim N(0, 1)$

$z = a$ standardized value (z-score)

mean = 0; standard deviation = 1

To find the K^{th} percentile of X when the z-scores is known:

$k = \mu + (z)\sigma$

z-score: $z = \dfrac{x - \mu}{\sigma}$

$Z =$ the random variable for z-scores

$Z \sim N(0, 1)$

Glossary

Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as \(Z \sim N(0, 1)\.

***z*-score**

the linear transformation of the form $z = \frac{x - \mu}{\sigma}$; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma$ the result is the standard normal distribution $Z \sim N(0, 1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently.

References

1. "Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewre...reportid=11960 (accessed May 14, 2013).
2. "The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).
3. "2012 College-Bound Seniors Total Group Profile Report." CollegeBoard, 2012. Available online at http://media.collegeboard.com/digita...Group-2012.pdf (accessed May 14, 2013).
4. "Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009." National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d...s/dt09_147.asp (accessed May 14, 2013).
5. Data from the *San Jose Mercury News*.
6. Data from *The World Almanac and Book of Facts*.
7. "List of stadiums by capacity." Wikipedia. Available online at https://en.wikipedia.org/wiki/List_o...ms_by_capacity (accessed May 14, 2013).
8. Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.2: The Standard Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.2E: The Standard Normal Distribution (Exercises)

Exercise 6.2.7

A bottle of water contains 12.05 fluid ounces with a standard deviation of 0.01 ounces. Define the random variable X in words. $X =$ _____.

Answer

ounces of water in a bottle

Exercise 6.2.8

A normal distribution has a mean of 61 and a standard deviation of 15. What is the median?

Exercise 6.2.9

$X \sim N(1, 2)$

$\sigma =$ _____

Answer

2

Exercise 6.2.10

A company manufactures rubber balls. The mean diameter of a ball is 12 cm with a standard deviation of 0.2 cm. Define the random variable X in words. $X =$ _____.

Exercise 6.2.11

$X \sim N(-4, 1)$

What is the median?

Answer

–4

Exercise 6.2.12

$X \sim N(3, 5)$

$\sigma =$ _____

Exercise 6.2.13

$X \sim N(-2, 1)$

$\mu =$ _____

Answer

–2

Exercise 6.2.14

What does a z-score measure?

Exercise 6.2.15

What does standardizing a normal distribution do to the mean?

Answer

The mean becomes zero.

Exercise 6.2.16

Is $X \sim N(0, 1)$ a standardized normal distribution? Why or why not?

Exercise 6.2.17

What is the z-score of $x = 12$, if it is two standard deviations to the right of the mean?

Answer

$z = 2$

Exercise 6.2.18

What is the z-score of $x = 9$, if it is 1.5 standard deviations to the left of the mean?

Exercise 6.2.19

What is the z-score of $x = -2$, if it is 2.78 standard deviations to the right of the mean?

Answer

$z = 2.78$

Exercise 6.2.20

What is the z-score of $x = 7$, if it is 0.133 standard deviations to the left of the mean?

Exercise 6.2.21

Suppose $X \sim N(2, 6)$. What value of *x* has a *z*-score of three?

Answer

$x = 20$

Exercise 6.2.22

Suppose $X \sim N(8, 1)$. What value of x has a z-score of –2.25?

Exercise 6.2.23

Suppose $X \sim N(9, 5)$. What value of x has a z-score of –0.5?

Answer

$x = 6.5$

Exercise 6.2.24

Suppose $X \sim N(2, 3)$. What value of x has a z-score of –0.67?

Exercise 6.2.25

Suppose $X \sim N(4, 2)$. What value of x is 1.5 standard deviations to the left of the mean?

Answer

$x = 1$

Exercise 6.2.26

Suppose $X \sim N(4, 2)$. What value of x is two standard deviations to the right of the mean?

Exercise 6.2.27

Suppose $X \sim N(8, 9)$. What value of x is 0.67 standard deviations to the left of the mean?

Answer

$x = 1.97$

Exercise 6.2.28

Suppose $X \sim N(-1, 12)$. What is the z-score of $x = 2$?

Exercise 6.2.29

Suppose $X \sim N(12, 6)$. What is the z-score of $x = 2$?

Answer

$z = -1.67$

Exercise 6.2.30

Suppose $X \sim N(9, 3)$. What is the z-score of $x = 9$?

Exercise 6.2.31

Suppose a normal distribution has a mean of six and a standard deviation of 1.5. What is the z-score of $x = 5.5$?

Answer

$z \approx -0.33$

Exercise 6.2.32

In a normal distribution, $x = 5$ and $z = -1.25$. This tells you that $x = 5$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.33

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

0.67, right

Exercise 6.2.34

In a normal distribution, $x = -2$ and $z = 6$. This tells you that $z = -2$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.35

In a normal distribution, $x = -5$ and $z = -3.14$. This tells you that $x = -5$ is ____ standard deviations to the ____ (right or left) of the mean.

Answer

3.14, left

Exercise 6.2.36

In a normal distribution, $x = 6$ and $z = -1.7$. This tells you that $x = 6$ is ____ standard deviations to the ____ (right or left) of the mean.

Exercise 6.2.37

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

Answer

about 68%

Exercise 6.2.38

About what percent of the x values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

Exercise 6.2.39

About what percent of x values lie between the second and third standard deviations (both sides)?

Answer

about 4%

Exercise 6.2.40

Suppose $X \sim N(15, 3)$. Between what x values does 68.27% of the data lie? The range of x values is centered at the mean of the distribution (i.e., 15).

Exercise 6.2.41

Suppose $X \sim N(-3, 1)$. Between what x values does 95.45% of the data lie? The range of x values is centered at the mean of the distribution (i.e., –3).

Answer

between –5 and –1

Exercise 6.2.42

Suppose $X \sim N(-3, 1)$. Between what x values does 34.14% of the data lie?

Exercise 6.2.43

About what percent of x values lie between the mean and three standard deviations?

Answer

about 50%

Exercise 6.2.44

About what percent of x values lie between the mean and one standard deviation?

Exercise 6.2.45

About what percent of x values lie between the first and second standard deviations from the mean (both sides)?

Answer

about 27%

Exercise 6.2.46

About what percent of x values lie between the first and third standard deviations(both sides)?

Use the following information to answer the next two exercises: The life of Sunshine CD players is normally distributed with mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts.

Exercise 6.2.47

Define the random variable X in words. $X =$ _____.

Answer

The lifetime of a Sunshine CD player measured in years.

Exercise 6.2.48

$X \sim$ _____(_____,_____)

11.2E: The Standard Normal Distribution (Exercises) is shared under a CC BY license and was authored, remixed, and/or curated by LibreTexts.

11.3: Using the Normal Distribution

The shaded area in the following graph indicates the area to the left of x. This area is represented by the probability $P(X < x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X < x)$.

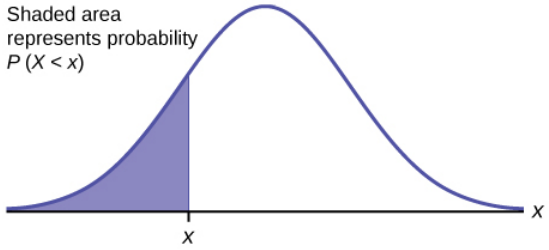
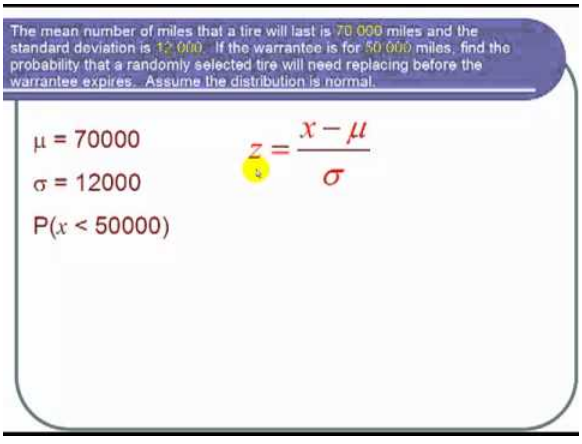


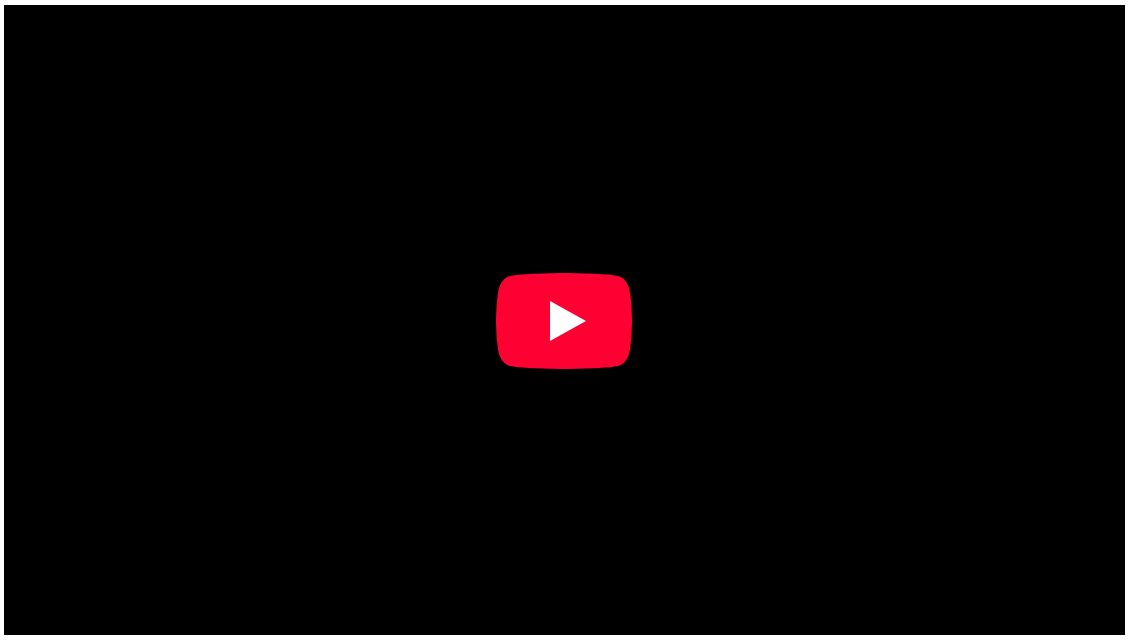
Figure 11.3.1.

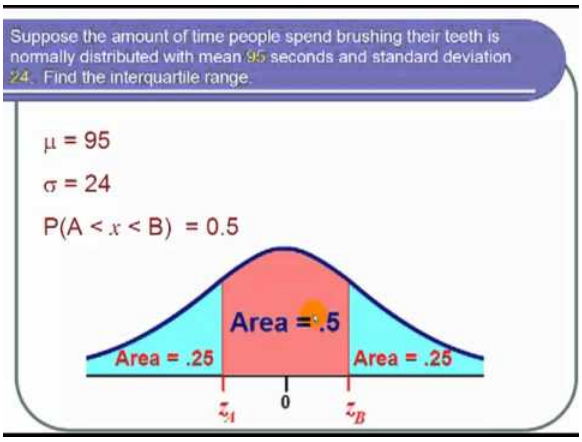
The area to the right is then $P(X > x) = 1 - P(X < x)$. Remember, $P(X < x) =$ Area to the left of the vertical line through x. $P(X < x) = 1 - P(X < x) =$ Area to the right of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

Probabilities are calculated using technology. There are instructions given as necessary for the TI-83+ and TI-84 calculators.







Online Normal Probability Calculator

Fill in all of the values except one below and hit Calculate then the last value will be given to you.

Low:	High:	Mean:	Std. Dev.:	p=

Calculate

Example 11.3.1

If the area to the left is 0.0228, then the area to the right is $1 - 0.0228 = 0.9772$

Exercise 11.3.1

If the area to the left of x is 0.012, then what is the area to the right?

Answer

$1 - 0.012 = 0.988$

Example 11.3.2

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 = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$$P(x > 65) = 0.3446$$

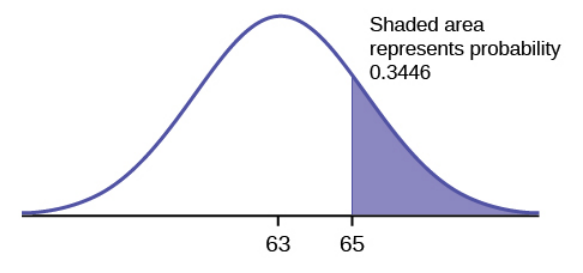


Figure 11.3.2.

The probability that any student selected at random scores more than 65 is 0.3446.

Go into `2nd DISTR` .

After pressing `2nd DISTR` , press `2:normalcdf` .

The syntax for the instructions are as follows:

normalcdf(lower value, upper value, mean, standard deviation) For this problem: normalcdf(65,1E99,63,5) = 0.3446. You get 1E99 (= 10^{99}) by pressing `1` , the `EE` key (a 2nd key) and then `99` . Or, you can enter `10^ 99` instead. The number 10^{99} is way out in the right tail of the normal curve. We are calculating the area between 65 and 10^{99}. In some instances, the lower number of the area might be –1E99 (= -10^{99}). The number -10^{99} is way out in the left tail of the normal curve.

Historical Note

The TI probability program calculates a z-score and then the probability from the z-score. Before technology, the z-score was looked up in a standard normal probability table (because the math involved is too cumbersome) to find the probability. In this example, a standard normal table with area to the left of the z-score was used. You calculate the z-score and look up the area to the left. The probability is the area to the right.

$$z = \frac{65 - 63}{5} \frac{65 - 63}{5} = 0.4$$

Area to the left is 0.6554.

$$P(x > 65) = P(z > 0.4) = 1 - 0.6554 = 0.3446$$

Calculate the z-score:

*Press `2nd Distr`

*Press `3:invNorm` (*Enter the area to the left of z followed by)
*Press `ENTER` .
For this Example, the steps are

2nd Distr
3:invNorm (.6554) ENTER

The answer is 0.3999 which rounds to 0.4.

Answer

b. Draw a graph.

Then find $P(x < 85)$, and shade the graph.

Using a computer or calculator, find $P(x < 85) = 1$.

$\text{normalcdf}(0, 85, 63, 5) = 1$(rounds to one)

The probability that one student scores less than 85 is approximately one (or 100%).

Answer

c. Find the 90th percentile. For each problem or part of a problem, draw a new graph. Draw the x-axis. Shade the area that corresponds to the 90th percentile.

Let $k =$ the 90th percentile. The variable k is located on the x-axis. $P(x < k)$ is the area to the left of k. The 90th percentile k separates the exam scores into those that are the same or lower than k and those that are the same or higher. Ninety percent of the test scores are the same or lower than k, and ten percent are the same or higher. The variable k is often called a critical value.

$k = 69.4$

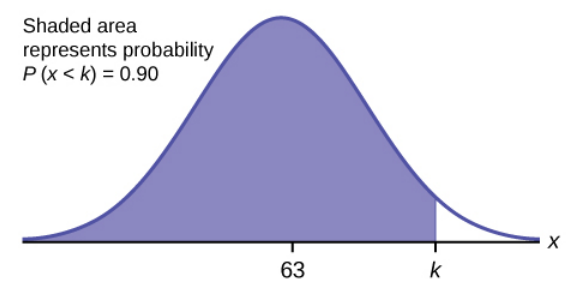


Figure 11.3.3.

The 90th percentile is 69.4. This means that 90% of the test scores fall at or below 69.4 and 10% fall at or above. To get this answer on the calculator, follow this step:

invNorm in 2nd DISTR . invNorm(area to the left, mean, standard deviation)

For this problem, $\text{invNorm}(0.90, 63, 5) = 69.4$

Answer

d. Find the 70th percentile.

Draw a new graph and label it appropriately. $k = 65.6$

The 70th percentile is 65.6. This means that 70% of the test scores fall at or below 65.6 and 30% fall at or above.

$$\text{invNorm}(0.70, 63, 5) = 65.6$$

Exercise 11.3.2

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a randomly selected golfer scored less than 65.

Answer

$\text{normalcdf}(10^{99}, 65, 68, 3) = 0.1587$

Example 11.3.3

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

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 $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

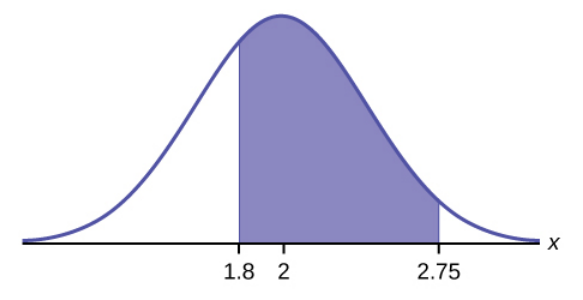


Figure 11.3.4.

$$\text{normalcdf}(1.8, 2.75, 2, 0.5) = 0.5886$$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b.

To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.

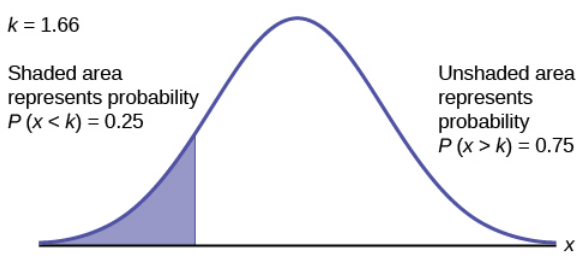


Figure 11.3.5.

$$\text{invNorm}(0.25, 2, 0.5) = 1.66$$

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Exercise 11.3.3

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

Answer

$\text{normalcdf}(66, 70, 68, 3) = 0.4950$

Example 11.3.4

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

a. $\text{normalcdf}(23, 64.7, 36.9, 13.9) = 0.8186$

b. $\text{normalcdf}(-10^{99}, 50.8, 36.9, 13.9) = 0.8413$

c. $\text{invNorm}(0.80, 36.9, 13.9) = 48.6$

The 80th percentile is 48.6 years.

80% of the smartphone users in the age range 13 – 55+ are 48.6 years old or less.

Use the information in Example to answer the following questions.

Exercise 11.3.4

a. Find the 30th percentile, and interpret it in a complete sentence.
b. What is the probability that the age of a randomly selected smartphone user in the range 13 to 55+ is less than 27 years old and at least 0 years old?

70.

Answer

Let $X =$ a smart phone user whose age is 13 to 55+. $X \sim N(36.9, 13.9)$

To find the 30th percentile, find k such that $P(x < k) = 0.30$.
$\text{invNorm}(0.30, 36.9, 13.9) = 29.6$ years
Thirty percent of smartphone users 13 to 55+ are at most 29.6 years and 70% are at least 29.6 years. Find $P(x < 27)$
(Note that $\text{normalcdf}(-10^{99}, 27, 36.9, 13.9) = 0.2382$ The two answers differ only by 0.0040.)

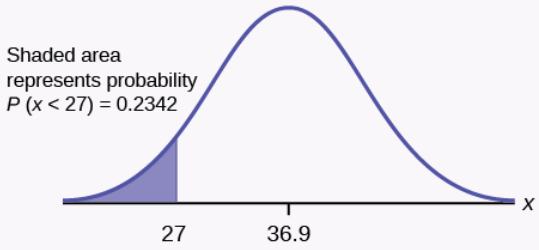


Figure 11.3.6.

$\text{normalcdf}(0, 27, 36.9, 13.9) = 0.2342$

Example 11.3.5

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. Using this information, answer the following questions (round answers to one decimal place).

a. Calculate the interquartile range (IQR).
b. Forty percent of the ages that range from 13 to 55+ are at least what age?

Answer

a.

$$IQR = Q_3 - Q_1$$

Calculate $Q_3 = 75^{\text{th}}$ percentile and $Q_1 = 25^{\text{th}}$ percentile.

$$\text{invNorm}(0.75, 36.9, 13.9) = Q_3 = 46.2754$$

$$\text{invNorm}(0.25, 36.9, 13.9) = Q_1 = 27.5246$$

$$IQR = Q_3 - Q_1 = 18.7508$$

b.

Find k where $P(x > k) = 0.40$ ("At least" translates to "greater than or equal to.")

$0.40 =$ the area to the right.

Area to the left $= 1 - 0.40 = 0.60$.

The area to the left of $k = 0.60$.

$\text{invNorm}(0.60, 36.9, 13.9) = 40.4215$

$k = 40.42$.

Forty percent of the ages that range from 13 to 55+ are at least 40.42 years.

Exercise 11.3.5

Two thousand students took an exam. The scores on the exam have an approximate normal distribution with a mean $\mu = 81$ points and standard deviation $\sigma = 15$ points.

a. Calculate the first- and third-quartile scores for this exam.
b. The middle 50% of the exam scores are between what two values?

Answer

a. $Q_1 = 25^{\text{th}}$ percentile $= \text{invNorm}(0.25, 81, 15) = 70.9$
   $Q_3 = 75^{\text{th}}$ percentile $= \text{invNorm}(0.75, 81, 15) = 91.1$
b. The middle 50% of the scores are between 70.9 and 91.1.

Example 11.3.6

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.
b. The middle 20% of mandarin oranges from this farm have diameters between _____ and _____.
c. Find the 90$^{\text{th}}$ percentile for the diameters of mandarin oranges, and interpret it in a complete sentence.

Answer

a. $\text{normalcdf}(6, 10^{99}, 5.85, 0.24) = 0.2660$

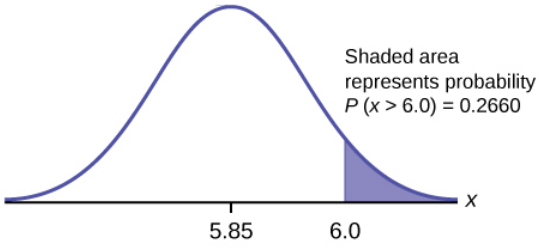


Figure 11.3.7.

Answer

b.

$1 - 0.20 = 0.80$

The tails of the graph of the normal distribution each have an area of 0.40.

Find $k1$, the 40th percentile, and $k2$, the 60th percentile ($0.40 + 0.20 = 0.60$).

$k1 = \text{invNorm}(0.40, 5.85, 0.24) = 5.79\text{cm}$

$k2 = \text{invNorm}(0.60, 5.85, 0.24) = 5.91\text{cm}$

Answer

c. 6.16: Ninety percent of the diameter of the mandarin oranges is at most 6.15 cm.

Exercise 11.3.6

Using the information from Example, answer the following:

a. The middle 40% of mandarin oranges from this farm are between _____ and _____.
b. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 5 cm.

Answer a

The middle area $= 0.40$, so each tail has an area of 0.30.

$-0.40 = 0.60$

The tails of the graph of the normal distribution each have an area of 0.30.

Find $k1$, the 30th percentile and $k2$, the 70th percentile ($0.40 + 0.30 = 0.70$).

$k1 = \text{invNorm}(0.30, 5.85, 0.24) = 5.72\text{cm}$

$k2 = \text{invNorm}(0.70, 5.85, 0.24) = 5.98\text{cm}$

Answer b

$\text{normalcdf}(5, 10^{99}, 5.85, 0.24) = 0.9998$

References

1. “Naegele’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).
2. “403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radi...sode/403/nummi (accessed May 14, 2013).
3. “Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/publi...partment40.cfm (accessed May 14, 2013).
4. “Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).
5. “Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebo...tics/(accessed May 14, 2013).

Chapter Review

The normal distribution, which is continuous, is the most important of all the probability distributions. Its graph is bell-shaped. This bell-shaped curve is used in almost all disciplines. Since it is a continuous distribution, the total area under the curve is one. The parameters of the normal are the mean μ and the standard deviation σ. A special normal distribution, called the standard normal distribution is the distribution of z-scores. Its mean is zero, and its standard deviation is one.

Formula Review

- Normal Distribution: $X \sim N(\mu, \sigma)$ where μ is the mean and σ is the standard deviation.
- Standard Normal Distribution: $Z \sim N(0, 1)$.
- Calculator function for probability: normalcdf (lower x value of the area, upper x value of the area, mean, standard deviation)
- Calculator function for the k^{th} percentile: $k = \text{invNorm}$ (area to the left of k, mean, standard deviation)

Exercise 6.3.7

How would you represent the area to the left of one in a probability statement?

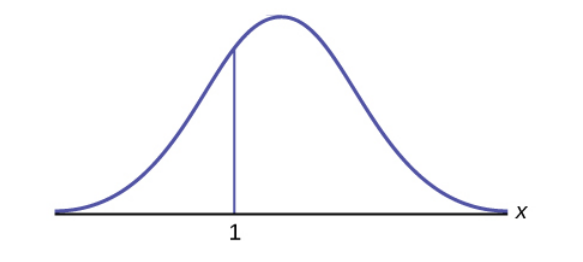


Figure 11.3.8.

Answer

$P(x < 1)$

Exercise 6.3.8

Is $P(x < 1)$ equal to $P(x \leq 1)$? Why?

Answer

Yes, because they are the same in a continuous distribution: $P(x = 1) = 0$

Exercise 6.3.9

How would you represent the area to the left of three in a probability statement?

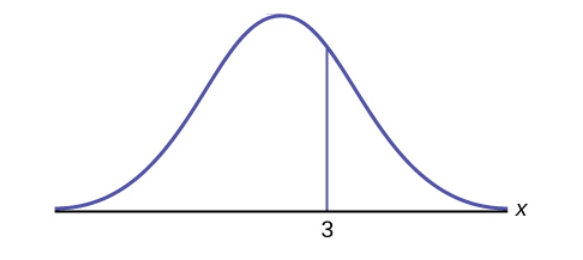


Figure 11.3.10.

Exercise 6.3.10

What is the area to the right of three?

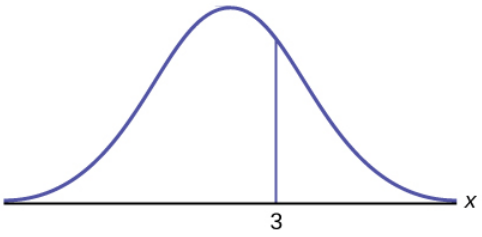


Figure 11.3.11.

Answer

$1 - P(x < 3)$ or $P(x > 3)$

Exercise 6.3.11

If the area to the left of x in a normal distribution is 0.123, what is the area to the right of x?

Exercise 6.3.12

If the area to the right of x in a normal distribution is 0.543, what is the area to the left of x?

Answer

$1 - 0.543 = 0.457$

Use the following information to answer the next four exercises:

$X \sim N(54, 8)$

Exercise 6.3.13

Find the probability that $x > 56$.

Exercise 6.3.14

Find the probability that $x < 30$.

Answer

0.0013

Exercise 6.3.15

Find the 80th percentile.

Exercise 6.3.16

Find the 60th percentile.

Answer

56.03

Exercise 6.3.17

$X \sim N(6, 2)$

Find the probability that x is between three and nine.

Exercise 6.3.18

$X \sim N(-3, 4)$

Find the probability that x is between one and four.

Answer

0.1186

Exercise 6.3.19

$X \sim N(4, 5)$

Find the maximum of x in the bottom quartile.

Exercise 6.3.20

Use the following information to answer the next three exercise: The life of Sunshine CD players is normally distributed with a mean of 4.1 years and a standard deviation of 1.3 years. A CD player is guaranteed for three years. We are interested in the length of time a CD player lasts. Find the probability that a CD player will break down during the guarantee period.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

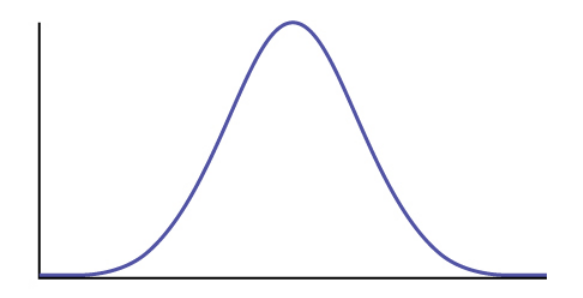


***Figure* 11.3.12.**

$P(0 < x <$ _____$)=$ _____ (Use zero for the minimum value of x.)

Answer

a. Check student's solution.
b. 3, 0.1979

Exercise 6.3.21

Find the probability that a CD player will last between 2.8 and six years.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the probability.

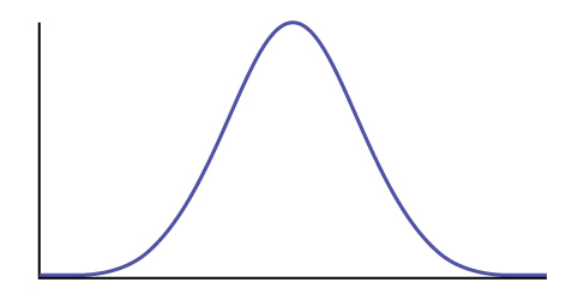


***Figure* 11.3.13.**

$P($_____$< x <$ _____$)$ = _____

Exercise 6.3.22

Find the 70th percentile of the distribution for the time a CD player lasts.

a. Sketch the situation. Label and scale the axes. Shade the region corresponding to the lower 70%.

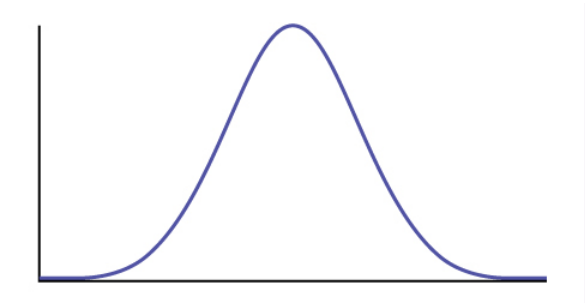


***Figure* 11.3.14.**

$P(x < k) =$ _____ Therefore, $k =$ _____

Answer

a. Check student's solution.
b. 0.70, 4.78 years

Contributors

- Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.

This page titled 11.3: Using the Normal Distribution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.4: Normal Distribution - Lap Times (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcome

- The student will compare and contrast empirical data and a theoretical distribution to determine if Terry Vogel's lap times fit a continuous distribution.

Directions

Round the relative frequencies and probabilities to four decimal places. Carry all other decimal answers to two places.

Collect the Data

1. Use the data from Appendix C. Use a stratified sampling method by lap (races 1 to 20) and a random number generator to pick six lap times from each stratum. Record the lap times below for laps two to seven.

_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____
_____	_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

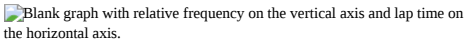


Figure 6.4.1.

3. Calculate the following:
   a. $\bar{x}$ = _____
   b. s = _____
4. Draw a smooth curve through the tops of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram to help, what is the approximate theoretical distribution of the data?

- $X \sim$ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Use the data you collected to complete the following statements.

- The IQR goes from _____ to _____.
- IQR = _____. ($IQR = Q_3 - Q_1$)
- The 15^{th} percentile is _____.

- The 85th percentile is _____.
- The median is _____.
- The empirical probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. You should use a normal approximation based on your sample data.

- The *IQR* goes from _____ to _____.
- *IQR* = _____.
- The 15th percentile is _____.
- The 85th percentile is _____.
- The median is _____.
- The probability that a randomly chosen lap time is more than 130 seconds is _____.
- Explain the meaning of the 85th percentile of this distribution.

Discussion Questions

Do the data from the section titled Collect the Data give a close approximation to the theoretical distribution in the section titled Analyze the Distribution? In complete sentences and comparing the result in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

This page titled 11.4: Normal Distribution - Lap Times (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.5: Normal Distribution - Pinkie Length (Worksheet)

Name: _____

Section: _____

Student ID#:_____

Work in groups on these problems. You should try to answer the questions without referring to your textbook. If you get stuck, try asking another group for help.

Student Learning Outcomes

- The student will compare empirical data and a theoretical distribution to determine if data from the experiment follow a continuous distribution.

Collect the Data

Measure the length of your pinky finger (in centimeters).

1. Randomly survey 30 adults for their pinky finger lengths. Round the lengths to the nearest 0.5 cm.

_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

2. Construct a histogram. Make five to six intervals. Sketch the graph using a ruler and pencil. Scale the axes.

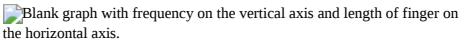


Figure 6.5.1.

3. Calculate the following.
   a. $\bar{x}$ = _____
   b. s = _____
4. Draw a smooth curve through the top of the bars of the histogram. Write one to two complete sentences to describe the general shape of the curve. (Keep it simple. Does the graph go straight across, does it have a v-shape, does it have a hump in the middle or at either end, and so on?)

Analyze the Distribution

Using your sample mean, sample standard deviation, and histogram, what was the approximate theoretical distribution of the data you collected?

- $X \sim$ _____(_____,_____)
- How does the histogram help you arrive at the approximate distribution?

Describe the Data

Using the data you collected complete the following statements. (Hint: order the data)

REMEMBER

($IQR = Q_3 - Q_1$)

- IQR = _____
- The 15^{th} percentile is _____.
- The 85^{th} percentile is _____.

- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Theoretical Distribution

Using the theoretical distribution, complete the following statements. Use a normal approximation based on the sample mean and standard deviation.

- *IQR* = _____
- The 15th percentile is _____.
- The 85th percentile is _____.
- Median is _____.
- What is the theoretical probability that a randomly chosen pinky length is more than 6.5 cm?
- Explain the meaning of the 85th percentile of this data.

Discussion Questions

Do the data you collected give a close approximation to the theoretical distribution? In complete sentences and comparing the results in the sections titled Describe the Data and Theoretical Distribution, explain why or why not.

This page titled 11.5: Normal Distribution - Pinkie Length (Worksheet) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

11.E: The Normal Distribution (Exercises)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

6.1: Introduction

6.2: The Standard Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Q 6.2.1

What is the median recovery time?

a. 2.7
b. 5.3
c. 7.4
d. 2.1

Q 6.2.2

What is the *z*-score for a patient who takes ten days to recover?

a. 1.5
b. 0.2
c. 2.2
d. 7.3

S 6.2.2

c

Q 6.2.3

The length of time to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes. If the mean is significantly greater than the standard deviation, which of the following statements is true?

I. The data cannot follow the uniform distribution.
II. The data cannot follow the exponential distribution..
III. The data cannot follow the normal distribution.

a. I only
b. II only
c. III only
d. I, II, and III

Q 6.2.4

The heights of the 430 National Basketball Association players were listed on team rosters at the start of the 2005–2006 season. The heights of basketball players have an approximate normal distribution with mean, μ = 79 inches and a standard deviation, σ = 3.89 inches. For each of the following heights, calculate the *z*-score and interpret it using complete sentences.

a. 77 inches
b. 85 inches
c. If an NBA player reported his height had a *z*-score of 3.5, would you believe him? Explain your answer.

S 6.2.4

a. Use the z-score formula. $z = -0.5141$. The height of 77 inches is 0.5141 standard deviations below the mean. An NBA player whose height is 77 inches is shorter than average.
b. Use the z-score formula. $z = 1.5424$. The height 85 inches is 1.5424 standard deviations above the mean. An NBA player whose height is 85 inches is taller than average.

c. Height $= 79 + 3.5(3.89) = 90.67$ inches, which is over 7.7 feet tall. There are very few NBA players this tall so the answer is no, not likely.

Q 6.2.5

The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. Systolic blood pressure for males follows a normal distribution.

a. Calculate the *z*-scores for the male systolic blood pressures 100 and 150 millimeters.
b. If a male friend of yours said he thought his systolic blood pressure was 2.5 standard deviations below the mean, but that he believed his blood pressure was between 100 and 150 millimeters, what would you say to him?

Q 6.2.6

Kyle's doctor told him that the *z*-score for his systolic blood pressure is 1.75. Which of the following is the best interpretation of this standardized score? The systolic blood pressure (given in millimeters) of males has an approximately normal distribution with mean $\mu = 125$ and standard deviation $\sigma = 14$. If $X =$ a systolic blood pressure score then $X \sim N(125, 14)$.

a. Which answer(s) **is/are** correct?
   i. Kyle's systolic blood pressure is 175.
   ii. Kyle's systolic blood pressure is 1.75 times the average blood pressure of men his age.
   iii. Kyle's systolic blood pressure is 1.75 above the average systolic blood pressure of men his age.
   iv. Kyles's systolic blood pressure is 1.75 standard deviations above the average systolic blood pressure for men.
b. Calculate Kyle's blood pressure.

S 6.2.6

a. iv
b. Kyle's blood pressure is equal to $125 + (1.75)(14) = 149.5$.

Q 6.2.7

Height and weight are two measurements used to track a child's development. The World Health Organization measures child development by comparing the weights of children who are the same height and the same gender. In 2009, weights for all 80 cm girls in the reference population had a mean $\mu = 10.2$ kg and standard deviation $\sigma = 0.8$ kg. Weights are normally distributed. $X \sim N(10.2, 0.8)$. Calculate the *z*-scores that correspond to the following weights and interpret them.

a. 11 kg
b. 7.9 kg
c. 12.2 kg

Q 6.2.8

In 2005, 1,475,623 students heading to college took the SAT. The distribution of scores in the math section of the SAT follows a normal distribution with mean $\mu = 520$ and standard deviation $\sigma = 115$.

a. Calculate the z-score for an SAT score of 720. Interpret it using a complete sentence.
b. What math SAT score is 1.5 standard deviations above the mean? What can you say about this SAT score?
c. For 2012, the SAT math test had a mean of 514 and standard deviation 117. The ACT math test is an alternate to the SAT and is approximately normally distributed with mean 21 and standard deviation 5.3. If one person took the SAT math test and scored 700 and a second person took the ACT math test and scored 30, who did better with respect to the test they took?

S 6.2.8

Let $X =$ an SAT math score and $Y =$ an ACT math score.

a. $X = 720 \frac{720-520}{15} = 1.74$ The exam score of 720 is 1.74 standard deviations above the mean of 520.
b. $z = 1.5$
   The math SAT score is $520 + 1.5(115) \approx 692.5$ The exam score of 692.5 is 1.5 standard deviations above the mean of 520.
c. $\frac{X-\mu}{\sigma} = \frac{700-514}{117} \approx 1.59$, the *z*-score for the SAT. $\frac{Y-\mu}{\sigma} = \frac{30-21}{5.3} \approx 1.70$, the *z*-scores for the ACT. With respect to the test they took, the person who took the ACT did better (has the higher *z*-score).

https://stats.libretexts.org/@go/page/24894

6.3: Using the Normal Distribution

Use the following information to answer the next two exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

Q 6.3.1

What is the probability of spending more than two days in recovery?

a. 0.0580
b. 0.8447
c. 0.0553
d. 0.9420

Q 6.3.2

The 90^{th} percentile for recovery times is?

a. 8.89
b. 7.07
c. 7.99
d. 4.32

S 6.3.2

c

Use the following information to answer the next three exercises: The length of time it takes to find a parking space at 9 A.M. follows a normal distribution with a mean of five minutes and a standard deviation of two minutes.

Q 6.3.3

Based upon the given information and numerically justified, would you be surprised if it took less than one minute to find a parking space?

a. Yes
b. No
c. Unable to determine

Q 6.3.4

Find the probability that it takes at least eight minutes to find a parking space.

a. 0.0001
b. 0.9270
c. 0.1862
d. 0.0668

S 6.3.4

d

Q 6.3.5

Seventy percent of the time, it takes more than how many minutes to find a parking space?

a. 1.24
b. 2.41
c. 3.95
d. 6.05

Q 6.3.6

According to a study done by De Anza students, the height for Asian adult males is normally distributed with an average of 66 inches and a standard deviation of 2.5 inches. Suppose one Asian adult male is randomly chosen. Let $X =$ height of the individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person is between 65 and 69 inches. Include a sketch of the graph, and write a probability statement.
c. Would you expect to meet many Asian adult males over 72 inches? Explain why or why not, and justify your answer numerically.
d. The middle 40% of heights fall between what two values? Sketch the graph, and write the probability statement.

S 6.3.6

a. $X \sim N(66, 2.5)$
b. 0.5404
c. No, the probability that an Asian male is over 72 inches tall is 0.0082

Q 6.3.7

IQ is normally distributed with a mean of 100 and a standard deviation of 15. Suppose one individual is randomly chosen. Let $X =$ IQ of an individual.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the person has an IQ greater than 120. Include a sketch of the graph, and write a probability statement.
c. MENSA is an organization whose members have the top 2% of all IQs. Find the minimum IQ needed to qualify for the MENSA organization. Sketch the graph, and write the probability statement.
d. The middle 50% of IQs fall between what two values? Sketch the graph and write the probability statement.

Q 6.3.8

The percent of fat calories that a person in America consumes each day is normally distributed with a mean of about 36 and a standard deviation of 10. Suppose that one individual is randomly chosen. Let $X =$ percent of fat calories.

a. $X \sim$ _____(_____,_____)
b. Find the probability that the percent of fat calories a person consumes is more than 40. Graph the situation. Shade in the area to be determined.
c. Find the maximum number for the lower quarter of percent of fat calories. Sketch the graph and write the probability statement.

S 6.3.8

a. $X \sim N(36, 10)$
b. The probability that a person consumes more than 40% of their calories as fat is 0.3446.
c. Approximately 25% of people consume less than 29.26% of their calories as fat.

Q 6.3.9

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet.

a. If $X =$ distance in feet for a fly ball, then $X \sim$ _____(_____,_____)
b. If one fly ball is randomly chosen from this distribution, what is the probability that this ball traveled fewer than 220 feet? Sketch the graph. Scale the horizontal axis X. Shade the region corresponding to the probability. Find the probability.
c. Find the 80th percentile of the distribution of fly balls. Sketch the graph, and write the probability statement.

Q 6.3.10

In China, four-year-olds average three hours a day unsupervised. Most of the unsupervised children live in rural areas, considered safe. Suppose that the standard deviation is 1.5 hours and the amount of time spent alone is normally distributed. We randomly select one Chinese four-year-old living in a rural area. We are interested in the amount of time the child spends alone per day.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. Find the probability that the child spends less than one hour per day unsupervised. Sketch the graph, and write the probability statement.
d. What percent of the children spend over ten hours per day unsupervised?
e. Seventy percent of the children spend at least how long per day unsupervised?

S 6.3.10

a. $X =$ number of hours that a Chinese four-year-old in a rural area is unsupervised during the day.
b. X $N(3, 1.5)$
c. The probability that the child spends less than one hour a day unsupervised is 0.0918.
d. The probability that a child spends over ten hours a day unsupervised is less than 0.0001.
e. 2.21 hours

Q 6.3.11

In the 1992 presidential election, Alaska's 40 election districts averaged 1,956.8 votes per district for President Clinton. The standard deviation was 572.3. (There are only 40 election districts in Alaska.) The distribution of the votes per district for President Clinton was bell-shaped. Let $X =$ number of votes for President Clinton for an election district.

a. State the approximate distribution of X.
b. Is 1,956.8 a population mean or a sample mean? How do you know?
c. Find the probability that a randomly selected district had fewer than 1,600 votes for President Clinton. Sketch the graph and write the probability statement.
d. Find the probability that a randomly selected district had between 1,800 and 2,000 votes for President Clinton.
e. Find the third quartile for votes for President Clinton.

Q 6.3.12

Suppose that the duration of a particular type of criminal trial is known to be normally distributed with a mean of 21 days and a standard deviation of seven days.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. If one of the trials is randomly chosen, find the probability that it lasted at least 24 days. Sketch the graph and write the probability statement.
d. Sixty percent of all trials of this type are completed within how many days?

S 6.3.12

a. $X =$ the distribution of the number of days a particular type of criminal trial will take
b. $X \sim N(21, 7)$
c. The probability that a randomly selected trial will last more than 24 days is 0.3336.
d. 22.77

Q 6.3.13

Terri Vogel, an amateur motorcycle racer, averages 129.71 seconds per 2.5 mile lap (in a seven-lap race) with a standard deviation of 2.28 seconds. The distribution of her race times is normally distributed. We are interested in one of her randomly selected laps.

a. In words, define the random variable X.
b. $X \sim$ _____(_____,_____)
c. Find the percent of her laps that are completed in less than 130 seconds.
d. The fastest 3% of her laps are under _____.
e. The middle 80% of her laps are from _____ seconds to _____ seconds.

Q 6.3.14

Thuy Dau, Ngoc Bui, Sam Su, and Lan Voung conducted a survey as to how long customers at Lucky claimed to wait in the checkout line until their turn. Let $X =$ time in line. Table displays the ordered real data (in minutes):

0.50	4.25	5	6	7.25
1.75	4.25	5.25	6	7.25
2	4.25	5.25	6.25	7.25
2.25	4.25	5.5	6.25	7.75

2.25	4.5	5.5	6.5	8
2.5	4.75	5.5	6.5	8.25
2.75	4.75	5.75	6.5	9.5
3.25	4.75	5.75	6.75	9.5
3.75	5	6	6.75	9.75
3.75	5	6	6.75	10.75

a. Calculate the sample mean and the sample standard deviation.
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____)
f. Use the distribution in part e to calculate the probability that a person will wait fewer than 6.1 minutes.
g. Determine the cumulative relative frequency for waiting less than 6.1 minutes.
h. Why aren't the answers to part f and part g exactly the same?
i. Why are the answers to part f and part g as close as they are?
j. If only ten customers has been surveyed rather than 50, do you think the answers to part f and part g would have been closer together or farther apart? Explain your conclusion.

S 6.3.14

a. mean $= 5.51$, $s = 2.15$
b. Check student's solution.
c. Check student's solution.
d. Check student's solution.
e. $X \sim N(5.51, 2.15)$
f. 0.6029
g. The cumulative frequency for less than 6.1 minutes is 0.64.
h. The answers to part f and part g are not exactly the same, because the normal distribution is only an approximation to the real one.
i. The answers to part f and part g are close, because a normal distribution is an excellent approximation when the sample size is greater than 30.
j. The approximation would have been less accurate, because the smaller sample size means that the data does not fit normal curve as well.

Q 6.3.15

Suppose that Ricardo and Anita attend different colleges. Ricardo's GPA is the same as the average GPA at his school. Anita's GPA is 0.70 standard deviations above her school average. In complete sentences, explain why each of the following statements may be false.

a. Ricardo's actual GPA is lower than Anita's actual GPA.
b. Ricardo is not passing because his z-score is zero.
c. Anita is in the 70th percentile of students at her college.

Q 6.3.16

Table shows a sample of the maximum capacity (maximum number of spectators) of sports stadiums. The table does not include horse-racing or motor-racing stadiums.

40,000	40,000	45,050	45,500	46,249	48,134
49,133	50,071	50,096	50,466	50,832	51,100

51,500	51,900	52,000	52,132	52,200	52,530
52,692	53,864	54,000	55,000	55,000	55,000
55,000	55,000	55,000	55,082	57,000	58,008
59,680	60,000	60,000	60,492	60,580	62,380
62,872	64,035	65,000	65,050	65,647	66,000
66,161	67,428	68,349	68,976	69,372	70,107
70,585	71,594	72,000	72,922	73,379	74,500
75,025	76,212	78,000	80,000	80,000	82,300

a. Calculate the sample mean and the sample standard deviation for the maximum capacity of sports stadiums (the data).
b. Construct a histogram.
c. Draw a smooth curve through the midpoints of the tops of the bars of the histogram.
d. In words, describe the shape of your histogram and smooth curve.
e. Let the sample mean approximate μ and the sample standard deviation approximate σ. The distribution of X can then be approximated by $X \sim$ _____(_____,_____).
f. Use the distribution in part e to calculate the probability that the maximum capacity of sports stadiums is less than 67,000 spectators.
g. Determine the cumulative relative frequency that the maximum capacity of sports stadiums is less than 67,000 spectators. Hint: Order the data and count the sports stadiums that have a maximum capacity less than 67,000. Divide by the total number of sports stadiums in the sample.
h. Why aren't the answers to part f and part g exactly the same?

S 6.3.16

a. $\text{mean} = 60,136$, $s = 10,468$
b. Answers will vary.
c. Answers will vary.
d. Answers will vary.
e. $X \sim N(60136, 10468)$
f. 0.7440
g. The cumulative relative frequency is $\frac{43}{60} = 0.717$.
h. The answers for part f and part g are not the same, because the normal distribution is only an approximation.

Q 6.3.17

An expert witness for a paternity lawsuit testifies that the length of a pregnancy is normally distributed with a mean of 280 days and a standard deviation of 13 days. An alleged father was out of the country from 240 to 306 days before the birth of the child, so the pregnancy would have been less than 240 days or more than 306 days long if he was the father. The birth was uncomplicated, and the child needed no medical intervention. What is the probability that he was NOT the father? What is the probability that he could be the father? Calculate the *z*-scores first, and then use those to calculate the probability.

Q 6.3.18

A NUMMI assembly line, which has been operating since 1984, has built an average of 6,000 cars and trucks a week. Generally, 10% of the cars were defective coming off the assembly line. Suppose we draw a random sample of n = 100 cars. Let X represent the number of defective cars in the sample. What can we say about X in regard to the 68-95-99.7 empirical rule (one standard deviation, two standard deviations and three standard deviations from the mean are being referred to)? Assume a normal distribution for the defective cars in the sample.

S 6.3.18

- $n = 100; p = 0.1; q = 0.9$
- $\mu = np = (100)(0.10) = 10$

- $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

a. $z = \pm : x_1 = \mu + z\sigma = 10 + 1(3) = 13$ and $x_2 = \mu = z\sigma = 10 - 1(3) = 7.68$ of the defective cars will fall between seven and 13.
b. $z = \pm : x_1 = \mu + z\sigma = 10 + 2(3) = 16$ and $x_2 = \mu = z\sigma = 10 - 2(3) = 4.95$ of the defective cars will fall between four and 16
c. $z = \pm : x_1 = \mu + z\sigma = 10 + 3(3) = 19$ and $x_2 = \mu = z\sigma = 10 - 3(3) = 1.997$ of the defective cars will fall between one and 19.

Q 6.3.19

We flip a coin 100 times ($n = 100$) and note that it only comes up heads 20% ($p = 0.20$) of the time. The mean and standard deviation for the number of times the coin lands on heads is $\mu = 20$ and $\sigma = 4$ (verify the mean and standard deviation). Solve the following:

a. There is about a 68% chance that the number of heads will be somewhere between ___ and ___.
b. There is about a ____chance that the number of heads will be somewhere between 12 and 28.
c. There is about a ____ chance that the number of heads will be somewhere between eight and 32.

Q 6.3.20

A $1 scratch off lotto ticket will be a winner one out of five times. Out of a shipment of $n = 190$ lotto tickets, find the probability for the lotto tickets that there are

a. somewhere between 34 and 54 prizes.
b. somewhere between 54 and 64 prizes.
c. more than 64 prizes.

S 6.3.21

- $n = 190; p = 1515 = 0.2; q = 0.8$
- $\mu = np = (190)(0.2) = 38$
- $\sigma = \sqrt{npq} = \sqrt{(190)(0.2)(0.8)} = 5.5136$

a. For this problem: $P(34 < x < 54) = \text{normalcdf}(34, 54, 48, 5.5136) = 0.7641$
b. For this problem: $P(54 < x < 64) = \text{normalcdf}(54, 64, 48, 5.5136) = 0.0018$
c. For this problem: $P(x > 64) = \text{normalcdf}(64, 10^{99}, 48, 5.5136) = 0.0000012$(approximately 0)

Q 6.3.22

Facebook provides a variety of statistics on its Web site that detail the growth and popularity of the site.

On average, 28 percent of 18 to 34 year olds check their Facebook profiles before getting out of bed in the morning. Suppose this percentage follows a normal distribution with a standard deviation of five percent.

a. Find the probability that the percent of 18 to 34-year-olds who check Facebook before getting out of bed in the morning is at least 30.
b. Find the 95th percentile, and express it in a sentence.

6.4: Normal Distribution (Lap Times)

6.5: Normal Distribution (Pinkie Length)

This page titled 11.E: The Normal Distribution (Exercises) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

- **6.E: The Normal Distribution (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/introductory-statistics.

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Pre-Statistics

Webpages: 89

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 83.1% (74 pages)
- Undeclared: 11.2% (10 pages)
- CC BY-NC-SA 4.0: 5.6% (5 pages)

By Page