CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-content, web-based collaborative model termed the FlexBook®, CK-12 intends to pioneer the generation and distribution of high-quality educational content that will serve both as core text as well as provide an adaptive environment for learning, powered through the FlexBook Platform®.

Copyright © 2011 CK-12 Foundation, www.ck12.org

The names “CK-12” and “CK12” and associated logos and the terms “FlexBook®”, and “FlexBook Platform®", (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections must include the referral attribution link http://www.ck12.org/saythanks (placed in a visible location) in addition to the following terms.

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the Creative Commons Attribution/Non-Commercial/Share Alike 3.0 Unported (CC-by-NC-SA) License (http://creativecommons.org/licenses/by-nc-sa/3.0/), as amended and updated by Creative Commons from time to time (the “CC License”), which is incorporated herein by this reference.

Complete terms can be found at http://www.ck12.org/terms.

Printed: June, 2016
Authors
Ms. Heather Haney, Mr. Ernest Johnson

Contributors
Mr. Bruce DeWitt, Mr. Michael Engelhaupt, Mr. Matthew Henderson,
Ms. Anne Roehrich, Mr. Tom Skoglund, Ms. Renee Voltin

Editors
Ms. Katie Bruck, Ms. Elizabeth Dorsing, Ms. Wendy Durant, Mr. Mark Hedin
Mr. Charles Nowariak, Ms. Julie Rydberg, Ms. Meghann Witchger
Table of Contents

Table of Contents iii
Foreword v
Preface vi

1 Counting Methods 1
  1.1 Sample Spaces, Events, and Outcomes ................................................................. 1
  1.2 Fundamental Counting Principle ................................................................. 4
  1.3 Permutations .................................................................................................. 10
  1.4 Combinations .............................................................................................. 15
  1.5 Mixed Combinations and Permutations .................................................. 19
  1.6 Chapter 1 Review ...................................................................................... 23

2 Calculating Probabilities 26
  2.1 Calculating Basic Probabilities ........................................................................... 26
  2.2 Compound and Independent Events .......................................................... 33
  2.3 Mutually Exclusive Outcomes .......................................................................... 41
  2.4 Tree Diagrams and Probability Models ..................................................... 48
  2.5 Conditional Probabilities and 2-Way Tables .............................................. 54
  2.6 Chapter 2 Review ...................................................................................... 62

3 Expected Values & Simulation 67
  3.1 Probability Models & Expected Value .......................................................... 67
  3.2 Applied Expected Value Calculations ......................................................... 76
  3.3 Simulation and Experimental Probability .................................................. 82
  3.4 Chapter 3 Review ...................................................................................... 90

4 Data Collection 93
  4.1 DATA ........................................................................................................ 93
  4.2 Sample Survey and Census ......................................................................... 101
  4.3 Random Selection ....................................................................................... 114
  4.4 Statistical Conclusions ............................................................................... 119
  4.5 Experiments and Observational Studies .................................................. 124
  4.6 Chapter 4 Review ...................................................................................... 131
Foreword

Anoka-Hennepin Schools is fortunate to have many experienced math teachers contribute to this project. The primary authors, along with the editing team, have worked tirelessly over several years to write and keep this flexbook up to date, often with very tight deadlines.

Meet the Authors

Heather Haney has taught high school mathematics for 24 years and currently teaches at Coon Rapids High School in Coon Rapids, MN. She received her BS in Mathematics from St. Cloud State, MN (1991) and her M. Ed. in Curriculum and Instruction from Texas Wesleyan University (2003). Heather teaches a variety of math courses, including AP Statistics and general Probability and Statistics.

Ernest Johnson currently teaches mathematics at Andover High School in Andover, MN. He has taught mathematics for 24 years teaching courses varying from Algebra, to Statistics, to AP Calculus. Ernest graduated from the University of Minnesota in 1992 with a B.S. in Mathematics and received a M.Ed. in Instructional Systems and Technology in 1998 from the University of Minnesota.

About the Videos

Helpful video links have been added to the Flexbook by Matthew Henderson. Matt teaches at Andover HS in Andover, MN. He received a B.S. in Secondary Mathematics Education from Northwestern College in St. Paul, MN (1996) and his M.Ed. in Curriculum & Instruction: Instructional Systems and Technology at the University of Minnesota (2002). Matt teaches AP Calculus and AP Computer Programming as well as both online and offline Probability and Statistics.

https://bit.ly/probstatsUnit1

When you see a QR code it indicates that at least one video tutorial is available for the related material. Any of three methods can be used to access the video resource. When using a mobile device, a QR reader can be used to read the QR code to view the videos. If accessing this textbook in an electronic format, by either clicking on the live hyperlink below the QR code, the user will be directed to the video or by typing the web address into any browser, the user will be able to navigate to the resource indicated. If you do not have internet or have a slow connection, talk to your teacher about other options for accessing the videos.
Preface

About the Book
Anoka-Hennepin Schools is thrilled to release the fifth publication of its very own Probability and Statistics textbook. *Anoka-Hennepin Probability and Statistics* (Fifth Edition) represents the work of a large team of dedicated writers and editors who have produced a truly unique and flexible “ebook.” Available in multiple electronic formats, the content demonstrates 21st century math learning at its finest. Students can access the book from anywhere at any time through a variety of electronic devices in multiple formats.

Technology
While paper copies will be available for student use, the electronic version of the book is interactive and includes web site links, simulations, videos, and real world statistical examples. Students can access the textbook through the district Learning Management Site Moodle where large amounts of supplemental and enrichment content can also be found.

The book incorporates the use of the TI-83/84 graphing calculators along with spreadsheet software to display and manipulate statistical data. Additional content is available through Kahn Academy, which offers individualized problem activities with instructional videos. You can find the electronic version of the book at http://moodle.anoka.k12.mn.us/District Courses → District Math → Probability and Statistics Textbook).

Coverage
This foundational course covers the Minnesota Data, Analysis, and Probability benchmarks. The course also meets Anoka-Hennepin math graduation requirements.

Goals
From the Minnesota Twins to the weather forecast, probability and statistics are used everywhere in our lives. *Anoka-Hennepin Probability and Statistics* demonstrates the connection between probability and statistics and the real world. Students, read and immerse yourself in this interactive textbook. Challenge yourself to dig deeper into the content or find solutions to your questions online. This textbook is alive and responsive to your needs. Give feedback to your teacher for potential incorporation into future revisions. Your input is valued going forward.

Thank you.
Chapter 1 – Counting Methods

1.1 Sample Spaces, Events, and Outcomes

Learning Objectives

• Determine the sample space for a given event or series of events
• Produce an organized list of outcomes within a sample space

A sample space is a list of all the possible outcomes that may occur. What might happen when you flip a coin? You will either get heads or tails. What will happen when you roll a single die? You will either get a 1, 2, 3, 4, 5, or 6. The sample space for flipping a coin is $S = \{\text{heads, tails}\}$. The sample space for rolling a die is $S = \{1,2,3,4,5,6\}$

On a coin flip, there are two outcomes, heads and tails. There are six different outcomes when considering the event of rolling a single die.

Example 1

Suppose you roll two dice. Build a 6 by 6 grid to show the different outcomes that might happen when you add the two dice together.

a) What is the sample space for the different sums that you might get?

b) What is the event for this situation?

c) Based on your grid, which outcome occurs most often?

Solution

a) The sample space is $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

b) The event is the rolling of the two dice.

c) Notice that a total of 7 can occur 6 different ways. A total of 7 is the most likely outcome.
Example 2

A child orders breakfast at a restaurant. The restaurant has two choices of drinks: milk and orange juice. The restaurant also has three choices of meat: sausage, ham, and bacon. Suppose the child orders one drink and one type of meat.

a) Give the sample space that shows all the different outcomes for what the child might order.

b) How many different outcomes are possible?

Solution

a) For the drinks, use M = Milk and O = Orange Juice. For the meat, use S = Sausage, H = Ham, and B = Bacon. The child might order MS, MH, MB, OS, OH, or OB. The sample space is

\[ S = \{ MS, MH, MB, OS, OH, OB \} \]

This list can also be generated using a simple grid as shown to the right.

<table>
<thead>
<tr>
<th></th>
<th>Milk</th>
<th>Orange Juice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sausage</td>
<td>MS</td>
<td>OS</td>
</tr>
<tr>
<td>Ham</td>
<td>MH</td>
<td>OH</td>
</tr>
<tr>
<td>Bacon</td>
<td>MB</td>
<td>OB</td>
</tr>
</tbody>
</table>

b) There are six possible outcomes. This can be found simply by counting the number of results within the sample space.

Sometimes, situations can get a bit too complex to simply make a list or build a grid. A tree diagram is a visual organizer that is very effective in handling situations with larger numbers of outcomes. We will introduce this concept here, but we will revisit tree diagrams in greater detail in section 1.2.

Example 3

A dart player is trying to hit the bulls-eye with each of three darts that he will throw. Each dart will either hit the bulls-eye or miss the bulls-eye. Use a tree diagram to give the sample space for the different outcomes that may occur.

Solution

Build the tree diagram shown to the right to track what might happen.

The sample space is

\[ S = \{ HHH, HHM, HMH, HMM, MHH, MHM, MMH, MMM \} \]
Problem Set 1.1

1) A single coin is flipped two times.
   a) Construct the sample space for this situation.
   b) How many different outcomes are possible?

2) A single coin is flipped three times.
   a) Use a tree diagram to construct the sample space for this situation.
   b) How many different outcomes are possible?

3) A single coin is flipped four times.
   a) Use an organizational strategy to construct the sample space for this situation.
   b) How many different outcomes are possible?

4) Suppose a 4-sided die is rolled one time. What is the sample space for the result of the roll?

5) Suppose two 4-sided dice are rolled and we keep track of the total for the two dice.
   a) Draw a four by four grid that demonstrates the different results for the total of the two dice.
   b) What is the sample space for the possible totals of the two dice?

6) Suppose two 4-sided dice are rolled two times and we keep track of the product when the result from the first die is multiplied by the result from the second die.
   a) Draw a four by four grid that demonstrates the different results for the product of the two dice.
   b) What is the sample space for the possible product of the two dice?
   c) How many different outcomes are possible for the product of the two dice?
   d) What outcome occurs most often?
1.2 Fundamental Counting Principle

Learning Objectives

- Apply the Fundamental Counting Principle to determine the number of outcomes
- Create tree diagrams to represent outcomes for a series of events

The **Fundamental Counting Principle** states that if you wish to find the number of outcomes for a given situation, simply multiply the number of possible outcomes for each step of the event. In Example 2 in section 1.1, the child had two different choices of drink and three different choices of meat. If we multiply 2 times 3, we get 6 which is the total number of outcomes possible. The Fundamental Counting Principle expands to any events with more than just two steps. For example, suppose it turned out that the child also wanted to order eggs and had to choose between the eggs being scrambled or sunny-side up. The Fundamental Counting Principle states that there are $2 \times 3 \times 2$ or 12 ways to order this breakfast.

**Fundamental Counting Principle**

\[
\text{Total # of Outcomes} = (\# \text{ of Choices}) \cdot (\# \text{ of Choices}) \cdot (\# \text{ of Choices}) \cdot \ldots
\]

A **tree diagram** will allow us to visually see what is happening in situations like this.

**Example 1**

Build a tree diagram that shows the different outcomes for what the child might order for breakfast if they have two choices for drinks, three choices for meats, and two choices for eggs.

**Solution**

The first set of branches of the tree diagram will represent the type of drink, the second set of branches will represent the type of meat, and the third set of branches will represent the type of egg. A diagram of what this will look like is shown on the top of the next page.
We have labeled the ends of two of the branches in the figure above to show what each branch means. For example, one of the labeled branches shows that the child might have ordered milk, bacon, and scrambled eggs.

The Fundamental Counting Principle is critically important especially when considering complex tree diagrams. Our tree diagram above has many branches and it tracks a great deal of material. It ultimately shows us the 12 different possible breakfast orders, but it takes a large amount of organization to successfully complete. Multiplying 2 by 3 by 2 is a much quicker way to find out the total number of possible outcomes.

**Example 2**

A couple is planning to have 3 children. Consider the different results that might occur in terms of sex. For example one outcome might be Boy, Boy, Girl (BBG).

a) Using the Fundamental Counting Principle, calculate the number of different outcomes for the children in this family.

b) Build a tree diagram that shows the different orders of children the couple might have.

c) Build the sample space that shows all the different orders of children the couple might have.
Solution

a) There are 2 choices for the first child, 2 for the second, and 2 for the third. Therefore, there are \(2 \times 2 \times 2 = 8\) outcomes for the gender order of the 3 children.

b)

c) In order to be organized, the list will be alphabetized. BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG. There are a total of 8 outcomes.

There are many other ways to apply the Fundamental Counting Principle. A standard deck of cards has 52 cards as shown to the right. If you are dealt just one card, there are 52 different outcomes.

<table>
<thead>
<tr>
<th>Standard Deck of 52 Playing Cards</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Black cards</strong></td>
</tr>
<tr>
<td>Clubs</td>
</tr>
<tr>
<td>A ♠</td>
</tr>
<tr>
<td>2 ♠</td>
</tr>
<tr>
<td>3 ♠</td>
</tr>
<tr>
<td>4 ♠</td>
</tr>
<tr>
<td>5 ♠</td>
</tr>
<tr>
<td>6 ♠</td>
</tr>
<tr>
<td>7 ♠</td>
</tr>
<tr>
<td>8 ♠</td>
</tr>
<tr>
<td>9 ♠</td>
</tr>
<tr>
<td>10 ♠</td>
</tr>
<tr>
<td>Jack ♠</td>
</tr>
<tr>
<td>Queen ♠</td>
</tr>
<tr>
<td>King ♠</td>
</tr>
</tbody>
</table>
Example 3
Suppose you are dealt two cards from a standard deck of 52 cards. How many different outcomes are possible for what your two cards might be?

Solution
We could certainly try drawing a tree diagram but that could get very large quite quickly. The first split alone would have 52 branches on it! On the other hand, if we use the Fundamental Counting Principle, we can simply calculate how many different ways we could be dealt 2 cards from a standard deck. There would be 52 choices for the 1st card and 51 choices for the 2nd card. (Once the first card is dealt, the deck only has 51 cards left in it.) There are \(52 \times 51 = 2652\) possible different results for what our two cards might be.

Example 4
How many different 10-digit phone numbers are possible if our only restrictions are that no phone number may begin with a zero or a one and that the 4th digit may not be a zero?

Solution
There are a total of 10 digits available \(\{0, 1, 2, \ldots, 7, 8, 9\}\). We can’t use zero or one for the first digit so there are only 8 choices for the 1st digit. In addition, we can’t use a zero for the 4th digit so the 4th digit only has 9 choices. All of the remaining digits will have 10 choices. This gives us \(8 \times 10 \times 10 \times 9 \times 10 \times 10 \times 10 \times 10 = 72 \times 10^8 = 7,200,000,000\) different possibilities for a 10-digit phone number.

Example 5
A teenager is given 5 different jobs that they must do before they may go out to a movie with friends. The jobs are washing the car, starting a load of laundry, vacuuming the family room, taking out the garbage, and putting away the dishes. In how many different orders could the teenager complete these jobs?

Solution
There are five choices the teenager could pick for the first job. Once that job is finished, there are only 4 jobs remaining. Once the 2nd job is completed, there are only 3 choices for the 3rd job. Once the 3rd job is finished, there are only 2 choices for the 4th job and finally there will only be one choice left for the 5th job. There are \(5 \times 4 \times 3 \times 2 \times 1 = 120\) different orders that these jobs could be completed. Note that there is a quick way to do this ordered multiplication using factorials. \(5 \times 4 \times 3 \times 2 \times 1 = 5!\) The 5! is read “Five Factorial”. Be sure to locate the factorial key on your calculator.

Factorial Formula
\[
n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (3) \cdot (2) \cdot (1)
\]
Problem Set 1.2

Exercises

1) A woman has three skirts, five shirts, and four hats from which to pick an outfit. How many different outfits can she wear if she picks one skirt, one shirt, and one hat?

2) How many different five-digit ZIP codes are possible if the digits can be repeated?

3) How many different five-digit ZIP codes are possible if the digits cannot be repeated?

4) In how many ways can a baseball manager arrange a batting order of nine players?

5) A store manager wishes to display six different brands of laundry soap by lining them up in a row on a shelf. In how many ways can this be done?

6) There are 8 different statistics books, 6 different geometry books, and 3 different trigonometry books being considered for next year. In how many ways can a textbook committee select one of each book?

7) At a film festival, there are eight different films that will be shown. In how many different orders can these films be shown?

8) The call letters of a radio station must have four letters. The first letter must be a K or a W. How many different call letter combinations are possible if letters may not be repeated?

9) The call letters of a radio station must have four letters. The first letter must be a K or a W. How many different call letter combinations are possible if letters may be repeated?

10) How many different four-digit ID tags can be made if repeats are allowed?

11) How many different four-digit ID tags can be made if it must start with a 7 and no repeats are allowed?

12) In how many different ways can the Harry Potter series of books (7 books total) be arranged in a row on a shelf?

13) In how many different ways can a manager select a pitcher - catcher combination if the manager has 5 pitchers and 2 catchers from which to choose?

14) A coin is tossed 8 times and the result of each flip is recorded. How many different outcomes are possible for this series of 8 flips?

15) A tile contractor is going to make a tile pattern that contains four tiles. He has six different colors of tile. How many possible different 4-color patterns are possible if no color may be repeated?
16) A tile contractor is going to make a tile pattern that contains four tiles. He has six different colors of tile. How many possible different 4-color patterns are possible if colors may be repeated?

17) Four cards are dealt from a standard deck of 52 cards. In how many different orders of suit could the cards be dealt? For example, one order is Club, Heart, Club, Diamond.

18) A pizza restaurant offers 6 different toppings for their pizzas. How many different pizzas are possible?

19) Use a tree diagram to find all possible outcomes for the result of a series of coin flips if the coin is flipped two times. Write a list of the possible results when complete.

20) The Super-Cool Ice Cream Shoppe sells sundaes, cones, and ice cream bars. The available flavors are either butterscotch or chocolate and you may choose to have it with nuts or without nuts.
   a) Draw a tree diagram to illustrate the different types of ice cream treats that you could order.
   b) How could you find the number of outcomes using the Fundamental Counting Principle?
   c) How many different outcomes are possible?

21) A quiz has four true/false questions on it. Use a tree diagram to show all the different possible answer keys.

22) A box contains a $1 bill, a $5 bill, and a $10 bill. Two bills are selected one after the other without replacing the first bill once it is drawn. Draw a tree diagram to show all the possible ways that these two bills may be picked.

23) The Eagles and Hawks play each other in a hockey tournament. The first team to win two games is the champion. Use a tree diagram to show all different possible outcomes for the tournament.

Review Exercises

24) Consider a situation in which a baseball manager must decide which one of 4 players will pitch (P1, P2, P3, or P4) and which one of 2 players will catch (C1 or C2).
   a) What is the sample space for the possible pitcher/catcher combinations for this situation?
   b) How many different outcomes are possible?

25) A person taking a poll stops two people on the street and asks each of them if they are leaning Democratic, leaning Republican, or undecided for an upcoming election.
   a) Build a grid that shows the different responses that may occur.
   b) How many different outcomes are possible?
1.3 Permutations

Learning Objectives

• Know the definition of a permutation
• Be able to calculate the number of permutations using the permutations formula and with technology
• Understand the connection between the Fundamental Counting Principle and permutations

The Fundamental Counting Principle provides us with a tool that allows us to calculate the number of outcomes possible in many situations. What if the situation is a bit more complex? For many situations, the order that we complete a task does not matter. Ordering milk, bacon, and scrambled eggs in that order is the same as ordering bacon, scrambled eggs, and milk. In that case the order that we make our choices wouldn’t matter. However, there are many situations in which the order that we do things does make a difference.

A permutation is a specific order or arrangement of a set of objects or items. What if I wish to call someone on the phone? If I make the call, the order that I punch in the numbers matters so this is an example of a permutation. A good question to ask when deciding if your arrangement is a permutation is “DOES ORDER MATTER?” If yes, then you are dealing with a permutation. For example, suppose you order an ice cream sundae and a cherry is put on the bottom, then chocolate sauce on top of that, and finally it is topped off with ice cream. You probably would not be too happy with that particular ice cream sundae. You would likely prefer that the ice cream comes first, followed by the chocolate sauce, and finally the cherry goes on top. Clearly each sundae had the same three ingredients, but they were quite different from one another. Each order that we can make the ice cream sundae is called a permutation.

There is a simple formula for figuring out how many permutations exist when ‘r’ objects are selected from a set of ‘n’ objects. The left side of the equation can be read “n P r”, just as it looks or “n Permutations of size r”.

\[ n \text{P} r = \frac{n!}{(n-r)!} \]
Recall that the exclamation point is a factorial. For example, $5! = 5 \times 4 \times 3 \times 2 \times 1$. Also, be sure to find the permutations command on your calculator.

In our ice cream sundae discussion, ‘$n$’ would be 3 because there are 3 items to select from and ‘$r$’ would also be 3 because we are going to select all three items. Using the permutations formula, this would be $3! \over (3-3)! = 6$. In other words, there are 6 different orders that the ice cream sundae could be made. (Note that 0! is equal to 1.)

**Example 1**

Suppose you are going to order an ice cream cone with two different flavored scoops. You are going to take a picture of your ice cream cone for use in the school newspaper. The ice cream shop has 5 flavors to choose from; chocolate, vanilla, orange, strawberry, and mint. How many different ice cream cone photos are possible?

**Solution**

The first question to ask is “Does Order Matter?”. If it does, then we are dealing with a permutation question. In this case, the order does make a difference. A chocolate on top of vanilla cone looks different than a vanilla on top of chocolate cone. We have five flavors to pick from, so $n = 5$. We are going to select 2 flavors so $r = 2$. $5! \over (5-2)! = 120 \over 6 = 20$. There are 20 different permutations of ice cream cones we could order. The notation representing this situation, $5! \over (5-2)!$, can be read as “Five ‘P’ Two” or “Five permutations of size Two”. Be sure to perform this calculation on your calculator as well.

In the example above, you could have also found your answer using the Fundamental Counting Principle. There were 5 choices for the 1st flavor and then only 4 choices for the 2nd flavor. There are $5 \times 4 = 20$ ice cream cones possible.
Example 2

Give the value of $\binom{6}{3}$ by using the formula for permutations. Verify your solution on your calculator.

Solution

$$\binom{6}{3} = \frac{6!}{(6-3)!3!} = \frac{6!}{3!} \cdot \frac{720}{6} = 120$$

Example 3

Decide whether each of the situations below involves permutations.

a) A five-card poker hand is dealt from a deck of cards.

b) A cashier must give 3 pennies, 2 dimes, a 5 dollar bill, and a 10 dollar bill back as change for a purchase.

c) A student is going to open a padlock that has a three number combination.

d) A child has red, blue, green, yellow, and orange color crayons and will be coloring a rainbow using each color one time.

Solution

a) The order you get your five cards for a poker hand does not matter. If one of your cards was the ace of spades, it didn’t matter if it was the first card or the last card dealt.

b) The order that the cashier gives you $15.23 in change does not matter as long as the total is $15.23.

c) The order you put in the three numbers for the combination makes a difference. If the correct combination is 12-27-19, the padlock will not open if you enter 19-12-27 even though the same three numbers are used.

d) The order that the child colors the rainbow does make a difference. The color pattern red, blue, green, orange, yellow will look different than green, blue, red, yellow, orange.
Problem Set 1.3

Exercises

1) Use the formula for Permutations, \( n^P_r = \frac{n!}{(n-r)!} \) to find the value for each expression. Confirm each result by using your calculator.
   a) \( 8^P_3 \)
   b) \( 4^P_4 \)
   c) \( 5^P_3 \)
   d) \( 5^P_0 \)

2) How many 4 letter permutations can be formed from the letters in word rhombus?

3) A board of directors composed of eight people must select board officers. In how many ways can a president, vice president, and treasurer be selected from this 8-person board?

4) How many different ID cards can be made if there are six digits on a card and no digit may be used more than once?

5) In how many ways can seven different brands of laundry soap be displayed on a shelf in a store?

6) A child has built a model car and will decorate it with two different racing-stripe stickers, one on each side of the car. In how many ways can the child decorate their car if the child has four different stickers from which to choose?

7) An inspector must select three tests to perform in a certain order on a manufactured part. He has a choice of seven tests. In how many different ways can he perform the three tests?

8) In how many different ways can 4 raffle tickets be selected from 50 tickets if each of the 4 ticket holders wins a different prize?

9) A teacher has five groups in a class. Each group was given the same 5 problems to work on for last night’s homework. The teacher has decided that each group will present one of the homework problems on the board during class. In how many ways can the teacher assign problems for the groups for presentation on the board?
10) There are five violinists in an orchestra. Three of them will be selected to play in a trio with a different part for each musician. In how many ways can the music be assigned?

11) There are five violinists in an orchestra. Four of them will be selected to play in a quartet with a different part for each musician. In how many ways can the music be assigned?

12) There are five violinists in an orchestra. All five of them will be selected to play in a quintet with a different part for each musician. In how many ways can the music be assigned?

13) There are five violinists in an orchestra. A piece of music is written so that it can be played with 3, 4, or 5 violinists. Each musician selected to play this piece will play a different part. In how many ways can a group of at least three musicians be selected? Use your answers from problems 10), 11) and 12) to help answer the question.

14) Decide whether each situation below involves permutations. Briefly explain your answers.
   a) Sophia picks three color crayons from a box of 12 crayons to make a picture for her cat, Butterscotch.
   b) A five-digit code is needed to open up an electronic lock on a car.
   c) Mica has 83 pairs of shoes and will pack five of those pairs of shoes to bring with him on vacation.
   d) There are seven steps that a student must follow when preparing cookies during their Family and Consumer Sciences course.

Review Exercises

15) Use the Fundamental Counting Principle to determine the number of different ways a person could order a meal if they are to pick one entree from four choices, one side order from three choices, and one drink from four choices.

16) A student wishes to check out three books from the library. She will check out one historical fiction book, one biography, and one book on art history. Build a tree diagram to show how many ways this can be done if there are two historical fiction books, three biographies, and two books on art history that she is considering checking out.

17) How many different outcomes are possible for the total on a roll of two dice if one die has 6 sides and one die has 4 sides?
1.4 Combinations

Learning Objectives

- Know the definition of a combination
- Be able to calculate the number of combinations using the combinations formula and with technology

We just looked at situations in which order matters. What if order does not matter? Suppose you have a younger brother or sister and your family goes out to a restaurant. All of the children who come to the restaurant get a menu that doubles as a children’s activity page. The owner of the restaurant has decided that each child will receive two different colored crayons to use for the activities on the menu. The restaurant happens to carry five colors of crayons: orange, yellow, blue, green, and red.

This is a situation in which the order that the child gets their two color crayons does not matter. If you gave a child a red crayon and then a blue crayon, it would be the same as if you gave the child a blue crayon followed by a red crayon. As with permutations, the first question to ask is “Does Order Matter?”. When the order does not matter, you are dealing with a situation that involves combinations.

Example 1

Consider the color crayon problem in the previous situation. Make a list showing all of the different color crayon combinations that might occur. Be organized so as not to repeat any combinations.

Solution

To be organized, use the letters O, Y, B, G, and R to represent the five colors (Orange, Yellow, Blue, Green, and Red). Alphabetizing the list to ensure that we don’t skip any combinations gives us BG, BO, BR, BY, GO, GR, GY, OR, OY, RY. Notice that while we have BG, we don’t have GB as that would be a repeat because it represents the same two colors. In addition, the list also does not include results like BB because that would represent the same color twice. It appears that there are 10 possible combinations.
As with the Fundamental Counting Principle, we now must ask the question “How can we find the solution quickly?” Making a list works nicely, but it could get a bit messy if the restaurant had 24 colors to choose from instead of 5 because our list would get very long. Out of curiosity, you may have tried $5P_2$. However, you will find that this gives us a result of 20 instead of 10. We will modify the permutation formula for situations involving combinations.

Shown below is the formula for finding how many combinations are possible when order **does not** matter.

$$ n \binom{r} = \frac{n!}{r!(n-r)!} $$

As with the permutation formula, the ‘n’ stands for the number of objects available and the ‘r’ stands for the number of objects that will be selected.

**Example 2**

Consider the color crayon problem once again. Use the combination formula to find out the number of different color crayon combinations that are possible.

**Solution**

In our problem, ‘n’ is equal to 5 and ‘r’ is equal to 2. Our calculation would be

$$ 5 \binom{2} = \frac{5!}{2!(5-2)!} = \frac{5!}{2! \cdot 3!} = \frac{5 \cdot 4}{2} = \frac{120}{12} = 10. $$

Be very careful that you find the result for the denominator before you divide! Find the $n \binom{r}$ command on your calculator and verify that $5 \binom{2}$ is indeed equal to 10.

**Example 3**

Suppose that there are 12 employees in an office. The boss needs to select 4 of the employees to go on a business trip to California. In how many ways can she do this?

**Solution**

We first ask whether the order that the employees are selected matters. In this case, the answer is no because either you will be going on the trip or you won’t be going. Being the fourth name on the list of people who get to go is just as good as being the first name on the list. We have 12 people to select from and we will be selecting 4 or $12 \binom{4} = \frac{12!}{4!(12-4)!} = 495$. There are 495 possible combination of groups of 4 that might be selected to go on the trip to California.
Problem Set 1.4

Exercises

1) Use the formula for combinations to find the value of each expression. Use a calculator to verify each answer.
   a) $\binom{5}{5}$
   b) $\binom{6}{4}$
   c) $\binom{3}{0}$
   d) $\binom{7}{3}$

2) In how many ways can 3 cards be selected from a standard deck of 52 cards?

3) In how many ways can three bracelets be selected from a box of ten bracelets?

4) In how many ways can a student select five questions to answer from an exam containing nine questions?

5) In how many ways can a student select five questions to answer from an exam containing nine questions if the student is required to answer the first and the last question?

6) The general manager of a fast-food restaurant chain must select 6 restaurants from 11 for a promotional program. In how many different possible ways can this selection be done?

7) There are 7 women and 5 men in a department. In how many ways can a committee of 4 people be selected?

8) For a fundraiser, a travel agency has donated 5 free vacations to Mexico as grand prizes in a raffle. Suppose that 220 people purchased raffle tickets. In how many different ways can the vacation winners be selected?
9) A high school choir has 27 female and 19 male members. Two students will be selected from the choir to represent the school in the All-State Choir.

a) In how many ways can the director select two students if she decides both students will be female?

b) In how many ways can the director select two students if she decides both students will be male?

c) In how many ways can the director select two students?

d) Using your answers from a), b) and c), determine how many ways the choir director can select two students such that one student will be a male and one student will be a female?

Review Exercises

10) In how many ways can the team captain of a kickball team arrange the kicking order for the 7 players on the team?

11) An electronic car door lock has five buttons on it and each button has a different letter - A, B, C, D, and E. Suppose the code to unlock the door is 4 letters long.

a) How many different codes are possible if a letter may be repeated?

b) How many different codes are possible if a letter may not be repeated?

12) Give the sample space for the different results that may occur if a coin is flipped twice.

13) Decide whether each situation involves permutations.

a) A teacher must pick two students from a class of 30 to put their answers on the board for problem #11 from last night’s homework.

b) In order to be allowed outside to play in the rain, a 5-year old must correctly put on socks, shoes, and boots.

c) It is 6 pm right now and Julia has 5 hours of schoolwork to complete by the time she goes to bed. Julia’s parents enforce a strict bedtime of 11 pm. She has a paper that will take two hours to write, a math assignment that will take one hour to complete, and a science experiment that will take two hours to finish.
1.5 Mixed Combinations and Permutations

Learning Objectives

- Determine whether a situation involves permutations or combinations
- Understand the mathematical implications of the words ‘and’ & ‘or’

Having covered the basics of combinations and permutations, you are ready to have a mixture of problems with slight variations. A common variation involves an understanding of some key words used in mathematics. Commonly, the word “and” indicates multiplication and the word “or” indicates addition. Consider the examples below.

Example 1

In how many ways can committee of 3 people be chosen if there are 8 men and 4 women available for selection and we require that two men and one woman be on the committee?

Solution

The order that we place the people on a committee does not matter. It makes no difference if you are the first person or the last person selected for the committee. Either you are on the committee or you are not on the committee, therefore this is a combination question. Notice that we want two men and one woman. The word ‘and’ indicates multiplication. In other words, we will look for the product of how many ways we can select two men from eight and one woman from four.

\[ _8C_2 \times _4C_1 = 28 \times 4 = 112. \]

There are 112 ways to select this committee of 3 people.

Example 2

In how many ways can a committee of 5 people be chosen if there are 7 men and 5 women available for selection and we require at least 4 women on the committee?

Solution

We first ask “Does order matter?”. In this case, the order that someone is placed on a committee does not matter. Either you are on the committee or you are not. Once again, we are dealing with a combination question. The key phrase in this example is at least. This can be interpreted to mean that we either select 4 women and 1 man or 5 women and 0 men.

Remember that the word ‘and’ indicates multiplication and the word ‘or’ indicates addition. It looks like we are going to have some addition and some multiplication in this problem.

\[ _5C_4 \times _7C_1 + _5C_5 \times _7C_0 = 5 \times 7 + 1 \times 1 = 35 + 1 = 36. \]

There are 36 ways to put this committee together.
Example 3

In a certain country, there are two political parties that will be represented on a ballot. Each party is responsible for nominating both a presidential and vice-presidential candidate for the ballot. In the first party, there are 6 candidates available and in the second party there are 5 candidates available. How many different ballots are possible?

Solution

The order that we select the candidates does make a difference. Selecting party member ‘A’ for a presidential candidate and party member ‘B’ for a vice-presidential candidate is different than selecting party member ‘B’ for a presidential candidate and party member ‘A’ for a vice-presidential candidate. Therefore, this is a permutations question. Since we will select candidates from the first party and candidates from the second party, we expect there to be multiplication in this problem as well.

\[ 6P_2 \times 5P_2 = 30 \times 20 = 600. \] There are 600 different possible ballots.
Problem Set 1.5

Exercises

1) State how you can tell the difference between a combination problem and a permutation problem.

2) Your closet contains 10 different styles of shoes. In how many ways can you pick out five different styles of shoes for the school week if you don’t care which day of the week you wear each style?

3) Your closet contains 10 different styles of shoes. In how many ways can you pick out five different styles of shoes for the school week if you do care which day of the week you wear each style?

4) You are drawing a rainbow using five different colored crayons from your box of 24 colors. In how many ways can you draw a rainbow if the first color you pick will be the top layer and so on?

5) In how many ways can you pick 5 different colors from a box of 24 colors to draw a rainbow?

6) Suppose 5 cards are dealt from a standard deck of 52 cards.
   a) How many 5-card hands are possible?
   b) In how many different orders can 5 cards be dealt from a standard deck?

7) Suppose the majority party in a foreign country must select a prime minister and secretary of state from an eligible group of 36 party members. In how many ways can this be done?

8) There are 7 women and 5 men in a department. Four people are needed for a focus group.
   a) In how many ways can a group of 4 people be selected?
   b) In how many ways can this group be formed if it must include exactly 2 men and 2 women?
   c) In how many ways can this group be formed if it must include at least 2 women?

9) In how many ways can 3 cars and 4 trucks be selected from 8 cars and 11 trucks for inspection?

10) In a train yard there are 4 tanker cars, 12 boxcars, and 7 flatcars available for a train. In how many ways can the train cars be selected so the train has 2 tanker cars, 5 boxcars, and 3 flatcars?

11) Flakes-R-Us cereal comes in two types, Sugar Sweet and Touch O’Honey. If a researcher has ten boxes of each type, how many ways can she select two boxes of each for a quality control test?

12) In how many ways can a jury of 12 people be selected from a pool of 12 men and 10 women?

13) In how many ways can a jury of 6 men and 6 women be selected from 12 men and 10 women?
14) A corporation president must select a manager and assistant manager for each of two stores. In how many ways can this be done if the first store has 9 employees and the second store has 7 employees?

15) Suppose the required sophomore class schedule consists of 2 math classes, 2 social studies classes, and 1 reading class. In how many ways can a student select classes if there are 4 different math classes, 5 different social studies classes, and 2 different reading classes from which to choose?

16) In how many ways can 6 people be assigned to 3 offices if there will be two people in each office?

Review Exercises

17) Use the formula for combinations to find the value of $\binom{7}{3}$.

18) In how many ways can the letters in the word ‘magic’ be arranged?

19) How many different sums are possible when two 4-sided dice are rolled?

20) A teacher will select three students to work problems on the board from her class of 34 students. In how many ways can this be done if the three problems to be worked are #11, #14, and #26?
1.6 Chapter 1 Review

There are three primary counting methods that are commonly used in probability: the Fundamental Counting Principle, combinations, and permutations. The Fundamental Counting Principle states that to find the number of outcomes for a given situation, simply multiply the number of ways each step of an event may occur by each other. When deciding whether to use combinations or permutations, you must ask if the order matters. If order matters, use permutations, otherwise use combinations. When working with counting outcomes, it is often helpful to have an organizational strategy. Common strategies involve making organized lists, grids, or tree diagrams. Using an organizational strategy will make it much easier for you to come up with the correct sample space.

Review Exercises

1) Suppose that two 5-sided dice are rolled.
   a) Draw a grid showing all the outcomes for the different totals that may occur.
   b) Use (brackets) to write down the sample space.
   c) Suppose a friend offers to play a game in which you are paid $4 any time a total divisible by 4 occurs. Otherwise you pay your friend $2. If you decide to play, would you expect to win money or lose money? Use your grid from part a) to help explain your answer.

2) The lunch at The Diner has a choice of ham, turkey, or roast beef on rye or white bread with coffee or milk. Draw a tree diagram that illustrates what a person might have for lunch if they pick only one meat, one bread, and one drink.

3) Find the value for each expression below. Show your work by hand and use your calculator to verify your results.
   a) $5!$
   b) $\binom{6}{3}$
   c) $\binom{7}{5}$
   d) $(5 - 2)!$
   e) $4! - 2!$

4) There are four runners in a race. In how many ways can the runners finish the race?

5) A store has eighteen outfits available for a window display, but only six outfits can fit at one time in the display. In how many different ways can 6 outfits be arranged?
6) Paul has three baseballs and four bats. How many possible ball and bat combinations can he choose if he picks one of each?

7) How many license plates are possible if each plate must have three letters followed by three digits and repeats are allowed?

8) How many license plates are possible if each plate must have three letters followed by three digits and repeats are not allowed?

9) There are twenty candidates in the Mr. Minnesota contest. In how many ways could the judges choose the winner, first-runner up, and second-runner up?

10) The yearbook editor must select two photos out of 42 junior photos and two out of 45 senior photos for a page in the yearbook. In how many ways can the photos be selected?

11) A homeless shelter has decided to purchase all new kitchen appliances. They need one oven, one refrigerator, and one dishwasher. The appliance store has 7 brands of ovens, 6 brands of refrigerators, and 5 brands of dishwashers. In how many brand arrangements can the shelter purchase their appliances?

12) An ice cream shop has 8 different flavors of ice cream available. How many 2-scoop cones can be made if you are allowed to have the same flavor for both scoops?
13) An ice cream shop has 8 different flavors of ice cream available. How many 2-scoop cones can be made if you decide not to have the same flavor for both scoops?

14) Suppose a jury of 12 is being selected from a pool of 20 candidates. In how many ways can the jury be selected?

15) Suppose a jury of 12 is being selected from a pool of 13 men and 7 women. In how many ways can the jury be selected if the judge states that the jury must contain exactly 5 women?

16) Suppose a jury of 12 is being selected from a pool of 13 men and 7 women. In how many ways can the jury be selected if the judge states that the jury must contain at least 5 women?

17) In how many ways can I put together an outfit if I have 7 shirts, 5 pairs of pants, and 4 hats from which to choose assuming I select one of each type of item?

18) For $7.99, a restaurant will sell you their lunch special. The special is a hamburger or chicken sandwich, onion rings or fries, and soda or coffee.
   a) Make a tree diagram showing the different ways a customer may order the lunch special.
   b) How many outcomes are there? Use the Fundamental Counting Principle to justify your answer.

Image References

Breakfast  http://worldaffairspittsburgh.blogspot.com
Tetrahedral Dice  http://www.bbc.co.uk
Harry Potter Books  http://www.dipity.com
Ice Cream Cones  http://www.bunrab.com
Raffle Ticket  http://canuckamusements.com
Color Crayon  http://www.rosespet.com
Exam  http://www.iphlebotomycertification.com
Rep/Dem  http://thyblackman.com
Rainbow  http://tracynicholls.webs.com/
Jury  http://adriandayton.com
License Plate  http://www.15q.net
Appliances  http://homeappliancesblog.com/
Chapter 2 – Calculating Probabilities

2.1 Calculating Basic Probabilities

Learning Objectives

- Understand how to calculate and write a probability
- Understand what constitutes chance behavior
- Understand the concept of the Law of Large Numbers

Probabilities give us an idea of how likely it is for a certain event to happen. For example, when a coin is flipped, the chance that it comes up heads is 50%. Probabilities can be expressed as decimals, fractions, percentages, or ratios. We could have said the probability of flipping heads is, 0.5, \( \frac{1}{2} \), 50%, or 1:2. Each of these conveys the idea that we should expect to get heads half the time. Probabilities give us an idea of what to expect in the long run but they do not tell us what will happen with certainty in the short term.

Suppose we flip a coin 10 times in a row and get heads each time. The next coin flip is still a random event because while we cannot tell for certain what the next flip will be, we can be certain that about 50% of all tosses over a long set of tosses will be heads. Some people think that we are on a roll so we are more likely to get another heads. Others will say that getting tails is more likely because we are due to get tails. The truth is that we cannot tell what will happen on the next flip. The only thing we know for certain is that there is a 50% chance that the coin will be heads on its next flip. If we continue to flip this same coin hundreds of times, we would expect the percent of heads to get closer and closer to 50%.

Chance Behavior is not predictable in the short term; however, it has long term predictability. The Law of Large Numbers tells us that despite the results on a small number of trials, we will eventually get close to the theoretical probability if we perform many trials of the event. The outcomes in any random event will always get close to the theoretical probability if the event is repeated a large number of times. We might roll a die 4 times in a row and get a 6 each time, however, if we rolled this die hundreds of times, the percent of time that we get a 6 will get closer and closer to the theoretical probability of \( \frac{1}{6} \).

When calculating a probability, we divide the number of ways our favorable outcome occurs (the outcome we are interested in) by the total number of ways all outcomes can occur. In other words, the probability that outcome ‘A’ occurs is found by the formula \( P(A) = \frac{\text{# of favorable outcomes}}{\text{total # of outcomes}} \).
Consider a standard deck of 52 playing cards.

Suppose we asked the question “What is the probability of being dealt a face card (a jack, queen, or king) from a standard deck of playing cards?” We would need to count how many cards are face cards and then divide that number by the total number of cards in a deck. In this situation there are 12 face cards and 52 cards overall so our probability of getting a face card is

$$\frac{12}{52} = \frac{3}{13} \approx 0.23.$$ 

In probability, there are outcomes that are sure to happen and there are outcomes that are impossible. If we are once again dealing with a standard 52 card deck, the chance of being dealt either a red card or a black card if one card is dealt is 100%. The chance of being dealt a blue card is 0% since there are no blue cards in a standard deck. All random events have probabilities between 0 and 1. In other words, if an event occurs, there is a 100% chance that one of the possible outcomes will happen. The list below summarizes these rules.

- The probability of a sure thing is 1.
- The probability of an impossible outcome is 0.
- The sum of the probabilities of all possible outcomes is 1.
- The probability for any random event must be somewhere from 0 to 1.

As shown earlier, we notate the probability of event ‘A’ happening as $P(A)$. For example, the probability of rolling a three on a six-sided die can be written $P(3) = \frac{1}{6}$. Sometimes we are interested in the probability of an event not occurring. This is called the complement of the event. We can write the probability of the complement of event ‘A’ happening as $P(A')$, $P(\text{not } A)$, or $P(A^c)$. The formula for the complement of an event is $P(\text{not } A) = 1 - P(A)$. On our die rolling question, $P(3') = 1 - P(3) = 1 - \frac{1}{6} = \frac{5}{6}$. This can be interpreted as there is a $\frac{5}{6}$ chance of the die not landing on a 3. It is important to notice that the probability of an event happening and the probability of its complement always add up to 1.
Example 1
Which of the following situations are random events?

i) A student looks through their closet to decide what shirt to wear to school.

ii) A student labels each of their 6 pairs of shoes 1 through 6 and then rolls a single die to decide which pair to wear.

iii) The state legislature decides to increase funding to schools by 3%.

iv) A professional golfer makes a hole-in-one on a 200 yard hole.

Solution

Situations i) and iii) are not random events. In both cases, there are additional factors that are influencing the decision. The day of the week or the temperature outside might influence your shirt choice and how much money the state legislature happens to have might influence funding.

Both situations ii) and iv) are random events because while we can’t predict what will happen in this particular instance, we can make long term predictions. We can predict the percent of the time the student might end up with the shoes labeled #2 and we can predict the percent of the time that the golfer will make a hole-in-one based upon previous performance.

Example 2

In the game of pool, there are a total of 15 balls. Balls numbered 1-8 are solid and balls 9-15 are striped. There are two pool balls of each color, for example, there are two yellow pool balls. One of these is solid and one of these is striped. The only exception to this is that there is only 1 black pool ball, the eight ball, and it is solid.

Suppose the pool balls were put in a bag and a single pool ball is pulled out of the bag. What is the probability that the ball:

a) is yellow?

b) is striped?

c) has a number on it that is greater than 10?

d) is not striped?

Solution

a) \( P(Y) = \frac{2}{15} \approx 0.13 \)

b) \( P(\text{Striped}) = \frac{7}{15} \approx 0.47 \)

c) \( P(>10) = \frac{5}{15} = \frac{1}{3} \approx 0.33 \)

d) \( P(\text{Striped}') = 1 - P(\text{Striped}) = 1 - \frac{7}{15} = \frac{8}{15} \approx 0.53 \)

In addition to these types of questions, we can also calculate probabilities by incorporating our counting methods from Chapter 1. Recall that the probability of an event occurring is the number of favorable outcomes divided by the number of total possible outcomes.
Example 3
A jury of 12 people is to be selected from a group of 12 men and 8 women. What is the probability that the jury has exactly 6 women on it?

Solution
The total number of outcomes possible is based upon selecting 12 members from a pool of 20. Since order will not matter, there are \( \binom{20}{12} = 125,970 \) ways to pick a jury of 12. We now want to have a total of 6 women on the jury. If we have 6 women, then we must also have 6 men to make a total of 12 jurors. Since there are 8 women available to pick from and 12 men available to pick from, we have \( \binom{8}{6} \times \binom{12}{6} = 28 \times 924 = 25,872 \) ways to select our jury so it has 6 women and 6 men. There are 25,872 ways to have exactly 6 women on the jury out of a possible total of 125,970 different juries. In other words, \( P(\text{exactly 6 women on the jury}) = \frac{25,872}{125,970} \approx 0.21 \). There is about a 21% chance that the jury will have exactly 6 women on it.

Sometimes, data is organized in a Venn diagram, as shown in Example 4 below. We will examine these in greater depth in section 2.3 but for now, it is important to understand that a Venn diagram is an organizational tool that makes it easier to interpret a situation and answer basic probability questions.

Example 4
A class of 30 students is surveyed to see whether or not they had a science class and/or a math class this trimester. There are 18 students that have a math class, 14 students who have a science class, and 4 students who have neither. It also turns out that this includes 6 students who currently have both classes. The results of the survey are shown in the Venn diagram below.

a) How many total students are taking a math class this trimester?
b) What is the probability that a randomly selected student is taking a math class this trimester?
c) What is the probability that a randomly selected student is taking both a math and science class this trimester?
d) What is the probability that a randomly selected student is not taking either a math or science class this trimester?

Solution
a) There are 12 students who only have a math class and 6 students who have both a math and science class this trimester for a total of 18 students.
b) \( P(\text{Math}) = \frac{18}{30} = \frac{3}{5} = 0.6 \)
c) \( P(\text{Math & Science}) = \frac{6}{30} = \frac{1}{5} = 0.2 \)
d) \( P(\text{No Math or Science}) = \frac{4}{30} = \frac{2}{15} \approx 0.13 \)
Problem Set 2.1

For problems 1-5, express your answer both as a fraction (simplify if possible) and as a decimal to the nearest hundredth.

1) Suppose a single card is dealt from a standard deck of 52 cards. Find the probability that the card is:
   a) a red card.
   b) a face card.
   c) an ace.
   d) a three.
   e) a club.
   f) the three of clubs.
   g) a black king.
   h) not a spade.

2) A bag contains some jelly beans. There are a total of 6 red jelly beans, 4 green jelly beans, 2 black jelly beans, 5 yellow jelly beans, and 3 orange jelly beans in the bag. Suppose one jelly bean is drawn from the bag.
   a) Find P(purple).
   b) Find P(yellow).
   c) Find P(red°).

3) A single 6-sided die is rolled one time. Find the probability that the result is:
   a) a three
   b) a seven
   c) an even number
   d) a prime number
   e) a number greater than or equal to 5.

4) The game Scattegories® uses a 20-sided die. It has all the letters of the alphabet on it except Q, U, V, X, Y, and Z. Find each probability below if the die is rolled one time.
   a) P(Vowel)
   b) P(Vowel°)
   c) P(Q)
   d) P(Q°)
   e) P(a letter alphabetically after Q)
5) The month of October in a 2016 calendar had 31 days with October 1st being a Saturday as shown in the calendar to the right. Suppose a day is randomly selected. Find each probability.
   a) \( P(\text{weekend}) \)
   b) \( P(\text{not a weekend}) \)
   c) \( P(\text{October 31st}) \)
   d) \( P(\text{October 32nd}) \)
   e) \( P(\text{not October 31st}) \)
   f) \( P(\text{an odd-numbered day}) \)

6) A roulette wheel contains 38 slots. When the wheel is spun, a ball is dropped onto the wheel and the ball will stop on one of the slots. There are 18 black slots, 18 red slots, and 2 green slots. Suppose the ball on a roulette wheel has landed on red four times in a row. What is the chance that the ball will drop on red on the next spin?

7) A coin has been flipped 10 times. Suppose that it has come up heads on only 2 out of those ten flips.
   a) What percent of the time has the coin come up heads so far?
   b) Suppose we flip the coin 90 more times and 45 of those 90 flips come up heads. Of the 100 flips completed so far, what percent of the time has the coin come up heads?
   c) Suppose we continue to flip the coin an additional 900 times and that 450 of those 900 flips come up heads. Of the 1000 flips completed, what percent of the time has the coin come up heads?
   d) As we flipped the coin more and more, the percentage of heads got closer and closer to 50% despite the fact that only 2 of the first 10 flips were heads. What rule does this illustrate?

8) Two 6-sided dice are rolled and we keep track of the total on the two dice.
   a) Make a 6 by 6 grid showing the different totals that you can get when rolling the two dice.
   b) What is the probability that you get doubles?
   c) What is the probability that you get a total of 7?
   d) What is the probability that you get a total of at least 8?
9) The high school concert choir has 7 males and 15 females. The teacher needs to pick three soloists for the next concert but all of the members are so good she decides to randomly select the three students for the solos. The first student selected will sing the first solo and so on.
   a) In how many ways can the teacher select the 3 soloists?
   b) What is the probability that all three students selected will be females?
   c) What is the probability that at least one male will be selected?

10) A test begins with 5 multiple choice questions with four options on each question. It then has 5 true/false questions.
   a) How many answer keys are possible?
   b) What is the probability of getting every question correct if a student guesses on each question? Leave your answer as a fraction.

11) A lawn and garden store is moving locations and needs to move its riding lawn mowers to the new store. They have 8 mowers with 36-inch decks, 15 mowers with 42-inch decks, and 6 mowers with 48-inch decks that need to be moved. The trailer they are using can move a total of 8 mowers on each load so several trips will have to be made.
   a) In how many ways can 8 mowers be randomly selected for the first load?
   b) What is the probability that all the mowers with 48-inch decks get selected for the first load? Leave your answer as a reduced fraction.
   c) What is the probability that the first load has exactly two 36-inch deck mowers, four 42-inch deck mowers, and two 48-inch deck mowers?

Review Exercises

12) In how many ways can three students be selected for a committee if there are 11 students from which to select?

13) A hockey player needs new skates, a new helmet, and a new stick. Hockey Central has 5 brands of skates, 6 brands of helmets, and 8 brands of sticks for sale. In how many different ways can the player select one of each item?

14) Two standard 6-sided dice are rolled and the results of the roll are added together. Build a grid to determine which outcome is most likely to occur.

15) On a TV game show, three contestants must each pick a box which they believe contains the day’s grand prize. In how many different ways can this be done if there are 10 boxes from which to choose, each box contains a different prize, and each contestant must pick a different box?
2.2 Compound and Independent Events

Learning Objectives

- Understand how to perform the calculations for compound events
- Compute probabilities for situations with and without replacement
- Understand when two events are independent
- Understand how to compute the probability when two independent events occur

From section 2.1, you found that it is quite straightforward to calculate probabilities for simple situations. What happens when we calculate probabilities from multiple events? For example, suppose you roll a single die and then flip a coin. What are the chances that the die comes up with a 5 and the coin gives you a heads? A situation that asks you to calculate probabilities for a situation that involves two or more events or steps is called a compound event. We will try to find out how to handle these types of situations by examining several situations and then making a conclusion.

Example 1

Suppose a single die is rolled and a coin is flipped. What is the probability that the die comes up with a 5 and the coin comes up heads? Use a list to help you find out.

Solution

Start with a list of all the possible outcomes. 1H, 1T, 2H, 2T, 3H, 3T, 4H, 4T, 5H, 5T, 6H, 6T. There are 12 equally-likely outcomes. Of these, only 5H is a five with a heads. Therefore, the answer is \( \frac{1}{12} \).

What you might have noticed is that \( P(\text{five}) = \frac{1}{6} \) and \( P(\text{heads}) = \frac{1}{2} \). Curiously, \( \frac{1}{6} \times \frac{1}{2} = \frac{1}{12} \). Is this just a coincidence or is there something more here? You might recognize that flipping a coin does not impact what you roll on a die. When two events do not have an impact on each other, the events are called independent. Consider the two situations below.

Situation 1: Does a coin have a memory? As far as we can tell, the answer is no. This suggests then that a coin does not pay attention to whether it came up heads or tails on a previous flip. It does not ‘try’ to make sure that the same number of heads come up as the number of tails. Even if it comes up heads many times in a row, the next flip of that coin is not influenced whatsoever by the previous flips. Successive coin flips are independent of one another.


1 video link
**Situation 2**: Suppose your teacher picks students to do problems on the board. After each student does their problem, the teacher gives the student a piece of candy. Because your teacher wants make sure that every student gets a chance to do a problem and get a piece of candy, she keeps track of who has worked problems on the board. The selection of the next student is not independent of previous selections the teacher has made because she intentionally does not pick any student twice.

**Example 2**

Decide which pairs of events below are independent.

i) Two cards are dealt, one after the other, from a standard deck of 52 cards.
ii) A spinner with three colors is spun twice.
iii) A single die is rolled and a coin is flipped.
iv) You play on the school baseball team and you win a carnival game by throwing a baseball to try to break a plate.

**Solution**

Situations ii) and iii) represent pairs of independent events. The result from the first spin of the spinner does not impact the result of the second spin of the spinner. The result of the roll of the die does not impact what happens when the coin is flipped.

Situations i) and iv) are not independent. Once the first card from the deck is dealt, the probabilities for what the second card might be will change. For example, if the first card was the ace of spades, it is impossible for the second card to also be the ace of spades. Being a baseball player makes it more likely that you are accurate and can throw a ball harder than a typical person and, as a result, you would be more likely to break a plate.

Let’s do another example involving calculations and investigate if multiplying probabilities in a situation involving independent events gives us the correct result.

**Example 3**

A coin is flipped three times in a row. What is the probability that all three flips result in heads? Find your answer by either using a tree diagram or by making a list.

**Solution**

To be organized, we can make an alphabetically list. HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. There are 8 outcomes altogether and only one of those is HHH. \( P(HHH) = \frac{1}{8} \).

The probability of heads on one flip is \( \frac{1}{2} \). As in example 1, we have \( P(HHH) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \).

Note that the three coin flips are independent of each other.
In both Example 1 and Example 3, we could multiply the probabilities of individual events to get the probability of multiple events happening. This is always true of independent events. In other words, suppose we want the probability of both some outcome ‘A’ from one event and some outcome ‘B’ from a second event that is independent of the first event. If the probability of our first outcome is \( P(A) \) and the probability of our second outcome is \( P(B) \), then the probability of both A and B happening is \( P(A \text{ and } B) = P(A) \times P(B) \).

For Independent Events

\[ P(A \& B) = P(A) \cdot P(B) \]

**Example 4**

You are dealt one card from each of two separate decks of cards.

a) What is the probability that both cards are the king of clubs?

b) What is the probability that the two cards are identical?

**Solution**

In both situations, the events are independent.

a) We have \( P(\text{K♣} \& \text{K♣}) = \frac{1}{52} \times \frac{1}{52} = \frac{1}{2704} \approx 0.00037 \).

b) There are two good ways to solve this problem. We could imagine that we do what we did in part a) for all 52 cards in the deck. We simply could multiply our answer for part a) by 52 to get \( \frac{1}{52} \approx 0.019 \). Another way to think about this would be to ask about each card separately. What is the chance that the first card could be useful in making a match? (100%) What is the chance that the second card will be useful in making a match? If we already have one card picked, the chance that the card from the second deck will match the first card is \( \frac{1}{52} \).

\[ 1 \times \frac{1}{52} = \frac{1}{52} \approx 0.019. \]

Multiplication of probabilities expands to more than just two independent events. It also works with three or more independent events and it even works with many situations that do not have independent events. In general, when finding the probability of compound events, multiply the probabilities of each individual event. If we are interested in the probability of events A, B, and C happening, we can multiply \( P(A) \times P(B) \times P(C) \).

This same principle also works with compound events in which we distinguish whether or not we have **replacement**. Suppose we are asked to pick two cards out of a deck. If we are asked to do this without replacement, we will select the first card and record what it is. When we select our second card, we must remember that the deck has changed. Nonetheless, we can find the probability of drawing these two particular cards by multiplying the individual probabilities.
Example 5

Suppose a standard set of pool balls are in a bag. You pull two pool balls out of the bag, one after the other, without replacement. What is the probability that both pool balls are striped?

Solution

c) There are 7 striped pool balls out of the 15 pool balls. The chance that the first pool ball is striped is \( \frac{7}{15} \). Since we are not going to replace the first pool ball, what is in the bag has now changed. There are only 14 pool balls left of which 6 are striped since the first one removed from the bag was also striped. The chance that the second pool ball is striped is \( \frac{6}{14} \). To find the probability that both pool balls are striped, we multiply the individual probabilities. This gives

\[
\frac{7}{15} \times \frac{6}{14} = \frac{42}{210} = \frac{1}{5} = 0.2.
\]

There is a 20% chance that both balls will be striped if we do not use replacement.

Example 6

Suppose you have a set of pool balls in a bag. You pull two pool balls out of the bag, one after the other, with replacement. (This means that after you record what the first ball is, you put it back into the bag and remix the pool balls before you select the second pool ball.) What is the probability that both pool balls are striped?

Solution

There are 7 striped pool balls out of the 15 pool balls. The chance that the first pool ball is striped is \( \frac{7}{15} \). Since we are going to replace the first pool ball, what is in the bag has not changed. There are still 15 pool balls of which 7 still are striped. Therefore, the chance that the second pool ball is also striped is \( \frac{7}{15} \). To find the probability that both pool balls are striped, we multiply the individual probabilities to get

\[
\frac{7}{15} \times \frac{7}{15} = \frac{49}{225} \approx 0.22.
\]

There is approximately a 22% chance that both balls will be striped if we do use replacement.

We also run into situations where we are dealing with compound events involving very large populations. In these sorts of situations, we must be careful about how we interpret the mathematics.
Example 7

Approximately 20% of all Americans smoke. Suppose two Americans are selected at random. What is the probability that both Americans are smokers?

Solution

The chance that the first person is a smoker is 20%. Some students think that the chance that the second person is a smoker changes after the first person is selected, however, it does not. The population of America is so large that selecting a single person out from that population will not impact the overall percentage of Americans that smoke. The probability that the second person smokes is also 20%.

\[ P(2 \text{ smokers selected}) = 0.2 \times 0.2 = 0.04 = 4\% . \]

Example 8

Approximately 20% of all Americans smoke. Suppose five Americans are selected at random. What is the probability that all five are non-smokers?

Solution

Since 20% of Americans are smokers, 80% must be non-smokers. This gives us

\[ 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.8 = (0.8)^5 = 0.33. \]

The chance that all five Americans selected will be non-smokers is about 33%.
Problem Set 2.2

Exercises

1) What does it mean for two events to be independent?

2) Suppose you are dealt one card each from two separate decks of cards. What is the probability that both of your cards are:
   a) red?
   b) spades?
   c) jacks?
   d) face cards?

3) For each situation below, determine whether the two events are independent.
   a) Flip a coin and then draw a card from a standard deck of 52 cards.
   b) Draw a marble from a bag; do not replace it; and then draw a 2nd marble from the same bag.
   c) Get a raise at work and purchase a new car.
   d) Drive on ice and lose control of your car.
   e) Have a large shoe size and have a high IQ.
   f) Be a chain smoker and get lung cancer.
   g) Dad is left handed and son is left handed.

4) A spinner with three equal spaces of red, blue, and green is spun one time. A single six-sided die is rolled once. What is the probability that you get blue and a number greater than 3?

5) Suppose you are dealt two cards, one after another from a standard deck of cards. What is the probability that both of your cards are:
   a) spades?
   b) the same suit?
   c) kings?

6) Three cards are drawn from a standard deck without replacement. Find the probability that:
   a) all are jacks.
   b) all are clubs.
   c) all are red cards.
7) In a carnival game, players are given three darts and throw them at a set of balloons on a wall. Suppose there are eight balloons on the wall. Five of the eight balloons have slips of paper in them that say ‘Winner’ while three of the eight balloons have slips of paper that are blank. Suppose you pop a balloon with each of your three darts. If all three balloons have ‘Winner’ slips, you win the grand prize. If all three balloons have blank slips, you win the consolation prize. What is the probability that:
   a) you win the grand prize?
   b) you win the consolation prize?

8) A classroom contains 12 males and 18 females. Two different students will be randomly selected to give speeches. What is the probability that the two students who give speeches are:
   a) two females?
   b) two males?
   c) 1 male and 1 female (in either order)? (Hint: Use your answers from a) and b) along with some subtraction to calculate your solution.)

9) If 18% of all Americans are underweight, find the probability that two randomly selected Americans will both be underweight.

10) A survey found that 68% of book buyers are 40 years old or older. If two book buyers are selected at random, what is the probability that both are 40 years old or older?

11) The Gallup Poll reported that 82% of Americans used a seat belt the last time they got into a car. If four people are selected at random, find the probability that they all used a seat belt the last time they got into a car.

12) Eighty-three percent of diners favor the practice of tipping to reward good service. If three restaurant customers are selected at random, what is the probability that all three are in favor of tipping?

13) Suppose that 25% of U.S. federal prisoners are not U.S. citizens.
   a) Find the probability that a randomly selected federal prisoner is a U.S. citizen.
   b) Find the probability that three randomly selected prisoners are all U.S. citizens.
14) At a local university, 70% of all incoming freshmen have computers. If three students are selected at random, what is the probability that:
   a) none have computers?
   b) all three have computers?

15) The U.S. Department of Justice states that 6% of all murders occur without weapons. If three murder cases are selected at random, what is the probability that all three occurred with the use of a weapon?

**Review Exercises**

16) Which of the following are random events?
   a) You need to pick 2 people to be your partners in a group project so you select two of your friends.
   b) You make a rock skip across the surface of a lake 12 times.
   c) A baby elephant is born and it is a male.
   d) You spin the big wheel on the TV game show “The Price is Right” and you win $1000.

17) In how many ways can two 12th-graders be selected for speaking at graduation if there are 16 seniors that apply? One speaker will give a short introductory speech and one will give a longer speech that reflects upon the experiences of this particular senior class.

18) A family of 4 has just won the lottery and goes to an auto dealership to purchase a new vehicle for each member of the family. The parents each decide that they want a car while their two teenagers decide they would each like a truck. The family agrees that no one will purchase the same model of vehicle as anyone else. In how many ways can their purchase their 4 vehicles if the dealership has 17 car models and 23 truck models available?

19) The Strikers and the Kicks soccer teams are playing a best of five playoff series. The first team to win three games is the winner. Draw a tree diagram to show the different ways the series might play out.
2.3 Mutually Exclusive Events

Learning Objectives

- Understand when two events are mutually exclusive
- Understand the concepts of unions and intersections
- Be able to compute probabilities using Venn diagrams and formulas

Sometimes there are situations in which two different events cannot occur at the same time. For example, if you roll a single die one time and you wish to find the probability of getting an even number and a 3 on that one roll. These two events cannot occur at the same time, so we say the events are mutually exclusive or disjoint. If events ‘A’ and ‘B’ are mutually exclusive then it is impossible for events ‘A’ and ‘B’ to happen at the same time, or \( P(A \text{ and } B) = 0 \). However, if ‘A’ and ‘B’ are mutually exclusive then \( P(A \text{ or } B) = P(A) + P(B) \).

Using proper notation we have \( P(A \cup B) = P(A) + P(B) \). Remember, this is only true if the two events are mutually exclusive.

For Mutually Exclusive Events

\[
P(A \cup B) = P(A) + P(B)
\]

\[
P(A \text{ or } B) = P(A) + P(B)
\]

The \( \cup \) can be read as the union of events ‘A’ and ‘B’. The probability for the union of two events ‘A’ and ‘B’ can be thought of as the chance that either ‘A’ occurs, ‘B’ occurs, or both ‘A’ and ‘B’ occur.

Example 1

A single die is rolled one time. What is the probability of getting either an odd number or a 6?

Solution

The events of getting an odd number and getting a 6 are mutually exclusive since they cannot occur at the same time. \( P(\text{Odd U 6}) = P(\text{Odd}) + P(6) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \approx 0.67 \)

Of course, not every situation involves mutually exclusive events.
The diagrams in Figure 2.1 below are **Venn diagrams**. They are useful in showing us how different events are related. Events ‘A’ and ‘B’ are not mutually exclusive if they overlap and we can say that there is an **intersection** where events ‘A’ and ‘B’ overlap.

For example, if we roll a single 6-sided die, the events of getting an odd number and getting a number bigger than 3 intersect. They both include the number 5. The symbol for an intersection is \( \cap \). A logical extension of the formula for the union of two events that are not mutually exclusive is

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]

The next two examples illustrate this formula.

### Example 2

A single 6-sided die is rolled. Suppose the events that we are interested in are getting an odd number and getting a prime number (2, 3, or 5). Draw a Venn diagram for this situation.

#### Solution

Notice that since the numbers 3 & 5 belong to both the odd numbers and the prime numbers, they are placed into the intersection of the ‘Odds’ circle and the ‘Primes’ circle. Notice also that the numbers 4 & 6 do not belong to either set and are placed outside both circles.

### Example 3

A single 6-sided die is rolled. What is the probability of getting either an odd number or a prime number? Note that this is the same as asking for \( P(\text{Odd} \cup \text{Prime}) \).

#### Solution

Using the figure from Example 2 we see that there are four values out of six that are either odd or prime. Therefore, \( P(\text{Odd or Prime}) = \frac{4}{6} + \frac{2}{3} = \frac{2}{3} \approx 0.67 \). If we use the formula,

\[
P(\text{Odd or Prime}) = P(\text{Odd}) + P(\text{Prime}) - P(\text{Odd} \cap \text{Prime})
\]

This gives \( \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{2}{3} = 0.67 \).
Example 4

Suppose there is a 60% chance it will rain today and that there is a 70% chance that it will be over 90°F. Suppose also that there is a 45% chance that it will both rain and be above 90 degrees. What is the chance that it will neither rain nor be above 90 degrees? Solve using both a Venn diagram and using a formula.

Solution

Start by drawing a Venn diagram and noting that we have two circles, one for rain and one temperature. These two events are not mutually exclusive because they can occur at the same time therefore the circles should overlap. We can quickly fill in the intersection of the two events as 45%.

![Venn Diagram](https://bit.ly/probstatsSection2-3b)

Probability with Venn Diagrams

If we remember that there is a 60% chance of rain and we already have 45% filled in for the rain circle, the remainder of the rain circle must be 15% so it adds up to 60%. Likewise, the remainder of the >90° circle must be 25%.

We can now see that we have a total of 15%+45%+25%=85%. This means that the ‘Neither’ category must be 15% to give us a total of 100%.

Using the formula, \( P(Rain \cup 90) = P(Rain) + P(90) - P(Rain \cap 90) \) gives

\[
P(Rain \cup 90) = 60\% + 70\% - 45\% = 85\%.
\]

\[
100\% - 85\% = 15\%.
\]
Example 5

Which pairs of events are mutually exclusive?

a) You go to the pet store to buy a pet. Event A = You buy a pet that flies, Event B = You buy a pet that has no legs.

b) You order a pizza. Event A = Your pizza has pepperoni on it, Event B = Your pizza has mushrooms.

c) You select a football player to take a picture of for the yearbook. Event A = The player is a 4-year varsity starter, Event B = The player is 14 years old.

d) Radio stations have 4-letter station names such as KDWB. You decide to pick a radio station to listen to. Event A = The station’s 4-letter name starts with a W, Event B = The station’s 4-letter name contains three E’s.

Solution

a) Is mutually exclusive. You cannot buy a pet that both flies and also has no legs.

b) Is not mutually exclusive. You can order a pizza with both mushrooms and pepperoni.

c) Is mutually exclusive. If a 4-year varsity starter is 14 years old, then they would have been a varsity starter when they were 10 years old. That simply does not happen.

d) Is not mutually exclusive. This could happen if the radio station’s call sign is WEEE.

Example 6

The Rockin’ Rollers performance company has 30 musicians who play either bass guitar, lead guitar, or rhythm guitar. Some of these musicians play more than one instrument. Suppose 4 musicians can play lead, rhythm, or bass guitar. Fourteen can play lead or rhythm but not bass, 2 can play bass or rhythm but not lead, 3 can play bass only, and 4 can play rhythm only. There are no musicians who play lead and bass only. Draw a Venn diagram to determine how many musicians play lead only.

Solution

From the information given, we can fill in a Venn diagram as shown to the right.

We’ve used up 27 of the 30 musicians so there must be 3 musicians who play lead guitar only.
Problem Set 2.3

Exercises

1) What does it mean for two events to be mutually exclusive?

2) Give an example of two events from a situation in which the two events are mutually exclusive.

3) Give an example of a situation with two events that are not mutually exclusive.

4) Consider each situation. Decide whether the pairs of events are mutually exclusive.
   a) Roll a die: Get an even number and get a number less than 3.
   b) Roll a die: Get a prime number (2, 3, or 5) and get a six.
   c) Roll a die: Get a number greater than 3 and get a number less than 3.
   d) Select a student: Get a student with blue eyes and get a student with blond hair.
   e) Select a college student: Get a sophomore and get a student that is a math major.
   f) Select a course: Get an Algebra course and get an English course.
   g) Select a voter: Get a Republican and get a Democrat.

5) There are 200 male students at a particular school. Of these, 58 play football, 40 play basketball, and 8 play both.
   a) Draw and label a Venn diagram for this situation.
   b) How many play both sports?
   c) How many play basketball but not football?
   d) How many play football but not basketball?
   e) How many do not play football or basketball?

6) An architectural firm is putting out bids to design two large governmental buildings. Suppose they believe they have 35% chance of getting the contract for the first building, an 80% chance of getting the contract for the second building and a 10% chance of getting neither job.
   a) Draw a Venn diagram for this situation and use your diagram to find the chance that they get both contracts.
   b) Use a formula for this situation to find the chance that they get both contracts.
A single card from a standard deck can have many descriptions. For example, the King of Spades could be described as a black card, a face card, a king, or a spade. Suppose we pull a single card out of a deck and we pay attention to the events of getting a red card, getting a jack, and getting a spade.

a) Draw a Venn diagram to illustrate this situation paying attention to whether or not the card is red, a jack, or a spade.

b) Shade the portion of your diagram with vertical lines that represents the intersection of getting a red card and getting a jack.

c) Shade the portion of the diagram with horizontal lines that represents the union of getting a jack or getting a spade.

A student tells their teacher that they want to build a cabinet in wood shop. Students sometimes build this project with oak only, sometimes with cherry only, sometimes with both and sometimes with neither. There is a 40% chance the project will be built using oak, a 50% chance the project will be built using cherry, and a 30% chance that the project will be built using both types of wood. What is the chance that the student will not use either oak or cherry?

Consider a set of 15 pool balls. Balls numbered 1 through 8 are solid and balls 9 through 15 are striped. Suppose the balls are placed into a bag and you randomly select one ball. Find the probability that:

a) you selected either a solid ball or a ball numbered greater than 12.

b) you selected an even numbered ball or a solid ball.

c) you selected a solid ball or a striped ball.

d) you selected a ball that was striped and even.

Suppose you again have a standard set of 15 pool balls. This time, you pull two pool balls out of the bag, replacing the first ball before you select the second ball. What is the probability that:

a) your two pool balls are both solid?

b) you pick exactly the same ball twice?

c) your first ball was solid and your second was odd?

At a particular school, there are 20 teachers. Three of them teach math, 5 teach science, and 3 teach computer science. It turns out that among these teachers, there is one teacher who teaches all three classes and one teacher who teaches both science and computer science. Draw a Venn diagram to illustrate the situation and determine how many of the 20 teachers teach courses other than math, science, or computer science. Note that you will need 3 circles to build this diagram.
Review Exercises

12) Two cards are selected from a standard deck of 52 cards, one after the other without replacement. What is the probability that the two cards are both face cards?

13) Suppose 90% of all Americans have attended a religious ceremony at least one time in the past year. What is the probability that 4 randomly selected Americans will all have attended at least one religious ceremony in the past year?

14) A single 6-sided die is rolled once and a single card is drawn from a standard deck of 52 cards. What is the probability that the die shows a result greater than 3 and the card is a heart?

15) A young girl has a box of 8 color crayons but has decided she needs only 3 colors to make a picture for her grandfather. In how many ways can the girl select the three crayons?

16) In how many ways can a committee of 4 people be selected if there must be at least 1 man and 1 woman on the committee and there are 6 men and 7 women from which to pick?
2.4 Tree Diagrams and Probability Models

Learning Objectives

- Understand how to build and properly notate a tree diagram
- Understand how to calculate probabilities using a tree diagram
- Understand how to verify if a tree diagram is correct
- Be able to build a probability model by using a tree diagram

As we advance through probability, it becomes very apparent that we need to be quite organized with our problems as they become more complex. In this section we will use tree diagrams to help us calculate probabilities for given situations. Tree diagrams are a visual aid that can help us break down a situation and calculate probabilities. There are two key principles that we must observe for all tree diagrams. First of all, to find the total probability for any given branch on a tree, multiply the individual probabilities along that branch. Secondly, the sum of the probabilities from the ends of each branch must total to 1. We will examine several examples of probabilities using tree diagrams in order to solidify our understanding of this concept.
Example 1

At a restaurant, there are two breakfast platters that are served, one featuring pancakes and one featuring eggs. There are also two choices for drinks, milk or juice. Thirty percent of customers choose the pancake platter while 70 percent choose the egg platter. Forty percent of customers choose milk while 60 percent choose juice. Assume the drink choice is independent of platter choice. Build a probability model for this situation by using a tree diagram.

Solution

Step one is to build the tree diagram as shown to the right. Be sure to label each branch with what it represents and the associated probability.

Step two is to calculate the probabilities at the end of each branch. To do this, we multiply the probabilities along each branch. For example, the top branch’s value of 0.28 was found by multiplying 0.7 by 0.4.

Table 2.1, shown below, summarizes the data in the tree diagram. There are two critical ideas to pay attention to here. First of all, the probability at the end of each branch is the product of the probabilities on that branch. Secondly, notice that the sum of the four probabilities at the ends of the branches add up to 1. Table 2.1 summarizes these results from our tree diagram and is called a probability model. Notice that the probabilities in the table also sum to 1.

<table>
<thead>
<tr>
<th>Order:</th>
<th>Eggs &amp; Milk</th>
<th>Eggs &amp; Juice</th>
<th>Pancakes &amp; Milk</th>
<th>Pancakes &amp; Juice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability:</td>
<td>0.28</td>
<td>0.42</td>
<td>0.12</td>
<td>0.18</td>
</tr>
</tbody>
</table>
Example 2

The Diamonds and the Dusters baseball teams are playing a best-of-three playoff series. The first team to win two games is the winner of the series. Suppose the Diamonds have a 60% chance to win any game they play against the Dusters. Build a tree diagram and a probability model to determine the probability of each team winning the series.

Solution

Notice in the tree diagram below that not all branches go three games. There are two sets of branches that only go two games. A third game does not need to be played in all situations.

Table 2.2 below gives the probability model based upon our tree diagram. The probability of the Diamonds winning the series is $0.36 + 0.144 + 0.144 = 0.648$ while the probability of the Dusters winning the series is $0.096 + 0.096 + 0.16 = 0.352$. In our solution, notice that each branch is labeled and includes a probability. Multiply the probabilities along each branch to get the probability at the end of the branch. Once again, note that the total of all the probabilities in the probability model sums to 1.

Table 2.2:

<table>
<thead>
<tr>
<th>Outcome:</th>
<th>Diamonds Win in 2 games</th>
<th>Dusters Win in 2 games</th>
<th>Diamonds win in 3 games</th>
<th>Dusters win in 3 games</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability:</td>
<td>0.36</td>
<td>0.16</td>
<td>0.288</td>
<td>0.192</td>
</tr>
</tbody>
</table>
Example 3
You are dealt two cards from a standard deck of 52 cards. What is the probability that the two cards can be classified as a red card and a face card (in either order)?

Solution

Begin by considering the first card. We are concerned primarily with getting a face card and then a red card or a red card and then a face card. In order to reach our goal, the first card could be a red card, a face card, or both a red and a face card. Any other card would result in us not reaching our goal. As a result our tree diagram will have four initial branches as shown below; red (not face), face (not red), red & face, or other.

Once our first set of branches are complete, we look at the second stage. We will examine the red, not face branch in detail (top branch). There are 20 red cards out of our 52 card deck that are not face cards. Once we have that card, we now want to know the chances of getting a face card (assuming we just drew a red card). There are still 12 face cards in the deck but there are now only 51 cards remaining in the deck. We multiply this branch out to get \( \frac{20}{52} \times \frac{12}{51} = \frac{240}{2652} \).

There are two other branches we work in a similar fashion as they are the only other branches that help us achieve our goal. Adding these gives us \( \frac{240}{2652} + \frac{156}{2652} + \frac{186}{2652} = \frac{582}{2652} \approx 0.22 \). There is about a 22% chance of getting a red card and a face card.

The key to success when working with tree diagrams and probability models is to work in a neat and organized fashion. Many errors are often due to sloppy work. In addition, another useful suggestion is to make your tree diagrams large enough so that you have plenty of room to work.
Problem Set 2.4

Exercises

1) Fourteen red marbles and sixteen green marbles are in a bag. Two marbles are picked out one at a time and replaced after they are picked. Build a tree diagram and probability model to show the different combinations of marbles that could be pulled out of the bag and their associated probabilities.

2) A bag contains a standard set of pool balls. Two balls are pulled out, one after another, and not replaced. What is the probability that the two balls are a solid and a striped ball in either order? (Recall that there are 8 solid pool balls and 7 striped pool balls.)

3) A bag contains a $100 bill and two $20 bills. A second bag contains 1 gold marble and 2 silver marbles. You get to pick one bill out of the first bag. After this, you pick a marble out of the second bag. If you get the gold marble, you get to triple the amount of money you pulled from the first bag. If you get a silver marble, you get to double the amount of money you picked from the first bag. Build a probability model for all the different amounts of money that you might win.

4) A basketball player is practicing shooting free throws. Suppose she makes 75% of her free throw attempts. Make a tree diagram and probability model for what might happen if she decides to shoot three free throws. In other words, what is the probability that she makes zero shots, one shot, two shots, or all three shots?

5) A coin is flipped and then two dice are rolled. Build a probability model that shows how likely it is to get heads followed by doubles, heads and a non-doubles, tails and doubles, and tails and a non-doubles.

6) A spinner with four evenly-spaced wedges of red, blue, green, and orange is spun and a coin is flipped.
   a) How many different outcomes are possible?
   b) Build a probability model that shows the probabilities for each outcome.

7) A baseball player is a .400 hitter. This means that he gets a hit (single, double, triple, or home run) 40% of the time he has an at-bat. Use a tree diagram to build a probability model that shows the probability of the player having 0, 1, 2, or 3 hits if he has 3 at-bats in one game.
8) In some sports, the home team wins a higher percentage of games played. Suppose the Dunkers and the Hoopsters are playing a best-of-three game series against each other. When the Dunkers are home, they have a 60% chance of winning a game against the Hoopsters. When the Hoopsters are home, they have a 55% chance of winning a game against the Dunkers. The Dunkers will be the home team in games 1 and 3 while the Hoopsters will be the home team in game 2. Use a tree diagram to build a probability model for this situation. The model should show the chances that the Dunkers win in 2 games or in 3 games and the chances that the Hoopsters win in 2 games or in 3 games.

9) A patient is scheduled to have two surgeries. The results of each surgery are independent of each other. Suppose the first surgery has a 90% success rate and the second surgery has an 85% success rate. Build a probability model by using a tree diagram that shows all the different results that might occur.

10) A bag contains ten red cubes numbered 1 through 10 and five blue cubes numbered 1 through 5. You pull two cubes out of the bag without replacement. What is the probability that the two cubes will be an odd cube and a red cube (in either order)?

Review Exercises

11) Suppose that events ‘A’ and ‘B’ are mutually exclusive and that \(P(A) = 0.35\) and \(P(B) = 0.14\).
   
   a) What is \(P(A \cup B)\)?
   
   b) What is \(P(A \cap B)\)?

12) How many unique three-letter ‘words’ can be formed by selecting three letters from the alphabet if no letter may be repeated?

13) How many unique three-letter ‘words’ can be formed by selecting three letters from the alphabet if letters may be repeated?

14) 20% of all households in the Twin Cities get the Star Tribune newspaper delivered to their home while only 15% get the Pioneer Press delivered to their home. If 70% of homes do not get either newspaper delivered, what percent of homes get both newspapers delivered?
2.5 Conditional Probabilities and 2-Way Tables

Learning Objectives
- Understand how to calculate conditional probabilities
- Understand how to calculate probabilities using a contingency or 2-way table

It is quite easy to calculate simple probabilities. What is the chance of rolling a 4 with a single die? What is the chance of being dealt a queen from a deck of cards? We are now going to focus on conditional probabilities. A conditional probability is a probability in which a certain prerequisite condition has already been met.

We can start by thinking about cards being dealt from a standard deck of 52 cards. Suppose a specific piece of information is given to you about a particular card that has been dealt from the deck face down. For example, suppose that we tell you that the card is red. We now might ask what the probability is that the card is a heart. Since we already know it is red, the probability of it being a heart must be 50%. This is because there are equal numbers of hearts and diamonds in the deck of cards. The formal notation for this is P(Heart|Red). This is read as “The probability of a heart given that the card is red”. The mathematics for these types of situations is typically very logical. In our case, we know there are 26 red cards and that 13 of them are hearts. Therefore, we make the conclusion that P(Heart|Red) = 0.50.

Example 1
A single card is dealt from a standard deck of 52 cards. Find each conditional probability.

a) \(P(2\spadesuit|\text{Black})\)

b) \(P(\text{Black}|\text{Diamond})\)

c) \(P(\text{Queen}|\text{Face})\)

Solution
a) There are 26 black cards in the deck and one of them is the two of clubs.

\[
P(2\spadesuit|\text{Black}) = \frac{1}{26} = 0.04
\]

b) It is impossible for the card to be black if we know it is a diamond. \(P(\text{Black}|\text{Diamond}) = 0.\)

c) There are 4 queens out of the 12 face cards in a deck. \(P(\text{Queen}|\text{Face}) = \frac{4}{12} = \frac{1}{3} = 0.33\)
Example 2

In a common poker game, 5 cards are dealt to a player. The best possible hand is called a royal flush. This occurs if a player gets the ten, jack, queen, king, and ace, all of the same suit. What is the chance of being dealt a royal flush? Leave your answer as a fraction.

Solution

We will solve this by looking at one card at a time. What is the chance that the first card might be part of a royal flush? Before any cards are dealt, there are four 10’s, four jack’s, four queen’s, four kings, and four aces available. Twenty of the 52 cards can help you on your way to a royal flush.

Once you receive this card, what are the chances that the second card will also help on your way to the royal flush? We might answer this by simply imagining us getting a useful card on the first card. Suppose our first card was the jack of spades. There are only 4 other cards of the remaining 51 cards that will help now, the 10, queen, king, and ace of spades. Suppose we get one of those cards, perhaps the king of spades.

There are now only 3 cards of the remaining 50 that can help us complete our royal flush. Suppose our third card was the queen of spades. Only 2 of the remaining 49 cards will help as our fourth card on our quest for a royal flush. Likewise, there is only one card of the last 48 that can help us on card number five. Putting this all together, we have

\[
\frac{480}{311,875,200} = \frac{1}{649,740}
\]

Putting this in perspective, if you dealt 1000 poker hands every single day, it would take nearly two years to deal 649,740 hands, of which we would only expect about one to be a royal flush.

Good Luck!

Another way we can look at conditional probabilities is through the use of two-way tables or contingency tables. These are often referred to as two-way tables because there are two distinct categories of information gathered in these tables. For example, we may record how many siblings you have and in how many activities you participate in school. Two-way tables can be filled in either using counts or probabilities.

We will start by answering simple questions such as “What is the probability that a student participates in exactly 2 activities?”. After we understand how to work with these tables, we will begin asking more complex questions such as “What is the chance a student participates in 3 activities given that they have 1 sibling?”. Let’s begin with an easy example to help us understand how to read these tables.
Example 3

Suppose we survey all the students at school and ask them how they get to school and also what grade they are in. The chart below gives the results. Suppose we randomly select one student.

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Walk</th>
<th>Car</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>9th or 10th grade</td>
<td>106</td>
<td>30</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>11th or 12th grade</td>
<td>41</td>
<td>58</td>
<td>184</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>147</td>
<td>88</td>
<td>254</td>
<td>11</td>
</tr>
</tbody>
</table>

a) Give all the row and column totals.

b) What is the probability that the student walked to school?

c) What is the probability that the student was a 9th or 10th grader?

d) What is the probability that a student either rode the bus or is in 11th or 12th grade?

Solution

a)

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Walk</th>
<th>Car</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>9th or 10th grade</td>
<td>106</td>
<td>30</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>11th or 12th grade</td>
<td>41</td>
<td>58</td>
<td>184</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>147</td>
<td>88</td>
<td>254</td>
<td>11</td>
</tr>
</tbody>
</table>

b) There were 88 walkers out of 500 total students or \( \frac{88}{500} = \frac{22}{125} \approx 0.18 \).

c) There were 210 9th or 10th graders out of 500 total students or \( \frac{210}{500} = \frac{21}{50} = 0.42 \).

d) There are 147 kids who rode the bus and there are 290 kids who are 11th or 12th graders. However, notice that these two categories intersect and we must be careful not to count the 41 kids who are in both categories twice. We will take the 290 11th or 12th graders and just add the 106 bus riders who are not 11th and 12th graders for a total of 396 students. The probability of selecting an 11th or 12th grader or a bus rider is \( \frac{396}{500} = \frac{99}{125} \approx 0.79 \).

In the example above, note that the total across the bottom, 147 + 88 + 254 + 11, and the total for the last column, 210 + 290, both add up to 500. This is true of all 2-way tables. Now that we have the basic ideas down in a contingency table, let’s move to a couple of more challenging questions.
Example 4

Consider the completed chart in the solution of part a) of Example 3.

a) What is the probability that a student is in 11th or 12th grade given that they rode in a car to school?

b) What is P(Walk|9th or 10th grade)?

Solution

a) The trick to dealing with conditional probabilities in two-way tables is to make sure that you only use what you are given. We are given that they rode in a car to school. We will only look at the Car column. We first note that there were a total of 254 kids who rode in a car to school. We then see that 184 of these kids were 11th and 12th graders. This gives us \( \frac{184}{254} = \frac{92}{127} \approx 0.72 \).

b) We want the probability that a student walked to school given that they were in 9th or 10th. We will only look only at the 9th and 10th grade row. There are 210 students who are 9th and 10th graders. Of these, only 30 walked to school. This gives us \( \frac{30}{210} = \frac{1}{7} \approx 0.14 \).

Example 5

The manager of an ice cream shop is curious as to which customers are buying certain flavors of ice cream. He decides to track whether the customer is an adult or a child and whether they order vanilla ice cream or chocolate ice cream. He finds that of his 224 customers in one week that 146 ordered chocolate. He also finds that 52 of his 93 adult customers ordered vanilla. Build a contingency table that tracks the type of customer and type of ice cream.

Solution

Start by filling in the values we are given and then work from there. The table to the right shows what we are given in the initial problem.

<table>
<thead>
<tr>
<th></th>
<th>Adult</th>
<th>Child</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>52</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chocolate</td>
<td></td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>131</td>
<td>224</td>
</tr>
</tbody>
</table>

Our next step is to fill in the Adult/Chocolate space with 41, the Child/Total box with 131, and the Total/Vanilla box with 78 by using subtraction. It is now easy to fill in the remaining boxes. For example, we can quickly determine that the Child/Chocolate box must be 146 – 41 = 105. You can verify that this table is correct by checking each row and column total.

<table>
<thead>
<tr>
<th></th>
<th>Adult</th>
<th>Child</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>52</td>
<td>26</td>
<td>78</td>
</tr>
<tr>
<td>Chocolate</td>
<td>41</td>
<td>105</td>
<td>146</td>
</tr>
<tr>
<td>Total</td>
<td>93</td>
<td>131</td>
<td>224</td>
</tr>
</tbody>
</table>
Example 6

A survey asked students which types of music they listen to? Students could select more than one type of music. Out of 200 students, 75 indicated they listened to pop music and 45 indicated they listened to country music with 22 students indicating that they listened to both. Use a Venn diagram to find the probability that a randomly selected student listens to pop music given that they listen country music.

Solution

Consider our Venn diagram below. First, fill in the “both” section with 22 students. We can now logically deduce how many students are left to fill up the Pop circle and the Country circle. Since we are given that they listen to country music we may only use the information that is in the Country circle. There are only 45 students that landed in this circle. Of the 45 students who listen to country music, 22 of them also listen to pop music or \( \frac{22}{45} \approx 0.49 \).
Problem Set 2.5

Exercises

1) The table below shows the counts of earned degrees for several colleges on the East Coast. The level of degree and the gender of the degree recipient were tracked. Row & Column totals are included.

<table>
<thead>
<tr>
<th></th>
<th>Bachelor’s</th>
<th>Master’s</th>
<th>Professional</th>
<th>Doctorate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>542</td>
<td>128</td>
<td>26</td>
<td>18</td>
<td>714</td>
</tr>
<tr>
<td>Male</td>
<td>438</td>
<td>165</td>
<td>38</td>
<td>20</td>
<td>661</td>
</tr>
<tr>
<td>Total</td>
<td>980</td>
<td>293</td>
<td>64</td>
<td>38</td>
<td>1375</td>
</tr>
</tbody>
</table>

a) What is the probability that a randomly selected degree recipient is a female?
b) What is the probability that a randomly chosen degree recipient is a male?
c) What is the probability that a randomly selected degree recipient is a woman, given that the degree recipient received a Master’s Degree?
d) For a randomly selected degree recipient, what is \( P(\text{Bachelor’s Degree} | \text{Male}) \)?

2) In poker, 5 cards are dealt to a player. One of the stronger poker hands is a flush. This means that all 5 cards are of the same suit, for example, all hearts. What is the probability of being dealt a flush?

3) The table below shows the probability breakdown of ages and genders for the typical American college student. Each value in the table is given as a probability. For example, there is a 12% chance that a randomly selected college student will be a male between 25 and 34 years old. Copy the table into your homework, find row and column totals, and then answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>Age 14-17</th>
<th>Age 18-24</th>
<th>Age 25-34</th>
<th>Age &gt;34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.01</td>
<td>0.30</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>Female</td>
<td>0.01</td>
<td>0.30</td>
<td>0.13</td>
<td>0.09</td>
</tr>
</tbody>
</table>

a) What is the probability that a randomly selected American college student is female?
b) What is the probability that a randomly selected American college student is female given that the student is more than 34 years old?
c) What is the probability that a randomly selected college student is either a female or more than 34 years old?

4) Suppose that 40% of adults like eating bananas while 60% like eating apples. Suppose also that 32% of adults like eating both. What is the conditional probability that a randomly selected adult likes apples given that they like bananas? Use a Venn Diagram to answer this question.
5) Another good poker hand is called a straight. This means that your five cards will be numerically in order such as an 8, 9, 10, jack, and queen. The cards do not need to match suit in a straight. Suppose you receive the first four cards of a five card poker hand. You have 5♥, 7♦, 8♣, and 9♠. What is the probability that the next card will give you a straight?

6) Suppose you receive the first four cards of a five card poker hand. You have 3♥, 4♦, 5♣, and 6♠. What is the probability that your next card will give you a straight?

7) A statistics class has 18 juniors and 10 seniors in it. 6 of the seniors are females and 12 of the juniors are males. Build a contingency table to find the probability that a randomly selected student is:
   a) a junior or a female.
   b) a senior or a female.
   c) a junior or a senior.
   d) a female given that the student was a senior.

8) At a used-book sale, there are 120 children’s books and 80 adult books available. 50 of the adult books are nonfiction while 40 of the children’s books are nonfiction. All other books are fiction. Build a contingency table to find the probability that a randomly selected book is:
   a) fiction.
   b) not a children’s nonfiction.
   c) an adult book or children’s nonfiction.
   d) a children’s book given that it was nonfiction.

9) Animals on the endangered species list are given in the table below by type of animal and whether it is domestic or foreign to the United States. Copy the table into your homework, find row and column totals, and then answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>Mammals</th>
<th>Birds</th>
<th>Reptiles</th>
<th>Amphibians</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>63</td>
<td>78</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Foreign</td>
<td>251</td>
<td>175</td>
<td>64</td>
<td>8</td>
</tr>
</tbody>
</table>

An endangered animal is selected at random. What is the probability that it is:
   a) a bird found in the United States?
   b) foreign or a mammal?
   c) a bird given that it is found in the United States?
   d) a bird given that it is foreign?
10) Suppose a standard set of pool balls (1-8 are solid and 9-15 are striped) are in a bag. A single pool ball is picked out of the bag without replacement.
   a) Find the probability that the pool ball has an odd number if we know that it is solid.
   b) Find P(Striped|Even).

11) Cable channels 6, 8, & 10 show quiz shows, comedies, & dramas. The table below shows the distributions of these shows. Copy the table, find row and column totals, and then answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>Channel 6</th>
<th>Channel 8</th>
<th>Channel 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz Show</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Comedy</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Drama</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

If a show is selected at random, find the probability that the show is:
   a) a quiz show or shown on Channel 8.
   b) a drama or a comedy.
   c) a comedy given that it is shown on Channel 8.
   d) shown on Channel 6 given that it is a drama.

12) Suppose you receive your first three cards of a five card poker hand. You have 5♦, 6♦, 7♥. What is the probability that your next two cards will result with you having a straight?

Review Exercises

13) Suppose that 25% of all 9th graders have an unweighted GPA after their 9th grade year of 3.5 or higher. Also suppose that 60% of all 9th graders are involved in a sport at some time during their 9th grade year. Assume that a student’s GPA and whether or not they are in a sport are independent of one another. Draw and label a tree diagram for this situation and build a probability model that summarizes the different probabilities possible for 9th grade students in regards to GPA and being in a sport.

14) A special deck of cards contains only the face cards and aces from a standard deck of cards.
   a) If one card is dealt, what is the probability that the card is an ace?
   b) If one card is dealt, what is the probability that the card is a black ace?
   c) If two cards are dealt, what is the probability that both cards are face cards?

15) Suppose for a moment that all months have exactly 30 days and the chance of you being born in any particular month is \( \frac{5}{365} \). What is the probability that neither of two randomly selected people will have been born in the same month as you?

16) The standard California license plates made in 2011 or later must begin with a digit anywhere from 6 to 9 followed by a letter anywhere from T to Z. They then have any two letters followed by any three digits. How many of these license plates are possible?
2.6 Chapter 2 Review

Probability is a simple question of how likely it is for a particular outcome to occur. We always look to divide the number of favorable outcomes by the total number of outcomes. We cannot predict a specific outcome for a random event but the Law of Large Numbers allows us to make long term predictions of chance behavior. The rules of probability dictate that we pay attention to whether events are independent, outcomes are mutually exclusive, or whether replacement is used. We also must deal with conditional probabilities for situations in which a particular outcome is assumed to have occurred. To help organize situations, use Venn diagrams, tree diagrams, or contingency (2-way) tables. Having a clear understanding of situations and being organized when dealing with probability is critical to successful calculations.

Review Exercises

1) Suppose that 57 of 110 students at a school are underclassmen (freshmen or sophomores) while the rest of the students are upperclassmen (juniors or seniors). Suppose three students are selected at random.
   a) What is the probability that all three of the students are underclassmen?
   b) What is the probability that all three students are upperclassmen?
   c) What is the probability that there is at least one underclassman and at least one upperclassman in the group of three students?

2) Two 6-sided dice are rolled, one after the other. Find each probability.
   a) $P($total of 10 or more$)$
   b) $P($doubles$)$
   c) $P($total is even or total is less than 6$)$
   d) $P($an odd product$)$
   e) $P($first die is greater than second die$)$
   f) $P($a 6 or a 3 is showing on at least one die$)$
   g) $P($total is odd or at least one of the dice is a 2$)$

3) A pet store surveys his customers during the day and finds that 15 customers own dogs and 9 own cats. Included in these were 4 customers who owned both.
   a) Draw a Venn diagram for this situation
   b) How many total customers were surveyed?
   c) Suppose one of these customers was selected at random. What is $P($owned a dog$)$?
   d) Suppose one of these customers was selected at random. What is $P($own a dog | own a cat$)$?
   e) Suppose one of these customers was selected at random. What is $P($own a cat | own a dog$)$?
4) Suppose that 40% of all adults in a certain town are females and that 60% are males. In addition, 60% of the females hold full-time jobs while 80% of the males hold full-time jobs.
   a) Draw and label a tree diagram to represent this situation.
   b) What is the chance that a randomly selected person holds a full-time job?

5) For Halloween at my house, kids spin a spinner that has three equally marked spaces labeled 1, 2, and 3. The number they spin is the number of pieces of candy they get. In my bag, I start with 20 chocolate bars and 30 sugar bombs - all with identical packaging. Trick-or-Treaters pick randomly out of my bag after they spin. I only restock my candy bag after each child finishes picking all their candy.
   a) What is the chance that a trick-or-treater gets to pick three pieces of candy?
   b) Suppose a trick-or-treater spins a three. What is the chance that they pick three sugar bombs?
   c) Suppose a trick-or-treater spins a three. What is the chance that they pick 3 chocolate bars?
   d) What is the chance that the trick-or-treater gets only one chocolate bar and nothing else?
   e) What is the chance that the trick-or-treater gets exactly one chocolate bar and one sugar bomb?

6) Suppose the table below gives a breakdown of the ages and genders of the teachers at your school. Copy the table into your homework, find row and column totals, and then answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>Age ≤ 29</th>
<th>Age 30-39</th>
<th>Age 40-49</th>
<th>Age ≥ 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>5</td>
<td>6</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>Female</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>4</td>
</tr>
</tbody>
</table>

Find the probability that a randomly selected teacher is:
   a) a male.
   b) 39 years old or younger.
   c) either a male or at least 50 years old.
   d) from 30 to 39 years old given that they are a female.
   e) a female given that they are at least 40 years old.
7) The 2-way table shown below shows the number of different types of automobiles produced by major manufacturers.

<table>
<thead>
<tr>
<th></th>
<th>GM</th>
<th>Ford</th>
<th>Chrysler</th>
<th>Toyota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars</td>
<td>14</td>
<td>11</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Trucks</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Vans</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

What is the probability that a randomly selected vehicle is:

a) a Ford?

b) a truck?

c) a van or a Toyota?

d) a car given that the vehicle is built by GM?

e) a Ford given that the vehicle is a truck?

8) Two cards are dealt from a standard 52 card deck without replacement. What is the probability that neither of the two cards are face cards?

9) A baseball player has a batting average of .250 which means that he averages one hit for every four times he comes to the plate. What is the probability that this player will end up with exactly 2 hits if he comes to the plate 3 times in a single game?

10) Two bags have an assortment of marbles in them. The first bag contains 11 black, 12 white, and 7 gold marbles. The second bag contains 9 black and 11 white marbles. One marble is randomly selected out of each bag.

   a) Draw a tree diagram to represent this situation.

   b) What is the probability that the two marbles are both black?

   c) What is the probability that the two marbles are the same color?

11) A special deck of cards has exactly 8 cards that consist only of the aces and face cards that happen to be red. Two cards are dealt off the top of the deck.

   a) What is the probability that the two cards dealt are both kings?

   b) What is the probability that the two cards have different values?

   c) What is the probability that the two cards have the same value (two kings, two queens, etc...)?

   d) What is the probability that the two cards are the same suit?
12) A bag contains ten red cubes numbered 1 to 10 and five green cubes numbered 1 to 5. Two cubes are pulled from the bag at random without replacement. What is the probability that the cubes are:
   a) both red?
   b) both odd?
   c) the same color?
   d) the same value?

13) For a carnival game, a bag contains one $100 bill and nine $20 bills. You roll a single 6-sided die one time. If you roll a one or two you get to pull one bill out of the bag. If you roll a three, four, five, or six, you get to pull two bills out of the bag.
   a) Draw a tree diagram for this situation.
   b) Build a probability model for this situation.
   c) What is the probability that you win exactly $120?

14) A burglar alarm system has three separate detection mechanisms it uses to detect an intruder.
    Suppose a skilled burglar has a 30% chance to get around the first part of the detection system, a 60% chance of getting around the second part of the system, and a 55% chance of getting around the third part of the system. Assume each part of the detection system is independent of the other parts of the system.
    a) What is the chance that the system does not detect the burglar?
    b) Based upon your answer to part a), what must be the chance that the system does detect the burglar?
    c) What is the chance that the burglar can get around exactly 2 of the 3 parts of the system?

15) On a basketball team, players can play at least one of three positions; guard, forward, or center. Suppose that 30 girls try out for the basketball team. During tryouts 13 girls indicate they can play guard only, 3 state they can play center only, 6 state they can play center or forward and the rest state they can play forward only. A player is selected at random.
    a) Draw a Venn diagram for this situation.
    b) What is the probability that the randomly selected player says they can play forward?
    c) Given that the player indicates they can play forward, what is the probability they can also play center?

16) A girl is deciding what jewelry to wear as she gets ready for school. She has 5 bracelets, 6 rings, and 8 necklaces from which to choose.
    a) In how many ways can she choose exactly one of each item to wear from the 19 available items?
    b) If she decides to randomly select three pieces of jewelry, what is the probability that all three of the items she picks are exactly the same type of jewelry?
    c) What is the probability that she picks exactly one bracelet, one ring, and one necklace if she randomly selects three pieces of jewelry?
17) Your statistics teacher needs to select 3 students to help demonstrate an activity. Your class has 12 sophomores, 19 juniors, and 5 seniors in it. Your teacher makes a random selection of three students.

a) In how many ways can your teacher select three students from this class of 36 students?

b) What is the probability that all three students will be juniors?

c) What is the probability that exactly one student from each grade will be selected?

18) All football plays that an offense can run can be classified as a pass, run, or a kick. No play can ever be put into two categories.

a) If an offense completes two plays, will these two plays be independent of each other? Why or why not?

b) If the offense runs one play, are the possible outcomes (pass, run, or kick) mutually exclusive? Why or why not?

Image References

Coins http://coinauctionshelp.com
Pool Balls http://plutonium.aibrean.com
Scattegories Die http://ehow.com
October Calendar http://printablecalendars.resources2u.com
Roulette Wheel https://bitcoinvideocasino.com/static/images/wheel.png
Kids at Board http://teachers.greenville.k12.sc.us
Two Striped Pool Balls http://demo.physics.uiuc.edu
Ace and King of Spades http://www.123rf.com
Seat Belt http://sawmengzhi.blogspot.com
Graduation http://www.prlog.org
Aerosmith http://www.obit-mag.com
Cabinet http://www.renovation-headquarters.com
Cathedral in St. Paul, MN http://www.scenicreflections.com
Tree http://www.onenewsnow.com
$100 Bill http://onlinecurrencytradingfxcm.blogspot.com
Royal Flush http://www.artpoker.net
Straight http://www.findabet.co.uk
Turtle http://www.maine.gov
Trick or Treaters http://www.myremoteradio.com
Bag of Marbles http://www.worldwiseimports.com
Burglar http://www.emovingstorage.com
Jewelry Box http://www.123rf.com
Chapter 3 – Expected Values & Simulation

3.1 Probability Models & Expected Value

Learning Objectives

- Understand how a probability distribution table can be used to calculate the mean (expected value) for various situations
- Be able to construct a probability distribution model (expected value table) given all possible outcomes and the associated probabilities
- Be able to calculate expected values for various scenarios.
- Be able to find missing values in a probability model

Imagine walking into a casino. You would see all sorts of games varying from blackjack and poker to slot machines. It would not take long for you to notice that there are some players who are winning some money, sometimes a substantial amount. You might wonder how the casino makes money when they are clearly giving some large amounts of money away.

Casinos have a clear understanding of expected value. The expected value for a situation is the average result over the long run. In other words, it can be thought of as the expected winnings or average payout for a game of chance after many plays. Consider the thinking of the owner of a casino. While there are some people who win a little, and occasionally a few people who win a lot, most people end up losing some money at the casino. The casino actually expects people to occasionally win big. In fact it makes for great advertisement! As long as the mathematic calculations show that the expected value is in the casino’s favor, the casino will continue to make money in the long run. In this section, we will focus on how to calculate the expected value.

The expected value is the average result over the long run. Whether you are asked to find the mean, the average, the expectation or the expected value for a given scenario, the calculations will be the same. To find the average value of a series of numbers, we simply add up the numbers and divide by however many numbers there are. For example, the average of the numbers 3, 4, 5, and 6 is 4.5 because \( \frac{3 + 4 + 5 + 6}{4} = 4.5 \). Notice that the average value of 4.5 is not one of the numbers in the original set of numbers. This is often true with expected values. The expected value for a situation is rarely one of the possible outcomes.

Use the concept of averages to find the expected value for the example below.
Example 1

A game is played in which a coin is flipped one time. If the coin lands on tails, the player wins $5. If the coin lands on heads, the player wins $10. What is the expected value for a player who plays this game one time?

Solution

The expected value is $7.50. This is strange because it is actually impossible for a player to win $7.50. They could only win either $5 or $10 but the average winnings will be $7.50. One way to see this is to imagine playing the game two times. If the flips come out matching their theoretical probabilities, one of the flips will be heads for $10 and the other will be tails for $5. The player will have won $15 in two games so the average winnings or expected value would be $\frac{5+10}{2} = \frac{15}{2} = 7.50$.

This method works quite well in simple situations where the probability of each outcome is the same, but it gets more cumbersome as the situations get more complex. Consider the following example.

A really helpful tool when trying to organize the information in a more complex problem is a probability distribution model. A probability distribution model is a table that lists out the entire Sample Space (all possible outcomes) of a scenario and each respective probability. The values listed should include the entire Sample Space and the probabilities should add up to 1 or 100%.

Example 2

The student council at SRHS is raising money to support a program called “Shoes for the Homeless”. A booth was set up in the lunchroom at which students could pledge a donation of $1, $5, or $10 for money towards a large shoe purchase. A total of one-hundred twenty-five students pledged money for this fundraiser. Eighty students pledged $1, twenty-five students pledged $5, and twenty students pledged $10.

a) Build a probability model for this situation.

b) What was the average donation per student, for students who donated to this cause?
Solution

a) First list all possible outcomes: $1, $5, and $10 are all possible donations in this situation. Second determine the probability of each outcome. Remember that a probability model needs probabilities, not just counts. The probabilities can be given as decimals or fractions.

\[
\begin{array}{c}
\frac{80}{125} = 0.64 \\
\frac{25}{125} = 0.20 \\
\frac{20}{125} = 0.16
\end{array}
\]

Be sure to verify that your probabilities do in fact add up to one. In this case: \(0.64 + 0.20 + 0.16 = 1\).

<table>
<thead>
<tr>
<th>Shoe Donation Amounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>

b) One way to determine the average donation per student is to use the fact that there were 80 students who pledged $1 each for a total of $80, there were 25 students who pledged $5 each for a total of $125, and there were 20 students who pledged $10 each for a total of $200. Therefore, the total amount pledged was $80 + $125 + $200 = $405. We now divide this by the 125 students who made donations to get \(\frac{405}{125} = 3.24\) per student. The average donation per student who donated to this cause was $3.24.

Another way to determine the average donation is to use the numbers in your probability distribution model as they are already organized: Average Donation = \((\$1) (0.64) + (\$5) (0.2) + (\$10) (0.16) = \$3.24\). Simply multiply the amount of each donation by its associated probability and add those results together. This leads us to our expected value formula which is given below. The formula is shown in words, and then in symbols.

**Probability Distribution Table:**

<table>
<thead>
<tr>
<th>Value</th>
<th>X</th>
<th>Value of Outcome 1</th>
<th>Value of Outcome 2</th>
<th>Value of Outcome 3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability P(X)</td>
<td>P(outcome 1)</td>
<td>P(outcome 2)</td>
<td>P(outcome 3)</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Expected Value Formula:**

\[ EV = (value\ 1) \cdot P(value\ 1) + (value\ 2) \cdot P(value\ 2) + (value\ 3) \cdot P(value\ 3) + \cdots \]

**Probability Distribution Table in Symbols:**

<table>
<thead>
<tr>
<th>Value</th>
<th>X</th>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability P(X)</td>
<td>P(x₁)</td>
<td>P(x₂)</td>
<td>P(x₃)</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Expected Value Formula in Symbols:**

\[ EV = (x₁) \cdot P(x₁) + (x₂) \cdot P(x₂) + (x₃) \cdot P(x₃) + \cdots \]
Example 3

Many games involve rolling two dice at the same time and counting the total number of pips that are showing. What is the expected value (or average) for the total number of pips showing when two 6-sided dice are rolled?

Solution

We will address this two ways. The first method will be done by using averaging and the second method will be done by using the expected value formula.

First method using averaging: Begin by building the sample space for the sum of two dice. As in section 1.1, we get the dice chart shown below. Notice that there are exactly 36 equally likely spaces on the grid. Instead of playing just one time, suppose we play 36 times. If everything matches the theoretical probabilities, each of these outcomes would happen exactly one time. Add the values for each of the 36 spaces and divide by 36. For simplicity, we will add diagonally to get \[\sum_{i=2}^{12} \text{value}_i \times \frac{1}{36} = \frac{252}{36} = 7.\] The expected value is 7, which means that 7 is the average value of a roll for two 6-sided dice.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

Second method by building a probability model: Build a probability distribution table. Then use the expected value formula to calculate the average.

\[EV = (\text{value}_1)(\text{prob. } 1) + (\text{value}_2)(\text{prob. } 2) + (\text{value}_3)(\text{prob. } 3) + \ldots\]

The probability model for this situation is given below. Does this make sense and did you verify that the total of the probabilities in the table add up to 1?

Number of Pips showing when two dice are rolled:

<table>
<thead>
<tr>
<th>Value</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\frac{1}{36})</td>
<td>(\frac{2}{36})</td>
<td>(\frac{3}{36})</td>
<td>(\frac{4}{36})</td>
<td>(\frac{5}{36})</td>
<td>(\frac{6}{36})</td>
<td>(\frac{5}{36})</td>
<td>(\frac{4}{36})</td>
<td>(\frac{3}{36})</td>
<td>(\frac{2}{36})</td>
<td>(\frac{1}{36})</td>
</tr>
</tbody>
</table>

Using our expected value formula we have

\[EV = \left(2 \left(\frac{1}{36}\right) + (3) \left(\frac{2}{36}\right) + (4) \left(\frac{3}{36}\right) + (5) \left(\frac{4}{36}\right) + (6) \left(\frac{5}{36}\right) + (7) \left(\frac{6}{36}\right) + (8) \left(\frac{5}{36}\right) + (9) \left(\frac{4}{36}\right) + (10) \left(\frac{3}{36}\right) + (11) \left(\frac{2}{36}\right) + (12) \left(\frac{1}{36}\right)\right) = 7.\]

The expected value for the total number of pips showing when two dice are rolled is 7.
Example 4

At a carnival, there is a Duck Pond Game that has exactly 100 ducks. Each duck has a prize amount written on the bottom of its belly. Each player wins the amount of money on the duck they select from the pond. The duck is then returned to the pond. The table below shows the probability model for this game.

a) What is the probability that someone wins $10 on a single play of this game?

b) Calculate the expected value and explain what it means.

<table>
<thead>
<tr>
<th>Probability Model for Duck Pond Carnival Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value of Prize</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>

Solution

a) The probabilities in a probability model must add up to 1. So, 0.01 + 0.03 + ??? + 0.9 = 1 must be true. We can use algebra to determine that the missing value must be 0.06. So, there is 6% chance that a person will win $10 on a single play of Duck Pond.

b) Using the expected value formula, we get:

\[
EV = (\$30)(0.01) + (\$20)(0.03) + (\$10)(0.06) + (\$1)(0.9) = \$0.30 + \$0.60 + \$0.60 + \$0.90 = \$2.40.
\]

Our expected value is $2.40. This means that if this game were played many times, the average payout would be $2.40 per play. Note that the expected value of $2.40 is not an actual possible prize that a player can win.
Example 5

Suppose a casino game has an expected payout of $1 per play. The game involves a Giant Wheel that each player spins once. The spinner is divided such that a player wins nothing 45% of the time, one dollar 35% of the time, three dollars 15% of the time, and some other amount the remainder of the time.

a) Build a probability model for this situation. Be sure to calculate the percent of time the remaining payout occurred.

b) How much should this payout be so that the expected value is equal to exactly $1?

Solution

a) Start by noticing we have used 45% + 35% + 15% = 95% of all outcomes. This means that the remaining outcome has a 5% chance of occurring. This allows us to build a probability model that is mostly complete.

Probability Model for Giant Wheel Game

<table>
<thead>
<tr>
<th>Amount</th>
<th>$0</th>
<th>$1</th>
<th>$3</th>
<th>???</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.45</td>
<td>0.35</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>

b) Our calculation is now based upon the expected value formula.

We will use the variable x to represent the missing amount. Then use algebra to solve for x.

\[(0.45)(0) + (0.35)(1) + (0.15)(3) + (0.05)(x) = 1.\]

\[0 + 0.35 + 0.45 + (x)(0.05) = 1.\]

\[0.80 + (x)(0.05) = 1.\]

\[(x)(0.05) = 0.20\]

\[x = \frac{0.20}{0.05} = 4.\] The missing payout is a prize of $4.

Example 6

A carnival game has prizes and probabilities as shown in the table below. How much should the game cost if the carnival owner wants to average a $2 profit per play?

<table>
<thead>
<tr>
<th>Value</th>
<th>$3</th>
<th>$5</th>
<th>$20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.65</td>
<td>0.30</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Solution

First calculate the expected value to get \( EV = (3)(0.65) + (5)(0.30) + (20)(0.05) = 4.45. \) This means that the average player will be paid $4.45 when they play. Therefore, the owner should charge $2 more than this in order to make an average of $2 profit per play. They should charge $6.45 for each play of this carnival game.
Problem Set 3.1

Exercises

1) The probability distribution table below reports the number of vehicles owned per household and the associated probability of having each number of vehicles. What is the expected, or average, number of vehicles in a typical household?

<table>
<thead>
<tr>
<th>Vehicle Ownership</th>
<th># Owned</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.04</td>
</tr>
</tbody>
</table>

2) Juliet sold products as part of a fundraiser to raise money for a choir trip to New York. She sold 75 items total which included 50 rolls of cookie dough for $6 each, 15 packages of butter braids at $10 each, and 10 bake-at-home bread packs for $12 each.

   a) Build a probability model for this situation.
   
   b) Find the average amount for each of Juliet’s sales.

3) The owner of Friendly’s Casino decides that she will set up her payouts in their ‘Fast Cash’ game so that the average gambler neither wins nor loses money. For a gambler who plays this game, the chance of getting paid nothing is 30%, the chance of getting paid $5 is 40%, the chance of getting paid $10 is 25%, and the chance of getting paid $30 is 5%. Build a probability model for Fast Cash and determine how much the owner of Friendly’s should charge for this game?

4) The owner of Greedy’s Casino decides he wants to make an average of $1.50 every time a gambler plays the game called ‘Funny Money’. The chance of getting paid $2 is 20%, the chance of getting paid $5 is 40%, the chance of getting paid $10 is 30%, and the chance of getting paid $15 is 10%. Build a probability model and determine how much money Greedy should charge to play this game?

5) In a certain racing video game, players try to go around a track as many times as possible. If a racer completes a lap in time, they continue on to the next lap. If they don’t complete a lap in time, their race is complete at the end of the lap they are currently finishing. The probability model below gives the probabilities of the maximum number of laps completed by people who play this video game. What is the expected number of laps completed for each racer?

<table>
<thead>
<tr>
<th>Laps Completed During Racing Game</th>
<th># of Laps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

| Probability | 0.29 | 0.38 | 0.17 | 0.11 | 0.05 |
6) In a certain casino game, the average payout (expected value) for a player is $2.53. A partially completed probability model for this game is given below.

<table>
<thead>
<tr>
<th>Casino Game Payouts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amount Paid</strong></td>
</tr>
<tr>
<td><strong>Probability</strong></td>
</tr>
</tbody>
</table>

a) What is the missing payout amount?

b) If the casino was going to set a price for this game, do you think they would choose to charge $2 to play or $3 to play? Explain your choice.

c) If the casino was going to set a price for this game, do you think they would choose $3 to play or $6 to play? Explain your choice.

7) What is the average number of pips showing when a single 6-sided die is rolled?

8) A game is played in which a coin is flipped one time. If it lands on heads, you win $20. If the coin lands on tails, you win $30. Build a probability model and calculate the expected value for this game by using the expected value formula.

9) Two students are given the partially completed probability model below as part of a project. The teacher tells them that the expected value for this situation is $6.95.

<table>
<thead>
<tr>
<th>Probability Model Given to Students</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value</strong></td>
</tr>
<tr>
<td><strong>Probability</strong></td>
</tr>
</tbody>
</table>

a) Assuming that the expected value of $6.95 is correct, use the expected value formula and algebra to determine the value of the missing probability? Explain why this is impossible.

b) Now that we see that the expected value of $6.95 is impossible, what should be the value of the missing probability?

c) Using the correct probabilities, calculate the actual expected value for this situation?

10) I want to come up with a game that has 5 prizes. There will be a 20% chance of getting paid $1, a 25% chance of getting paid $3, a 15% chance of getting paid $4, and a 30% chance of getting paid $7.

a) Build a probability model and find the probability of winning the 5th prize?

b) What is the amount of the 5th prize if I want the expected payout for the game to be $4.75?

11) For Halloween next year, Mr. Crabtree has decided that he will distribute an average of 1.6 pieces of candy per child who comes to his door. To help him do this, he has set up a game of chance whereby each trick-or-treater gets to play a game that determines how many pieces of candy they get to pick. He started building a probability model that shows the probabilities of being able to select 0, 1, 2, 3, or 4 pieces of candy. Unfortunately, he does not know how to assign the remaining probabilities so that the average number of candies is exactly 1.6 pieces. Use the table on the following page to answer the questions.
Mr. Crabtree’s Halloween Candy Distribution

<table>
<thead>
<tr>
<th># of Pieces</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.03</td>
<td>0.45</td>
<td>x</td>
<td>y</td>
<td>0.02</td>
</tr>
</tbody>
</table>

a) Give the expected value equation using the variables x and y and the expected value of 1.6. Simplify your equation by combining like terms.

b) Give an equation using the variables x and y that uses the fact that the probabilities in a probability model must add up to one. Simplify your equation by combining like terms.

c) What should the probabilities of receiving 2 or 3 pieces of candy be? Using your answers from parts a) and b), write a system of equations and solve for the variables x and y.

Review Exercises

12) A sample of 325 students were asked which electronic device they use most frequently, their cell phone, a computer (including wireless devices), or a television (including video games). Whether the student was male or female was also recorded. The table below shows the results.

<table>
<thead>
<tr>
<th></th>
<th>Cell Phone</th>
<th>Computer</th>
<th>Television</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>60</td>
<td>30</td>
<td>55</td>
<td>145</td>
</tr>
<tr>
<td>Female</td>
<td>115</td>
<td>45</td>
<td>20</td>
<td>180</td>
</tr>
<tr>
<td>Total</td>
<td>175</td>
<td>75</td>
<td>75</td>
<td>325</td>
</tr>
</tbody>
</table>

a) What is the probability that a randomly selected student was a male?

b) What is the probability that a randomly selected student said that they used a cell phone most frequently?

c) What is the probability that a student was male given that they indicated they used the television most frequently?

d) What is the probability that a student indicated that they used a computer most frequently given that they were a female?

13) One floor of an office building is being remodeled and redecorated and an employee is responsible for picking out three different styles of chair and two different styles of table for the new office furniture. Suppose the furniture store has 10 different chair styles and 4 different table styles that would be appropriate office furniture for the new decor. In how many different ways can the employee select the three chair styles and two table styles?

14) Suppose 1 card is drawn randomly from a standard deck of 52 cards. Find each probability. There is information about a deck of cards in the appendix if needed.

a) P(Red Card)

b) P(Spade)

c) P(Face)

d) P(Heart | Red)
3.2 Applied Expected Value Calculations

Learning Objectives
- Understand the concept of a fair game
- Understand how probability distribution models can be used to organize information and solve problems
- Be able to use various methods to calculate probabilities
- Be able to analyze a game of chance by building a probability model and calculating expected values from scratch

Casino owners have a very delicate balancing act they must manage. First of all, they want to make a profit. However, most people simply don’t like to lose money. In order to make money, the casino games have to be in favor of the house and not the player. Why don’t casinos tilt the games even more to the house’s favor? If they did, their expected value would certainly go up. On the other hand, if the games cost too much and the average winnings are low, then attendance at the casino would go down.

No matter what the odds, casinos cannot make money unless they can keep people coming through the doors. Setting the games up so that there are still winners, some occasionally big winners, is good for attendance. You might even know someone who has made a large amount of money at a casino.

In this section, we will combine what we have learned about calculating probabilities from Chapter 2 with the concept of expected value from Section 3.1. This will allow us to analyze various situations, including games of chance. A fair game is a game in which neither the player nor the house has an advantage. In other words, when all is said and done, the average player will not have made or lost any money whatsoever. A fair game is one in which everyone breaks even in the long run.

Example 1
A bag has 10 red marbles and 8 blue marbles in it. A player reaches into the bag pulling out 2 marbles, one after the other without replacement. If the colors of the two marbles match, the player wins $10. If they don’t match, the player wins nothing. The game costs $5 to play.

a) Use a tree diagram to find the Sample Space and respective probabilities. Use your results to build a probability model.

b) Is this game a fair game? If so, explain why. If not, give the value that the game should cost in order to be fair.
Solution

a) Begin with a tree diagram that shows what might happen when two marbles are pulled without replacement. Then multiply the probabilities to find

\[
P(\text{Red}) = \frac{9}{17} \quad \text{and} \quad P(\text{Blue}) = \frac{8}{17}
\]

Summarizing these results into a table, we get the probability distribution table below.

<table>
<thead>
<tr>
<th>Marble Distribution</th>
<th>Red, Red</th>
<th>Red, Blue</th>
<th>Blue, Red</th>
<th>Blue, Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>$10</td>
<td>$0</td>
<td>$0</td>
<td>$10</td>
</tr>
<tr>
<td>Value (amount won)</td>
<td>\frac{90}{306}</td>
<td>\frac{80}{306}</td>
<td>\frac{80}{306}</td>
<td>\frac{56}{306}</td>
</tr>
<tr>
<td>Probability</td>
<td>\frac{90}{306}</td>
<td>\frac{80}{306}</td>
<td>\frac{80}{306}</td>
<td>\frac{56}{306}</td>
</tr>
</tbody>
</table>

b) Use the expected value formula to calculate the average expectation per play.

\[
EV = \left(\$10 \right) \left(\frac{90}{306}\right) + \left(\$0 \right) \left(\frac{80}{306}\right) + \left(\$0 \right) \left(\frac{80}{306}\right) + \left(\$10 \right) \left(\frac{56}{306}\right) = \frac{1460}{306} \approx \$4.77
\]

The expected value is $4.77. Notice, however, that $4.77 is the expected amount paid out to the player each game. Thus, the expected gain for the person running the game is $0.23 because every player must pay $5 to play. At $5, the game is not fair because it favors the person running the game by an average of 23 cents every time the game is played. To be a fair game, it should cost exactly $4.77 per play.
The game of *GREED* is a game of chance in which players try to decide when they have accumulated enough points on a turn to stop. Two 6-sided dice are rolled. The player gets to keep the total that shows on the two dice. After every roll, the player can either decide to roll again and try to add to their current total for that turn or stop and put their points in the bank. The only catch in this game is that if a total of 5 is rolled, all points accumulated on that turn are lost. For example, suppose the first roll has a total of 9 and the player decides to go again. The next roll has a total of 7. The player now has 16 points accumulated on this turn and must decide to either put those 16 points in the bank or risk them. If they decide to risk the 16 points and a total of 5 comes up next, the score for that turn will be 0.

**Example 2**

Suppose a person is playing *GREED* and has accumulated 26 points so far. Is it to their advantage to roll one more time? Evaluate what his or her expected score will be after another roll of the dice.

**Solution**

We will build a probability model and calculate the expected value based upon what might happen with one more roll. (See Example 3 from Section 3.1.) For example, there is a $\frac{1}{36}$ chance that the total will be 2. If this happened, it would mean the player would now have a total of 28 points. The highest a player could have after this turn would be 38 points if he or she happened to roll a total of 12. The risk is that the player will roll a total of 5 and lose their 26 points.

**Total score when two more dice are rolled, if starting with 26 points:**

<table>
<thead>
<tr>
<th>Value</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob.</td>
<td>$\frac{1}{36}$</td>
<td>$\frac{2}{36}$</td>
<td>$\frac{3}{36}$</td>
<td>$\frac{4}{36}$</td>
<td>$\frac{5}{36}$</td>
<td>$\frac{6}{36}$</td>
<td>$\frac{5}{36}$</td>
<td>$\frac{4}{36}$</td>
<td>$\frac{3}{36}$</td>
<td>$\frac{2}{36}$</td>
<td>$\frac{1}{36}$</td>
</tr>
</tbody>
</table>

The one item to be careful about here is that it is impossible to get a total of 31 in the chart. Remember, if you roll a total of 5, you lose all of your points. When we use the expected value formula for the probability model above, we get approximately 29.6. In other words, if we roll exactly one more time, our average result will be almost 30 points. This is definitely better than stopping with 26 points. It is to the advantage of the player to roll again.

**Example 3**

An investor is going to make a long-term investment in a company. If all goes well, an investment of $100,000 will be worth $900,000 in twenty years. The risk is that the company may go bankrupt within twenty years in which case the investment is worthless. Suppose there is a 25% chance that the company will go bankrupt within 20 years. What is the expected value of this investment?

**Solution**

Start by building the probability model shown below that shows that there is a 25% chance of making nothing and a 75% chance of making $900,000.

<table>
<thead>
<tr>
<th>Value</th>
<th>$0$</th>
<th>$900,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.25</td>
<td>0.75</td>
</tr>
</tbody>
</table>

$EV = (0)(0.25) + (900,000)(0.75) = 675,000$. Taking into account that this investment cost $100, the investor expect that an investment of this sort has an expected value of $575,000.
Problem Set 3.2

Exercises

1) In the carnival game Wiffle Roll, a player will roll a wiffle ball across some colored cups. If the ball stops in a blue cup, the player wins $20. If it stops in a red cup they win $10, and if it stops in a white cup, the player wins nothing. There are 25 white cups, 4 red cups, and 1 blue cup. Assume the chances of stopping in any cup is the same. Build a probability model and determine how much this game should cost if it is to be a fair game?

2) In a simple game, you roll a single 6-sided die one time. The amount you are paid is the same as the value rolled. For example, if you roll a one, you get paid $1. If you roll a two you get paid $2 and so on. The only exception to this is if you roll a 6 in which case you get paid $12. Build a probability model and determine the amount this game costs to play if it is a fair game?

For problems 3, 4 and 5, refer to the description of the game GREED before Example 2.

3) Suppose you are playing the game of GREED as described in Example 2. You have accumulated a total of 55 points so far on one turn. Is it to your advantage to roll one more time? Evaluate what his or her expected score will be after another roll of the dice.

4) Suppose you are playing the game of GREED again. This time you have accumulated a total of 60 points so far in one turn. Is it to your advantage to roll one more time? Evaluate what his or her expected score will be after another roll of the dice.

5) Using your results from problems 3) and 4) and a little more investigation, for what number of points in a turn in the game of GREED does it make no difference if you roll one more time or stop? In other words, at what point total does the expected value with one more roll give the same total as if you had stopped?

6) In the Minnesota Daily 3 lottery, players are given a lottery ticket based upon 3 digits that they pick. Digits may be repeated. If their 3 digits match the winning digits in the correct order, then the player wins $500. If the digits don’t match, then the player loses. The game costs $1 to play. Build a probability distribution table and calculate the expected value for a player of this lottery game?

7) A bucket contains 12 blue, 10 red, and 8 yellow marbles. For $5, a player is allowed to randomly pick two marbles out of the bucket without replacement. If the colors of the two marbles match each other, the player wins $12. Otherwise the player wins nothing. Build a probability distribution table and determine the expected gain or loss for the player?
8) An insurance company insures an antique stamp collection worth $20,000 for an annual premium of $300. The insurance company collects $300 every year but only pays out the $20,000 if the collection is lost, damaged, or stolen. Suppose the insurance company assesses the chance of the stamp collection being lost, stolen, or damaged at 0.002. Build a probability distribution table and determine what the expected annual profit for the insurance company will be? Copy and complete the table below onto your paper.

<table>
<thead>
<tr>
<th>Result</th>
<th>Lost, stolen, damaged etc.</th>
<th>No problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value (to insurance company)</td>
<td>-$20,000</td>
<td>$0</td>
</tr>
<tr>
<td>Probability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9) A prospector purchases a parcel of land for $50,000 hoping that it contains significant amounts of natural gas. Based upon other parcels of land in the same area, there is a 20% chance that the land will be highly productive, a 70% chance that it will be moderately productive, and a 10% chance that it will be completely unproductive. If it is determined that the land will be highly productive, the prospector will be able to sell the land for $130,000. If it is determined that the land is moderately productive, the prospector will be able to sell the land for $90,000. However, if the land is determined to be completely unproductive, the prospector will not be able to sell the land. Based upon the idea of expected value, did the prospector make a good investment? A probability distribution table would be a great way to organize and evaluate this information.

10) A woman who is 35 years old purchases a term life insurance policy for an annual premium of $360. Based upon US government statistics, the probability that the woman will survive the year is 0.999057. Find the expected profit this year for the insurance company for this particular policy if it pays $250,000 upon the woman's death. Copy and complete the table below onto your paper.

<table>
<thead>
<tr>
<th>Result</th>
<th>The woman lives</th>
<th>The woman passes away</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value (to insurance company)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11) A bucket contains one gold, three silver, and sixteen red marbles. A player randomly pulls one marble out of this bucket. If they pull a gold marble, they get to pick one bill at random out of a money bag containing one $100 bill, five $20 bills, and fourteen $5 bills. If they pull a silver marble out of the bag, they get to pick one bill at random out of a bag containing one $100 bill, two $20 bills, and seventeen $2 bills. If your marble is red, you automatically lose. The game costs $5 to play.

   a) Build a tree diagram to determine the Sample Space and respective probabilities for this situation.

   b) Build a probability model for this situation.

   c) Calculate the expected gain or loss for the player.
12) A spinner has four colors on it: red, blue, green, and yellow. Half of the spinner is red and the remaining half of the spinner is split evenly among the three other colors. A player pays some money to spin one time. If the spinner stops on red, the player receives $2. If it stops on blue, the player receives $4. If it stops on either green or yellow, the player wins $5. What should this game cost in order to be a fair game?

13) A bag contains one gold, three silver, and six red marbles. A second bag contains one $20 bill, three $10 bills, and six $1 bills. A player pulls out one marble from the first bag. If it is gold, they get to pick two bills from the money bag (without replacement). If it is a silver marble, they get to pick one bill from the money bag, and if the marble is red, they lose. The game costs $3 to play. Should you play? Explain why or why not.

Review Exercises

14) Lulu is designing a carnival game in which players can win $2, $4, $7 or $11. She is trying to decide how much to charge to play. Use the partially complete probability model given below to answer the questions that follow.

<table>
<thead>
<tr>
<th>Lulu’s Carnival Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>

a) What is the probability of winning $2?

b) What is the expected value for this situation?

c) How much should Lulu charge if she wants to make an average of two dollars profit per play?

15) The student council is starting to prepare for prom and decides to name a committee of 6 members. Suppose that they decide the committee will have 2 juniors and 4 seniors on it. In how many ways can the committee be selected if there are 8 juniors and 8 seniors from which to select?

16) Three cards are dealt off the top of a well-shuffled standard deck of cards. What is the probability that all three cards will be the same color? Information about a deck of cards is provided in the appendix.

17) A student does not have enough time to finish a multiple choice test so they must guess on the last two questions. Each question has only choices a, b, and c.

a) How many possible ways might this student answer those last two questions?

b) List the sample space of the possible guesses for the last two questions.
3.3 Simulation and Experimental Probability

Learning Objectives

- Understand how to generate random numbers using a random digit table or a calculator
- Be able to properly assign digits to simulate a random situation
- Be able to interpret results from a simulation and understand the connection to the Law of Large Numbers

For many of the problems we have addressed, putting together a theoretical model is very reasonable for us to do. A theoretical model gives a picture of what should happen in the long run for any situation involving probability. It will give us a very clear idea of what to expect out of a particular situation. For example, if you have been dealt an ace, you can quickly figure out the probability that the next card will be a face card. That probability is a theoretical probability.

However, the truth is that in many situations it is beyond the scope of the mathematics of this course to calculate theoretical probabilities. Also, it is often not practical or even unsafe to carry out an actual experiment to see what will happen. In these situations, we can estimate probabilities by performing a simulation, which is an imitation of a real-world situation. You may have used a driving simulator in driver’s education. This is a safer way to let new drivers try “driving” before they are actually on the road. As another example, researches may want to predict how a disease will spread in order to have enough vaccinations prepared. It would be unethical to set up an experiment to see how quickly it spreads, so they use a probability simulation to make a good estimate without actually putting people in danger. We often do simulations through the use of experimental models. Some of our simulations can be done quite easily using actual probability tools like dice or spinners. Some situations, though, will require us to use a random number generator on a computer or calculator or a table of random digits. Many calculators have a random number generator.

We will first learn how to use a random list of numbers to carry out a simulation. You can find a table of random digits at the end of this book in Appendix A. A table of random digits contains a random mix of digits from 0 through 9. These digits can be used to simulate many situations involving chance behavior from rolling dice and drawing cards, to simulating the spread of a disease. It is important that you are detailed enough in the explanation of how you assign digits and carry out your simulations, so that others may model your simulation procedure exactly. The steps for designing a simulation using a random digit table are explained on the following page.
Steps to include when designing a simulation that uses a random digit table:

- **Step 1:** Assign digits (must be an equal number of digits when using a random digit table). Clearly describe how you will assign digits to each of the different possible outcomes. Be sure that your digit assignment matches the probability of each of the outcomes.

- **Step 2:** Report starting line from the random digit table. Choose a line number (or use the one given in the problem) from the random digit table.

- **Step 3:** State the number of digits you will select at a time. If your largest value is less than or equal to 10, you will be able to select one digit at a time (0 is your tenth digit). If it is less than or equal to 100 you will have to use two digits at a time (00 is your one-hundredth number). If it is less than or equal to 1000, you will have to use 3 digits at a time and so on.

- **Step 4:** Report anything that will be ignored. State what digit combinations you will ignore. These typically are values that are larger than your largest value. Also state whether repeats will be allowed or ignored.

- **Step 5:** State when to stop. Pay attention to the number of trials you must complete.

- **Step 6:** Report what exactly it is that you will be recording. Say something like, “I will record the number of...” Carry out your simulation by selecting your random numbers and recording your results.

- **Step 7:** Report your results. Interpret and summarize your results in context.

**Example 1**
Julio’s mother has informed him that he can have exactly three days every month to “have the day off,” to do whatever he wants to do with his friends. On these three days he will not have to do any chores, watch his younger siblings or even clean up after himself! The catch is that these three days will be randomly picked each month and he cannot trade for other days. Design a simulation that uses a random digit table to pick Julio’s three “days off” for the month of October. Start on line #111 from the random digit table in Appendix A.

**Solution**

- Step 1 – I will let the numbers 01-31 represent each day of the month in October. October has 31 days in it so: 01 = October 1st, 02 = October 2nd, etc.
- Step 2 – I will use line #111 from the random digit table.
- Step 3 – I will look at 2-digits each time.
- Step 4 – I will ignore 32 – 99 and 00, because the largest value is 31. I will also ignore repeats so that I don’t get the same day twice.
- Step 5 – I will carry this out one time.
- Step 6 – I will record which days in October that Julio gets to “have the day off.”

| Line 111 | 81 | 48 | 6 | 16 | 94 | 87 | 60 | 51 | 3 | 0 | 92 | 97 | 00 | 41 | 2 | 7 | 12 | 38 | 27649 |

Notice that we crossed out 81, 48, etc., because these were all beyond our largest value of 31.

**Julio’s three “days off” in October are the 30th, the 27th, and the 12th.**
Example 2
Suppose you wish to simulate rolling two dice a total of 5 times, keeping track of the totals. You don’t have any dice, but you do have access to a table of random digits. Explain how you could simulate the totals from rolling two dice using the random digits and then perform the simulation. Use line #144.

Solution
- Step 1 – I will let the numbers 1-6 represent each side of a die.
- Step 2 – I will use line #144 from the random digit table.
- Step 3 – I will look at 1-digit each time. I will choose two single-digit numbers to represent two dice.
- Step 4 – I will ignore 7-9 and 0, because the largest value is 6. I will not ignore repeats because I can get the same results on other rolls of a die.
- Step 5 – I will carry this out five times.
- Step 6 – I will record the total from my two “dice”.

<table>
<thead>
<tr>
<th>Line 144</th>
<th>6 1 2 9 1 6 4 8 1 1 4 5</th>
<th>8 3 0 8 3 1 6 9 1 4 5 3</th>
<th>4 6 1 0 9</th>
<th>5 9 5 0 5</th>
</tr>
</thead>
</table>

The first roll was 6 + 2 = 8. The second roll was 6 + 4 = 10. The third roll was 1 + 4 = 5. The fourth roll was 5 + 3 = 8. The fifth roll was 3 + 6 = 9.

**Our five results are 8, 10, 5, 8, and 9.**

Remember that it is unwise to make assumptions after only a very small set of rolls. For example, it would be incorrect to say that a total of 7 is unlikely to happen since it did not occur in our simulation. We only simulated this situation 5 times which is not nearly enough to make a conclusion. The Law of Large Numbers states that as we increase the number of trials we should get closer and closer to the theoretical probability. Theoretically, there is a $\frac{1}{6}$ chance that the total is 7. If we did our simulation for thousands of rolls, we would expect that a total of 7 would occur about $\frac{1}{6}$ of the time.

Example 3
Suppose that at the start of this season, several Major League Baseball fans were randomly selected and asked which American League Central team they thought would be most likely to win the division this year. The table below gives the results of the poll.

<table>
<thead>
<tr>
<th>Fans' Opinions: Most Likely to Win AL Central</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>

Using line # 135 from the random digit table, design and carry out a simulation that asks 10 fans who they think will win the AL Central. Clearly describe your process. Carry out your simulation six times. What percent of the simulated fans said that Minnesota would win in each case?
Solution

- Step 1 – I will assign the numbers to each simulated fan’s opinion so that it matches the reported probabilities. See table below.

<table>
<thead>
<tr>
<th>Team</th>
<th>Chicago</th>
<th>Cleveland</th>
<th>Detroit</th>
<th>Kansas City</th>
<th>Minnesota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assigned #s</td>
<td>01 – 06</td>
<td>07 – 18</td>
<td>19 – 33</td>
<td>34 – 80</td>
<td>81 – 99, 00</td>
</tr>
</tbody>
</table>

The probabilities add up to 100%, so we can use two digits every time. Since 6%, or 6 out of 100 fans support Chicago, assign 01-06 as Chicago fans. Assign 07-18 as fans who think Cleveland will win (this is equal to 12%). After Chicago and Cleveland, we have used up 18% of results. So, Detroit will start at 19 and go through 33, Kansas City will be 34-80, and Minnesota will be 81-99 and 00. Remember that we cannot use 100 in this case because it is three digits and we are only selecting two-digit numbers. The digit combination ‘00’ can be used to represent 100.

- Step 2 – I will use line #135 from the random digit table.
- Step 3 – I will look at 2-digits each time. I will look at 10 2-digit numbers per trial.
- Step 4 – I will not need to ignore anything because I need all one-hundred numbers and each 2-digit number represents the probability of someone having an opinion and repeats are allowed.
- Step 5 – I will simulate asking ten fans and I will carry this out six times.
- Step 6 – I will record the number of fans who said that Minnesota would win. I am looking for 2-digit numbers from 81-99 and 00.

| Line 135 | 66|92|5 | 5|15|6|58|1 | 39|10|0 | 7|84|58|1 | 11|20|6 | 1|98|76|1 | 87|51|1 | 3|12|60|1 |
| Line 136 | 08|42|1 | 4|47|53|1 | 77|37|1 | 7|87|44|1 | 75|59|1 | 2|85|63|1 | 79|14|1 | 9|24|54 |
| Line 137 | 53|64|1 | 5|68|1 | 12|42|1 | 4|78|36|1 | 12|60|9 | 1|53|73|1 | 98|48|1 | 1|45|92 |

*When you get to the end of a line in a random digit table, continue on the next line below.

In the first trial, I found two fans who said that Minnesota would win. In the second trial I found two fans who said that Minnesota would win. In the third trial, one fan said that Minnesota would win. In the fourth trial, one fan said Minnesota would win. In the fifth trial, no fans said that Minnesota would win. In the sixth trial, three fans said that Minnesota would win.

The results of our six trials were: 20%, 20%, 10%, 10%, 0%, and 30% saying that Minnesota would win. None of these is 15%, but in this case the average of the six trials is equal to 15%.

Once again, if we had done hundreds of trials instead of just six trials of 10 fans, our percentages would tend to get very close to the theoretical probability according to the Law of Large Numbers.
Example 4

Every person is born on a different day of the month. Some people are born on the 1st of a given month and some people are born as late as the 31st. How many people must you go through until you find two who were born on the same day of the month? Simulate this one time using the random digits below. (Ignore the fact that people are not equally likely to be born on all days. For instance, it is more likely you were born on the 17th than the 31st since all months have a 17th but not all months have a 31st.)

45467  71709  77558  00095  32863  29485  82226  90056
52711  38889  93074  60227  40011  85848  48767  52273

Solution

We will select 2 digits at a time as our largest value, 31, requires two digits. We will use 01, 02, ... 30, 31 and ignore 32-99 and 00.

The numbers we get are 45, 46, 77, 17, 09, 77, 55, 80, 00, 95, 32, 86, 32, 94, 85, 82, 22, 69, 00, 56, 52, 71, 13, 88, 89, 93, 07, 46, 02, 27, 40, 01, 18, 58, 48, 48, 76, 75, 22, and 73. The only ‘keepers’ are 17, 09, 22, 13, 07, 02, 27, 01, 18, and 22. We did not get a match until we got our second 22. It took us 10 people to find a pair that were born on the same day of the month.
Problem Set 3.3

Exercises

1) Suppose that 80% of a school’s student population is in favor of eliminating final exams. Design and carry out a simulation for asking 20 students if they would like to eliminate final exams. Describe your process clearly and record how many of the 20 students asked are in support of eliminating final exams. Use line 147 from the random digit table in Appendix A.

2) Suppose that the students at Best Ever College are asked about their class rank when they were in high school. The table below shows what they said.

<table>
<thead>
<tr>
<th>Class Rank</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10%</td>
<td>0.2</td>
</tr>
<tr>
<td>Top 10% to 25%</td>
<td>0.4</td>
</tr>
<tr>
<td>Top 25% to 50%</td>
<td>0.3</td>
</tr>
<tr>
<td>Bottom 50%</td>
<td>???</td>
</tr>
</tbody>
</table>

a) What must the probability be for the bottom 50%?

b) How will you assign the digits to carry out a simulation for this situation?

c) Design and carry out a simulation for asking twenty students from BEC what their class rank was in high school. Explain your process clearly and record your results. Use line 103 from the random digit table in Appendix A.

3) Charlotte is having a great basketball season. This year she has a 74% free-throw average. Her coach wants to estimate the probability that she will make at least two of any three free-throws that she attempts. Assume that each free-throw that Charlotte attempts is independent.

   a) Design and carry out a simulation for Charlotte attempting three free-throws. Starting on line 129 in the random digit table, run 20 trials and record the number she makes out of each three attempts.

   b) Based on your results, what is the experimental probability that Charlotte will make at least two of three free-throws in a row?

   c) Based on your results, what is the experimental probability that Charlotte misses at least one of the three free-throws that she attempts?
4) Suppose the grades for students in your Probability & Statistics course were distributed as shown in the table below.

<table>
<thead>
<tr>
<th>Probability &amp; Statistics Grades</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>

a) Explain how you could assign digits to simulate the grades of randomly chosen students.

b) Design and carry out a simulation for the grades of 30 randomly selected Probability & Statistics students. Use line 106 from the random digit table. Build a tally chart to track your results.

c) How closely did your simulation match the actual distribution? Explain.

5) How many five card poker hands must you be dealt in order to get a hand with exactly one pair? A pair is two cards that have matching values. (For example, the Jack of hearts and Jack of diamonds have matching values.)

a) Explain how you will assign digits for this situation. Keep in mind that all 52 cards are different from each other.

b) Perform the simulation one time and state how many five-card hands it took for you to get your first hand with two cards that match. Start with a fresh deck for each deal. Begin on line 138 from the random digit table.

6) There are some basic concepts that should be clearly understood about a random digit table. Answer the questions below.

a) Is it possible to have four 6’s next to each other in a random digit table?

b) Approximately what percent of the digits in a random digit table will be 9’s?

c) What should you do if you come to the end of a line of random digits and you still need more numbers?

7) Suppose we have a class of 30 students and you are wondering what the chances are that there is at least one pair of students who have the same birthday. Assume that there are 365 days in a year.

a) Explain how could you assign digits from a random digit table to simulate this situation?

b) Perform this simulation one time and record whether or not there was a match in our class of 30 students. Use line 121 from the random digit table.
Review Exercises

8) Suppose you are dealt two cards from a well-shuffled standard deck of 52 cards. What is the probability that your two cards are a king and an ace (in either order)?

9) Consider a set of 15 pool balls. Pool balls numbered 1 - 8 are solid and pool balls numbered 9 - 15 are striped.
   a) You pull two pool balls randomly out of a bag without replacement. What is the probability that your second pool ball will be solid if your first pool ball had an even number?
   b) Suppose you pull one pool ball out of the bag. What is P(Even | Striped)?

10) Ninety-seven randomly selected junior boys were asked which sports they planned to play next year. Use the Venn Diagram below to answer the questions.
   a) How many of these junior boys are planning on playing hockey or football?
   b) What is the probability that a randomly selected student from this sample plans to play neither hockey nor football next year?
   c) What is the probability that a randomly selected student from this sample plans to play hockey, given that he intends to play football?

![Venn Diagram]

- **Football**: 36
- **Hockey**: 8
- **Intersection**: 11
### 3.4 Chapter 3 Review

In this chapter we learned that the expected value gives us the average result over the long term. In order to calculate the expected value of an event, we learned to use probability distribution tables and the expected value formula:

\[
EV = (\text{Value 1}) \cdot P(\text{Value 1}) + (\text{Value 2}) \cdot P(\text{Value 2}) + (\text{Value 3}) \cdot P(\text{Value 3}) + \cdots
\]

We saw how we can use our probability calculations such as tree diagrams and counting methods to determine probabilities. We learned how probability distribution tables are a useful way to organize information in order to do a full analysis of many situations that involve probability. For example, casinos are cognizant of what the expected value is on any of their games and are confident, despite having to occasionally give away some substantial prizes, that their games will make them money in the long run. We realize that we can never predict with absolute certainty what is going to happen in a given situation, but we can always run a simulation to approximate what is likely to happen. We will often use a random number generator or a table of random digits to help us run a simulation.

#### Review Exercises

1) Ten red marbles and 15 blue marbles are in a bag. A game is played by first paying $4 and then picking two marbles out of the bag without replacement. If both marbles are red you win $10. If both marbles are blue, you win $5. If the marbles don’t match, you are paid nothing.

   a) Build a probability distribution table for this game.

   b) What is the expectation for a single play of this game?

   c) Analyze this game and determine whether or not it is to your advantage to play. Explain.

2) The lacrosse booster club is running a raffle to raise money for new uniforms. There will be one prize of $500, two prizes of $100, five prizes of $50, and ten prizes of $20. They will be selling one-thousand tickets for $5 each. Assume that all tickets will be sold.

   a) Build a probability distribution table for this raffle.

   b) What is the average prize for a single ticket purchased for this raffle?

   c) What is the most likely prize for a single ticket purchased for this raffle?

3) Suppose there are 38 kids in your physics class. Your teacher Ms. Smartypants has decided to randomly select 4 students each day to do a problem on the board. Design and carry out a simulation of her doing this for one week. Start on line 137 from the random digit table.

   a) Clearly describe the method for your simulation.

   b) Carry out your simulation and report the numbers of the four students who are selected each day for the entire week.
4) When two dice are rolled, you can get a total of anything from 2 to 12.
   a) Use the table of random digits in Appendix A to simulate rolling two dice 36 times. Begin on line 119. Make a chart displaying the different results that you get and how many times you get each result.
   b) How close was your simulation to the theoretical probability of what should happen in 36 rolls? Explain.

5) A bag contains one $100 bill and two $20 bills. A person plays a game in which a coin is flipped one time. If it is heads, then the player gets to pick two bills out of the bag. If it is tails, the player only gets to pick one bill out of the bag.
   a) Construct a tree diagram to find the Sample Space and the respective probabilities.
   b) Build a probability distribution table for this game.
   c) How much should this game cost to play if it is to be a fair game?

6) A spinner with three equally sized spaces on it are labeled 1, 2, and 3. A bag contains a $1 bill, a $5 bill, and a $10 bill. A player gets paid the amount they pull out of the bag times the number that they spin. What should this game cost in order to be a fair game?

7) The table below shows the probabilities for how kids get to school in the morning.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bus</th>
<th>Walk</th>
<th>Car</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.31</td>
<td>0.14</td>
<td>0.39</td>
<td>???</td>
</tr>
</tbody>
</table>

   a) What must the Other category have as a probability?
   b) Describe how you would assign digits from a random digit table to run a simulation for this situation.
   c) Design and carry out a simulation for randomly selecting 10 students and asking how they get to school in the morning. Run three trials and record your results. Use line 104 from the random digit table.
8) In an archery competition, competitors shoot at a total of 20 targets. The table below shows
the probabilities associated with hitting the center of the target a given number of times (out of
twenty shots). Some shooters are perfect and hit the center of all 20 targets. However, the
poorest shooters still hit the center of at least 15 targets.

   a) What is the most likely number of
      centers that a shooter will hit?

   b) What is the expected number of centers
      that a shooter will hit?

<table>
<thead>
<tr>
<th># of Centers</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.04</td>
<td>0.12</td>
<td>0.35</td>
<td>0.28</td>
<td>0.18</td>
<td>0.03</td>
</tr>
</tbody>
</table>

9) In a game of chance, players pick one card from a well-shuffled deck of 52 cards. If the card is red,
they get paid $2. If the card is a spade they get paid $3. If the card is a face card, they get paid $5
and if the card is an ace they get paid $10. A player gets paid for all the categories they meet. For
example, the King of Spades would be worth $8 because it is a spade and a face card. Build a
probability distribution table for this game of chance. How much should this game cost in order to
be a fair game?

Image References
- Slot Machine  http://www.gamedev.net
- Duck Pond  http://scoutermom.com
- Quarter  http://www.marshu.com
- $10 Bill  http://wingedliberation.tumblr.com
- Welcome to Las Vegas  http://pilipon.wordpress.com
- Race Cars  http://www.thunderboltgames.com
- Electronic Devices  http://www.topnews.in
- Poker Chips  http://www.ppppoker.com
- Minnesota State Lottery  http://www.mnlottery.com
- $100 Bills  http://www.sciencebuzz.org
- MN Twins Logo  www.twins.mlb.com
- Final Exams Yes  http://www.york.org
- Pair of Jacks  http://xdeal.com
- Dice  http://goblin-stock.deviantart.com
- Targets  http://www.theasbc.org
Chapter 4 – Data Collection

Introduction

What is data? Why collect data? How is data collected? Who cares anyway?

- How many walleye are in Lake Mille Lacs?
- Does aspirin prevent heart attacks?
- What is the approval rating for the President?
- How have the schools in Minnesota been doing to prepare students for success in college?
- What percent of people would return money when given too much change?

All of these questions (and infinitely many more) can be answered through statistics. Statisticians begin by posing a question. They then plan a method for collecting information, called data, about that question. Next they collect the data and analyze it. The statisticians will ‘look’ at the data in the form of graphs or tables. They then ‘analyze’ the data with numerical statistics. Finally they will ‘explain’ what they have learned, what conclusions can be made, and what is still unknown, in a written or verbal report.

4.1 DATA

Learning Objectives

- Know the terminology of data collection, variables, and measurement
- Understand how measurements are used in statistics
- Distinguish between the various methods for data collection

Data and Variables

When a topic needs to be studied or a question needs to be answered, researchers often collect data in an effort to find the answer. Data is a collection of facts, measurements, or observations about a set of individuals (data is plural; the word datum refers to a single observation). There are a variety of ways to collect data in order to study topics of interest. Researchers can analyze and compare test scores for various Minnesota High Schools. Scientists can conduct an experiment to determine the effectiveness of a new medication. Union leaders can conduct a census of every union member before deciding to strike. Market researchers can survey a randomly selected sample of teenage girls to determine what qualities they look for when purchasing a new cell phone.

When a topic is being studied, there are often several variables, or characteristics about the individuals, that the researchers are interested in. Each person, animal, or object being studies is one individual (or subject). The variables are generally either categorical or numerical. A categorical variable (or qualitative variable) can be put into categories, like favorite color, type of car, etc. A numerical variable (or quantitative variable) can be assigned a numerical value, such as height, distance, temperature, etc.
Example 1

Suppose 1,845 teenage girls are to be surveyed by a cell phone company that wants to design a new cell phone that they can market to females under 20 years old. The questionnaire will likely include questions related to age, birth date, race, area code where they live, number of texts sent per month, data usage, amount of money willing to spend per month, services they want offered, features they want included, length of time they have had a cell phone, favorite colors, etc. All of these are variables, because they will vary from individual to individual. However, only some of these variables are numerical. Identify the individuals and the numerical variables.

Solution

Individuals: each girl who completes a questionnaire

Numerical variables: age, number of texts per month, data usage, amount of money willing to spend per month, and length of time they have had a cell phone.

When determining which variables are numerical, it might help to decide whether or not it would be appropriate to calculate a numerical statistic, such as an average or the range for the reported data. Age is numerical, because we can certainly report an average age of those surveyed. Even though birth date and area code may be reported as numbers, it would make no sense to report an ‘average birth date’ or ‘mean area code’. Numbers such as these, social security numbers, or student ID numbers, divide the data into a bunch of categories of one item each. They are simply used for identification and are not considered numerical variables.

Measurement in Statistics

When a topic is to be studied the researchers decide what it is they want to know about each individual. These variables of interest can be measured using different instruments and need to be reported in specific units. The instrument is the tool used to make the measurement. This instrument could be something obvious like a scale, tape measure, thermometer, or speedometer. But, it could also be something like a questionnaire, a rubric, or an exam. The units explain what the numbers represent, and might be feet, points, pounds, degrees Celsius, miles per hour, etc. Keep in mind that data is useless unless it is in context. For example, the number 12 could mean anything. Is it $12, or 12 inches, or 12 as in thousands of dollars, or 12 apple pies? Without knowing the units, all you have is a meaningless list of numbers.
Validity, Reliability and Bias

The way in which any given variable is to be measured needs to be valid and reliable. Validity refers to the appropriateness of the instrument and units used. Reliability means that the instrument can be depended upon to consistently give the same results or nearly the same results when repeated. If an instrument gives different results when measuring the same thing, it is not reliable, and it has a high degree of variability because the results vary a lot. Another potential problem with measurements is bias. When a measurement is repeatedly too high or too low, it is said to be biased. In other words, a biased measurement is ‘consistently wrong in the same direction’.

Researchers would like to limit bias in measurements as much as possible. Ideally, we hope for measurements that are valid, low in bias, and highly reliable. No measurement is perfect. Averaging repeated measurements can be a way to limit variability. Be aware though, averaging will only reduce variability (or increase reliability). Averages will not make an invalid measurement suddenly valid. The average of biased measurements will still be biased.

For example, if the variable being studied is the weight of all of the members of the school wrestling team, then using a scale as the instrument and pounds as the units will be valid. As long as the scale being used is in working order, the measurements reported should be reliable.

However, what if someone had set the scale being used to weigh the wrestlers to start at 10 pounds rather than zero? Each person who stepped on the scale would think that they were 10 pounds heavier than they actually were, resulting in biased measurements. If that were the case, using the scale as the instrument and pounds as the units would still be valid because it makes sense as a way to measure weight. It would also be reliable because if the same person steps on the scale again and again, they will have nearly the same result. However, it would be biased because each measurement is 10 pounds too heavy. So, even though something is wrong with this measurement, it doesn’t mean that everything is wrong with it. We want valid, reliable, and unbiased measurements.

Example 2

Suppose that a teacher intends to base grades in a math class on the students’ heights. She plans to use a tape measure as her instrument and inches as her units. Her grading system will be as follows: the shortest student will receive the lowest grades and the tallest will receive the highest grades. Comment on the validity, reliability, and potential bias of this.

Solution

Validity? This clearly is not a valid way to measure a student’s success and assign grades, because height has absolutely nothing to do with someone’s understanding of math, or their grade in a math course.

Reliability? The tape measure should be reliable. If used properly, each time a particular student’s height is measured we will expect to get the same answer.
Bias? This should not be biased. Some tall people will deserve lower grades, while some will
deserve higher grades. The same will be true for students of all heights.

Therefore, this teacher’s method for assigning grades would be unbiased and it would be
reliable (both good things), but it would also not be valid (a bad thing). She should come up with
a better way to measure students’ grades. Perhaps she should use a combination of test scores
and homework completion.

In conclusion, keep in mind that just because a statistical measurement is bad, does not mean
that everything will be wrong with it. It is important to think through each question separately:
Is the measurement valid?; Is the measurement reliable?; Is the measurement unbiased?

Rates versus Counts

Something to watch out for is whether numbers should be changed to rates or percentages in order to
make appropriate comparisons. For example, it would not make any sense to compare ‘the number of
people living in poverty’ for each of the fifty states in the United States because of the variety in
population sizes. Think of the number of people who live in the state of Rhode Island versus the number
who live in California. It would be much more appropriate to compare ‘the percentage of people living in
poverty’ for each state instead.

Example 3

Luigi got a pair of jeans that are normally $64.95, for $52.50. Javier paid $48.75 for a pair of jeans that
normally cost $58.25. Which jeans had a higher rate of discount?

Solution

Luigi’s jeans were marked down $12.45 (64.95 - 52.50). Divide the amount of discount by the
original cost (12.45/64.95) and get 0.1917. So, Luigi’s jeans were marked down 19.17%.

Javier’s jeans were marked down $9.50 (58.25 - 48.75). Divide the amount of discount by the
original cost (9.50/58.25) and get 0.1631. So, Javier’s jeans were marked down 16.31%.

Luigi’s jeans had a higher rate of discount.

Methods for Collecting Data

Once a question of interest is posed, there are different ways of collecting data. This is a quick overview
of the methods for collecting data that will be studied in this chapter: sample surveys, census,
observational studies, and experiments. Each will be covered in more detail in the following sections.
For now, we just want to be able to recognize which method was used or described.
Sample surveys are often used as a way to collect data from just some of the people or objects being studied. Some examples of sample surveys are: mailed out questionnaires, online surveys, phone interviews, or quality control checks. Another way to collect data is through a census. This means that every single person or item in the population is checked, tested, or asked. When trying to determine whether something was a sample or a census, ask yourself if the researchers asked everyone (or tested everything). If yes, then it was a census.

Sometimes it will be most appropriate to conduct an experiment which occur when the researchers actually ‘do something’ to the subjects. Observational studies are another common way to collect data. In observational studies, the researchers do not ‘do anything’ to the subjects, they simply collect data that has already happened or happens naturally. A sample survey and a census are actually types of observational studies. All of these methods of data collection can yield interesting results and often answer questions. However, the only method that can actually prove that one variable causes another is an experiment. When trying to determine whether a research method was an experiment, ask yourself if the researchers changed anything or did anything to the people or objects that were being studied. If yes, then it was an experiment.

Example 4

For each of the following scenarios, determine whether the situation described is an experiment, observational study, census, or a sample survey. Explain how you know.

a) Researchers suspected that aspirin could help reduce the risk of having a heart attack. Seven hundred people, aged 40 or older, were willing to participate in a study. Half of these participants were randomly selected to take an aspirin each day. The remaining participants were given a pill that looked like the aspirin, but contained no actual medicine. The study went on for five years and the participant’s health was monitored.

b) In an effort to study how the high schools in Minnesota have been preparing students for college, an extensive questionnaire was developed. Ten percent of the high school juniors at every high school in the state were selected randomly to complete this questionnaire.

c) Researchers suspected that tanning beds caused skin cancer. Each time a person was diagnosed with skin cancer, they were asked a series of questions including whether or not they had used a tanning bed. If they had, further questions were asked regarding how often, what type, and at what age, etc.

d) In an effort to determine how many fish were in Lake George, the lake was drained and the fish were counted.

Solution

a) This is an experiment because the researchers changed something. They had the people take aspirin (or fake aspirin).

b) This is a sample survey because only a part of all high school students were questioned.

c) This is an observational study because no change was made. The researchers simply asked about past behavior. Additionally, this could also be viewed as a census because all skin cancer patients were asked about tanning booth use.

d) This is a census because every fish was counted. Let’s hope they can find a better way to determine how many fish are in a lake next time!
Problem Set 4.1

Exercises

1) Lucas is writing an article about the baseball teams for the school paper. He collects data about each player’s position, batting average, number of at-bats, hits, stolen bases and whether each player is on the junior varsity or varsity team. Who are the individuals? Which variables are categorical? Which are numerical?

2) Malia has been put in charge of analyzing the employees at her company. She collects information regarding annual salary, years with the company, highest degree earned, job title, yearly contribution toward 401K, number of children, home address and phone number. Who are the individuals? Which variables are categorical? Which are numerical?

3) Determine whether each of the following variables is categorical, numerical, or neither.
   a) The heights of all of the volleyball players.
   b) The position played by all of the football players.
   c) The brand of mascara preferred by those surveyed.
   d) The numbers of texts sent per month.
   e) Each person’s social security number.
   f) Each person’s cell phone provider.

4) The fourth graders at Sand Creek Elementary are doing a unit on weather. There is a thermometer on the building just outside the classroom window. The students will record and analyze the temperature at 8:00 a.m. and 2:00 p.m. every school day for 5 weeks, create a graph of data, and then write a report based on their findings.
   a) Identify the variable of interest, the instrument used, and the units.
   b) Comment on the validity, reliability, and potential bias for this study.

5) The first graders at Sand Creek Elementary are doing a unit on measurement. Each student has traced her or his own foot and cut it out. Each student will use his or her ‘foot’ to measure various objects around the room and school. Some of the measurements they will make are height of themselves and at least two other friends, width of the classroom door, and length of a lunch table.
   a) The variables of interest are the lengths, widths and heights of various objects. Identify the instrument used, and the units.
   b) Comment on the validity, reliability, and potential bias for this study.
6) Determine whether each of the measurements described below would have a problem with validity, reliability, or bias. A measurement may have a problem in more than one area. For each problem you identify, suggest a better way to make the measurement. Your answers should be similar to those in Example #2.
   a) Johnny is driving a car with a speedometer that is totally unpredictable.
   b) Cholesterol levels are determined by each patient filling out a survey regarding their diet.
   c) Time is measured by using the clock on a cell phone.
   d) Grades in a Physics class are determined by students assessing themselves on a scale of 1 to 10.
   e) Grades in a statistics class are determined by students’ scores on one cumulative test.
   f) Sobriety is determined by a breathalyzer that is calibrated to be too sensitive.

7) Super Duper High School has a total of 143 teachers. Suppose that you are a researcher who is interested in studying Teacher Effectiveness at SDHS. You intend to evaluate the effectiveness of all of the teachers for your report.
   a) What type of data collection method is this?
   b) Suggest at least two valid variables that you might study. Include an instrument that could be used to measure your variables and state the units.
   c) Suggest at least two invalid variables that you might study. Include an instrument that could be used to measure your variables and state the units.

8) For each of the following scenarios, determine whether the situation described is an experiment, observational study, census, or a sample survey. Explain how you know.
   a) The Super Spaz Energy drink company randomly selects 2% of the cans filled each day and tests them for volume, ingredient content, and taste.
   b) A government lobbyist analyzes the crime reports for the 4 counties in her community.
   c) New advertisements are generally tried out on focus groups before investing a lot of money to pay for airtime on national TV.
   d) Each student in Probability and Statistics will take the District Common Assessment as a final exam.
   e) A teenager decides to evaluate how serious her parents are about her curfew by coming home 15 minutes late just to see what happens.

9) Pasquale’s Big and Tall Shop sold 127 suits during the first quarter of this year, and 17 were returned. Marco’s XXL Shop sold 268 suits during the same time period, and 27 were returned.
   a) What were the number of returns for each shop? Which shop had a higher number of returns?
   b) What were the rates of returns for each shop? Which shop had a higher rate of returns?
   c) Which of these statistics gives a more clear representation of customer satisfaction? Explain.
10) Jolene makes $12.45 per hour at her job. Last year she made $10.85. What percent of a raise did Jolene receive?

11) Michaela’s favorite shoes are normally $42.99. Today she found a sale in which they were marked down to $27.99. What percent discount is this?

12) The number of incidents of hazing reported at Some Random High School was 84 during the 2015-2016 school year. The following year there were 37 incidents of hazing reported at SRHS. What is the rate of change in reported hazing incidents between these two school years? Is it an increase or a decrease? (Assume that the population size of the school remained the same.)

13) SRHS has had a huge problem getting students to class on time, so the administrators have implemented a new tardy policy. In an effort to determine whether or not it is working to deter students from being tardy to class, data has been collected and analyzed. The following table shows some of the data:

<table>
<thead>
<tr>
<th>School Year</th>
<th>2015-2016</th>
<th>2016-2017</th>
<th>% of change (+ or -)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of tardies</td>
<td>5186</td>
<td>4295</td>
<td></td>
</tr>
<tr>
<td>Number of students with more than 10 tardies</td>
<td>175</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Number of students with more than 20 tardies</td>
<td>112</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

a) Calculate the percent of change for each category and complete the table (round to the nearest tenth of a percent).

b) Which category saw the most significant change?

c) Based on these calculations, do you feel that the tardy policy is working? Explain your reasoning.

Review Exercises

14) Three 6-sided dice are rolled.

a) What is the probability that the total is 4 or less?

b) What is the probability that the total is greater than 4?

15) How many three and four digit numbers are possible using the digits 3, 4, 5, and 6 if the number must be greater than 500 but less than 5000 and no digit may be repeated?

16) How many three and four digit numbers are possible using the digits 3, 4, 5, and 6 if the number must be greater than 500 but less than 5000 and digits may be repeated?

17) A game has prizes of $10, $5, and $2. Suppose the chance of winning both the $2 prize and $5 prize is 40% while the chance of winning the $10 prize is 20%. What should this game cost to play if it is to be a fair game?
4.2 Sample Survey and Census

Learning Objectives

- Differentiate between population and sample
- Understand the terminology of sampling methods
- Identify various sampling methods
- Recognize and name sources of bias or errors in sampling

Population vs. Sample

What is the approval rate of the President? If we really wanted to know the true approval rating of the president, we would have to ask every single adult in the United States her or his opinion. If a researcher wants to know the exact answer in regard to some question about a population, they would have to conduct a census. In a census, every individual in the population being studied is measured or surveyed. In this example our population, the entire group of individuals that we are interested in, is every adult in the United States of America.

A census like this (asking the opinion of every single adult in the United States) would be impractical, if not impossible. First, it would be extremely expensive for the polling organization. They would need a large workforce to try and collect the opinions of every single adult in the United States. Once the data is collected, it would take many workers many hours to organize, interpret, and display this information. There are other practical problems that might arise. Some adults may be difficult to locate, some may refuse to answer the questions or not answer truthfully, some people may turn 18 before the results are published, others may pass away before the results are published, or an event may happen that changes peoples’ opinions drastically. Even if this all could be done within several months, it is highly likely that peoples’ opinions will have changed. By the time the results are published, they may already be obsolete.
Another reason why a census is not always practical is because a census has the potential to be destructive to the population being studied. For example, it would not be a good idea for a biologist to find the number of fish in a lake by draining the lake and counting them all. Also, many manufacturing companies test their products for quality control. A padlock manufacturer, for example, might use a machine to see how much force it can apply to the lock before it breaks. If they did this with every lock, they would have none to sell. In both of these examples it would make much more sense to simply test or check a sample of the fish or locks. The researchers hope that the sample that they select represents the entire population of fish or locks.

This is why sampling is often used. **Sampling** refers to asking, testing, or checking a smaller sub-group of the population. A **sample** is a representative subset of the population, whereas the population is every single member of the group of interest. The purpose of a sample is to be able to generalize the findings to the entire population of interest. Rather than do an entire census, samples are generally more practical. Samples can be more convenient, efficient, and cost less in money, labor and time.

A number that describes a sample is a **statistic**, while a number that describes an entire population is a **parameter**. Researchers are trying to approximate parameters based on statistics that they calculate from the data that they have collected from samples. However, results from samples cannot always be trusted.

**Example 1**

A poll was done to determine how much time the students at SDHS spend getting ready for school each morning. One question asked, “Do you spend more than 20 minutes styling your hair for school each morning?” Of the 263 students surveyed, 61 said that they spend more than 20 minutes styling their hair before school. Identify the population, the parameter, the sample, and the statistic for this specific question.

**Solution**

- Population (of interest): **All students at SDHS**
- Parameter (of interest): **The true proportion of students that spend more than 20 minutes styling their hair for school each morning.**
- Sample: **The 263 SDHS students who were surveyed**
- Statistic:  \[ \hat{p} = \frac{61}{263} = 0.2319 = 23.19\% \]
Randomization

One common problem in sampling is that the sample chosen may not be representative of the entire population. In such cases, the statistics found from these samples will not accurately approximate the parameters that the researchers are seeking. Samples that do not represent the population are biased. If someone was interested in the average height of all male students at his or her high school, but somehow the sample of students measured included the majority of the varsity basketball team, the results would certainly be biased. In other words, the statistics that were calculated would most certainly overestimate the average height of male students at the school. Samples should be selected randomly in order to limit bias. Also, if only three students’ heights are measured, it is very possible that the average height of these three will not be close to the average height of all of the male students. The average of the heights of 40 randomly chosen male students would be more likely to result in a number that will match the average of the entire population than those of just three students. The average of larger sample sizes will have less variability, so small sample sizes should be avoided if possible. In other words, to decrease the amount of variability, one should increase the sample size.

There are many ways to select a random sample. The way many raffles are conducted is that every ticket is put into a hat (or box), then they are shaken or stirred up, and finally someone reaches into the hat without looking and selects the winning ticket(s). Flipping a coin to decide which group someone belongs in is another way to choose randomly. Computers and calculators can be used to make random selections as well. The purpose of choosing randomly is to avoid any personal bias from influencing the selection process. Randomization will limit bias by mixing up any other factors that might be present.

Think of the heights of those male students, if we assigned every male at that school a number and then had a computer program select 40 numbers at random, it is most likely that we would end up with a mixture of students of various heights (rather than a bunch of basketball players). A random selection will prevent someone from just measuring their friends’ heights, the first 40 males they see staying after school, or everyone in first lunch who is willing to participate. A computer program has no personal stake in the outcome and is not limited by its comfort level or laziness.

If the goal of our sample is to truly estimate the population parameter, then some planning should be done as to how the sample will be selected. First of all, the list of the population should actually include every member of the population. This list of the population is called the sampling frame. For example, if the population is supposed to be all adults in a given city and someone is working from the phone book to make selections, then everyone who is unlisted and those who do not have a land line telephone will not have any chance of being selected. Therefore, this is not an accurate sampling frame.

Good Sampling Methods

Simple Random Sample

When the selection of which individuals to sample is made randomly from one big list, it is called a simple random sample (or SRS). An example of this would be if a teacher put every single student’s name in a hat and then drew 5 names from the hat, without looking, to receive a piece of candy. In an SRS every single member of the population has an equal probability of being selected - every student has an equal chance of getting the candy. In an SRS, every combination of individuals also has an equal chance of being selected - any group of 5 students might end up getting candy. It might be all 5 girls, it
might be the 5 students who sit in the back row, or it might even end up being the 5 students who misbehave the most. Anything is possible with an SRS!

**Stratified Random Sample**

A simple random sample is not always the best choice though. Suppose you were interested in students’ opinions regarding the homecoming theme, and you wanted to make certain that you heard from students from all four grades. In such a case it would make more sense to have four separate lists (freshmen, sophomores, juniors and seniors), and then to randomly select 50 students from each list to participate in the survey. A selection done in this way is called a stratified random sample. A **stratified random sample** is when the population is divided into deliberate groups called **strata** first, and then individual SRS’s are selected from each of the strata. This is a great method when the researchers want to be sure to include data from specific groups. Divisions may be done by sex, age group, race, geographic location, income level, etc. (This is not true. Suppose there were only 50 freshmen and 500 seniors. The chance of any freshman getting selected would be 100% while only a 10% chance exists for each senior.) With stratified random samples, every member of the population has a chance of being selected, but not every combination of individuals is possible.

**Systematic Random Sample**

Another way to choose a sample is systematically. A **systematic random sample** makes the first selection randomly and then uses some type of ‘system’ to make the remaining selections. A system could be: every 15th customer will be given a survey, or every 30 minutes a quality control test will be run. A systematic random sample might start with a single list like an SRS, randomly choose one person from the list, and then every 25th person after that first person will also be selected. Systematic random samples still give every member of the population an equal chance of being chosen, but do not allow for all combinations of individuals. Some groups are impossible, such as a group including several people who are in order on the list.

**Multi-Stage Random Sample**

When seeking the opinions of a large population, such as all registered voters in the United States, a multi-stage random sample is often employed. A **multi-stage random sample** involves more than one stage of random selection and does not choose individuals until the last step. A pollster might start by randomly choosing 10 states from a list of the 50 states in the U.S.A. Then she might randomly choose 10 counties in each of those states. And, finally she can randomly choose 50 registered voters from each of those counties to interview over the telephone. When she is done, she will have $10 \times 10 \times 50 = 5000$ individuals in her sample. This is another sampling method that gives individuals an equal chance of being chosen, but does not allow for all possible combinations of individuals. For example, there is no possible way that all 5000 of these voters will be from Texas.

**Random Cluster Sample**

Sometimes cluster samples are used to collect data. Splitting the population into representative **clusters**, and then randomly selecting some of those clusters, can be more practical than making only individual selections. In cluster sampling, a census is done on each cluster or group selected. When appropriately used, cluster sampling can be very useful and efficient. One should be careful that the
clusters are in fact selected randomly and that this method is the best choice. When a study of teenagers across the country is to be done, a random cluster method can be the best choice. An SRS of all teens would be nearly impossible. Imagine the reality of having a list that includes all teens! A multi-stage random sample might be theoretically ideal, but the practicality of surveying one teenager from a high school in Little Rock, and one from another high school in Duluth, and so on would be quite a nightmare. The best choice might be to randomly select 10 metropolitan areas, 10 suburban areas, and 10 urban areas from across the country. And then to randomly select one high school in each of these areas and then finally to randomly select 4 second hour classes from each of those high schools. Then survey the entire classes selected (clusters). This would be a combination of multi-stage random selection and cluster sampling. Another use for random cluster sampling is quality control at a popcorn factory. If every hour, a bucket of popcorn is scooped out, the entire bucket of popcorn can be checked for salt content, appearance, number of kernels not popped, number of kernels not burnt, etc. This is an example of a systematic random cluster sample, the system being ‘every hour’ a sample is taken, and the clusters being each bucket of popcorn.

**Bad Sampling Methods**

**Voluntary Response Sample**

Beware of call in surveys, and online surveys! Suppose that a radio hosts on KDWB says something like, "Do you think texting while driving should be legalized? Call in and have your opinion heard!" It is highly likely that many people will call in and vote "Yes!" However, the people who do take the time to call will not represent the entire population of the Twin Cities and so the results cannot possibly be trusted to be equal to what all members of the population think. The ‘statistic’ that this ‘survey’ calculates will be biased. The only people who will take the time to call in are those who feel strongly that texting while driving should be legal (or illegal). Such a sampling method is called a voluntary response sample. In voluntary response samples, participants get to choose whether or not to participate in the survey. Online, text-in, call-in, mail-in, and surveys that are handed out to people with an announcement of where to turn them in when completed, are all examples of voluntary response surveys. Voluntary response samples are almost always biased because they result in no response whatsoever from most people who are invited to complete the survey. As a result, most opinions are never even heard, except for those who have really strong opinions for or against the topic in question. Also, those who have strong opinions can call or text multiple times. A new problem that comes with the Internet is that many companies are offering to pay people to complete surveys, which makes any results questionable. For these reasons, the results of voluntary response samples are always suspect because of the potential for bias.
Convenience Sample

Another commonly used but dangerous method for choosing a sample is to use a convenience sample. A convenience sample just asks those individuals who are easy to ask or who are conveniently located by the pollster. The big problem here is that the sample is unlikely to be representative of the entire population. The fact that this group was convenient implies that they most likely have at least something in common. This will almost always result in biased results. An interviewer at the mall only asks people who shop at the mall, and only at some given time of day. As a result, many people in the community will never have the opportunity to be interviewed. When the population of interest is only mall-shoppers, we will have a result that will be somewhat better than when the population of interest is community members. Even then, the interviewers choose whom to go up to and the interviewees can easily refuse to participate.

With both of these bad sampling methods, the word random is nowhere to be found. That lack of randomness should serve as a big hint that some type of bias will likely be present. The scary thing is that most of the results we see published in the media are the results of convenience samples and voluntary response samples. One should always ask questions about where and how the data was collected before believing the reported statistics.

Example 2

Suppose that a survey is to be conducted at the new Viking’s Stadium. A five question survey is developed. The population of interest is all of the 31,045 fans present that day. The sample size will be 2,500 randomly selected fans. Identify specifically the sampling method that is being proposed in each scenario. Also, comment on any potential problem or bias that will likely occur.

a) The first 2,500 fans to arrive are asked five questions.

b) Fifty sections are randomly selected. Then ten rows are randomly selected from each of those sections. Then five seats are randomly selected from each of those rows. The people in these seats are interviewed in person during the game.

c) A computer program selects 2,500 seat numbers randomly from a list of all seats occupied that day. The people in these seats are interviewed in person during the game.

d) 2,500 seats are randomly selected. The surveys are taped to those 2,500 seats with instructions as to where to return the completed surveys.

e) The number 8 was randomly selected earlier. The 8th person through any gate is asked five questions. Then, every 12th person after that is also asked the same five questions.

f) The seats are divided into 25 sections based on price and view. A computer program randomly selects 100 seats from each of these sections. The people in these seats are interviewed in person during the game.
Solution

a) This is a convenience sample. It will not represent everyone present that day. This will suffer from bias because all of these people have at least one thing in common—they arrived early.

b) This is a multi-stage random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

c) This is a simple random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

d) This is a voluntary response sample. It is very likely that most of those surveys will end up on the ground or in the garbage. This will likely suffer from many people not responding. It is also probable that anyone who had an extremely negative experience will be more likely to complete their surveys.

e) This is a systematic random sample. It will probably represent the entire population. As long as the people are willing to answer the questions honestly, it could be a good plan.

f) This is a stratified random sample. It will probably represent the entire population. As long as the people are in their seats and willing to answer the questions honestly, it could be a good plan.

Errors in Sampling

Sampling Errors

Some errors have to do with the way in which the sample was chosen. The most obvious is that many reports result from a bad sampling method. Convenience samples and voluntary response samples are used often and the results are displayed in the media constantly. We have seen that both of these methods for choosing a sample are prone to bias. Another potential problem is when results are based on too small of a sample. If a statistic reports that 80% of doctors surveyed say something, but only five doctors were even surveyed this does not give us a good idea of what all doctors would say.

Another common mistake in sampling is to leave an entire group (or groups) out of the sample. This is called undercoverage. Suppose a survey is to be conducted at your school to find out what types of music to play at the next school dance. The dance committee develops a quick questionnaire and distributes it to 12 randomly selected 5th period classes. However, what if they did this on a day when the football teams and cheerleaders had all left early to go to an ‘out of town’ game. The results of the dance committee’s survey will suffer from undercoverage, and will therefore not represent the entire population of your school.

There is also the fact that each sample, randomly selected or not, will result in a different group of individuals. Thus, each sample will end up with slightly different statistics. This expected variation is called random sampling error and is fairly small from sample to sample. However, every now and then the sample selected can be a ‘fluke’ and just simply not represent the entire population. A randomly selected sample might accidentally end up with way too many males for example. A survey to determine the average GPA of students at your school might accidentally include mostly honors students. There is no way to avoid random sampling error. This is one reason that many important surveys are repeated with a new sample. The odds of getting such a ‘fluke’ group more than once are very low.
Non Sampling Errors

One of the biggest problems in polling is that most people just don’t want to bother taking the time to respond to a poll of any kind. They hang up on a telephone survey, put a mail-in survey in the recycling bin, or walk quickly past an interviewer on the street. Even when the researchers take the time to use an appropriate and well-planned sampling method, many of the surveys are not completed. This is called non-response and is a source of bias. We just don’t know how much the beliefs and opinions of those who did complete the survey actually reflect those of the general population, and, therefore, almost all surveys could be prone to non-response bias. When determining how much merit to give to the results of a survey, it is also important to look for the response rate \( \frac{\text{number returned}}{\text{total number sent}} \).

The wording of the questions can also be a problem. The way a question is worded can influence the responses of those people being asked. For example, asking a question with only two answer choices forces a person to choose one of them, even if neither choice describes his or her true belief. When you ask people to choose between two options, the order in which you list the choices may also influence their response. Also, it is possible to ask questions in leading ways that influence the responses. A question can be asked in different ways which may appear to be asking the same thing, but actually lead individuals with the same basic opinions to respond differently.

Consider the following two questions about gun control.

“Do you believe that it is reasonable for the government to impose some limits on purchases of certain types of weapons in an effort to reduce gun violence in urban areas?”

“Do you believe that it is reasonable for the government to infringe on an individual’s constitutional right to bear arms?”

The first question will result in a higher rate of agreement because of the wording ‘some limits’ as opposed to ‘infringe’. Also, ‘an effort to reduce gun violence’ rather than ‘infringe on an individual’s constitutional right’ will bring more agreement. Thus, even though the questions are intended to research the same topic, the second question will render a higher rate of people saying that they disagree. Any person who has strong beliefs either for or against government regulation of gun ownership will most likely answer both questions the same way. However, individuals with a more tempered, middle position on the issue might believe in an individual’s right to own a gun under some circumstances, while still feeling that there is a need for regulation. These individuals would most likely answer these two questions differently.

You can see how easy it would be to manipulate the wording of a question to obtain a certain response to a poll question. This type of bias may be done intentionally in an effort to sway the results. But it is not necessarily always a deliberate action. Sometimes a question is poorly worded, confusing, or just plain hard to understand, and this will still lead to non-representative results. Another issue to consider when critiquing the results of a survey is to know who paid for or who is reporting the results. Do the sponsors of this survey have an agenda they are trying to push through?
A major problem with surveys is that you can never be sure that the person is actually responding truthfully. When an individual responds to a survey with an incorrect or untruthful answer, response bias can occur. This can occur when asking questions about extremely sensitive, controversial or personal issues. Some responses are actual lies, but it is also common for people just to not remember correctly. Also, sometimes someone who is completing a survey or answering interview questions will ‘mess with the data’ by lying or making up ridiculous answers.

Response bias is also common when asking people to remember what they watched on TV last week, or how often they ate at a restaurant last month, or anything from the past. Someone may have the best intentions as they complete the questionnaire, but it is very easy to forget what you did last week, last month, or even yesterday. Also, people are often hurrying through survey questions, which can lead to incorrect responses. The results on questions regarding the past should be viewed with caution.

It is difficult to know whether or not response bias is present. We can look at how questions were worded, how they were asked, and who asked them. For example, person-to-person interviews on controversial topics carry a definite potential for response bias. It is sometimes helpful to see the actual questionnaire that the subjects were asked to complete when considering how much weight to give the reported results.

There are sometimes mistakes in calculations or typos present in results. These are called processing errors (or human errors). For example, it is not uncommon for someone to enter a number incorrectly when working with large amounts of data, or to misplace a decimal point. These types of mistakes happen frequently in life, and are not always caught by those responsible for editing. If a reported statistic just doesn’t seem right, then it is a good idea to recheck calculations when possible. Also, if the numbers appear to be ‘too good to be true’, then they just might be!

Example 3

The department of health often studies the use of tobacco among teens. The following is a description by the Minnesota Department of Health describing how they chose the sample for the 2008 Minnesota Youth Tobacco and Asthma Survey. In 2008, they had 2,267 high school students complete surveys and 2,322 middle school students complete surveys. Each student in the sample completed an extensive questionnaire consisting of many questions related to tobacco use. Answer the questions that follow. To see the entire report go to: http://www.health.state.mn.us/divs/hpcd/tpc/

Students were selected for the survey in two stages. First, 48 public middle schools (grades 6-8) and 51 public high schools (grades 9-12) were randomly selected, with probability of selection based on size of enrollment. Alternative schools and charter schools were included. The sample schools were randomly chosen by CDC using enrollment data provided by the Minnesota Department of Health. Next, three or four classrooms within each participating school were randomly selected, and all students in these classrooms were invited to participate. The number of schools and classrooms selected was reduced substantially in 2008 in order to reduce the burden on schools. The sample size is still adequate to provide reasonable statewide estimates.
a) Which types of sampling methods were used for this study?
b) Identify the population, the parameter of interest, the sample, and the individuals for this study.
c) When asking teens about tobacco use, what types or causes of bias will likely be present? What could be done to limit these biases?
d) This graph to the right shows how the percent of teens using tobacco has changed from 2000 to 2008. Identify specifically the statistics that were found for this question in 2008.

Solution

a) This study used a complicated combination of sampling methods. They used a stratified, multi-stage, random cluster sample method to select individuals. It was a stratified random sample (by high school and middle school), it was a multi-stage random sample (first random schools were selected, second random classes were selected), and it was a cluster sample (every student in each class was given the survey).

b) Population (of interest): All middle and high school students in Minnesota
   Parameter (of interest): The true rate of teen tobacco use
   Sample: 2267 high school, and 2322 middle school students in Minnesota from 48 public middle schools and 51 public high schools in Minnesota
   Individuals: each student who completed a survey

c) Response bias: Tobacco use is not legal for people under 18, so teens will not want to tell the truth if they think they may get in trouble.
   Non-response bias: Some people were absent the day survey was given.
   Undercoverage: Only public school students were included, so those who attend private schools were left out.
   Wording of the questions: This could be a problem, but we do not know the exact wording so cannot be sure.
   To avoid the response bias factor, surveys regarding controversial topics should all be anonymous. If you read further into this report, you will see that the students were assured all results would be anonymous (no names or ID numbers included).
   To avoid the non-response bias factor, students who were absent could be given the survey when they returned to school.
   To avoid the undercoverage of private school students, private schools could be included in the sample.

d) In 2008, 27.0% of the high school students asked and 6.9% of the middle school students asked had used tobacco in the last 30 days.
Problem Set 4.2

Exercises

1) For each of the following, determine whether the bold number is a parameter or a statistic. (Hint: Remember that a parameter is a value that represents an entire population and a statistic is a value that represents a sample.)

a) The average height of all oak trees is 42.3 feet.

b) Ms. Anderson’s class average on the final exam was 71.4%.

c) The average number of songs that the students surveyed had on their iPhone was 791 songs.

d) iTunes reports that the average number of songs people have on their iPhone is 503 songs.

e) The sticker on the Super Speedster Sport Sedan says 17.82 mpg.

f) Martin had to keep track of how much time he spent watching TV for a whole week. He found that last week he averaged 3.4 hours of TV per day.

2) Minnesota’s Best High School found that last year they did not have enough seats or room for all of the family members who wished to attend the graduation ceremony. The administrators at MBHS need to decide where to hold the graduation ceremony this year, so they sent a questionnaire home with each of this year’s 543 seniors early in September. They asked for the surveys to be completed and returned by September 27th. Of the 148 surveys returned, the average number of seats that will be needed is 6.2. To be safe, the administrators use 7 and determine that they will need a hall that can hold 3800 people (543 students X 7 seats = 3801 seats needed). Using this number they find an appropriately sized hall and reserve it. Identify each of the following as specifically as possible.

a) Population (of interest)

b) Parameter (of interest)

c) Sample

d) Statistic

e) Sampling method that was used?

f) What is the response rate (the percent of surveys returned)?

g) What is wrong with what these administrators have done? What type of bias or error is likely present?

h) Will the statistic most likely by too high or too low? What is a likely consequence of this biased result?
3) Suppose that a survey is to be conducted at Minnesota’s Best High School. Population of interest: 2640 MBHS students. Sample size: 240 MBHS students. Identify specifically the sampling method that is being proposed in each scenario. Also, comment on any potential problem or bias that will likely occur.

a) Every freshman’s name is put on a slip of paper and put into a giant bucket. Sixty names are pulled out of the hat. This process is repeated for each grade level.

b) A list of all students is obtained from the counselors. Julie randomly selects a number between 1 and 2640 and then finds the student that matches this number on the list. She then selects every eleventh person on the list after that one (cycling back to the beginning of the list) until 240 names are chosen.

c) Surveys are handed out with lunches. Students are asked to complete them and turn them in on a table in the front of the cafeteria.

d) A computer randomly selects 240 names from the entire list of students in the school database.

e) Twelve teachers are randomly selected. Two of each of their classes are then randomly selected. Ten students from each of these classes are then selected.

f) Three teachers, Mr. Niceguy, Mr. Greatguy, and Mr. Happyguy, each volunteer to survey the students in all of their classes.

4) Consider each of the following situations:

a) What is the name of the type of bias in the cartoon?

b) As the 2010 Census was being conducted, many people did not return their forms. What type of bias is this?

c) What type of bias would most likely be present if high school students are interviewed about their drinking and drug use habits? Would the statistics most likely over- or under-estimate the true parameters?

d) What is the type of sampling error that we expect to happen, but cannot do about?

e) When calculating the statistics from a survey, a typo is made. What type of error is this?

f) A radio talk show host asks, “Do you think that the driving age should be changed to 18?” What type of bias will most likely be present? Why is this?

g) A survey is conducted by door-to-door interviews and the interviewers skip a few neighborhoods that ‘make them nervous’. What type of bias is this called?

h) If an interviewer asks each person, “Do you prefer Pizza Ickaroooni, or the delicious fresh flavors of Pizza Delicioso?” what type of bias will likely be present?
Review Exercises

5) One die is rolled. What is the chance that a number greater than four or an even number is showing?

6) One die is rolled. What is the chance that a number greater than four and an even number is showing?

7) Two dice are rolled. What is the probability that the sum of the number of dots showing is nine or greater?

8) If three dice are rolled, what is the probability of getting three of a kind (all 3 dice show the same number of dots)?
4.3 Random Selection

Learning Objectives

- Obtain a random sample using a random digit table
- Describe the process followed to obtain an SRS
- Outline an appropriate sampling method

Random Selection

We have discussed that it is important to choose samples randomly in order to reduce bias, but we haven’t discussed how to actually carry out the process. There are many ways to make random selections. A common way to choose things at random is to use a ‘big hat’, or box, or bowl, etc. For example, suppose that a teacher wanted to randomly select 5 students every day, from a class of 34 students, to hand in their homework to be graded. Each day she has all of the students’ names in a big fish bowl. She will mix the names up and select 5 names. These students will turn their homework papers in right then, and the other students will not need to turn in their homework. The five selected names will be put back in the fishbowl so they may be selected again tomorrow. This is an example of an SRS of size 5 of this class. Every student has an equal probability (5/34 or 14.7% chance) of being required to turn in his or her homework on any given day and any combination of five students may be chosen. One student may end up turning in her assignment several days in a row, while another student may never need to turn hers in all year long. The idea of a ‘big hat’ is a good method for random selection when working with small populations, but it is not always practical.

Random selections can be made by flipping coins, rolling dice, or spinning a spinner. These days, most random selections can be done using technology such as a computer program or a random number generator on a graphing calculator. Another way that random selections are made in statistics is by using a random digit table. A random digit table is a long list of randomly generated digits from 0 to 9. The digits are listed in groups of five simply to make it easier to read and not lose your place. Imagine that someone has a ten-sided die with each digit from 0 to 9 marked on a side. They sit down, roll the die and write down the digit that appears, then they roll it again and write down the next digit that appears, then they do this again and again. As you can imagine, this would take quite a while, but would result in a long list of random digits. This is basically what a random digit table is. There is a random digit table in Appendix A for you to use.

How to Use a Table of Random Digits

There is a process to follow when using a random digit table to make your selection. You need to report your process with enough detail so that if someone else were to follow your steps, they would end up with the exact same randomly selected numbers. The purpose of this is to prove, if needed, that your selection process was truly random so that no one can accuse you otherwise. The following example illustrates the steps you will need to follow (and explain) when using a random digit table to make your random selection. The random digit table can be found in Appendix A.
Example 1

Five boxes, each containing 24 cartons of strawberries, are delivered in a shipment to a grocery store. The produce manager always selects a few cartons randomly to inspect. He knows better than to just look at some of the cartons on the top or only in one box, because sometimes the rotten ones are on the bottom. Today he wishes to select a total of 6 cartons to inspect. He has the boxes arranged in order and has a set way to count the cartons inside each box. Explain the process used to make the random selection using a random digit table.

Solution

Step 1: Assign numbers to the list (must all be an equal number of digits long)

Since he has 120 cartons total, he will assign the numbers 001 to 120 to represent the cartons in order.

Step 2: Choose a starting line on the random digit table. If the problem states a line to start at, use that line. Otherwise, pick any line you want and record the line number. If you run out of digits, simply move to the next line down.

He will use line 119 to make the selections.

Step 3: Decide how many digits to look at each time. The number of digits in your largest number is required.

He will need to look at 3-digit numbers every time.

Step 4: Decide if any numbers will need to be ignored and whether or not repeats will be allowed.

He will not want to inspect the same carton twice, so he will ignore any repeats. And, any numbers above 120 will not apply in this case, so he will ignore numbers 121-999 and 000.

Step 5: State when to stop.

He will stop once six numbers are selected. He will then find the cartons that the numbers represent and inspect those cartons.

Step 6: Report the numbers that were selected. When given a specific list, go back and determine which specific individuals have been selected.

Here is a part of the random digit table so that you can see how the selection was made. Note that dividers have been placed between each group of 3-digits for this example. When we reach the end of a line, we simply continue on the following line.

<table>
<thead>
<tr>
<th>Line #</th>
<th>random digits in groups of five:</th>
<th>selection:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 119</td>
<td>958</td>
<td>57</td>
</tr>
<tr>
<td>Line 120</td>
<td>35</td>
<td>476</td>
</tr>
<tr>
<td>Line 121</td>
<td>7</td>
<td>148</td>
</tr>
<tr>
<td>Line 122</td>
<td>138</td>
<td>73</td>
</tr>
<tr>
<td>Line 123</td>
<td>54</td>
<td>580</td>
</tr>
</tbody>
</table>

As you can see, the strawberries in cartons numbered 042, 119, 025, 052, 087, and 072 will be inspected. The entire delivery will be accepted or rejected based on this random sample of 6 cartons.
Example 2

Five of the employees at the Stellar Boutique are going to be selected to go to a training seminar in Las Vegas for four days. Everyone wants to go of course, so the owner has decided to make the selection randomly. She has decided to send two managers and three sales representatives. The employees’ names are listed in the tables below.

a) What type of sampling method is this?

b) Explain the process she can follow to use a random digit table, starting at line #108, to select the employees who will get to go to the training. Select the managers first, then select the sales representatives.

Solution

a) This is a stratified random sample.

b) For the managers:
   - Assign numbers to the list 1 to 8
   - Use random digit table, starting at line #108
   - Look at one digit at a time
   - Ignore 9, 0, and any repeats
   - Stop when two have been selected
   - State the names

<table>
<thead>
<tr>
<th>Managers</th>
<th>Sales Representatives</th>
<th>Sales Representatives</th>
<th>Sales Representatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angela</td>
<td>Alfie</td>
<td>Ilma</td>
<td>Ray Anne</td>
</tr>
<tr>
<td>Barbara</td>
<td>Bettie Lou</td>
<td>Jo Jo</td>
<td>Sandy</td>
</tr>
<tr>
<td>Elise</td>
<td>Cari</td>
<td>Katarina</td>
<td>Shirley</td>
</tr>
<tr>
<td>Gigi</td>
<td>Carry</td>
<td>Lin</td>
<td>Suzi</td>
</tr>
<tr>
<td>Malena</td>
<td>Darcy</td>
<td>Marcie</td>
<td>Tawanda</td>
</tr>
<tr>
<td>Rosie</td>
<td>Fan Fan</td>
<td>Nancy</td>
<td>Wendy</td>
</tr>
<tr>
<td>Tammy</td>
<td>Heidi</td>
<td>Oprah</td>
<td>Zulu</td>
</tr>
<tr>
<td>Veronica</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

So, Rosie (#6) and Gigi (#4) will be the managers who get to go to Las Vegas.

For the sales representatives:
   - Assign numbers to the list 01-21
   - Use random digit table, starting on the next line, #109
   - Look at two digits at a time
   - Ignore 22-99, 00, and any repeats
   - Stop when three have been selected
   - State the names

<table>
<thead>
<tr>
<th>Line #</th>
<th>random digits in groups of five:</th>
<th>selection:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line 108</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

So, Ray Anne (#15), Sandy (#16), and Tawanda (#19) will be the sales representatives who get to go to Las Vegas.
Problem Set 4.3

Exercises

Use the table of random digits in Appendix A to complete each problem.

1) The manager at Big-N-Nummy-Burger wishes to know his employees’ opinions regarding the work environment. He has 56 employees and plans to select 12 employees at random to complete a survey.
   a) Explain the process he can follow to use a random digit table, starting at line 108, to select an SRS of size 12.
   b) Which employees’ numbers were selected?

2) Use a random digit table to select an SRS of five of the fifty U.S. States. Explain your process thoroughly and report the five states that you chose. Repeat this a second time, but begin on a different line on the random digit table. Compare your lists to another classmate’s lists. Did you end up with any of the same states in your samples?

Table 4.1:

<table>
<thead>
<tr>
<th>Alabama</th>
<th>Alaska</th>
<th>Arizona</th>
<th>Arkansas</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>Colorado</td>
<td>Connecticut</td>
<td>Delaware</td>
</tr>
<tr>
<td>Florida</td>
<td>Georgia</td>
<td>Hawaii</td>
<td>Idaho</td>
</tr>
<tr>
<td>Illinois</td>
<td>Indiana</td>
<td>Iowa</td>
<td>Kansas</td>
</tr>
<tr>
<td>Kentucky</td>
<td>Louisiana</td>
<td>Maine</td>
<td>Maryland</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>Michigan</td>
<td>Minnesota</td>
<td>Mississippi</td>
</tr>
<tr>
<td>Missouri</td>
<td>Montana</td>
<td>Nebraska</td>
<td>Nevada</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>New Jersey</td>
<td>New Mexico</td>
<td>New York</td>
</tr>
<tr>
<td>North Carolina</td>
<td>North Dakota</td>
<td>Ohio</td>
<td>Oklahoma</td>
</tr>
<tr>
<td>Oregon</td>
<td>Pennsylvania</td>
<td>Rhode Island</td>
<td>South Carolina</td>
</tr>
<tr>
<td>South Dakota</td>
<td>Tennessee</td>
<td>Texas</td>
<td>Utah</td>
</tr>
<tr>
<td>Vermont</td>
<td>Virginia</td>
<td>Washington</td>
<td>West Virginia</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>Wyoming</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3) Washington High School has had some recent problems with students using steroids. The district decides that it will randomly test student athletes for steroids and other drugs. The boy’s hockey team is to be tested. There are 13 players on the varsity team and 21 players on the junior varsity team. Use a table of random digits starting at line 122 to choose a stratified random sample of 3 varsity players and 5 junior varsity players to be tested. Remember to clearly describe your process.

<table>
<thead>
<tr>
<th>Varsity Team Roster (by last name):</th>
<th>Juniour Varsity Team Roster (by last name):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander</td>
<td>Andersen</td>
</tr>
<tr>
<td>Rix</td>
<td>Manzel</td>
</tr>
<tr>
<td>Baker</td>
<td>Andersen</td>
</tr>
<tr>
<td>Radamacher</td>
<td>Peterson</td>
</tr>
<tr>
<td>Brooks</td>
<td>Baker</td>
</tr>
<tr>
<td>Ritchie</td>
<td>Randal A.</td>
</tr>
<tr>
<td>Finch</td>
<td>Christian</td>
</tr>
<tr>
<td>Smithe</td>
<td>Randal J</td>
</tr>
<tr>
<td>Gustaf</td>
<td>Donovan</td>
</tr>
<tr>
<td>Thomas</td>
<td>Reeder</td>
</tr>
<tr>
<td>Linder</td>
<td>Greene</td>
</tr>
<tr>
<td>West</td>
<td>Rice</td>
</tr>
<tr>
<td>Mullen</td>
<td>Hansen</td>
</tr>
<tr>
<td></td>
<td>Sams</td>
</tr>
<tr>
<td></td>
<td>James</td>
</tr>
<tr>
<td></td>
<td>Sentel</td>
</tr>
<tr>
<td></td>
<td>Klein</td>
</tr>
<tr>
<td></td>
<td>Thorne</td>
</tr>
<tr>
<td></td>
<td>Linder</td>
</tr>
<tr>
<td></td>
<td>West</td>
</tr>
<tr>
<td></td>
<td>Lutz</td>
</tr>
</tbody>
</table>

Review Exercises

4) Sketch a Venn Diagram that shows two events that are mutually exclusive.

5) Suppose that a survey was conducted at SRHS and it found that 86% of students have their own cell phones, and that 64% of students have their own Twitter account. Furthermore, 9% of the students at SRHS say that they have neither one of these.

   a) Construct and label a Venn Diagram that fits this scenario.

   b) What is the probability that a randomly selected student has both a cell phone and a Twitter account?

   c) What is the probability that a randomly selected student has either a cell phone or a Twitter account?
4.4 Statistical Conclusions

Learning Objectives

- Understand when valid statistical conclusions can be made.
- Calculate an estimated margin of error and 95% confidence interval
- Make confidence statements

Statistical Conclusions

Remember that when you collect information from every individual in a population, it is called a census. In doing a census, we can be certain that the numbers we have calculated really do represent the entire population. But, because a census is often impractical, we generally take a representative sample of the population and use that sample to try to make conclusions about the entire population. The downside to sampling is that we can never be 100% sure that we have captured the truth about the entire population.

For example, imagine taking a random sample of 100 individuals from a large population. Put those back and choose another sample of 100 individuals, repeating many times. Each of these samples of size 100 will include a different combination of 100 members of the population. Thus, each sample will result in different statistics. This natural difference between various samples is an expected random sampling error. To take this into account, researchers generally report their findings with a margin of error or to be within a certain range of possible values. This range is called a confidence interval. For example the President’s approval rating might be reported as, “The approval rating for the President is 43.2%, with a margin of error of ±3%.” Which could also be reported as, “The approval rating for the President is between 40.2% and 46.2%.”

Using a statistic to make a conclusion about a population is called statistical inference. This course is an introductory course, so we will only briefly touch on this idea. In a future statistics class, you will learn much more about statistical inference and calculations. It is important to note that statistical conclusions are meaningless when poor sampling techniques have been used. If the data was collected from a voluntary response sample, or you had a low response rate, or an incomplete sampling frame was used, then don’t waste your time performing inference on your statistics. Random sampling error is the only type of error or bias that the margin of error accounts for.
95% Confidence Intervals

Once a statistic is calculated for a sample, it is used as an estimate for what the actual parameter might be. We do not know whether our statistic is close to the population parameter, or if it is too high, or if it is too low, so we build our interval around the statistic. We add the margin of error to, and subtract the margin of error from, our statistic. We then report this range of values as our confidence interval which is the interval within which we are fairly confident that the true parameter must be. In a more formal course you will learn how to calculate the margin of error more precisely, and for various levels of confidence (such as 90% or 99% etc.). In this course we will use a simple formula that estimates the margin of error for a 95% confidence interval. We will also make a 95% confidence statement, which explains our conclusion regarding the population parameter in context. The formulas for an estimated 95% margin of error and confidence interval for a proportion are:

\[
\text{Margin of error formula: } m.e. = \pm \frac{1}{\sqrt{n}}
\]

\[
\text{Confidence interval: } \hat{p} \pm m.e. \quad \text{or} \quad \text{statistic} \pm \text{margin of error}
\]

\[
n = \text{sample size}
\]

\(\hat{p}\) (p-hat) is the symbol for the proportion of a sample that has a certain characteristic.

*Note: In order to make a smaller margin of error, and therefore a more narrow confidence interval, one must increase the size of the sample.

Once you have found the range of numbers for your confidence interval, you are going to state your conclusion in context. Such a statement is called a confidence statement. The confidence interval refers to the population - not the sample. We are 100% certain of our sample statistic. It is the population parameter that we are estimating. To write a confidence statement, use the following template:

“We are 95% confident that the true proportion of \((\text{parameter of interest})\) will be between \((\text{low value of CI})\) and \((\text{high value of CI})\).”

https://bit.ly/probstatsSection4-4b

MOE and CI
Example 1

A random sample of 125 union members was conducted to see whether or not the union members would support a strike. Sixty-four of those surveyed said that they would support a strike unless safety conditions were improved. Identify the...

a) Population (of interest)

b) Parameter (of interest)

c) Sample

d) Statistic

e) Estimated margin of error

f) Estimated 95% confidence interval

g) Confidence statement

Solution

a) Population (of interest): All members of this union

b) Parameter of (interest): The true percent of all the union members who would support a strike

c) Sample: The 125 union members who were surveyed

d) Statistic: (p-hat) \( \hat{p} = \frac{64}{125} = 0.512 \)

e) Margin of Error: m.e. \( \pm \frac{1}{\sqrt{125}} = \pm 0.0894 \)

f) 95% Confidence Interval: \( 0.512 + 0.0894 = 0.6014 \) and \( 0.512 - 0.0894 = 0.4226 \)

\[ 0.4226 \text{ to } 0.6014 \] or \[ 42.26\% \text{ to } 60.14\% \]

g) Confidence Statement: “We are 95% confident that the true proportion of union members who would support a strike is between 42.26% and 60.14%”
Problem Set 4.4

Exercises

1) A survey was done to determine the texting habits of MBHS students. An SRS of 270 students was asked several questions related to texting and cell phone usage. Of particular interest to the researchers was the proportion of students who text while in class. Of those surveyed, 178 said that they text during class at least ten times per week. Identify each of the following as specifically as possible:
   a) Population (of interest)
   b) Parameter (of interest)
   c) Sample
   d) Statistic
   e) Margin of Error
   f) 95% Confidence Interval
   g) Confidence Statement
   h) Do you personally feel that this is too high or too low of an estimate of the proportion of teens at your high school who text during class?

2) To predict the outcome of an upcoming Mayoral election, a random sample of 814 voters was selected. These people were asked several questions regarding the election. One question asked whether they were “...leaning Republican, Democratic, Independent, or other/undecided?” Based on this question, 38.2% of respondents said that they were “leaning Democratic...”. Identify each of the following as specifically as possible:
   a) Population (of interest)
   b) Parameter (of interest)
   c) Sample
   d) Statistic
   e) Margin of Error
   f) 95% Confidence Interval
   g) Confidence Statement

3) In the same survey as problem 2), 42.3% said that they were “...leaning Republican...”.
   a) Calculate an estimated 95% confidence interval
   b) Is this enough evidence to “call” the election in favor of the Republicans? Why or why not?
      (Hint: Compare your results from 2f and 3a.)
4) The quality control officer at Spaz Cola uses a systematic random sampling method to select cans of Spaz Cola to determine whether the machines are maintaining the correct recipe. Among the 480 cans analyzed today, 43 cans contained less sugar than the Spaz recipe requires! Identify each of the following as specifically as possible:

a) Population (of interest)
b) Parameter (of interest)
c) Sample
d) Statistic
e) Margin of Error
f) 95% Confidence Interval
g) Confidence Statement
h) Do you think that the company should be concerned? Why or why not?

Review Exercises

5) Marcus got 18 points correct out of 42 possible points on his science test. On his history test, Marcus got 31 points out of 55 possible points. On which test did Marcus do better? Explain or show how you know.

6) Lydia got 15 points correct out of 23 possible on her probability quiz. She then earned 37 points of the 48 possible points on her probability test. On which of these assessments did Lydia do better? Explain or show how you know.

7) The figure below is a dartboard. Suppose that a dart is thrown at it randomly. What is the probability that the dart will land on the shaded area?

8) Sketch two different “dart boards” such that the probability of hitting the shaded is equal to one-third.
4.5 Experiments and Observational Studies

Learning Objectives

- Know the terminology of basic experimental design
- Identify the elements of an experiment
- Distinguish between observational studies and experiments
- Outline experiments
- Understand the effects of lurking variables

Observational Studies and Experiments

When researchers collect data about subjects without imposing any type of treatment, they are doing an observational study. Many conclusions have been based on observational studies. The discovery that smoking can cause lung cancer was initially theorized based on observational studies. Many consumers of cigarettes and tobacco companies questioned the validity of such studies, suggesting that it could have been some other variables that caused the cancers, not the cigarettes. Retrospective studies, based on past history of lung cancer patients showed that a high proportion of them were smokers. This did not convince those who either enjoyed smoking, or were making money off of tobacco. They claimed there could be some lurking variables to blame. Lurking variables are extra variables that were not taken into account, but may actually be the cause. Prospective studies, following people in the future, were undertaken in an effort to see whether or not there was a link between cigarette smoking and lung cancer. The statistics were still called into question because statisticians know that the only way to truly show causation is through a controlled experiment.

An experiment imposes some ‘treatment’ on the subjects. A controlled experiment involves having more than one treatment group, where the only variable that is different between the groups is the treatment being tested. Additionally, subjects will need to be assigned at random to the various treatment groups to control for lurking variables. With regard to cigarettes and lung cancer, researchers would need to find a group of non-smokers and randomly divide them into two groups. The randomization will divide up lurking variables that the researchers cannot control for. Also, there needs to be a fairly large number of subjects in each group so that the results do not appear to be some kind of a fluke. The researchers would then need to force one group to smoke cigarettes, while making sure that those in the control group did not smoke. This would go on for several years and both groups would need to be checked for lung cancer regularly. Clearly, there is no ethical way to do such an experiment. We cannot force people to do something that we suspect may cause cancer! Scientists were able to experiment on rats to see whether or not cigarettes caused cancer, and it did. Eventually, the compilation of all of these studies convinced everyone that smoking does cause cancer.
The Three Elements of Good Experimental Design

[1] **Randomization**—Subjects must be randomly assigned to treatment groups in an effort to divide up any lurking variables.

[2] **Control**—There should be a control group. This is a group that does not receive the same treatment as the group receiving the treatment being tested. Having more than one treatment group, where the only difference is the treatment being tested, allows for comparisons to be made. The intention is to limit lurking variables.

[3] **Replication**—There should be a large enough number of subjects so it seems reasonable that the results simply didn’t happen by chance. Also, the experiment should be able to be replicated on a different group of subjects.

Experimental Design

In an experiment, the people, animals, or objects, that are being experimented on are called the **subjects**. The treatment that is being tested is the **explanatory variable**. The result, outcome, or change that happens (or doesn’t happen) is the **response variable**. Keep in mind that sometimes it is necessary to give a pre-test prior to imposing the treatment. For example, if we are testing a medication that claims to lower cholesterol levels, we will certainly need to know the cholesterol levels of all of our subjects prior to giving them the treatment. At the end of the experiment we will again test them and look for any change in cholesterol levels.

The control group may be given no treatment at all. Or, you may want to use the control group as a way to compare a new treatment to an old treatment. For example, if someone has developed a new medication that they believe will cure headaches; they will want to compare it to aspirin, acetaminophen, and ibuprofen. Researchers will likely form four randomly assigned groups (Groups A, B, C, and D), and assign the subjects in each respective group to take their specific treatment whenever they have a headache and to record how quickly and whether or not it worked. After some length of time, the researchers will collect the data from the four groups and compare the results. With the only difference being which treatment was taken, researchers can make conclusions determining which treatment (if any) worked better than the others.
There are some other potential problems here though. For instance, would you want the subjects to know which medication they are receiving? It is very possible that they may have some preconceived notions regarding the effectiveness of one or more of these medicines. Such unconscious bias can influence how they perceive the treatment to work. What researchers often do to avoid any bias that the subjects will bring with them is to not tell them what treatment they are receiving. Such an experiment is said to be blind. It is also possible that the researcher may have preconceived notions, or hopeful expectations, regarding the effectiveness of one or all of the treatments. To avoid this, a third party can package the various treatments in similar looking containers, each marked only with a code, before the researcher distributes them to the subjects. In this case neither the subjects nor the researcher distributing the treatments know who is getting what. This is a double blind experiment, and is used often in clinical trials to limit bias.

Another issue is that often a patient’s symptoms may improve just at the ‘idea’ of getting a medication. This is called the placebo effect. Imagine a child who is crying dramatically over a scraped knee, but stops immediately once mom puts a bandage on. The bandage acts as the placebo. It is also common for a participant, who believes that she or he is receiving a potentially promising medication, to have symptoms improve simply because of her or his expectation that their symptoms will improve. To account for this placebo effect, researchers will often give the control group a fake treatment called a placebo. A placebo is sometimes called a “sugar pill”—it looks like the real treatment, but has no active ingredients. Placebos make blind and double-blind experiments possible. An experiment could involve a placebo shot, or even a placebo surgery (aka sham surgery).

We will demonstrate how to outline an experiment through the following examples. See the sample outline above as a reference.
Example 1

Suppose that a group of scientists have developed a medication that they believe will cure people of being mean. They are calling it Kind At Last (KAL). There are 520 mean people who are willing to participate in this study (300 males and 220 females). This pill needs to be taken twice daily and it may take a few weeks to be fully absorbed into a person’s system. Identify the following:

a) Subjects
b) Explanatory Variable
c) Response Variable
d) Will it be blind? Double-blind? Placebo controlled? Is a pre-test necessary?
e) Outline a completely randomized experiment

Solution

a) Subjects: the 520 mean people (300 male & 220 female)
b) Explanatory Variable: the KAL pills
c) Response Variable: any change in how mean the subject is
d) Will it be blind? Double-blind? Placebo controlled? Is a pre-test necessary? This could definitely be placebo controlled and double-blind. Neither the patients, nor the person distributing the medicine should know which people are receiving which medication. The KAL pills and the placebos will look identical and be in similar packages. A pretest might be used to determine how mean each of the people in the experiment actually are before receiving a treatment.

e) Outline a completely randomized experiment:

The previous example is a completely randomized experiment because all of the subjects started in one group. All subjects were then randomly assigned to treatment groups, with any combination of people being possible. What if it was theorized that this medication actually has different effects on males than on females? With a completely randomized design it is very possible that we would not end up with an equal number of males and females in each treatment group. If that were to happen, we would not be able to tell whether the treatment affected the different sexes in the same way or not.
Randomized Block Designs

In such a case, it is a good idea to involve blocking in your experimental design. When it is suspected that different subgroups may respond differently to the treatment, the statisticians separate them at the beginning into subgroups called blocks. The subjects in an experiment may be blocked by age, sex, race, previous medical history, etc. Be sure that you do not say that you will randomly assign to the blocks. You cannot randomly choose who is male or female, and you cannot randomly choose who is which race, etc. Each block is then randomly divided among the various treatment groups. This assures a more equal distribution of the subjects among the treatments. It also directly addresses the effects of this suspected lurking variable. Experimental designs in which blocking is used are called randomized block designs.

Example 2

Outline a randomized block design to test the KAL pills that blocks by sex. (Continued from example 1)

Solution

*Once you have done the comparisons within blocks, you will also want to compare across blocks to see if there are differences. For example, perhaps this KAL medicine works really well on males, but doesn't do a thing for females. It is even possible that one sex experiences negative side effects from the medication.
Problem Set 4.5

Exercises

1) Researchers want to determine how effective a new allergy drug called Scratch-Be-Gone is at reducing pet allergies. One pill should be taken daily with a meal. 450 pets suffering from allergies will participate in a clinical study comparing this new drug with an existing market drug and a placebo. Identify each of the following:
   a) Subjects
   b) Explanatory Variable(s)
   c) Response Variable
   d) Is it possible for this experiment to be double-blind? Explain.
   e) Outline a completely randomized experiment.

2) Ms. Rokinroll has a theory that listening to music while working on probability problems will help students retain knowledge. She has a set of earphones for each student and intends to compare the effects of classical music, country music, and heavy metal music. Her first period probability class has 36 students and her last period probability class has 34 students. Identify each of the following:
   a) Subjects
   b) Explanatory Variable(s)
   c) Response Variable
   d) Do you feel that a control group of no music is necessary? Why or why not?
   e) Do you feel that any of the following should be a part of this experiment: blind, double-blind, pre-test, or placebo controlled?
   f) Ms. Rokinroll also suspects that listening to music in the morning will have a different effect than listening to music in the afternoon. Outline a randomized block design experiment that tests this idea.

3) Researchers want to test a new eye drop against Blink Brand Eye Drops to see if it is better at reducing dry eye symptoms for contact wearers. The researchers are also interested in whether males and females will respond differently. The subjects available are 480 male and 502 female contact wearers who suffer from frequent dry eyes. Identify the following:
   a) Subjects
   b) Explanatory Variable(s)
   c) Response Variable
   d) Outline an appropriate experiment: (will it be blind? double-blind? blocked? placebo controlled?)
   e) Clearly explain how a table of random digits can be used to do the randomization. Using line #129, select the first five males who will be in the first treatment group.
4) A new type of cell phone is being developed by The Millionaire Phone Makers Corporation. This phone, called Make-Us-More-Money (MUMM), has a target audience of college students and young professionals (18-35 year olds). The company has developed three different ad campaigns and a commercial for each has been made that will be tried on the test subjects. The company wants to determine which ad campaign will be most effective prior to flooding the market, so they will test the various commercials on 744 University of Wisconsin students and 3,057 people who attend this year’s Young Professional’s Conference in Los Angeles. After viewing a commercial, each subject will fill out a questionnaire that assesses how likely they would be to purchase the MUMM phone. Identify each of the following:

a) Subjects
b) Explanatory Variable(s)
c) Response Variable
d) Outline an appropriate experiment design blocked by the two different locations.
e) This scenario is different than the previous examples, because in the other examples we were able to do the randomization in advance. That would not be possible for something like this in either of the locations because the subjects will walk up to the researcher and need to be assigned to a ‘treatment group on the spot’. Explain how the randomization can be done in a case like this.

Review Exercises
5) A survey was conducted to find out if Minnesota teenagers like SpongeBob SquarePants. Suppose 1300 surveys were mailed out but only 516 of those were returned. Of those 516 who responded, 434 said they liked SpongeBob SquarePants.

a) What was the response rate?
b) What is the population of interest?
c) What is the parameter of interest?
d) What is the sample?
e) What is the statistic?
f) What is the estimated margin of error?
g) What is the estimated 95% confidence interval?
h) Write a 95% confidence statement?
i) Why should we be cautious about using these results to make a conclusion?

6) Suppose that license plates in a certain city must begin with any 2-digit number followed by any 5 letters. How many license plates are possible in this city?

7) A single card is drawn from a standard deck of 52 cards. What is the probability that the card is a queen, given that it is a face card?
4.6 Chapter 4 Review

This chapter covers the topics of data collection methods and potential sources of bias. We learned about experiments, observational studies, sample surveys, and censuses. Several potential errors and sources of bias were introduced. We also learned how to use a random digit table to make random selections, how to outline experimental designs, and how to calculate and state estimated 95% confidence intervals. You should go back and read through each of the sections in this chapter, paying careful attention to all of the new terms in bold. This will help you to do problem 1 from your homework assignment.

Chapter 4 Review Exercises

1) Study your new vocabulary!
   a) If you have not already, make flashcards for the terms from this chapter. You may use an App to create your flashcards or make them on notecards. Write the term on one side of the card. On the other side, write a brief definition and include an example. Terms that appear in bold through section 4.1 through 4.5 are the new terms.
   b) Study your flashcards.

2) Each statement below claims that the ACT’s are not a fair measurement for college readiness, but for a different reason. For each student’s statement, determine whether he or she is questioning the validity, the reliability, or claiming that it will be biased. Explain your answers.
   a) The ACT’s are not fair because it is timed and I cannot work fast enough. Consequently I am not really doing as well as I could; I always get a lower score than I should receive.
   b) The ACT’s are not fair because the vocabulary is not clear and I do not even understand what the questions are asking me. I always study really hard and do my homework and I am totally ready for college, but that doesn’t show up on some stupid test.
   c) The ACT’s are not fair because the first time I took them I scored a 21, but the second time I scored a 16. How can that be right?

3) Suppose you want to take a simple random sample of 350 women from a population of 4700 females on the University of Coolness campus.
   a) Explain the steps you would follow if you were going to make the selection using a table of random digits (be thorough).
   b) Starting at line 137, select the first five numbers.
4) Suppose that after carrying out your survey of 350 women (from problem 3), you found that 74 of the women said that they “did not feel safe walking on campus after dark”. Identify each of the following.
   a) Population (of interest)
   b) Parameter (of interest)
   c) Sample
   d) Statistic
   e) Margin of error
   f) The 95% confidence interval
   g) The 95% confidence statement

5) A high school social studies teacher wants to see if giving a completely multiple choice test versus a traditional free response test will improve student scores. She has designed two versions of the chapter 6 assessment to test her question. She has two classes – a 1st hour class with 32 students, and a 5th hour class with 37 students. The two classes are very different in both behavior and academic performance, so she decides to carry out her experiment using blocking. Identify the following:
   a) Subjects
   b) Explanatory variable(s)
   c) Response variable
   d) Will this experiment be blind? Double-blind? Placebo controlled?
   e) Outline a randomized block experiment.

6) Suppose that you are trying to determine whether kindergarten students who have gone to child care centers show more aggressive behaviors than children who have not attended child care centers. You have the data regarding whether or not each child attended a child care center and for what length of time. You are now going to study aggressive behaviors. For each of the following, decide which type of data collection method is being proposed: observational study, sample survey, census, or experiment.
   a) Observers will watch the kindergartners on the playground, recording aggressive behaviors.
   b) A survey will be given to 20 randomly selected parents asking each to rate his or her child’s behavior.
   c) A survey will be given to all kindergarten parents asking each to rate his or her child’s behavior.
   d) During center time, a teacher will take a toy away from a child and record whether they act aggressively.
7) Use the information for the study in question 6) to answer the following questions.
   a) Suggest something that may go wrong with, or may be a source of bias, for each of the proposed data collection methods.
   b) Which of the methods do you feel will yield the best results? Explain.

8) Your teacher wants to find out whether chocolate helps students concentrate on their tests. In one class, she gives each of the students a piece of chocolate before the test begins. In another class, she does nothing. Is this an example of an observational study, a sample survey, a census, or an experiment? Give reasons to support your answer.

9) The cost of gas in 2001 was $1.45 per gallon. The average cost in 2011 was $3.75 per gallon.
   a) What was the amount of increase?
   b) What was the percent of increase?

10) A researcher at the University of Minnesota believes that a certain component of ant venom can be used to lessen the amount of swelling in the knuckles of people suffering from arthritis. The ant venom treatment has been made into a capsule form that can be swallowed, and is designed to be taken one time per day. Suppose that you have 200 people suffering from arthritis who have volunteered to participate in this study. Identify the following:
   a) Subjects
   b) Explanatory variable(s)
   c) Response variable
   d) Will this experiment be blind? double-blind? placebo controlled?
   e) Outline a completely randomized design

11) You bought a sweater on discount that was originally marked at $30. When you got to the register, it rung up as $23. What was the percent discount?

12) The table below shows the number of seniors and the number of seniors graduating for three high schools.

<table>
<thead>
<tr>
<th>School</th>
<th>Number of Seniors</th>
<th>Number Graduating</th>
<th>Graduation Rate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>McArthur</td>
<td>423</td>
<td>354</td>
<td></td>
</tr>
<tr>
<td>Meade</td>
<td>125</td>
<td>110</td>
<td></td>
</tr>
<tr>
<td>Eisenhower</td>
<td>392</td>
<td>379</td>
<td></td>
</tr>
</tbody>
</table>

   a) Which school has the most students graduating?
   b) Determine the graduation rate for each school.
   c) Which school has the highest graduation rate?
13) Suppose that a poll was commissioned to determine whether people in the U.S. believe that pro wrestling is a sport. Identify the potential problem or type(s) of bias that will likely be present in each of the following scenarios. Some will have more than one. Explain your answers.

a) An online poll was sent to all visitors of the WWE website.

b) Telephone interviews are done to randomly selected phone numbers between 4 p.m. and 8 p.m.

c) Some people were embarrassed to admit that they liked wrestling and think it is a sport.

d) One of the questions asked was, “Do you believe that the pro wrestlers are actors or should they be considered serious athletes?”

e) One of the researchers spilled coffee on a big stack of surveys and several had to be thrown away.

f) A second poll was conducted and it had slightly different results.

g) WWE had fans fill out a survey as they left a pro wrestling event.

14) Determine whether the sampling method used in each situation was an SRS, stratified random sample, systematic random sample, multi-stage random sample, random cluster sample, voluntary response sample, or convenience sample.

a) Every fifth person boarding a plane is searched thoroughly.

b) At a local community College, five math classes are randomly selected out of 20 and all of the students from each class are interviewed.

c) A researcher randomly selects and interviews forty male and forty female teachers, from a university with 122 female and 135 male instructors.

d) A researcher for an airline interviews all of the passengers on five randomly selected flights.

e) Based on 12,500 responses from 42,000 surveys sent to its alumni, a major university estimated that the annual salary of its alumni was $92,500.

f) A community college student interviews everyone in his biology class to determine the percentage of students who own a car.

g) A market researcher randomly selects 200 drivers under 35 years of age and 100 drivers over 35 years of age, from those insured with Quality Car Insurance, to interview.

h) A researcher selects 12 states randomly. From each state, she randomly selects 20 middle schools. From each middle school, she randomly selects 15 teachers. The 3,600 teachers were then interviewed by phone.

i) To avoid working late, the quality control manager inspects the last 10 items produced that day.

j) The names of 70 contestants are written on 70 cards. The cards are placed in a bag, and three names are picked from the bag to win a door prize.
15) The athletic director wants to know how taxpayers in the community feel about funding for athletics at the high school. He surveys his coaches and the parents of athletes at his school. Describe what is wrong with his methodology.

**Image References:**

- Pop can. [http://popartmachine.com July 25, 2011](http://popartmachine.com)
- Clipboard. [http://boylston.bbrsd.schoolfusion.us July 25, 2011](http://boylston.bbrsd.schoolfusion.us)
Chapter 5 – Analyzing Univariate Data

Introduction

Now that we have discussed some methods for collecting data, we can look at what to do with those findings. Whether you have collected categorical or numerical data, you will want to choose an appropriate type of graphical display so that you can visualize the data. Charts and graphs of various types, when created carefully, can provide important information about a data set. You will also need to analyze the data with numerical and summary statistics. Once you have constructed a graphical display and have calculated numerical statistics, it will be necessary to describe your findings verbally. Statisticians can then make appropriate conclusions and comparisons based on the data and statistics while avoiding opinion and judgment statements. This chapter will focus on some of the more common visual presentations of data, numerical analyses of data, and verbal descriptions of data.

5.1 Categorical Data

Learning Objectives

- Organize categorical data in tables
- Construct bar graphs and pie charts by hand and with technology
- Describe, summarize, and compare categorical data

Each student in the class should complete the following survey. The data collected will be used in your homework problems. Notice that the variables in each question are categorical.

1. What is your gender? Choose one
   - Female
   - Male

2. What is your favorite season? Choose one
   - Winter
   - Spring
   - Summer
   - Fall

3. Which of these is your favorite type of food? Choose one
   - Italian
   - Asian
   - Mexican
   - American

4. What type of pet(s) do you have? Choose all that apply
   - Dog
   - Cat
   - Fish
   - Reptile
   - Rodent
   - Other
   - None

https://bit.ly/probstatsSection5-1
(3 video links included)
Frequency Tables and Bar Graphs

When analyzing categorical data (also called qualitative data), bar graphs are commonly used. A bar graph is a graph in which each bar shows how frequently a given category occurs. It is usually helpful to organize the data in a frequency table, a table that shows the number of occurrences for each category, before constructing the bar graph. The bars can go either horizontally or vertically, should be of consistent width, and need to be equally spaced apart. The categories are separate and can be put in any order along the axis. It is common to put them in alphabetical order, but not required. As with all the graphs you will construct, be sure to use a consistent scale, include a title, labels for axes, numbers to mark axes as necessary, and a key whenever needed.

Example 1

A bar graph could show the number of different types of pets for a group of students. The number and types of pets owned by a class of 33 geometry students are shown to the right.

a) What could cause the numbers to add up to more than 33?
b) Construct a bar graph to display this data set.
c) Describe what the graph shows.

Solution

a) They add up to more than 33 because some students likely own more than one type of pet and are being counted in more than one category.
b) Here is a bar graph that was created using Microsoft Excel.
c) For this class, the most common pet is a dog. Fourteen students, or 42% of the class, own a dog. Having a cat or no pet at all are the next most common results. Five students own some type of rodent, two have reptiles for pets, and three have fish. There are also two students who own some other type of pet.
Example 2

A great deal of electronic equipment ends up in landfills as people update their computers, TVs, cell phones, etc. This is a concern because the chemicals from batteries and other electronics add toxins to the environment. This electronic waste has been studied in an effort to decrease the amount of pollution and hazardous waste. The following frequency table shows the amount of tonnage of the most common types of electronic equipment discarded in the United States in 2005. Construct a bar graph and comment on what it shows.

<table>
<thead>
<tr>
<th>Electronic Equipment</th>
<th>Thousands of Tons Discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cathode Ray Tube (CRT) TV's</td>
<td>7591.1</td>
</tr>
<tr>
<td>CRT Monitors</td>
<td>389.8</td>
</tr>
<tr>
<td>Printers, Keyboards, Mice</td>
<td>324.9</td>
</tr>
<tr>
<td>Desktop Computers</td>
<td>259.5</td>
</tr>
<tr>
<td>Laptop Computers</td>
<td>30.8</td>
</tr>
<tr>
<td>Projection TV's</td>
<td>132.8</td>
</tr>
<tr>
<td>Cell Phones</td>
<td>11.7</td>
</tr>
<tr>
<td>LCD Monitors</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Electronics Discarded in the US (2005)
Source: National Geographic, January 2008. Volume 213 No.1, pg 73

Solution

The type of electronic equipment is a categorical variable, and therefore, this data can easily be represented using the bar graph below.

According to this 2005 data, the most commonly disposed of electronic equipment was CRT TV’s, by more than 19 times than that of the next most common type of electronic discard.
**Pie Charts**

Pie charts (or circle graphs) are used extensively in statistics. These graphs are used to display categorical data and appear often in newspapers and magazines. A pie chart shows each category (sectors) as a part of the whole (circle). The relationships between the parts, and to the whole, are visible in a pie chart by comparing the sizes of the sectors (slices).

Constructing a pie chart uses the fact that the whole of anything is equal to 100%. All of the sectors equal the whole circle. Remember from geometry that the central angles of a circle total 360°. In regards to pie charts, 360° = 100% of the circle. The sections should have different colors or patterns to enable an observer to clearly see the difference in size of each section.

Pie charts are an appropriate choice when you are working with categorical data that can be viewed as covering 100% of all results. It is not an appropriate choice when you aren’t working with 100% of the data, when choices may include overlaps, or results come from different categories. For example, when we asked every student in this class to list the pets they currently have, we found some students who had more than one pet. A pie chart would not be an appropriate way to display the data in this case. The sectors in a circle graph do not allow for overlaps such as this. Another time when pie charts are not appropriate is when the choices do not cover all possibilities. For example, the electronic waste example above does not include every possibility, so the categories would not add to 100%. In such cases, a bar graph would be a more appropriate choice because it allows for overlaps and does not need to cover exactly 100% of the choices.

**Example 3: How to Construct a Pie Chart**

The Red Cross Blood Donor Clinic had a very successful morning collecting blood donations. Within three hours twenty-five people had made donations. The types of blood donated are shown in table 5.2 below.

<table>
<thead>
<tr>
<th>Blood Type</th>
<th>A</th>
<th>B</th>
<th>O</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Donors</td>
<td>7</td>
<td>5</td>
<td>9</td>
<td>4</td>
</tr>
</tbody>
</table>

a) Construct a pie chart to represent the data.

b) Comment on what the graph shows.

**Solution**

a) **Step 1:** Determine the total number of donors. \(7 + 5 + 9 + 4 = 25\)

**Step 2:** Express each donor number as a percent of the whole by using the formula

\[
\text{Percent} = \frac{f}{n} \cdot 100\%
\]

where \(f\) is the frequency and \(n\) is the total number.

\[
\frac{7}{25} \cdot 100\% = 28\% \quad \frac{5}{25} \cdot 100\% = 20\% \quad \frac{9}{25} \cdot 100\% = 36\% \quad \frac{4}{25} \cdot 100\% = 28\%
\]
Step 3: Express each donor number as the number of degrees of a circle that it represents by using the formula

\[ \text{Degree} = \frac{f}{n} \cdot 360 \]

where \(f\) is the frequency and \(n\) is the total number.

\[
\begin{align*}
\frac{7}{25} \cdot 360^\circ &= 100.8^\circ \\
\frac{5}{25} \cdot 360^\circ &= 72^\circ \\
\frac{9}{25} \cdot 360^\circ &= 129.6^\circ \\
\frac{4}{25} \cdot 360^\circ &= 57.6^\circ 
\end{align*}
\]

Step 4: Using a protractor or technology, draw the central angles for each section of the circle.

Step 5: Write the label and correct percentage inside or next to the section. Color each section a different color. Be sure to include a title, and a key if needed.

In order to create a pie graph by using the circle, it is necessary to use the percent of a section to compute the correct degree measure for the central angle. The blood type graph labels each section with context and percent, not the degrees. This is because degrees would not be meaningful to an observer trying to interpret the graph. If the sections are not labeled directly as in this example, it becomes necessary to include a key so that the observers will know what each section represents.

b) From the graph, you can see that more donations were of Type O (36%) than any other type. The least amount of blood collected was of Type AB (16%).

Graphs on Computer Software

The above pie chart could be created by using a protractor and graphing each section of the circle according to the number of degrees needed for each section. However, bar graphs and pie charts are most frequently made with computer software programs such as Excel or Google Docs. You will be asked to create bar graphs and pie charts both by hand and by using computer software programs. Always remember to include titles, labels, and keys as needed. Be sure to ‘fix’ the graph generated by the software program so that it looks the way you want it to look and shows clearly whatever it is you are trying to convey.

Example 4

Comment on what the graph shows:
Solution
Several people were asked to choose their favorite fruits from a list of six options. Apples were the favorite choice with 35% of the participants choosing them. The second favorite fruit was cherries at 25%, followed by grapes with 20%. Ten percent of the people said that dates were their favorite fruit. However, only 7% chose bananas from the choices provided and the remaining 3% liked some fruit other than those listed.

Pictographs
Another type of graph that is sometimes used to display categorical data is a pictograph. A pictograph is basically a bar graph with pictures instead of bars. A problem with pictures in graphs is that the area that they take up can mislead the observer. The width and height both increase as the picture gets larger. Pictographs are often used in advertisements and magazines. They can be a fun way to make the graphs more interesting in appearance. However, pictographs can be misleading and can be distracting, so they are generally avoided in serious statistical representations.

Example 5
The following graph compares the number of wins for high school football teams during the 2010 seasons. Explain why the pictograph is misleading.

Solution
The pictures increased in both height and width. So when something should be doubled, it actually looks four times as big. For example, when comparing the number of wins between Eisenhower and Adams the graph should show 4 times as many wins. However, in this pictograph it looks as though Adams had 16 times as many wins (4 times as wide X 4 times as tall).
Problem Set 5.1

Exercises

1) Many students at SRHS were given a questionnaire regarding their interests outside of school. The results of one of the questions, ‘What is your favorite After-School Activity?’, are shown in the table below. Each student chose exactly one of the choices in the table.

<table>
<thead>
<tr>
<th>Students' Favorite After-School Activities</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Play Sports</td>
<td>45</td>
</tr>
<tr>
<td>Talk on Phone</td>
<td>53</td>
</tr>
<tr>
<td>Visit With Friends</td>
<td>99</td>
</tr>
<tr>
<td>Earn Money</td>
<td>44</td>
</tr>
<tr>
<td>Chat Online</td>
<td>66</td>
</tr>
<tr>
<td>School Clubs</td>
<td>22</td>
</tr>
<tr>
<td>Watch TV</td>
<td>37</td>
</tr>
</tbody>
</table>

Source: http://www.mathgoodies.com

a) Create a bar graph for this data.

b) Would a pie chart also be appropriate for this example?

c) Calculate the percent of total for each category and the central angle for each category.

d) Create a pie chart for this data.

2) Based on what you can see in the graph, write a brief description of what it is showing. This should be at least three sentences and be written in context.

3) Use the Type of Pet data collected from your class to complete each problem.
   a) Construct a frequency table to show the Type of Pet data from your class.
   b) Create a bar graph that shows the types of pets the students in our class have. This may be done by hand or with technology.
   c) Write a brief description of what your graph shows.

4) Use the Favorite Season data collected from your class to complete each problem.
   a) Construct a frequency table to show the Favorite Season data from your class.
   b) Create a pie chart that shows the favorite season of the year for the students in your class. This may be done by hand or with technology.
   c) Write a brief description of what your graph shows.

5) Look at the school lunch graph that was created by some students:

![School Lunch Graph]

   a) In what way is this graphical representation misleading? Explain.
   b) Create a better graphical representation for this same data.

6) Use the Favorite Food and Gender data from your class to complete each problem.
   a) Construct a frequency table to show the Favorite Food data separately for males and females from your class.
   b) Create two pie charts that compare the favorite food types for the boys and girls in our class. The charts should ‘match’ as much as possible. In other words, they should be the same size, use the same colors, use the same fonts, etc. This may be done by hand or with technology.
   c) Write a brief description comparing the male and female choices for favorite food. Look for similarities and differences.
Review Exercises

7) The following table has statistics for the Minnesota Wild hockey team for the 2015-2016 season for a selection of players. Thirteen variables are listed across the top of the page.
   a) Identify the individuals.
   b) Identify what each variable represents, for example, GP = games played. You may need to do some research or ask classmates.
   c) Classify each variable as numerical or categorical.

<table>
<thead>
<tr>
<th>#</th>
<th>POS</th>
<th>PLAYER</th>
<th>GP</th>
<th>G</th>
<th>A</th>
<th>P</th>
<th>+/-</th>
<th>PIM</th>
<th>PP</th>
<th>SH</th>
<th>GW</th>
<th>S</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>C</td>
<td>MIKKO KORVU</td>
<td>82</td>
<td>17</td>
<td>39</td>
<td>56</td>
<td>6</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>141</td>
<td>12.10</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>RYAN SUTER</td>
<td>82</td>
<td>8</td>
<td>43</td>
<td>51</td>
<td>10</td>
<td>30</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>188</td>
<td>4.30</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>CHARLIE COYLE</td>
<td>82</td>
<td>21</td>
<td>21</td>
<td>42</td>
<td>1</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>140</td>
<td>15.00</td>
</tr>
<tr>
<td>64</td>
<td>C</td>
<td>MIKAEL GRANLUND</td>
<td>82</td>
<td>13</td>
<td>31</td>
<td>44</td>
<td>-12</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>160</td>
<td>8.10</td>
</tr>
<tr>
<td>22</td>
<td>R</td>
<td>NINO NIEDERREITER</td>
<td>82</td>
<td>20</td>
<td>23</td>
<td>43</td>
<td>9</td>
<td>36</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>159</td>
<td>12.60</td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td>MATT DUMEAU</td>
<td>81</td>
<td>10</td>
<td>15</td>
<td>26</td>
<td>1</td>
<td>30</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>152</td>
<td>6.60</td>
</tr>
<tr>
<td>46</td>
<td>D</td>
<td>JARED SPURGEON</td>
<td>77</td>
<td>11</td>
<td>18</td>
<td>29</td>
<td>11</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>122</td>
<td>9.00</td>
</tr>
<tr>
<td>56</td>
<td>C</td>
<td>ERIK HAULA</td>
<td>76</td>
<td>14</td>
<td>20</td>
<td>34</td>
<td>21</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>99</td>
<td>14.10</td>
</tr>
<tr>
<td>29</td>
<td>R</td>
<td>JASON POMINVILLE</td>
<td>75</td>
<td>11</td>
<td>25</td>
<td>36</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>187</td>
<td>5.90</td>
</tr>
<tr>
<td>26</td>
<td>L</td>
<td>THOMAS VANEK</td>
<td>74</td>
<td>13</td>
<td>23</td>
<td>41</td>
<td>-10</td>
<td>22</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>146</td>
<td>12.30</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>MARCO SCANDELLA</td>
<td>73</td>
<td>5</td>
<td>16</td>
<td>21</td>
<td>6</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>126</td>
<td>4.00</td>
</tr>
<tr>
<td>16</td>
<td>L</td>
<td>JASON ZUCKER</td>
<td>71</td>
<td>13</td>
<td>10</td>
<td>23</td>
<td>-4</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>158</td>
<td>8.20</td>
</tr>
<tr>
<td>11</td>
<td>L</td>
<td>ZACH PARISE</td>
<td>70</td>
<td>25</td>
<td>28</td>
<td>53</td>
<td>-3</td>
<td>36</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>234</td>
<td>10.70</td>
</tr>
</tbody>
</table>

8) John forgot to study for his history quiz, so he will guess on each question. The quiz has 5 true-false questions and 5 multiple-choice questions (with 4 choices each). He will guess an answer for each question. In how many possible ways might John answer all of the questions?

9) What is the probability that John will get all of the questions correct?
5.2 Time Plots & Measures of Central Tendency

Learning Objectives
- Construct time plots
- Describe trends in time plots
- Calculate range and measures of central tendency: mean, median, mode
- Understand how a change in the data will effect the statistics

Line Graphs as Time Plots

We are often interested in how something has changed over time. The type of graphical display that shows this the most clearly is the time plot, or line graph. When one of the variables is time, it will almost always be plotted along the horizontal axis as the explanatory variable. A time plot is a continuous graph that allows us to examine if there is some type of trend in how the response variable behaves over a period of time.

Example 1

The total municipal waste generated in the US by year is shown in the data set below.

a) Construct a time plot to show the change in the amount of municipal waste generated in the United States during the 1990’s.

b) Comment on the trend that is shown in the graph.

c) Suggest factors (other than time) that may be leading to this trend.

<table>
<thead>
<tr>
<th>Year</th>
<th>Municipal Waste Generated (Millions of Tons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>269</td>
</tr>
<tr>
<td>1991</td>
<td>294</td>
</tr>
<tr>
<td>1992</td>
<td>281</td>
</tr>
<tr>
<td>1993</td>
<td>292</td>
</tr>
<tr>
<td>1994</td>
<td>307</td>
</tr>
<tr>
<td>1995</td>
<td>323</td>
</tr>
<tr>
<td>1996</td>
<td>327</td>
</tr>
<tr>
<td>1997</td>
<td>327</td>
</tr>
<tr>
<td>1998</td>
<td>340</td>
</tr>
</tbody>
</table>

Source: http://www.zerowasteamerica.org
Solution

a) In this example, the time (in years) is considered the explanatory variable, and is graphed along the horizontal axis. The amount of municipal waste is the response variable, and is graphed along the vertical axis. Time plots can be drawn by hand most easily using graph paper. They can also be created with computer software programs or graphing calculators. This graph was made using Microsoft Excel.

b) This graph shows that the amount of municipal waste generated in the United States increased at a fairly steady rate during the 1990s. Between 1991 and 1992 there was a decrease of 13 million tons of municipal waste, but every other year during the 1990s had an increase.

c) It should be noted that factors other than the passage of time cause our waste to increase. Population growth, economic conditions, and societal habits and attitudes may also be contributing factors.

Example 2

Here is a line graph that shows how the hourly minimum wage changed from when it was first mandated through 1999.

a) During which decade did the hourly wage increase by the greatest amount?

b) During which decade did it increase the most times?

c) When did it stay constant for the longest?

Solution

a) The greatest increase appears to have happened during the 1990’s, when it went from ≈$3.75 to ≈$5.20.

b) The 1970’s appear to have had 5 or 6 increases in the minimum wage.

c) The longest constant minimum wage was during the 1980’s.
Measures of Central Tendency & Spread

The mean, the median, and the mode are all measures of central tendency. They all show where the center of a set of data “tends” to be. Each one is useful at different times. Any one of these three measures of may be referred to as the center of a set of data.

Mean

The mean, often called the ‘average’ of a numerical set of data, is found by taking the sum of all of the numbers divided by the number of values in the data set. This value is sometimes called the arithmetic mean. Geometrically, the mean is the balance point of a distribution. The mean is a summary statistic that gives you a description of the entire data set and is especially useful with large data sets where you might not have the time to examine every single value. However, the mean can be dramatically impacted by outliers (unusual values), and can end up leaving the observer with the wrong impression of a data set.

Example: Suppose these are the hourly wages for the employees at Burger Boy: $9.25, $9.55, $10.15, $9.40, $9.25, $10.90, $18.75, $10.10. If you calculate the mean wage, you would get $10.92. If someone were to report the average wage at Burger Boy to be $10.92 it would give the impression that this is what the average employee makes. However, this is misleading because all employees other than the manager makes less than this amount. In this particular situation, the mean is misleading. The outlier (the manager’s salary) is causing a significant increase in the mean.

Median

The median is the number in the middle position once the data has been organized from smallest to largest. This is the only number for which there are as many values above it as below it in the set of organized data. The median is sometimes referred to as the equal areas point. The median, for a data set with an odd number of values, is the value that is exactly in the middle of the ordered list. It divides the data into two halves. The median for data set with an even number of values, is the mean of the two values in the middle of the ordered list. The median is a useful measure of center when there are outliers in the data set because the middle number will stay in the middle. The median often gives a good impression of the center because half of the values are above the median and half of the values below the median. It doesn’t matter how big the largest values are or how little the smallest values are.

Example: If you calculate the median salary for the Burger Boy employees you get $9.83. This is a much better description of what the typical employee at Burger Boy gets paid because half the employees make more than this amount and half make less than this amount. The manager’s higher salary does not affect the median.

Mode

The mode of a set of data is simply the number that appears most frequently in the data set. There are no calculations required to find the mode of a data set. You simply need to look for the most common result. Be aware that it is not uncommon for a data set to have no mode, one mode, two modes or even more than two modes. If there is more than one mode, simply list them all. If there is no mode, write ‘no mode’. No matter how many modes, the same set of data will have only one mean and only one median.
The mode is a measure of central tendency that is simple to locate but is not used much in practical applications. It is the only one of these three values that can be for either categorical or numerical data. Remember the example regarding pets from section 5.1? The mode was ‘dog’ because that was the most common response.

**Range**

The range of a data set describes how spread out the data is. It is one measure of variability. To calculate the range, subtract the smallest value from the largest value (maximum value – minimum value = range). This value provides information about a data set that we cannot see from only the mean, median, or mode. For example, two students may both have a quiz average of 75%, but one of them may have scores ranging from 70% to 82% while the other may have scores ranging from 24% to 90%. In a case such as this, the mean would make the students appear to be achieving at the same level, when in reality one of them is much more consistent than the other.

**Example 3**

Stephen has been working at Wendy’s for 15 months. The following numbers are the number of hours that Stephen worked at Wendy’s during the past seven months:

24, 24, 31, 50, 53, 66, 78

What is the mean number of hours that Stephen worked per month for the last seven months?

**Solution**

Stephen has worked at Wendy’s for 15 months but note we are only given data for the last seven months. Therefore, this set of data represents a sample of the population. The mean of a sample is denoted by $\bar{x}$ which is called “x bar” and is found using the formula below.

The number of data points for a sample is written as $n$. The formula to the right shows the steps that are involved in calculating the mean for a data sample.

The formula can now be written using symbols.

You can now use the formula to calculate the mean number of hours that Stephen worked.

The mean number of hours that Stephen worked during this time period was 47 hours per month.
Example 4

The ages of several randomly selected customers at a coffee shop were recorded. Calculate the mean, median, mode, and range for this data.

23, 21, 29, 24, 31, 21, 27, 23, 24, 32, 33, 19

Solution

\[ \text{mean: } \frac{23 + 21 + 29 + 24 + 31 + 21 + 27 + 23 + 24 + 32 + 33 + 19}{12} = \frac{307}{12} = 25.58 \]

\[ \text{median: } \text{Organize the ages in ascending order: } 19, 21, 21, 23, 23, 24, 24, 27, 29, 31, 32, 33 \]

Count in to find the middle value. Note that 24 & 24 are both in the middle. The middle value will be halfway between these two values or the average of 24 and 24.

\[ \text{mode: } \text{Look for the values that occur most frequently (21, 23, 24). This data set has three modes.} \]

\[ \text{range: } \text{Subtract the smallest value from the largest value (max - min = range) } 33 - 19 = 14. \]

\[ \text{Solution: Make your conclusion in context.} \]

At this coffee shop, the mean age of the people in our sample was 25.58 years old and the median age was 24 years old. There were three modes for age at 21, 23, and 24 years old and the range for ages was 14 years.

Example 5

Lulu is obsessing over her grade in health class. She just simply cannot get anything lower than an A- or she will cry! She knows that the grade will be based on her average (mean) test grade and that there will be a total of six tests. They have taken five so far, and she has received 85%, 95%, 77%, 89%, and 94% on those five tests. The third test did not go well, and she is getting worried. The cutoff score for an A is 93% and 90% is the cutoff score for an A-. She wants to know what she has to get on the last test.

a) What is the lowest grade Lulu will need to get on the last test in order to get an A in health?

b) What is the lowest grade Lulu will need to get on the last test in order to get an A- in health?

Solution

a) Set up an equation thinking about how Lulu would calculate her average test grade if she knew all six scores. Knowing that she wants the final average to equal 93%, she puts an ‘x’ in the place of the last test score, and then does some algebra to solve for x.

\[ \frac{85 + 95 + 77 + 89 + 94 + x}{6} = 93 \]

\[ (85 + 95 + 77 + 89 + 94 + x) = 93 \cdot 6 \]

\[ 85 + 95 + 77 + 89 + 94 + x = 558 \]

\[ 440 + x = 558 \]

\[ x = 118 \]

Oh no! There is no way she can get 118%. So, there is no possible hope for her to get an A.
b) It is time to try for an A-, but that 118% scared her, so she is going to think of the lowest possible score that will still be an A-. Because her teacher rounds grades, she knows that she can get an A- if her mean score is 89.5%. The algebra for this calculation is shown to the right.

There is hope! As long as she gets a 97% or higher on this last test, she can get an A-. She is going to study like crazy!

\[
\begin{align*}
\frac{85 + 95 + 77 + 89 + 94 + x}{6} &= 89.5 \\
440 + x &= 89.5 \cdot 6 \\
440 + x &= 537 \\
x &= 97
\end{align*}
\]
Problem Set 5.2

Exercises

1) Determine the mean, median, mode, and range for each of the following sets of values:
   a)  20, 14, 54, 16, 38, 64
   b)  22, 51, 64, 76, 29, 22, 48
   c)  40, 61, 95, 79, 9, 50, 80, 63, 109, 42
2) The mean weight of five men is 167.2 pounds. The weights of four of the men are 158.4 pounds, 162.8 pounds, 178.2 pounds, and 178.2 pounds. What is the weight of the fifth man?
3) The mean height of 12 boys is 5.1 feet. The mean height of 8 girls is 4.8 feet.
   a) What is the total height of the boys?
   b) What is the total height of the girls?
   c) What is the mean height of all 20 boys and girls together?
4) The following data represents the number of mailing advertisements received by ten families during the past month. Make a statement describing the ‘typical’ number of advertisements received by each family during the month. Be sure to include statistics to support your statement.
   43  37  35  30  41  23  33  31  16  21
5) Mica’s chemistry teacher bases grades on the average of each student’s test scores during the trimester. Mica has been kind of slacking this year, but hasn’t been too concerned because he knows that he will at least get the credit (60% = passing). However, his parents just informed him that he will not be allowed to use the car if he has any grades below a C which begin at 73%. Below are Mica’s chemistry test scores for the first eight chapters.
   10, 70, 71, 82, 65, 76, 58, 75
   a) Calculate the mean, median, mode, and range for Mica’s chemistry test scores. What grade will Mica receive in chemistry based on this?
   b) His teacher has decided that each student may retake any one of his or her tests in an effort to improve his or her grade. Mica jumps at this opportunity, studies chapter one for hours and retakes the test. To his, and his mother’s delight, his 10% turns into a 70%!! Woo-hoo! Calculate the mean, median, mode, and range for Mica after this change. Which of these values changed? Which did not? What grade will Mica receive now?
   c) Suppose after Mica turned the 10% into a 70%, he studied only a little bit and earned a 60% on the chapter 9 test and a 76% on the chapter 10 test. What would his final average be in this case?
   d) Suppose instead that after Mica turned the 10% into a 70%, he studied hard and earned an 85% on the chapter 9 test and a 90% on the chapter 10 test. What would his final average be in this case?
6) **Deals on Wheels:** The table below lists the retail price and the dealer’s costs for 10 cars at a local car lot this past year.

<table>
<thead>
<tr>
<th>Car Model</th>
<th>Retail Price</th>
<th>Dealer’s Cost</th>
<th>Amount of Mark-Up</th>
<th>Percent of Mark-Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nissan Sentra</td>
<td>$24,500</td>
<td>$18,750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ford Fusion</td>
<td>$26,450</td>
<td>$21,300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyundai Elantra</td>
<td>$22,660</td>
<td>$19,900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chevrolet Malibu</td>
<td>$25,200</td>
<td>$22,100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pontiac Sunfire</td>
<td>$16,725</td>
<td>$14,225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mazda 5</td>
<td>$27,600</td>
<td>$22,150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>$14,280</td>
<td>$13,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honda Accord</td>
<td>$28,500</td>
<td>$25,370</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volkswagen Jetta</td>
<td>$29,700</td>
<td>$27,350</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subaru Outback</td>
<td>$32,450</td>
<td>$28,775</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Calculate the amount each car was marked up.

b) Calculate the percent that each car was marked up and report answers rounded to the nearest tenth of a percent.

c) Calculate the mean, median, mode and range for the percent of mark-up column.

d) Do the “amount of mark-up column” and the “percent of mark-up column” put the cars in the same order for profit? Explain or give an example.

7) Write a brief description of what the line graph for platinum prices shows. Be sure that you do this in context using complete sentences and that you include at least three observations.

![Line Graph: Platinum Prices, 1960 to 2005](http://www.admc.hct.ac.ae)
8) According to the U.S. Census Bureau, “household median income” is defined as “the amount which divides the income distribution into two equal groups, half having income above that amount, and half having income below that amount.” The table shows the median U.S. household incomes every 3 years from 1975 until 2008, according to the U.S. Census Bureau.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>$11,800</td>
<td>$15,064</td>
<td>$19,074</td>
<td>$22,415</td>
<td>$26,061</td>
<td>$29,943</td>
<td>$31,241</td>
<td>$35,492</td>
<td>$40,696</td>
<td>$42,409</td>
<td>$46,326</td>
<td>$50,303</td>
</tr>
</tbody>
</table>

a) Construct a time plot for the median household data. You may do this by hand, on graph paper, or by using technology.

b) Write a brief description of what the line plot shows. This should be done using complete sentences in context and it should include at least three distinct observations.

Review Exercises

For each of the following problems, decide whether you will use a combination, a permutation, or the fundamental counting principle. Then, set up and solve the problem.

9) A camp counselor is in charge of 10 campers. The kids will be going horseback riding today. There are 5 horses, so they will go in two shifts. In how many ways can the camp counselor assign campers to the specific horses for the first shift?

10) In how many ways can the camp counselor select four of the ten campers to attend the afternoon archery class?

11) How many different three-topping pizzas are possible if there are 12 toppings from which to select?

12) Luigi has 3 pairs of shoes, 7 pairs of jeans, and 8 shirts that he likes to wear that happen to be clean. He is going to put together an outfit for his hot date tonight. If he will choose one of each item, how many different outfits are possible?

13) Eleven skiers are to be in a race. Prizes will be awarded for 1st, 2nd, and 3rd place. Assuming no ties, in how many ways can the prizes be awarded?
5.3 Numerical Data: Dot Plots & Stem Plots

Learning Objectives

- Construct dot plots, stem plots and split-stem plots
- Calculate numerical statistics for quantitative data
- Identify potential outliers in a distribution
- Describe distributions in context – including shape, outliers, center, and spread

Dot Plots

One convenient way to organize numerical data is a dot plot. A dot plot is a simple display that places a dot (or X, or another symbol) above an axis for each datum value (datum is the singular of data). The axis should cover the entire range of the data including numbers that will have no data marked above them. This will visually show outliers or gaps in the data set. There is a dot for each value, so values that occur more than once will be shown by stacked dots. Dot plots are especially useful when you are working with a small set of data across a reasonably small range of values. This type of graph gives the observer a clear view of the shape, mode, and range of the set of data. Outliers are also often easy to spot. Finally, since the numbers are already in order, locating the median is also a simple process.

Ages of all of the Sales People at Stinky’s Car Dealership.

Describing a Numerical Distribution

Once you have constructed a graphical representation of a data set, you should try to describe what the graph shows. There are several characteristics that should be mentioned when describing a numerical distribution and your description needs to explain what this specific data represents. Describe the shape of the graph, whether or not there are any outliers present in the data, the location of the center of the data, and how spread out the data is. All of this should be done in the specific context of the individuals and variable being studied. We will use an acronym to help you remember what to include in your descriptions (S.O.C.C.S.) - shape, outliers, context, center and spread. An explanation of each of these characteristics follows.
Shape

Once a graphical display is constructed, we can describe the distribution. When describing the distribution, we should be sure to address its shape. Although many graphs will not have a clear or exact shape, we can usually identify the shape as symmetrical or skewed. A **symmetrical** distribution will have a middle through which we can draw an imaginary line. The portions of the graphs on the two sides of this line should be fairly equal mirror images of one another. If you were to fold along the imaginary center line, the two sides would almost match up. Many symmetrical distributions are bell shaped; they will be tall in the middle with the two sides thinning out as you move away from the middle. The sides are referred to as tails. A **skewed** distribution is one in which the bulk of the data is concentrated on one end with the other side having less data and a longer tail. The direction of the longer tail is the direction of the skew. Skewed right data sets will have a longer tail to the right while skewed left data sets will have a longer tail to the left. Other shapes that you might see are uniform distributions which have nearly consistent heights all the way across the data set and bimodal distributions which have two peaks in the distribution.

![Symmetric Bell shaped](image)

![Skewed to the Left](image)

![Skewed to the Right](image)

Outliers

We should be sure to mention any outliers, gaps, groupings, or other unusual features of a distribution. An **outlier** is a value that does not fit with the rest of the data. Some distributions will have several outliers, while others will not have any. We should always look for outliers because they can affect many of our statistics. Also, sometimes an outlier is actually an error that needs to be corrected. If you have ever ‘bombed’ one test in a class, you probably discovered that it had a big impact on your overall average in that class. This is because the mean is impacted by outliers and will be pulled toward outliers. This is another reason why we should be sure to look at the data and not just at the statistics about the data. When an outlier occurs in the data set and we do not realize it, we can be misled by the mean to believe that the numbers are higher or lower than they really are.
Context

Do not forget that the graph, the numbers and the descriptions are all about something. There is a context. All of the elements of the distribution should be described in the specific context of the situation in question.

Center

The center of a distribution needs to be included in the verbal analysis as well. People often wonder what the ‘average’ is. The measure for center can be reported as the median, mean, or mode. Even better, give more than one of these in your description. Remember, an outlier will impact the mean but it will not impact the median. For example, while the median of a data set will stay in the center even when the largest value increases tremendously, the mean will change, sometimes significantly.

Spread

In our description of a data set, we should also mention the spread. The spread is a measure of variability and can be reported as the range of values of the data set. When analyzing a distribution, we often don’t want to simply report the range (saying that the range is equal to some number is not always enough information). It can be much more informative to say that the data ranges from _____ to _____ (minimum value to maximum value). For example, suppose the TV news reports that the temperature in St. Paul had a range of 20° during a given week. This could mean very different temperatures depending upon the time of year. It would be more informative to give specific information such as the temperature in St. Paul ranged from 68° to 88° last week.

S.O.C.C.S.

When you describe the distribution of a numerical variable, there are several key pieces of information to include. This text will use the acronym S.O.C.C.S (Shape, Outliers, Context, Center, Spread) to help us remember what characteristics to include in our descriptions.

Example 1

An anthropology instructor at the community college is interested in analyzing the age distribution of her students. The students in her Anthropology 102 class are: 21, 23, 25, 26, 25, 24, 26, 19, 18, 19, 26, 28, 24, 22, 24, 19, 23, 24, 24, 21, 23, and 28 years old. Organize the data in a dot plot. Calculate the mean, median, mode, and range for the distribution. Describe the distribution. Be sure to include the shape, outliers, center, context, and spread.

Solution

- Construct a dot plot:

```
   X X
   X X X X X X X X X
  18 19 20 21 22 23 24 25 26 27 28
```

Ages of Students in Anthropology 102
Solution (continued)

- **mean:**
  \[
  \frac{18 + 19 + 19 + 19 + 21 + 21 + 22 + 23 + 23 + 24 + 24 + 24 + 24 + 24 + 25 + 26 + 26 + 26 + 28 + 28}{20} = 23.27
  \]

  mean = \( \bar{x} = 23.27 \) years old

- **median:** With the numbers listed in order, count to locate the middle number. It is between 24 and 24 so calculate the mean of these two numbers. \((24 + 24) / 2 = 24\) The median = 24 years old.

- **mode:** The most frequent age is 24. The mode is 24 years old.

- **range:** The minimum age is 18 and the maximum is 28 so the range is 28 - 18 = 10 years and the ages range from 18 to 28 years old.

- **describe:** Address the shape, outliers, center, context, and spread of the distribution.

  The distribution of student ages in this Anthropology 102 class is fairly symmetrical with no clear outliers. Student ages range from 18 to 28 years old. The median and mode for age are both 24 years old and the mean is 23.27 years. Thus, the typical student in this class is 23-24 years of age.

**Stem Plots**

In statistics, data is represented in tables, charts or graphs. One disadvantage of representing data in these ways is that sometimes the specific data values are often not retained. Using a stem plot is one way to ensure that the data values are kept intact. A *stem plot* is a method of organizing the data that includes sorting the data and graphing it at the same time. This type of graph uses the stem as the leading part of the data value and the leaf as the remaining part of the value. The result is a graph that displays the sorted data in groups or classes. A stem plot is used with numerical data when it will be helpful to see the actual values organized in order.

To construct a stem plot you must first determine the range of your distribution. Build the stems so that they cover the entire range. Include every stem even if it will have no values after it. This will allow us to see the true shape of the distribution including outliers, whether it is skewed, and if there are any gaps. We then place all of the “leaves” after the appropriate stems. Place the numbers in ascending order and include all values. In other words, repeats will show up more than once. Some people like to put the numbers in order before they construct the stem plot, some like to try to put them in order as they make the plot, and others like to make a rough draft first without regard to order and then make a final copy with the numbers in the correct order. Any of these methods will result in a correct stem plot if completed carefully.
Example 2

A researcher was studying the growth of a certain plant. She planted 25 seeds and kept watering, sunlight, and temperature as consistent as possible. The following numbers represent the growth (in centimeters) of the plants after 28 days.

a) Construct a stem plot

b) Describe the distribution.

Solution

a) Construct a stem plot: Notice that the stem plot has the numbers in ascending order and includes a key and title.

b) Describe the distribution: Be sure to address shape, outliers, center, context, & spread.

The distribution of growth at 28 days ranged from 10 to 61 centimeters for these plants with the majority of plants growing to at least 30cm. The median height was 41cm after 28 days. The shape is bimodal and there is a gap in the distribution because there are no plants in the 20-29 cm class. There are some possible low outliers, but no high outliers for plant growth.

Example 3

Sometimes a stem plot ends up looking too crowded. When the data is concentrated in a few rows, or classes, it can be difficult to determine what the shape is or whether there are any outliers in the data. In the stem plot that follows, the ages of a group of people was concentrated in the 30's and 40's as shown in the plot on left. However, the statistician looking at this was not satisfied with the crowded appearance, so she decided to ‘split’ the stems. The resulting graph on the right, called a split-stem plot, shows very different results. Describe the distribution based on the split-stem plot.

Solution

To split the stems, each stem was written twice. The top one is for the first half of the leaves in that class, and the second one is for the leaves in the second half of that class. For example the first stem of 4 gets 40 to 44, and the second 4 gets 45 to 49.

When splitting stems into two separate groupings, the number 5 is the cutoff for moving into the second grouping, just like we normally round numbers.

The split-stem plot shows that the distribution of ages in this example is bimodal and also roughly symmetrical. It also shows that the ages of 20 and 22 appear to be low outliers. None of this was visible in the original stem plot. Both plots show that the ages range from 20 to 54 years, with a median age of 41 years old, a mean of 41.3 years old, and a mode of 47 years old.
Problem Set 5.3

Exercises

1) The following is data representing the percentage of paper packaging manufactured from recycled materials for a select group of countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>% of Paper Packaging Recycled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonia</td>
<td>34</td>
</tr>
<tr>
<td>New Zealand</td>
<td>40</td>
</tr>
<tr>
<td>Poland</td>
<td>40</td>
</tr>
<tr>
<td>Cyprus</td>
<td>42</td>
</tr>
<tr>
<td>Portugal</td>
<td>56</td>
</tr>
<tr>
<td>United States</td>
<td>59</td>
</tr>
<tr>
<td>Italy</td>
<td>62</td>
</tr>
<tr>
<td>Spain</td>
<td>63</td>
</tr>
<tr>
<td>Australia</td>
<td>66</td>
</tr>
<tr>
<td>Greece</td>
<td>70</td>
</tr>
<tr>
<td>Finland</td>
<td>70</td>
</tr>
<tr>
<td>Ireland</td>
<td>70</td>
</tr>
<tr>
<td>Netherlands</td>
<td>70</td>
</tr>
<tr>
<td>Sweden</td>
<td>70</td>
</tr>
<tr>
<td>France</td>
<td>76</td>
</tr>
<tr>
<td>Germany</td>
<td>83</td>
</tr>
<tr>
<td>Austria</td>
<td>83</td>
</tr>
<tr>
<td>Belgium</td>
<td>83</td>
</tr>
<tr>
<td>Japan</td>
<td>98</td>
</tr>
</tbody>
</table>


The dot plot for this data is shown below.

a) Calculate the mean, median, mode, and range for this set of data

b) Describe the distribution in context. Remember your S.O.C.C.S!
2) At the local veterinarian school, the number of animals treated each day over a period of 20 days was recorded.
   a) Construct a stem plot for the data
   b) Describe the distribution thoroughly. Remember your S.O.C.C.S!

   28 34 23 35 16
   17 47 05 60 26
   39 35 47 35 38
   35 55 47 54 48

3) The following table reports the percent of students who took the SAT for the 20 U.S. States with the highest participation rates for the 2004 SAT test.

<table>
<thead>
<tr>
<th>STATE</th>
<th>SAT Participation Rate 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>87%</td>
</tr>
<tr>
<td>Connecticut</td>
<td>85%</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>85%</td>
</tr>
<tr>
<td>New Jersey</td>
<td>83%</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>80%</td>
</tr>
<tr>
<td>D.C.</td>
<td>77%</td>
</tr>
<tr>
<td>Maine</td>
<td>76%</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>74%</td>
</tr>
<tr>
<td>Delaware</td>
<td>73%</td>
</tr>
<tr>
<td>Georgia</td>
<td>73%</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>72%</td>
</tr>
<tr>
<td>Virginia</td>
<td>71%</td>
</tr>
<tr>
<td>North Carolina</td>
<td>70%</td>
</tr>
<tr>
<td>Maryland</td>
<td>66%</td>
</tr>
<tr>
<td>Florida</td>
<td>67%</td>
</tr>
<tr>
<td>Vermont</td>
<td>66%</td>
</tr>
<tr>
<td>Indiana</td>
<td>64%</td>
</tr>
<tr>
<td>South Carolina</td>
<td>62%</td>
</tr>
<tr>
<td>Hawaii</td>
<td>60%</td>
</tr>
<tr>
<td>Oregon</td>
<td>56%</td>
</tr>
</tbody>
</table>

   Source: http://mathforum.org

4) This stem plot is one that looks too crowded.
   a) Create a split-stem plot for this example.
   b) Name at least two things that are visible in the second plot that were not apparent in the first plot.
   c) Invent a scenario that this data could represent.
5) Several game critics rated the Wow So Fit game, on a scale of 1 to 100 with 100 being the highest rating. The results are presented in the stem plot to the right.

a) Find the three measures of central tendency for the game rating data (mean, median, and mode).

b) Which of these three measures of central tendency gives the best impression of the ‘average’ (typical) rating for this game? Explain.

6) These dot plots do not have any numbers or context. For each of the following dot plots:

a) Identify the shape of each distribution and whether or not there appear to be any outliers.

b) For each plot, determine whether the mean or median would be greater, or if they would be similar.

c) Suggest a possible variable that might have such a distribution. (In other words, invent a context that fits the graph.)
The table below displays statistics for 23 Minnesota Wild hockey players for the 2015-2016 regular season. We will use this data from the players in problems 7 and 8.

<table>
<thead>
<tr>
<th>Forwards &amp; Defensemen</th>
<th>GP</th>
<th>G</th>
<th>A</th>
<th>P</th>
<th>+/-</th>
<th>PIM</th>
<th>PP</th>
<th>SH</th>
<th>GW</th>
<th>S</th>
<th>S%</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>POS</td>
<td>PLAYER</td>
<td>9</td>
<td>C</td>
<td>MIKKO KOIVU</td>
<td>82</td>
<td>17</td>
<td>39</td>
<td>56</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>RYAN SUTER</td>
<td>82</td>
<td>8</td>
<td>43</td>
<td>51</td>
<td>10</td>
<td>30</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>CHARLIE COYLE</td>
<td>82</td>
<td>21</td>
<td>21</td>
<td>42</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>C</td>
<td>MIKAEL GRANLUND</td>
<td>82</td>
<td>13</td>
<td>31</td>
<td>44</td>
<td>-12</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>R</td>
<td>NINO NIEDERREITER</td>
<td>82</td>
<td>20</td>
<td>23</td>
<td>43</td>
<td>9</td>
<td>36</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td>MATT DUNBA</td>
<td>81</td>
<td>10</td>
<td>16</td>
<td>26</td>
<td>1</td>
<td>30</td>
<td>5</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>45</td>
<td>D</td>
<td>JARED SPURGEON</td>
<td>77</td>
<td>11</td>
<td>18</td>
<td>29</td>
<td>11</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>55</td>
<td>C</td>
<td>ERIK HAULA</td>
<td>76</td>
<td>14</td>
<td>20</td>
<td>34</td>
<td>21</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>29</td>
<td>R</td>
<td>JASON POMINVILLE</td>
<td>75</td>
<td>11</td>
<td>25</td>
<td>36</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>L</td>
<td>THOMAS VANEK</td>
<td>74</td>
<td>18</td>
<td>23</td>
<td>41</td>
<td>-10</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>MARCO SCANDELLA</td>
<td>73</td>
<td>5</td>
<td>16</td>
<td>21</td>
<td>6</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>L</td>
<td>JASON ZUCKER</td>
<td>71</td>
<td>13</td>
<td>10</td>
<td>23</td>
<td>-4</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>L</td>
<td>ZACH PARISE</td>
<td>70</td>
<td>25</td>
<td>28</td>
<td>53</td>
<td>-3</td>
<td>36</td>
<td>7</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>25</td>
<td>D</td>
<td>JONAS BRODIN</td>
<td>68</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>-5</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>CHRIS PORTER</td>
<td>61</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>-6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>L</td>
<td>RYAN CARTER</td>
<td>60</td>
<td>7</td>
<td>5</td>
<td>12</td>
<td>-3</td>
<td>48</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>R</td>
<td>JUSTIN FONTAINE</td>
<td>60</td>
<td>5</td>
<td>11</td>
<td>16</td>
<td>3</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>D</td>
<td>NATE PROSSER</td>
<td>54</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>C</td>
<td>JARRET STOLL</td>
<td>51</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>-4</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>MIKE REILLY</td>
<td>29</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>-4</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>JORDAN SCHROEDER</td>
<td>26</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>CHRISTIAN FOLIN</td>
<td>26</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>-1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>R</td>
<td>DAVID JONES</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: http://wild.nhl.com/club/stats

7) Analyze the variable “GP”; which stands for **games played**.
   a) Create a stem plot for the number of games played by these Wild players.
   b) Calculate the mean, median, mode, range for the number of games played by these Wild players.
   c) Describe the distribution of the number of games played by these players. Remember your S.O.C.C.S!

8) Now, you will examine the +/- **statistic data**.
   a) Find out what +/- stands for?
   b) Construct a dot plot to show the +/- data.
   c) Describe the distribution.
Review Exercises

9) A random poll was conducted in Springfield to determine what percent of people enjoy watching the television show *The Simpsons*. Of the 1245 people surveyed, 1002 said that they do enjoy watching *The Simpsons*. Identify each of the following:

a) population of interest
b) parameter of interest
c) sample
d) statistic
e) estimated margin of error
f) estimated 95% confidence interval
g) confidence statement
5.4 Numerical Data: Histograms

Learning Objectives

- Construct histograms
- Describe distributions including shape, outliers, center, context, and spread.

Histograms

When it is not necessary to show every value the way a stem plot would do, a histogram is a useful graph. Histograms organize numerical data into ranges, but do not show the actual values. The histogram is a summary graph showing how many of the data points fall within various ranges. Even though a histogram looks similar to a bar graph, it is not the same. Histograms are for numerical data sets and each ‘bar’ covers a range of values. Each of these ‘bars’ is called a class or bin. Histograms are a great way to see the shape of a distribution and can be used even when working with a large set of data.

The bin width is the most important decision that needs to be made when constructing a histogram. The bins need to be of consistent width so that they cover the same range. A well-built histogram will not have fewer than 5 and not more than 15 bins. Find the range and divide by 10. This will give you an idea of how wide to make your bins. From there it becomes a judgment call as to what is a reasonable bin width. For example, it really does not make any sense to count by 11.24 just because that is what the range divided by 10 is equal to. In such a case, it might make more sense to count by 10’s or 12’s depending on the specific data.

Example 1

Suppose that the test scores of 27 students were recorded. The scores were: 8, 12, 17, 22, 24, 28, 31, 37, 39, 40, 42, 43, 47, 48, 51, 57, 58, 59, 60, 65, 74, 75, 84, 88, 91. The lowest score was an 8 and the highest was a 91. Construct a histogram.

Solution

Plan bin width: The first step is to look at the range which is 91 - 8 = 83. Divide the range by 10 to get 83/10 = 8.3. It doesn’t make any sense to count by bins of 8.3 points, so we may use 8, or 10, or 12. Next we look at where to start. The first number is 8. It doesn’t make any sense to start counting at 8 either, or to end at 91. We will probably want to start from 0 and end at 100. Counting by 10’s should work nicely.

*Where to begin, and what to count by, are not obvious to a calculator or many computer software programs. The graphing calculator would probably start at 8, and count by 8.3. Leaving you with bins of [8 -16.3); [16.3-24.6); [24.6 -32.9); etc. If you are using technology to create a histogram, you will generally need to ‘fix’ the window so that the bin widths make sense.*
Mark the horizontal axis: Mark your scale along the horizontal axis to cover your entire range and to count by the decided upon bin width. Include values where you marked your scale.

Count the number of values within each bin: We note that only one value falls between 0 and <10 so we will make the first bin one unit tall. There are two results between 10 and <20 so we make this bin two units tall. Continue counting in this fashion. A frequency table may be helpful here. You need to know how tall to make each bin. You especially need to know how tall to make the tallest of the bins so that your vertical axis will be scaled properly.

Mark the vertical axis: Your vertical axis needs to reach the height of the tallest bin. Mark your vertical axis by consistent steps so that it will reach the value needed. Include labels.

*For instance, if you need to get to 2,460; then you should probably count by steps of 250’s or even a larger number.

Make your histogram: Make the bins the correct heights, shade or color them in, add labels including any units, a title, and a key if needed.

TEST SCORES

The bins in this example are [0 to 10); [10 to 20); etc. This means that zero up to, but not including 10 are in the first bin. Note that 9.999 would be in bin #1, but 10 would be in bin #2.

You may be creating your histograms with paper and pencil. Graphing calculators are also a great way to create histograms as well because you have the opportunity to try out different bin widths without needing to erase and start all over. Also, you may want to try to create histograms in Excel or Google Sheets. When you use technology to create your graphs, you should sketch an approximate picture of what you see. Your sketch will look similar to the graph that the technology produces but you will still need to add labels and titles.
Example 2

a) Construct a histogram to look at the distribution of acceptance rates for these U.S. Universities.

b) Describe your findings.

Solution

a) **Try this on your calculator:** Enter the data in a list and set up a histogram.

**Plan bin width:** Determine the range (72 -11 = 61). Divide by 10 (61/10 = 6.1) to get a rough idea of a good bin width. We can try a variety of bin widths of 5, 7.5, 8, or 10, etc. We must start before the minimum of 11 (start at 0 or 10), and pass the maximum of 72 (80).

After trying a few of these options, we decide to use a bin width of 10, starting at 10 and ending at 80. Here is the window that was used on a TI-84 graphing calculator: (Xmin = 10, Xmax = 80, Xscl = 10, Ymin = -2, Ymax = 5, Yscl = 1)

**Mark the horizontal axis:** Mark your scale along the horizontal axis to cover your entire range and to count by your decided upon bin width. Include values.

**Count the number of values within each bin:** A frequency table may be helpful here. You need to know how tall to make each bin. You especially need to know how tall to make the tallest of the bins.

**Mark the vertical axis:** Your vertical axis needs to reach the height of the tallest bin. Mark your vertical axis by consistent steps so that it will reach the number needed. Include values.

**Make your histogram:** Make the bins the correct heights, shade or color them in, add labels, and include units, a title, and a key if needed.

b) The median and mean are difficult to identify from just a histogram. You will often only be able to estimate them. In this case, we were given all of the original data so we can find the exact values. When possible, identify outliers specifically.

The median acceptance rate for these Universities is 30%. The percent of students, who were accepted to these universities ranged from 11% to 72%. Note that 72% was a high outlier because the next highest rate was 49%. Most of these schools accepted 36% or fewer of those who applied. The distribution is skewed to the right with the high outlier of American University.

<table>
<thead>
<tr>
<th>College or University</th>
<th>Percent Accepted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard University</td>
<td>11</td>
</tr>
<tr>
<td>Yale University</td>
<td>16</td>
</tr>
<tr>
<td>Princeton University</td>
<td>12</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>32</td>
</tr>
<tr>
<td>New York University</td>
<td>29</td>
</tr>
<tr>
<td>M.I.T.</td>
<td>16</td>
</tr>
<tr>
<td>Duke University</td>
<td>26</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>36</td>
</tr>
<tr>
<td>George Washington University</td>
<td>49</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>33</td>
</tr>
<tr>
<td>American University</td>
<td>72</td>
</tr>
<tr>
<td>Cornell University</td>
<td>31</td>
</tr>
</tbody>
</table>

*Source: http://www.netmba.com*
Problem Set 5.4

Exercises

1) This graph shows the distribution of salaries (in thousands of dollars) for the employees of a large school district. Answer the questions that follow.

a) Approximately how many employees make $77,000 or more per year?

b) What is the bin width here? Be careful.

c) Without calculating anything, how would you describe the typical salary of an employee of this school district?

2) Jessica is a freshman at the University of Minnesota Duluth. She has been watching her weight because she is afraid of gaining that ‘freshman fifteen’ she keeps hearing about. She has weighed herself every Monday morning since school started. Here is a histogram showing the results in pounds of all of her Monday morning weight checks.

a) Describe the distribution. Remember your S.O.C.C.S!

b) What is the range for the bin that has 6 observations?

c) For her height, Jessica feels that 140 lbs. is her ideal weight. What percent of the time has she been within 5 lbs. of her ideal weight?
3) Pretend you are a journalist.
   a) What do you notice that is wrong with this graph?
   b) Based on only what you can see in the graph and labels, write several sentences that could go with this graph. (Think S.O.C.C.S!) Ignore the mistakes from part (a).

Pretend you are a journalist.

<table>
<thead>
<tr>
<th>Age</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-21</td>
<td>70</td>
</tr>
<tr>
<td>25-34</td>
<td>50</td>
</tr>
<tr>
<td>35-44</td>
<td>30</td>
</tr>
<tr>
<td>45-54</td>
<td>20</td>
</tr>
<tr>
<td>55-64</td>
<td>10</td>
</tr>
<tr>
<td>65-74</td>
<td>5</td>
</tr>
<tr>
<td>over 75</td>
<td>0</td>
</tr>
</tbody>
</table>

Source: Department of Health
Source: Men and exercise graph: http://www2.le.ac.uk
4) Here again are the statistics from several of the 2015-2016 Minnesota Wild players. We are going to analyze the **Penalties in Minutes (PIM)** data.

<table>
<thead>
<tr>
<th>#</th>
<th>POS</th>
<th>PLAYER</th>
<th>GP</th>
<th>G</th>
<th>A</th>
<th>+/-</th>
<th>PIM</th>
<th>PP</th>
<th>SH</th>
<th>GW</th>
<th>S</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>C</td>
<td>MIKKO KORVIU</td>
<td>82</td>
<td>17</td>
<td>39</td>
<td>6</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>141</td>
<td>12.10</td>
</tr>
<tr>
<td>20</td>
<td>D</td>
<td>RYAN SUTER</td>
<td>82</td>
<td>8</td>
<td>43</td>
<td>10</td>
<td>30</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>188</td>
<td>4.30</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>CHARLIE COYLE</td>
<td>82</td>
<td>21</td>
<td>21</td>
<td>1</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>140</td>
<td>15.00</td>
</tr>
<tr>
<td>64</td>
<td>C</td>
<td>MIKAEL GRANLUND</td>
<td>82</td>
<td>13</td>
<td>31</td>
<td>-12</td>
<td>20</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>160</td>
<td>8.10</td>
</tr>
<tr>
<td>22</td>
<td>R</td>
<td>NINO NIEDERREITER</td>
<td>82</td>
<td>20</td>
<td>23</td>
<td>9</td>
<td>36</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>159</td>
<td>12.50</td>
</tr>
<tr>
<td>24</td>
<td>D</td>
<td>MATT DUMBA</td>
<td>81</td>
<td>10</td>
<td>18</td>
<td>1</td>
<td>38</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>152</td>
<td>6.00</td>
</tr>
<tr>
<td>46</td>
<td>D</td>
<td>JARED SPURGEON</td>
<td>77</td>
<td>11</td>
<td>18</td>
<td>9</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>122</td>
<td>9.00</td>
</tr>
<tr>
<td>56</td>
<td>C</td>
<td>ERIK HUALA</td>
<td>76</td>
<td>14</td>
<td>20</td>
<td>34</td>
<td>21</td>
<td>24</td>
<td>0</td>
<td>2</td>
<td>99</td>
<td>14.10</td>
</tr>
<tr>
<td>29</td>
<td>R</td>
<td>JASON POMINVILLE</td>
<td>75</td>
<td>11</td>
<td>25</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>187</td>
<td>5.90</td>
</tr>
<tr>
<td>26</td>
<td>L</td>
<td>THOMAS VANEK</td>
<td>74</td>
<td>18</td>
<td>23</td>
<td>41</td>
<td>-10</td>
<td>22</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>146</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
<td>MARCO SCANDELLA</td>
<td>73</td>
<td>5</td>
<td>16</td>
<td>21</td>
<td>5</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>128</td>
</tr>
<tr>
<td>16</td>
<td>L</td>
<td>JASON ZUCKER</td>
<td>71</td>
<td>13</td>
<td>10</td>
<td>23</td>
<td>-4</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>158</td>
</tr>
<tr>
<td>11</td>
<td>L</td>
<td>ZACH PARISE</td>
<td>70</td>
<td>25</td>
<td>28</td>
<td>53</td>
<td>-3</td>
<td>36</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>234</td>
</tr>
<tr>
<td>25</td>
<td>D</td>
<td>JONAS BRODIN</td>
<td>68</td>
<td>2</td>
<td>5</td>
<td>7</td>
<td>-5</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>L</td>
<td>CHRIS PORTER</td>
<td>61</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>-6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>18</td>
<td>L</td>
<td>RYAN CARTER</td>
<td>60</td>
<td>7</td>
<td>5</td>
<td>12</td>
<td>-3</td>
<td>48</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>14</td>
<td>R</td>
<td>JUSTIN FONTAINE</td>
<td>60</td>
<td>5</td>
<td>11</td>
<td>15</td>
<td>3</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>39</td>
<td>D</td>
<td>NATE FROSSER</td>
<td>54</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td>19</td>
<td>C</td>
<td>JARRET STOLL</td>
<td>51</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>-4</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>MIKE REILLY</td>
<td>29</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>-4</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>JORDAN SCHROEDER</td>
<td>26</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
<td>CHRISTIAN FOLIN</td>
<td>26</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>-1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>R</td>
<td>DAVID JONES</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>17</td>
</tr>
</tbody>
</table>

a) Construct a histogram for PIM (Penalty Minutes) for the Wild players shown above.
b) Describe the distribution. Remember your S.O.C.C.S!

5) Sketch a histogram that fits each of the following scenarios: *(you will have 5 different histograms)*
   a) Symmetrical with a few high outliers and a few low outliers.
   b) Strongly skewed right with no outliers.
   c) Bimodal and symmetrical.
   d) Skewed left with a few outliers.
   e) Doesn’t fit any of the descriptions we have learned.
6) The table to the right lists the average life expectancy for people in several countries, as of 2010.
   a) Construct a histogram for the distribution of life expectancies for these countries (start at Xmin = 45 and use a bin width of 5).
   b) Based on the shape of your graph, do you expect the mean or median to be higher?
   c) Calculate the range and the three measures of central tendency for this data set.
   d) Which of these three measures of central tendency is most appropriate in this context? Explain.

### Review Exercises

7) The local booster club is holding a raffle. There will be one prize of $1000, two prizes of $250, five prizes of $50, and 10 prizes of $25. They are selling 500 tickets at $10 each.
   a) Construct a probability model that shows the different prizes and the probabilities of winning those prizes.
   b) What is the expected value of a single raffle ticket?
   c) Is this raffle considered a “fair game”? Explain why or why not.

8) A fish bowl on a counter contains 4 gold fish, 7 turquoise fish, and 5 pink fish. Simon the cat is playing a game where he closes his eyes, reaches into the bowl, grabs a fish and sees what color the fish is. He then puts the fish back and repeats the process because Simon is sometimes a very kind cat. Find each of the probabilities below.
   a) P(2 turquoise fish)
   b) P(exactly one of the fish is gold)
   c) P(a pink fish, then a gold fish)

9) If Simon changes the game so that he eats the fish after he takes them out of the bowl, find the following probabilities.
   a) P(2 pink fish)
   b) P(exactly one of the fish is turquoise)
   c) P(no gold fish)
5.5 Numerical Data: Box Plots & Outliers

Learning Objectives

- Calculate the five number summary for a set of numerical data
- Construct box plots
- Calculate IQR and standard deviation for a set of numerical data
- Determine which numerical summary is more appropriate for a given distribution
- Determine whether or not any values are outliers based on the 1.5*(IQR) criterion
- Describe distributions in context— including shape, outliers, center, and spread

Box Plots

A box plot (also called box-and-whisker plot) is another type of graph used to display data. A box plot divides a set of numerical data into quarters. It shows how the data are dispersed around a median, but does not show specific values in the data. It does not show a distribution in as much detail as does a stem plot or a histogram, but it clearly shows where the data is located. This type of graph is often used when the number of data values is large or when two or more data sets are being compared. The center and spread of the distribution are very obvious from the graph. It is easy to see the range of the values as well as how these values are distributed around the middle value. The smaller the box plot is, the more consistent the data values are with the median of the data. The shape of the box plot will give you a general idea of the shape of the distribution, but a histogram or stem plot will do this more accurately. Any outliers will show up as long ‘whiskers’. The box in the box plot contains the middle 50% of the data, and each ‘whisker’ contains 25% of the data.

The Five-Number Summary

In order to divide into fourths, it is necessary to find five numbers. This list of five values is called the five-number summary. The numbers in the list are {Minimum, Quartile 1, Median, Quartile 3, Maximum}. We have already learned how to find the median of a set of numbers by putting values in order and find the middle value. Clearly, the minimum and maximum are the smallest and largest values. We now will learn how to find the quartiles.

\[
5\# \text{ sum} = \{ \text{min}, \ Q_1, \ \text{Med}, \ Q_3, \ \text{max} \}
\]
Quartiles

The first step is to list all of the values in order from least to greatest. The minimum and maximum are now on the ends of the list and we can count in to find the median. It is a good idea to write down or circle these three values as you find them. Finding the quartiles is just like finding the median except you are only dealing with half of the data set. **Quartile 1** is the ‘median’ of all of the values to the left of the median. **Quartile 3** is the ‘median’ of all of the values to the right of the median. **Do not include the median when finding the Q1 and Q3.**

Constructing a Box Plot

Start by listing the five-number summary in order {Min, Q1, Med, Q3, Max}. The next step is to mark an axis that covers the entire range of the data. Mark the numbers along the axis before you make the box plot, so that the resulting plot shows the shape of the data. The last step is to place a dot above the axis for each of the 5 numbers from the five-number summary, and then to make a ‘box’ through the second and fourth dots, mark a line through the middle dot to show the median, and mark ‘whiskers’ from the box out to the first and fifth dots.

Example 1

You have a summer job working at Paddy’s Pond which is a recreational fishing spot where children can go to catch salmon which have been raised in a nearby fish hatchery and then transferred into the pond. The cost of fishing depends upon the length of the fish caught ($0.75 per inch). Your job is to transfer 15 fish into the pond three times a day. But, before the fish are transferred, you must measure the length of each one and record the results. Below are the lengths (in inches) of the first 15 fish you transferred to the pond this morning. Calculate the five number summary, and construct a box plot for the lengths of these fish.

<table>
<thead>
<tr>
<th>Length of Fish (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 14 6 9 10</td>
</tr>
<tr>
<td>21 17 15 15 7</td>
</tr>
<tr>
<td>10 13 13 8 11</td>
</tr>
</tbody>
</table>

Solution

Since box plots are based on the median and quartiles, the first step is to organize the data in order from smallest to largest.

6, 7, 8, 9, 10, 10, 11, 13, 13, 14, 15, 15, 17, 21

The minimum is the smallest number (min = 6), and the maximum is the largest number (max = 21). Next, we need to find the median. This has an odd number of values, so the median of all the data is the value in the middle position (Med = 13). There are 7 numbers before and 7 numbers after 13. The next step is the find the median of the first half of the data – the 7 numbers before the median, not including the median. This is called the lower quartile since it
marks the point above the first quarter of the data. On the graphing calculator this value is referred to as Q1.

6, 7, 8, 9, 10, 10, 11

Quartile 1 is the median of the lower half of the data (Q1 = 9).

This step must be repeated for the upper half of the data – the 7 numbers above the median of 13. This is called the upper quartile since it is the point that marks the third quarter of the data. On the graphing calculator this value is referred to as Q3.

13, 13, 14, 15, 15, 17, 21

Quartile 3 is the median of the upper half of the data (Q3 = 15).

Now that the five numbers have all been determined, it is time to construct the actual graph. The graph is drawn above a number line that includes all the values in the data set. Graph paper works very well since the numbers can be placed evenly using the lines of the graph paper. For this example we will need to mark from at least 6 to at least 21. Be sure to mark your axis before you start to construct the box plot. Next, represent the following values by placing dots above their corresponding values on the number line:

Minimum − 6  Quartile 1 − 9  Median − 13  Quartile 3 − 15  Maximum − 21

The five data values listed above are often called the five number summary for the data set and are necessary to graph every box plot.

Make the ‘box’ part around the Q1 and Q3 values, make ‘whiskers’ out to the min and max values, and make a vertical line to show the location of the median. This will complete the box plot.

### Length of fish (in inches) 5# summary = {6, 9, 13, 15, 21}

The five numbers divide the data into four equal parts. In other words, for this example:

- One-quarter of the data values are located between 6 and 9
- One-quarter of the data values are located between 9 and 13
- One-quarter of the data values are located between 13 and 15
- One-quarter of the data values are located between 15 and 21
More Measures of Spread

Range

We have already learned how to find the range of a set of data. The range represents the entire spread of all of the data.

The formula for calculating the range is $\text{max} - \text{min} = \text{range}$.

Interquartile Range

The quartiles give us one more measure of spread (variability) called the interquartile range. The interquartile range (IQR) is the range between the lower and upper quartile. To find the IQR, subtract the quartile 1 value from the quartile 3 value ($Q_3 - Q_1 = \text{IQR}$). The IQR represents the spread, or range, of the middle 50% of the data. The IQR is a measure of spread that is used when the median is the measure of central tendency.

The formula for calculating the IQR is $Q_3 - Q_1 = \text{IQR}$.

Note that while the range is impacted by outliers, the IQR is resistant to outliers.

Standard Deviation

Another measure of spread or variability that is used in statistics is called the standard deviation. The standard deviation measures the spread around the mean. This value is more difficult to calculate than range or IQR, but the formula used takes all of the data values in the distribution into account. Standard deviation is the appropriate measure of spread when the mean is the measure of center. However, the standard deviation is easily affected by outliers or skewness because every value is calculated in the formula. The symbol for standard deviation of a sample is $s$ (on the graphing calculators it is $S_x$) and for a population it is $\sigma$ (sigma).

The standard deviation can be any number zero or greater. It will only be equal to zero if there is no spread (i.e. all values are exactly the same). The more spread out the data is, the larger the standard deviation will be. The standard deviation is most appropriate when you have a very symmetrical, bell-shaped distribution called a normal distribution. We will study this type of distribution in chapter 7.

Which Numerical Summary Should We Use?

We have learned several statistics that are measures of central tendency and several that are measures of spread. How do we know which ones to use? The mean and standard deviation go together while the median will go with the IQR (or range). It is important to remember that the mean and the standard deviation are both affected by outliers and by skewness in a distribution. If either of these issues are present, then the mean and standard deviation are not appropriate. However, it is often interesting to calculate all of the statistics and compare them to one another. The general guidelines are given in the following diagram.
How to Calculate the Standard Deviation by using the Formula

In order to calculate the standard deviation you must have all of the values. Complete the steps below.

1) Calculate the mean of the values.
2) Subtract the mean from each data value. These are the individual deviations.
3) Each of these deviations is squared.
4) All of the squared deviations are added up.
5) The total of the squared deviations is divided by one less than the number of deviations. This is the variance.
6) Take the square root of the variance to get the standard deviation.

The formula for calculating the variance is:

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

The formula for calculating standard deviation is:

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

As you can probably tell, this formula is very time consuming when you have a large set of data. Also, it is easy to make a mistake in your calculations. We will show this process with a small set of data but generally we will use our calculator to find the standard deviation. See Appendix C for calculator instructions on how to find the standard deviation.
Example 2
There are five teenage girls on Buhl Street that the Millers often use to babysit their three rambunctious sons. The babysitters’ ages are 12, 15, 14, 17, and 19 years old. Find the mean and standard deviation for the ages of the Miller’s babysitters.

Solution

1. Calculate the mean of the values.
   \[
   \frac{12 + 15 + 14 + 17 + 19}{5} = 15.4
   \]

2. Subtract the mean from each data value. These are the individual deviations.

3. Each of these deviations is squared.

4. All of the squared deviations are then added up.

5. This total of the squared deviations is divided by one less than the number of deviations. This is the variance.

6. Take the square root of the variance. This is the standard deviation.

<table>
<thead>
<tr>
<th>Data values</th>
<th>Value – mean = deviation</th>
<th>Deviation squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>(x – (\bar{x}))</td>
<td>(x – (\bar{x}))^2</td>
</tr>
<tr>
<td>12</td>
<td>(12 – 15.4) = -3.4</td>
<td>(-3.4)^2 = 11.56</td>
</tr>
<tr>
<td>15</td>
<td>(15 – 15.4) = -0.4</td>
<td>(-0.4)^2 = 0.16</td>
</tr>
<tr>
<td>14</td>
<td>(14 – 15.4) = -1.4</td>
<td>(-1.4)^2 = 1.96</td>
</tr>
<tr>
<td>17</td>
<td>(17 – 15.4) = 1.6</td>
<td>(1.6)^2 = 2.56</td>
</tr>
<tr>
<td>19</td>
<td>(19 – 15.4) = 3.6</td>
<td>(3.6)^2 = 12.96</td>
</tr>
</tbody>
</table>

\[
\text{Sum of the squared deviations} = 29.2
\]

\[
\text{Variance} = \frac{\text{sum}}{n-1} = \frac{29.2}{5-1} = 7.3
\]

\[
\text{Standard Deviation} = \sqrt{s^2} = \sqrt{7.3} = 2.7019
\]

The mean age of the Miller family’s babysitters is 15.4 years old and the standard deviation is 2.7019 years.

The standard deviation is tedious to calculate. For any problem where you are asked to calculate the standard deviation, you may wish to use your calculator or a computer to find it.

Example 3
After one month of growing, the heights of 30 parsley seed plants were measured and recorded. The measurements (in inches) are shown in the table below.

<table>
<thead>
<tr>
<th>Table 5.6: Heights of Parsley (in.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

a) Calculate the five-number summary and construct a box plot to represent the data.
b) Describe the distribution.
c) Calculate the mean and standard deviation.
d) Calculate the median, and IQR
Solution

a) Order the values first. The data organized from smallest to largest is shown in the table below. (Note that you could use your calculator to quickly sort these values.)

| Table 5.7: Heights of Parsley (in.) |
| 6 | 8 | 11 | 11 | 12 | 14 |
| 16 | 17 | 18 | 18 | 22 | 23 |
| 23 | 25 | 26 | 26 | 26 | 27 |
| 28 | 30 | 33 | 34 | 37 | 37 |
| 38 | 38 | 39 | 40 | 46 | 49 |

Now find the 5-number summary. This time there is an even number of data values so the median will be the mean of the two middle values. Med = \( \frac{26 + 26}{2} = 26 \) Note that we will not use the median when finding the quartiles. The median of the lower half is the number in the 8th position which is 17. The median of the upper half is the number in the 22nd position (or 8th from the top) which is 37. The smallest number is 6 and the largest number is 49.

5-number summary = \{6, 17, 26, 37, 49\} (All values are measured in inches.)

b) We will remember to reference S.O.C.C.S. to guide us on our description.

The heights of these parsley plants ranged from 6 inches to 49 inches after one month. The distribution is very symmetrical and does not contain any outliers. The median height for these parsley plants was 26 inches tall. The middle 50% of the plants were all between 17 inches and 37 inches tall.

c) The mean and standard deviation can be calculated using technology. The mean is 25.93 inches and the standard deviation is 11.47 inches.

d) The median is part of the five-number summary and is 26 inches. The IQR = Q3 - Q1 = 37 - 17 = 20 inches.
Outliers

We have been noticing some values that appear to be outliers, but have not defined a specific criteria to be considered an outlier. The common outlier test, used to determine whether or not any of the values are outliers utilizes the IQR. This outlier test, often called the 1.5(IQR) Criterion, says that any value that is more than one and one-half times the width of the IQR box away from the box is an outlier.

Example 4

Test the sodium in the McDonald’s® sandwiches for outliers. The data can be found in the Section 5.5 Exercises, problem #1. Use the 1.5*(IQR) Criterion. Show your steps.

Solution

Calculate the five number summary for the Amount of Sodium (in mg):

Five-number summary = {480, 680, 1030, 1180, 1470}

Find the IQR: IQR = 1180 – 680 = 500

Test for low outliers: Q1 – 1.5(IQR) = 680 – 1.5(500) = -80

Test for high outliers: Q3 + 1.5(IQR) = 1180 + 1.5(500) = 1930

Check the data to see if we have any outliers:

We certainly have no sandwiches with less than -80 mg sodium so we have no low outliers. We also have no values that are greater than the cut off of 1930 mg so we also have no high outliers.
Problem Set 5.5

Exercises

Here is some nutritional information about a few of the sandwiches on the McDonald’s® menu.

1) Determine the median and the IQR for the following data regarding the McDonald’s® sandwiches:
   a) Calories from fat
   b) Cholesterol

2) Anaylze the calories for these McDonald’s® sandwiches.
   a) Calculate the five number summary and construct an accurate box plot for the calories for these sandwiches.
   b) Use the outlier test to determine whether there are any outliers for calories. Test for both high and low outliers. Show your steps.
   c) Describe the distribution in context- Remember your S.O.C.C.S!

3) Anaylze the sodium content further.
   a) Construct a box plot for sodium.
   b) Calculate the median and IQR for sodium by hand and compare your results to Example 4.
   c) Calculate the mean and standard deviation for sodium by using a calculator.
   d) It turns out that the Angus Bacon Cheeseburger has 2,070 mg of sodium. Would it be considered an outlier?
4) The following table shows the potential energy that could be saved by manufacturing each type of material using the maximum percentage of recycled materials, as opposed to using all new materials.

<table>
<thead>
<tr>
<th>Manufactured Material</th>
<th>Energy Saved (millions of BTU’s per ton)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aluminum Cans</td>
<td>206</td>
</tr>
<tr>
<td>Copper Wire</td>
<td>83</td>
</tr>
<tr>
<td>Steel Cans</td>
<td>20</td>
</tr>
<tr>
<td>LDPE Plastics (e.g. trash bags)</td>
<td>56</td>
</tr>
<tr>
<td>PET Plastics (e.g. beverage bottles)</td>
<td>53</td>
</tr>
<tr>
<td>HDPE Plastics (e.g. household cleaner bottles)</td>
<td>51</td>
</tr>
<tr>
<td>Personal Computers</td>
<td>43</td>
</tr>
<tr>
<td>Carpet</td>
<td>106</td>
</tr>
<tr>
<td>Glass</td>
<td>2</td>
</tr>
<tr>
<td>Corrugated Cardboard</td>
<td>15</td>
</tr>
<tr>
<td>Newspaper</td>
<td>16</td>
</tr>
<tr>
<td>Phone Books</td>
<td>11</td>
</tr>
<tr>
<td>Magazines</td>
<td>11</td>
</tr>
<tr>
<td>Office Paper</td>
<td>10</td>
</tr>
</tbody>
</table>

*Source: National Geographic, January 2008. Volume 213 No., pg 82*

a) Calculate the five number summary and construct an accurate box plot for the Energy Saved data.

b) Use the outlier test to determine whether there are any outliers. Show your steps.

c) Calculate the mean and standard deviation for the Energy Saved data. How do the mean and the median compare?

d) Delete any outliers. Recalculate the five number summary, mean and standard deviation. Which values changed?

5) The table shows the mean travel time to work (in minutes), for workers age 16+ for 16 cities in Minnesota. This is according to the U.S. Census website.

*Source: http://quickfacts.census.gov*

<table>
<thead>
<tr>
<th>City Name</th>
<th>Mean Travel Time To Work (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albertville</td>
<td>32.2</td>
</tr>
<tr>
<td>Andover</td>
<td>29.9</td>
</tr>
<tr>
<td>Anoka</td>
<td>24.0</td>
</tr>
<tr>
<td>Blaine</td>
<td>26.9</td>
</tr>
<tr>
<td>Brooklyn Center</td>
<td>23.8</td>
</tr>
<tr>
<td>Brooklyn Park</td>
<td>24.1</td>
</tr>
<tr>
<td>Champlin</td>
<td>26.2</td>
</tr>
<tr>
<td>Coon Rapids</td>
<td>24.7</td>
</tr>
<tr>
<td>Elk River</td>
<td>28.5</td>
</tr>
<tr>
<td>Maple Grove</td>
<td>25.3</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>22.1</td>
</tr>
<tr>
<td>Mounds View</td>
<td>21.6</td>
</tr>
<tr>
<td>North St. Paul</td>
<td>23.2</td>
</tr>
<tr>
<td>Roseville</td>
<td>21.1</td>
</tr>
<tr>
<td>Spring Lake Park</td>
<td>22.1</td>
</tr>
<tr>
<td>St. Paul</td>
<td>21.7</td>
</tr>
</tbody>
</table>

a) Construct a box plot for the mean travel time for residents of these Minnesota cities.

b) Make a statement, in context, about what the ‘box’ part of the box plot tells you.

c) Describe the distribution. Remember your S.O.C.C.S! Identify any unusual values specifically.
6) The Burj Khalifa, in Dubai, is the world’s tallest building. It is more than twice the height of the Empire State Building in New York. The chart below lists the 20 tallest buildings in the world.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Building &amp; Location</th>
<th>Year Completed</th>
<th>Architectural top (meters)</th>
<th>Architectural top (feet)</th>
<th>Floors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Burj Khalifa, Dubai, United Arab Emirates</td>
<td>2010</td>
<td>828</td>
<td>2,717</td>
<td>163</td>
</tr>
<tr>
<td>2</td>
<td>Abraj Al Bait, Mecca, Saudi Arabia</td>
<td>2011</td>
<td>601</td>
<td>1,972</td>
<td>120</td>
</tr>
<tr>
<td>3</td>
<td>One World Trade Center, New York City, USA</td>
<td>2013</td>
<td>541.3</td>
<td>1,776</td>
<td>104</td>
</tr>
<tr>
<td>4</td>
<td>Taipei 101, Taipei, Taiwan</td>
<td>2004</td>
<td>509</td>
<td>1,670</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>Shanghai World Financial Center, Shanghai, China</td>
<td>2008</td>
<td>492</td>
<td>1,614</td>
<td>101</td>
</tr>
<tr>
<td>6</td>
<td>International Commerce Centre, Hong Kong</td>
<td>2010</td>
<td>484</td>
<td>1,588</td>
<td>118</td>
</tr>
<tr>
<td>7</td>
<td>Petronas Towers, Kuala Lumpur, Malaysia</td>
<td>1998</td>
<td>452</td>
<td>1,483</td>
<td>88</td>
</tr>
<tr>
<td>8</td>
<td>Zifeng Tower, Nanjing, China</td>
<td>2009</td>
<td>450</td>
<td>1,480</td>
<td>89</td>
</tr>
<tr>
<td>9</td>
<td>Willis Tower, Chicago, United States</td>
<td>1974</td>
<td>442</td>
<td>1,450</td>
<td>108</td>
</tr>
<tr>
<td>10</td>
<td>Jin Mao Building, Shanghai, China</td>
<td>1998</td>
<td>421</td>
<td>1,381</td>
<td>88</td>
</tr>
<tr>
<td>11</td>
<td>Two International Finance Centre, Hong Kong</td>
<td>2003</td>
<td>415</td>
<td>1,362</td>
<td>88</td>
</tr>
<tr>
<td>12</td>
<td>CITIC Plaza, Guangzhou, China</td>
<td>1997</td>
<td>391</td>
<td>1,283</td>
<td>60</td>
</tr>
<tr>
<td>13</td>
<td>Shun Hing Square, Shenzhen, China</td>
<td>1996</td>
<td>384</td>
<td>1,260</td>
<td>69</td>
</tr>
<tr>
<td>14</td>
<td>Empire State Building, New York City, United States</td>
<td>1931</td>
<td>381</td>
<td>1,250</td>
<td>102</td>
</tr>
<tr>
<td>15</td>
<td>Central Plaza, Hong Kong</td>
<td>1992</td>
<td>374</td>
<td>1,227</td>
<td>78</td>
</tr>
<tr>
<td>16</td>
<td>Bank of China Tower, Hong Kong</td>
<td>1990</td>
<td>367</td>
<td>1,204</td>
<td>70</td>
</tr>
<tr>
<td>17</td>
<td>Bank of America Tower, New York City, United States</td>
<td>2008</td>
<td>366</td>
<td>1,201</td>
<td>54</td>
</tr>
<tr>
<td>18</td>
<td>Almas Tower, Dubai, United Arab Emirates</td>
<td>2008</td>
<td>360</td>
<td>1,180</td>
<td>74</td>
</tr>
<tr>
<td>19</td>
<td>Emirates Office Tower, Dubai, United Arab Emirates</td>
<td>2000</td>
<td>355</td>
<td>1,165</td>
<td>54</td>
</tr>
<tr>
<td>20</td>
<td>Tunlix Sky Tower, Kaohsiung, Taiwan</td>
<td>1997</td>
<td>348</td>
<td>1,142</td>
<td>85</td>
</tr>
</tbody>
</table>

Source: https://en.wikipedia.org

a) Calculate the five number summary for the heights (in feet) of the 20 buildings and construct an accurate box plot.

b) Use the outlier test to determine whether there are any outliers among the heights of these 20 buildings. Test for both high and low outliers. Show your steps. Identify any outliers by name.

c) Describe the shape of the distribution. Remember your S.O.C.C.S!

d) Within what range of heights are the middle 50% of these buildings?

e) Calculate the range and IQR for the number of floors for these 20 buildings.

f) Use the outlier test to determine whether or not there are any outliers for the number of floors. Do your results match your results in part (b)?
7) Several game critics rated the Wow So Fit game, on a scale of 1 to 100 (100 being the highest). The results are presented in this stem plot:

<table>
<thead>
<tr>
<th>Critics Ratings</th>
<th>Key: 6</th>
<th>7 = 67</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 2 3 5 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 2 2 4 6 7 7 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 0 0 2 5 6 7 9 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 1 1 2 2 3 5 6 6 7 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 0 2 2 2 6 7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Calculate the five number summary for the Wow So Fit data.

b) Construct a box plot for the data.

c) Describe this distribution. (S.O.C.C.S.)

d) Make a statement, in context, about what the "box" part of the box plot tells us.

---

Review Exercises

8) Read each of the criticisms below regarding game ratings and determine whether the person making the statement is questioning the validity, the reliability, or the presence of bias in the test. Explain.

a) “The game critics get free copies of the games for their families. So, these ratings are inflated.”

b) “The game critics have no set guidelines on which to use to critique the games. So, these ratings are meaningless.”

c) “The game critics may give different ratings to the same game, when asked at different times. So, these ratings are inconsistent.”

---

9) Construct a tree diagram that shows all possible outcomes, in regard to sex of the children, of a family with three children.

---

10) Assuming that P(boy) = P(girl) = 0.5, find the following probabilities for a family with three children.

a) P(boy, girl, then boy)

b) P(exactly two girls)

c) P(at least one boy)
5.6 Numerical Data: Comparing Data Sets

Learning Objectives

- Construct parallel box plots
- Construct back-to-back stem plots
- Compare more than one set of numerical data in context

Parallel Box Plots

Parallel box plots (also called side-by-side box plots) are very useful when two or more numerical data sets need to be compared. The graphs of parallel box plots are plotted, parallel to each other, along the same number axis. This can be done vertically or horizontally and for as many data sets as needed.

Example 1

The figure shows the distributions of the temperatures for three different cities. By graphing the three box plots along the same axis, it becomes very easy to compare the temperatures of the three cities. What are some conclusions that can be drawn about the temperatures in these three cities?

http://www.mathsheetscenter.com
Solution

Here are some conclusions, based on these graphs that might be made. Think S.O.C.C.S! Be sure to compare the distributions to one another and use statistics to support your observations.

- Quartile 1 for City 2 is higher than the quartile 3 in City 1 and the median in City 3. Also, the minimum temperature in City 2 is about the same as the median for the other two cities.
- City 2 is generally warmer than both of the other cities. Cities 1 and 3 have nearly the same median temperature, around 60° to 63° while the median temperature in City 2 is around 82°.
- City 3 has a much larger range in temperatures (35° to 85°), than City 1 (45° to 75°) or City 2 (62° to 95°). The temperature in City 3 varies the most of these three cities.
- The temperature distributions in all three cities are fairly symmetrical and none have any outliers.

Comparing Numerical Data Sets

When you are given numerical sets of data for more than one variable and asked to compare them, it will be necessary to construct graphical representations for each data set. In order to compare them to one another the scales must match. When comparing more than one box plot, we construct parallel box plots. When using histograms, we can match the horizontal and vertical scales so that the separate histograms can ‘line up’. Dot plots will work the same way as histograms. Such comparisons are also possible when working with stem plots. Two sets of numerical data can simply share the stems in the middle, with one set’s ‘leaves’ going to the right and the other set’s ‘leaves’ going to the left. On both sides of the plot, the ‘leaves’ will go in numerical order out. Plots like these are called back-to-back stem plots.

Once you have constructed any of these types of comparative graphical representations (using the same scale,) you can make observations about how the data sets are the same and how they are different. Just as we have been doing up to this point, those comparisons should be done in context. The observations made might address the shapes of the distributions and whether or not any outliers are present. It is important to compare the centers of the distributions (means, medians, or modes). The spread for each of the distributions should also be addressed (ranges, IQRs, or standard deviations).

Example 2

A teacher gave the same physics exam to her two sections of physics. She has been wondering whether the first period and fifth period classes are learning the same amount as one another. She constructed this back-to-back stem plot to compare the test scores for the two different classes.

a) Calculate the five number summary for both classes.
b) Calculate the mean and standard deviation for both classes.
c) Compare the two classes’ test scores in context.

<table>
<thead>
<tr>
<th>Class A Leaves</th>
<th>Stems</th>
<th>Class B Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 0</td>
<td>6</td>
<td>0 0</td>
</tr>
<tr>
<td>5 0</td>
<td>7</td>
<td>0 1 3 3 5 6 7</td>
</tr>
<tr>
<td>6 4 4</td>
<td>8</td>
<td>4 5 6</td>
</tr>
<tr>
<td>6 4 4 2 1 0</td>
<td>9</td>
<td>1 2</td>
</tr>
</tbody>
</table>
| 0 0            | 10    | Source: http://www.basic-mathematics.com
Solution

a) The numbers in the stem plots are already in order, so these statistics could be found either by hand or with a graphing calculator.

The five-number summary for Class A is {60, 75, 90.5, 94, 100}.

The five-number summary for Class B is {60, 71, 75.5, 85, 92}.

b) These statistics are most efficiently found using a graphing calculator.

For Class A, the mean is $\bar{x} \approx 85.71$ points and the standard deviation is $S \approx 12.64$ points.

For Class B, the mean is $\bar{x} \approx 76.64$ points and the standard deviation is $S \approx 10.05$ points.

c) Comparison:

Overall, Class A did better on this test than Class B did. Class A’s scores on this test are skewed to the left, but Class B’s scores are somewhat skewed to the right. Neither class has any outliers among the test scores. Class A has a mean score of about 9 points higher (85.7 compared to 76.6) and a median score of 15 points higher (90.5 compared to 75.5). The overall range for the two classes is fairly similar, but the Class A students’ scores were less consistent. The ranges (40 and 32), IQRs (19 and 14), and standard deviations (12.6 and 10.1), all show that Class B’s scores are less spread out than Class A’s scores.

Example 3

An oil company claims that its premium grade gasoline contains an additive that significantly increases gas mileage. They conducted the following experiment in an effort to prove their claim. They selected 15 drivers who all drove the same make, model, and year of car. Starting with an empty gas tank, each car was filled with 45L of one of the two types of gasoline (selected in a random order). The driver was asked to drive until the fuel warning light came on. The number of kilometers was recorded and then the car was filled with the other type of gasoline (whichever they had not yet used). The process was repeated and the number of kilometers was again recorded. The results below show the number of kilometers each car traveled.

<table>
<thead>
<tr>
<th>Regular Gasoline</th>
<th>Premium Gasoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>640  570  640   580  610</td>
<td>659  619  639  629  664</td>
</tr>
<tr>
<td>540  555  588   615  570</td>
<td>635  709  637  633  618</td>
</tr>
<tr>
<td>550  590  585   587  591</td>
<td>694  638  689  589  500</td>
</tr>
</tbody>
</table>

Display each set of data to examine whether or not the claim made by the oil company is true or false.
Solution

**Order the data** – List the values in order for each set of data.

<table>
<thead>
<tr>
<th>Regular Gasoline</th>
<th>Premium Gasoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>540</td>
<td>500</td>
</tr>
<tr>
<td>550</td>
<td>589</td>
</tr>
<tr>
<td>555</td>
<td>618</td>
</tr>
<tr>
<td>570</td>
<td>619</td>
</tr>
<tr>
<td>570</td>
<td>629</td>
</tr>
<tr>
<td>580</td>
<td>633</td>
</tr>
<tr>
<td>585</td>
<td>635</td>
</tr>
<tr>
<td>587</td>
<td>637</td>
</tr>
<tr>
<td>588</td>
<td>638</td>
</tr>
<tr>
<td>590</td>
<td>639</td>
</tr>
<tr>
<td>591</td>
<td>659</td>
</tr>
<tr>
<td>610</td>
<td>664</td>
</tr>
<tr>
<td>615</td>
<td>689</td>
</tr>
<tr>
<td>640</td>
<td>694</td>
</tr>
<tr>
<td>660</td>
<td>709</td>
</tr>
</tbody>
</table>

5-# summaries – Determine the five number summary for each set of data separately. Be sure to report your five number summary whether you are asked to report it or not.

<table>
<thead>
<tr>
<th>Five-Number Summary</th>
<th>Regular Gasoline</th>
<th>Premium Gasoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smallest #</td>
<td>540</td>
<td>500</td>
</tr>
<tr>
<td>Q₁</td>
<td>570</td>
<td>619</td>
</tr>
<tr>
<td>Median</td>
<td>587</td>
<td>637</td>
</tr>
<tr>
<td>Q₃</td>
<td>610</td>
<td>664</td>
</tr>
<tr>
<td>Largest #</td>
<td>660</td>
<td>709</td>
</tr>
</tbody>
</table>

**Box plots** – Mark your axis so that it covers the entire range needed. Here, the smallest value is 500 and the largest value is 709. Graph each box plot along the same axis so that they are parallel to each other. This allows for the two data sets to be easily compared to one another.

**Key:** blue (top graph) = regular gasoline  
gold (bottom graph) = premium gasoline

Conclusions – Make comparisons by looking for similarities and/or differences between the two distributions. Remember your S.O.C.C.S!

Based on this experiment, the number of kilometers that the cars were able to travel on the premium gasoline was generally greater than the number of kilometers that the same cars were able to travel with on regular gasoline. The median number of kilometers for premium gasoline was 637 compared to 587 for regular gas. The first quartile for premium was higher than the third quartile for regular. Also, 25% of those with the premium gasoline went further than all of those using regular gasoline. The distribution for the regular fuel is slightly skewed to the right, and it doesn’t have any outliers. The premium distribution is strongly pulled to the left toward one outlier on the low end (500 km). Based on these results, it appears that the additive in the premium gasoline does improve gas mileage for this make and model of car. Further tests should be done on other types of vehicles.
Example 4

The heights of a group of students are all included in the first histogram. The second histogram only contains the data from the male students and the third is a graph of the heights of only the girls. Explain what these histograms show.

Solution

The range of heights of all students in this group is approximately 20 inches. However, the female heights only have a range of about 11 inches while the male heights have a range of about 13 inches. The female height distribution is the most symmetrical of all three. There is one male whose height is a high outlier and there are no outliers for the females. The median height for the entire group is around 70 inches, for males it is slightly higher around 72 inches, and for females it is around 65 inches. In general, the female students tend to be shorter than the male students.
Problem Set 5.6

Exercises

1) Compare the %Daily Value for Total Fat to the %Daily Value for Saturated Fat for these McDonald’s® sandwiches.

a) Calculate the five-number summary for both %Daily Values.

b) Construct parallel box plots for both.

c) Make at least four observations to compare these two distributions.

<table>
<thead>
<tr>
<th>Nutrition Facts</th>
<th>Serving Size</th>
<th>Calories</th>
<th>Calories from Fat</th>
<th>Total Fat (g)</th>
<th>% Daily Value** Total Fat (g)</th>
<th>% Daily Value** Saturated Fat (g)</th>
<th>% Daily Value** Trans Fat (g)</th>
<th>Cholesterol (mg)</th>
<th>% Daily Value** Sodium (mg)</th>
<th>% Daily Value** Total Fat (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Mac</td>
<td>7.4 oz (211 g)</td>
<td>130</td>
<td>240</td>
<td>27</td>
<td>120</td>
<td>42</td>
<td>10</td>
<td>85</td>
<td>28</td>
<td>960</td>
</tr>
<tr>
<td>Quarter Pounder® with Cheese+</td>
<td>7.1 oz (202 g)</td>
<td>120</td>
<td>240</td>
<td>26</td>
<td>120</td>
<td>41</td>
<td>12</td>
<td>61</td>
<td>15</td>
<td>950</td>
</tr>
<tr>
<td>Bacon Clubhouse Burger</td>
<td>6 oz (162 g)</td>
<td>120</td>
<td>360</td>
<td>62</td>
<td>120</td>
<td>40</td>
<td>13</td>
<td>75</td>
<td>15</td>
<td>1100</td>
</tr>
<tr>
<td>Quarter Pounder Bacon &amp; Swiss® Ranch®</td>
<td>5.3 oz (153 g)</td>
<td>110</td>
<td>250</td>
<td>31</td>
<td>10</td>
<td>64</td>
<td>1.5</td>
<td>100</td>
<td>35</td>
<td>1160</td>
</tr>
<tr>
<td>Quarter Pounder Bacon &amp; Cheese+</td>
<td>5 oz (135 g)</td>
<td>100</td>
<td>260</td>
<td>39</td>
<td>10</td>
<td>65</td>
<td>1.5</td>
<td>100</td>
<td>29</td>
<td>960</td>
</tr>
<tr>
<td>Quarter Pounder Deluxe+</td>
<td>5.6 oz (184 g)</td>
<td>140</td>
<td>250</td>
<td>27</td>
<td>11</td>
<td>54</td>
<td>1.5</td>
<td>95</td>
<td>29</td>
<td>960</td>
</tr>
<tr>
<td>Double Quarter Pounder with Cheese+</td>
<td>10 oz (253 g)</td>
<td>280</td>
<td>380</td>
<td>43</td>
<td>19</td>
<td>96</td>
<td>2.6</td>
<td>160</td>
<td>63</td>
<td>1200</td>
</tr>
<tr>
<td>Hamburger</td>
<td>3.5 oz (98 g)</td>
<td>240</td>
<td>70</td>
<td>12</td>
<td>3</td>
<td>15</td>
<td>0.5</td>
<td>30</td>
<td>10</td>
<td>480</td>
</tr>
<tr>
<td>Cheeseburger</td>
<td>4 oz (113 g)</td>
<td>150</td>
<td>100</td>
<td>11</td>
<td>5</td>
<td>27</td>
<td>0.5</td>
<td>45</td>
<td>15</td>
<td>680</td>
</tr>
<tr>
<td>BBQ Ranch Burger</td>
<td>4 oz (113 g)</td>
<td>140</td>
<td>130</td>
<td>13</td>
<td>6</td>
<td>29</td>
<td>0.5</td>
<td>50</td>
<td>16</td>
<td>673</td>
</tr>
<tr>
<td>Grilled Onion Cheddar</td>
<td>4 oz (113 g)</td>
<td>200</td>
<td>110</td>
<td>13</td>
<td>6</td>
<td>29</td>
<td>0.5</td>
<td>45</td>
<td>16</td>
<td>640</td>
</tr>
<tr>
<td>Double Cheesburger</td>
<td>5.7 oz (161 g)</td>
<td>190</td>
<td>21</td>
<td>22</td>
<td>10</td>
<td>52</td>
<td>1</td>
<td>90</td>
<td>30</td>
<td>1040</td>
</tr>
<tr>
<td>McDouble</td>
<td>5.2 oz (147 g)</td>
<td>180</td>
<td>150</td>
<td>17</td>
<td>6</td>
<td>40</td>
<td>1</td>
<td>75</td>
<td>25</td>
<td>840</td>
</tr>
<tr>
<td>Jalapeno Double</td>
<td>5.6 oz (158 g)</td>
<td>210</td>
<td>23</td>
<td>25</td>
<td>1</td>
<td>90</td>
<td>27</td>
<td>1050</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Bacon McDouble</td>
<td>5.7 oz (161 g)</td>
<td>200</td>
<td>22</td>
<td>24</td>
<td>10</td>
<td>49</td>
<td>1</td>
<td>90</td>
<td>30</td>
<td>1110</td>
</tr>
</tbody>
</table>

2) The heights of the students in a statistics class were all measured to the nearest inch. The results are presented in this back-to-back stem plot. Notice that it is also a split stem plot. The girls’ heights are ordered out to the right side of the graph and the boys’ heights are ordered out to the left side of the graph.

a) Compute the standard deviation, the range, and the IQR for both girls and boys.

b) Compare the spread for the two groups, based on your answers to (a), in context.

c) Compute the mean, median, and mode for both boys and girls.

d) Compare the center for the two groups, based on your answers to (c), in context.

e) Compare the shape of the distributions, based on the graph, in context.

3) Compare the results of the Probability and Statistics District Common Assessment for two statistics classes.

**CLASS 3:**
45, 37, 14, 42, 24, 33, 41, 16, 39, 24, 38, 35, 35, 32, 51, 46, 30, 42, 25, 37, 37, 19, 26, 23, 28, 38, 16, 35, 21

**CLASS 4:**

a) Construct back-to-back stem plots (use split-stems) for these two classes.

b) Calculate the five number summaries for both classes.

c) Calculate the following statistics for both classes: mean, standard deviation, mode, range, and IQR.

d) Compare and contrast the two distributions. This should be in context and you should make at least four distinct observations.
4) The number of home-runs during a season is one of the statistics recorded about baseball players. The following table has the number of home-runs (over many seasons) for several of the best hitters in baseball. Compare the home-run hitting performance of these exceptional baseball players.

- **Babe Ruth**: 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22
- **Mark McGwire**: 49, 32, 33, 39, 22, 42, 9, 9, 39, 52, 58, 70, 65, 32, 29
- **Barry Bonds**: 16, 25, 24, 19, 33, 25, 34, 46, 37, 33, 42, 40, 37, 34, 49, 73, 46
- **Roger Maris**: 13, 23, 26, 16, 33, 61, 28, 39, 14, 8

a) Calculate the following statistics for all four players:
   \[ \bar{x} = \_\_\_\_\_\_ \quad s_x = \_\_\_\_\_\_\_ \quad \text{IQR} = \_\_\_\_\_\_ \quad \text{5# summary} = \{\_\_\_\_, \_\_\_\_, \_\_\_\_, \_\_\_\_, \_\_\_\_\_\_\_\} \]

b) Construct Parallel Box Plots for the four players. Be sure to use the same scale for all four graphs and to label each graph.

c) Test for outliers, for all four players, using the 1.5*IQR criterion. Show your work.

d) Compare and contrast the four distributions. This should be in context and you should make at least four distinct observations.

5) The following box plots show the average miles per gallon (city) for various types of vehicles. Comment on what these parallel box plots show. This should be in context and include at least 4 distinct observations. Any dots represent outliers for that data set.

---

**Box Plot – Average city MPG by Car Type**

6) Refer to the four dot plots to answer the questions that follow.

Graph I
Graph II
Graph III
Graph IV

(a) Identify the overall shape of each distribution.
(b) How would you characterize the center(s) of each of these distributions?
(c) Name at least two statistics that would most likely be the same for all four of these distributions.
(d) Which of these distributions has the smallest standard deviation? Which of these distributions has the largest standard deviation? Explain.
(e) For which of these distributions would it be appropriate to use the mean and standard deviation as numerical summaries? For which would the five-number summary be more appropriate?
5.7 Chapter 5 Review

Chapter 5 Summary

In this chapter, we have learned that when working with a set of data it is important to choose an appropriate type of graphical display so that we can see what the data looks like. Bar graphs and pie charts are useful ways to display categorical data. Time plots are line graphs that help us to see how a given variable has changed over a specified period of time. When working with numerical data we have learned how to make dot plots, stem plots, histograms, and box plots. It is also possible to make graphs so that comparisons can be made between more than one data set. Back-to-back stem plots and parallel box plots are two such types of graphs.

Our next step was to analyze the data sets by calculating numerical statistics. We began by looking at the mean, median, and mode. These statistics are measures of central tendency and give us an idea of where the center of the data set is located. The range, interquartile range (IQR), and the standard deviation are all measures of the spread or variability of a data set. We have calculated the five-number summary, which divides a set of data into quarters and allows us to construct a box plot.

Once the graphs were constructed and the statistics were calculated, we learned to describe what the data showed. When describing a numerical set of data, in addition to explaining where the center is located and what the spread is, we also describe the shape of the distribution and whether or not any outliers are present in the data. The shapes that we focused on are symmetrical distributions and skewed distributions, remembering that the direction of the skew is toward the tail or outliers. We learned to make appropriate conclusions and comparisons that are based on the data, graphs, and statistics. Statisticians should avoid opinions and judgment statements as much as possible.

We learned that the 1.5*(IQR) Criterion can be used to determine whether or not any data values are outliers. In addition, we found that the mean and standard deviation are easily influenced by unusual values or skewed data sets. We avoid using these statistics when working with data that contains outliers or is skewed.

Review Exercises

1) Multiple-Choice: Which of the following can be inferred from this histogram?
   a) The mode is 12.5
   b) mean < median
   c) median < mean
   d) The distribution is skewed left.
   e) None of the above can be inferred from this histogram.
2) The owner of a small company is trying to determine whether he should go with a different company for his shipping needs. He needs to analyze the weights of the packages that his company ships out. This graph shows the distribution of the weights of packages that were shipped during the last month.

a) Calculate the mean, standard deviation, mode, and range for this data. Use a calculator.

b) Determine the five number summary for this data. Construct a box plot for this data.

c) Which of these two graphs – the dot plot or the box plot – is more informative? Explain.

d) The owner figures that he will save money with the new company on any package that weighs less than 16.75 ounces. What percent of packages will he lose money on (those weighing more than 16.75 ounces)?

3) After some bullying issues were brought to light in a big high school, a committee was formed to study the issue. A questionnaire was designed that contained several questions related to bullying and safety. A stratified random sample was selected that included students from all four grade levels. The table that follows shows the responses to one of the questions on the questionnaire.

<table>
<thead>
<tr>
<th>Student Responses to the Question:</th>
</tr>
</thead>
<tbody>
<tr>
<td>“I Feel Safe at School”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n = sample size</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1003</td>
<td>133</td>
<td>274</td>
<td>529</td>
<td>67</td>
</tr>
</tbody>
</table>

a) Create a pie chart that shows the results of this survey question. Be sure to include labels, percentages, a title, and a key if needed. You may do this by hand or with technology.

b) Describe what the graph shows in context. Be sure to include percentages to support your observations.

c) Comment on whether or not the committee should be concerned. Explain.
In questions 4-7, match the distribution with the choice of the correct real-world situation that best fits the graph.

4)

5)

6)

7)

a) Andy collected and graphed the heights of all the 12th grade students in his high school.
b) Brittany asked each of the students in her statistics class to bring in 20 pennies selected at random from their pocket or piggy bank. She created a plot of the dates of the pennies.
c) Maya asked her friends what their favorite movie was this year and graphed the results.
d) Jeno bought a large box of doughnut holes at the local pastry shop, weighed each of them, and then plotted their weights to the nearest tenth of a gram.
Questions 8 - 17 are multiple-choice questions. Select the best answer from the choices given.

8) Which of the box plots on the right matches the histogram on the left?

9) Identify the 5 number summary for this set of numbers:

<table>
<thead>
<tr>
<th>12,356</th>
<th>16,564</th>
<th>15,684</th>
<th>12,358</th>
<th>15,987</th>
<th>13,556</th>
<th>18,564</th>
<th>18,965</th>
<th>19,683</th>
<th>18,432</th>
<th>18,563</th>
<th>19,352</th>
</tr>
</thead>
</table>

a) \{12,356; 14,600; 17,498; 18,000; 19,683\}
b) \{12,356; 14,620; 17,498; 18,764.5; 19683\}
c) \{12,356; 14,650.5; 17,498; 18,700.5; 19683\}
d) \{12,356; 14,683; 17,500; 18,800; 19683\}
e) \{12,356; 14,695.5; 17,900; 18,888; 19683\}

10) Thirty students took a statistics examination having a maximum of 50 points. The grade distribution is given in the stem-and-leaf plot:

The median grade is equal to:

<table>
<thead>
<tr>
<th>0</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>225</td>
</tr>
<tr>
<td>2</td>
<td>01335889</td>
</tr>
<tr>
<td>3</td>
<td>00136679</td>
</tr>
<tr>
<td>4</td>
<td>02244478</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

a) 30.5
b) 30.0
c) 25.0
d) 28.5
e) 44.0
11) Ms. Davis conducted a survey of the 44 students in her stats classes and asked how tall each student was in inches. The five-number summary of the data is (57, 64, 67, 69, 79).

Approximately how many people are shorter than 64 inches tall?

a) 8
b) 21
c) 22
d) 11
e) 18

12) In which scenario(s) would it be better to use the 5-number summary versus the mean and standard deviation?

a) a graph that is skewed
b) a graph that is fairly symmetric
c) a graph that is symmetric but has several high outliers
d) Both choice (b) and (c)
e) Both choice (a) and (c)
f) All of (a), (b) and (c)

13) Suppose the lowest score on an English exam was 35% and the highest score was 90%. If the teacher of the class was to examine her students’ test scores, which shape of the distribution would she prefer to see?

a) skewed to the right
b) skewed to the left
c) fairly symmetric
d) none of the above

14) Several people were surveyed as they were leaving a movie theatre. Among other things, they were asked how much money they had spent. Their answers were: $14, $17.50, $16, $16, $19.25, $12.75, $16, $37.75, $13.50 and $17. It was later discovered that the person who answered “$37.75” actually spent $17.75. Which of the following would not change as a result?

a) the box plot
b) the mean & the mode
c) the median & the mode
d) the standard deviation
e) they all change
15) What does the following five-number summary below tell you about the shape of the distribution?
\{15, 16.7, 18.7, 22.3, 32\}

a) It is skewed to the right.
b) It is skewed to the left.
c) It is symmetric.
d) It is uniform.
e) We cannot determine anything about the shape of the distribution.

16) According to the 1.5*(IQR) Criterion, what are the two cut-off values for determining whether the data set in question #15 contains any outliers?

a) 7.5 & 24.6
b) 7.7 & 27.9
c) 11.3 & 23.9
d) 4.5 & 24.1
e) 8.3 & 30.7

17) A class survey was conducted to determine students’ preferences. One question asked about each student’s favorite sport to watch on TV. Of those surveyed, 9 said football, 12 said hockey, 5 said basketball, 6 said baseball, and 3 said some other sport. What would the central angle be for “hockey” in a pie chart of this data?

a) 65°
b) 123°
c) 90°
d) 34°
e) 111°

18) The AHS Tornadoes and the BHS Bengals are big rivals! Every year students try to prove that their school is better at sports than the other school. The table to the right shows the number of points scored by each school’s basketball team during the last 15 games played against other teams this year.

a) Construct a back-to-back stem plot for the data.
b) Calculate the five number summary, mean and standard deviation for both teams.
c) Construct parallel box plots for the data.
d) Compare the two distributions. This should be done in context and include at least three distinct comparisons.
e) What other information would you like to know when comparing these two basketball teams? Explain.

Table 5.10:
<table>
<thead>
<tr>
<th>Tornadoes</th>
<th>Bengals</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>74</td>
</tr>
<tr>
<td>90</td>
<td>81</td>
</tr>
<tr>
<td>71</td>
<td>73</td>
</tr>
<tr>
<td>64</td>
<td>63</td>
</tr>
<tr>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>63</td>
<td>84</td>
</tr>
<tr>
<td>60</td>
<td>92</td>
</tr>
<tr>
<td>72</td>
<td>38</td>
</tr>
<tr>
<td>48</td>
<td>77</td>
</tr>
<tr>
<td>59</td>
<td>84</td>
</tr>
<tr>
<td>72</td>
<td>95</td>
</tr>
<tr>
<td>62</td>
<td>66</td>
</tr>
<tr>
<td>57</td>
<td>70</td>
</tr>
<tr>
<td>64</td>
<td>68</td>
</tr>
<tr>
<td>49</td>
<td>72</td>
</tr>
</tbody>
</table>
19) The following two graphs are based from the US Census Bureau in 2008. Note that ‘per capita’ means per person. The dots represent actual data values, and the red curves represent models that can be used to predict future trends. Study the two graphs and answer the questions that follow.

a) What type of graphs are these?

b) Describe the trend that each graph shows separately. This should be in context.

c) Notice that the horizontal scales are the same. Compare and contrast the trends that are shown in the two different graphs in context.

d) Approximately how many cell phones were there per person in 1997? In 2005? How many will there be, if the trend continues as the model indicates, in 2018?

e) Approximately how many landlines were there per person at the peak? What year did this occur? How many landlines did the model predict per person in 2015?

20) The table that follows shows the percent of people, 25 years and older, who are high school graduates for several states in the central United States according to the 2010 U. S. Census website.

a) Construct a histogram (use Xmin = 79%, and bin width = 1%).

b) Calculate the five number summary.

c) Identify any outliers. Use the outlier test.

d) Accurately sketch a box plot.

e) Calculate the range, IQR, and mode.

f) Calculate the mean and standard deviation.

g) Compare the mean and the median. (i.e. Which is larger? How different are they?)

h) In this case would the 5-number summary or the mean & standard deviation be more appropriate? Why?

i) Describe the distribution. Be thorough! Don’t forget your S.O.C.C.S! (Shape, Outliers, Center, Context, & Spread)

j) Among this Census data, where does Minnesota fall?
21) An employer in Minneapolis was interested in determining how much money his employees were spending on parking each week. An SRS of 50 employees was selected to complete a questionnaire about parking. Several questions were asked including where they park, how much they spend per week, how often they have difficulty finding spots, and if they pay daily, weekly, or monthly. The following table is the average weekly expenditure for parking for this sample of 50 employees.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>40</td>
<td>22</td>
<td>22</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>13</td>
<td>18</td>
<td>50</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>22</td>
<td>25</td>
<td>22</td>
<td>10</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>22</td>
<td>22</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>23</td>
<td>30</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>40</td>
<td>22</td>
<td>29</td>
<td>19</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>14</td>
</tr>
<tr>
<td>21</td>
<td>14</td>
<td>22</td>
<td>21</td>
<td>35</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>20</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Construct a split-stem plot.

b) Calculate the five number summary.

c) Identify any outliers. Use the outlier test.

d) Accurately sketch a box plot. Be sure to scale and label the graph.

e) What is the range? The IQR? The mode?

f) Calculate the mean and standard deviation.

g) Compare the mean and the median.

h) In this case would the 5-number summary or the mean & standard deviation be more appropriate? Why?

i) Describe the distribution. Be thorough! Remember your S.O.C.C.S!

**Image References:**
- Cars. http://www.icoachmath.com
- School Lunch pictogram. http://alex.state.al.us
- Stem plot example. http://www.basic-mathematics.com/
- Weight of Jessica graph. http://stat.psu.edu/
- Crowded stem plot. http://illuminations.nctm.org/
- Three histogram example. http://classes.cec.wustl.edu
- Package weight graph. http://flylib.com
Chapter 6 – Analyzing Bivariate Data

Introduction

In chapter 5 we learned how to analyze and describe univariate, or single-variable data. We explored ways to present our data visually with graphs and charts and how to analyze our data with numerical statistics. Also, we described our findings verbally and in context. Now we will be analyzing bivariate numerical data. This means that two numerical values have been collected about each individual. Such bivariate data is often given in a table or listed as ordered pairs. We will construct appropriate graphs, calculate numerical statistics and equations, and describe the relationship between the two variables in context. The purpose will be to explore whether or not a relationship or association exists between the two numerical variables. If an association does exist, statistics can be used to predict one variable based on the other variable.

6.1 Displaying Bivariate Data

Learning Objectives
- Construct and interpret scatterplots
- Identify explanatory and response variables
- Describe bivariate distributions in context—including strength, outliers, form and direction

Scatterplots

Scatterplots are graphs that represent a relationship between two variables. Two numerical values are measured about each individual being studied. When these two values become ordered pairs that are graphed on a coordinate plane, the resulting graph is called a scatterplot. We often suspect that one of these variables might explain, cause changes in, or help to predict the other variable. The explanatory variable is the variable that we believe may explain or affect the other variable. The explanatory variable is plotted along the x-axis. The response variable is the variable we believe may respond to, or be affected by, the explanatory variable. The response variable is plotted along the y-axis. The explanatory variable is often referred to as the independent variable and the response variable is referred to as the dependent variable. Even though we often look for an explanatory-response relationship between the two variables, we can create a scatterplot even if no such relationship exists.
Example 1

State whether or not you suspect that there will be an explanatory-response relationship between each of the following pairs of data. If yes, identify the explanatory and response variables.

a) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam. At the end of the exam, each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made a scatterplot to examine the data.

b) A different professor wanted to see whether or not there is an association between her students’ heights and their IQ scores. She gave each of her students an IQ test and had her TA (teaching assistant) measure each student’s height to the nearest inch. She constructed a scatterplot to examine the data.

Solution

a) It is reasonable to believe that the amount of studying does somehow have an effect on students’ exam scores. The explanatory variable is hours studying and the response variable is exam score. Often thinking in terms of a cause and effect relationship can help identify which variable is which. As a hint, try to determine if one of the variables comes first. If one variable does happen first, then it is most likely the explanatory variable. In our example, studying should come before the exam.

b) It is not reasonable to believe that there is an association between height and IQ scores. Neither of these variables comes before the other and neither would be useful in predicting the other. However, even though we do not believe that there is an explanatory-response relationship between these variables, we can still construct a scatterplot.
Example 2

The following table reports the recycling rates for paper and glass packaging for several individual countries. It would be interesting to see if there is a predictable relationship between the percentages of each material that countries recycle. Construct a scatter plot to examine the relationship. Treat percentage of paper packaging recycled as the explanatory variable.

<table>
<thead>
<tr>
<th>Country</th>
<th>% of Paper Packaging Recycled</th>
<th>% of Glass Packaging Recycled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonia</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>New Zealand</td>
<td>40</td>
<td>72</td>
</tr>
<tr>
<td>Poland</td>
<td>40</td>
<td>27</td>
</tr>
<tr>
<td>Cyprus</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>Portugal</td>
<td>56</td>
<td>39</td>
</tr>
<tr>
<td>United States</td>
<td>59</td>
<td>21</td>
</tr>
<tr>
<td>Italy</td>
<td>62</td>
<td>56</td>
</tr>
<tr>
<td>Spain</td>
<td>63</td>
<td>41</td>
</tr>
<tr>
<td>Australia</td>
<td>66</td>
<td>44</td>
</tr>
<tr>
<td>Greece</td>
<td>70</td>
<td>34</td>
</tr>
<tr>
<td>Finland</td>
<td>70</td>
<td>56</td>
</tr>
<tr>
<td>Ireland</td>
<td>70</td>
<td>55</td>
</tr>
<tr>
<td>Netherlands</td>
<td>70</td>
<td>76</td>
</tr>
<tr>
<td>Sweden</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>France</td>
<td>76</td>
<td>59</td>
</tr>
<tr>
<td>Germany</td>
<td>83</td>
<td>81</td>
</tr>
<tr>
<td>Austria</td>
<td>83</td>
<td>44</td>
</tr>
<tr>
<td>Belgium</td>
<td>83</td>
<td>98</td>
</tr>
<tr>
<td>Japan</td>
<td>98</td>
<td>96</td>
</tr>
</tbody>
</table>

Figure: Paper and Glass Packaging Recycling Rates for 19 countries

Solution

We will place the paper recycling rates on the horizontal axis because we are treating it as the explanatory variable. Glass recycling rates are then plotted along the vertical axis. Next, plot a point that shows each country’s rate of recycling for the two materials. Be sure to label your axes.

Notice that we do not always need to start at zero on either axis when making scatterplots.
Describing Bivariate Data

When we describe single variable data, we address several characteristics. We used the acronym S.O.C.C.S. to help us to remember to describe the shape, outliers, center, context and spread of a distribution. For bivariate situations, we will again need to remember to discuss several key characteristics of the data. The important characteristics to describe when looking at the relationship between two numerical variables will be strength, outliers, form and direction. We will do this in the context of the variables and individuals being compared. The acronym that will help us to remember what to include in our descriptions is: S.C.O.F.D. (Strength, Context, Outliers, Form and Direction).

When looking at a scatterplot, it is helpful to imagine drawing a line-of-best-fit through the data. A line-of-best-fit is a line that follows the trend of the data. It may go through some, all, or none of the actual points on the scatterplot. Do not actually draw such a line on your plot - just try to determine whether or not such a line would make sense, and if so, where it would fit. As you observe a scatterplot and imagine drawing such a line, you can ask yourself questions such as the following: How close to a line do the points lie? Would a curved pattern fit better? Are there points that would be far away from the line? Would the line have a positive or negative slope?

Strength

Once you have constructed a scatterplot, you can examine the strength of the relationship between the two variables. The strength refers to how closely the points form a pattern. The more closely the points fit a pattern, the stronger the relationship is between the variables. The more spread out and scattered the points are, the weaker the relationship. The first plot shows an extremely strong, linear pattern because the points form an obvious line. The second plot is more scattered so it is only moderately strong. The third plot does not show much of a pattern at all so it would be considered to have a weak association. Keep in mind that the association may be very strong, but not linear. We could find a very clear curved pattern in the data, for example. In the next section of this chapter we will learn about a statistic, called correlation, that measures the strength of the linear relationship between two variables.

In Example #2, the strength of the relationship between paper and glass recycling rates for these countries would be considered weak.
Context

Do not forget that the graph, the numerical values, and any equations, are all about some particular situation. Keep the context in mind when considering any bivariate situation. All of these elements should be described in the context of the variables and the individuals being examined. These graphs and statistics are not meaningless, they are about something!

In Example #2, the scatterplot explores the relationship between glass and paper recycling rates for several countries.

Outliers

When examining a scatterplot, look for any data values that do not fit the overall pattern of the rest of the data. An outlier will be a point that lies away from the rest of the data or one that seems to affect the strength of the relationship between the two variables. Some outliers will weaken the association between the variables. However, they often will not significantly change where a line-of-best-fit would be drawn. An influential point is an outlier that actually seems to influence the location of the line-of-best-fit. Imagine what the plot would look like without the point in question. If it would change the strength, then the point is an outlier. If it would change the slope of a line-of-best-fit, or where the line would be drawn, then the point is influential.

In Example #2, there seem to be some outliers. For example, Estonia and New Zealand each have a much lower paper recycling rate than their glass recycling rate. Without these data values, the relationship in the rest of the scatterplot would be stronger.

Form

Many scatterplots show a clear form or pattern. The first plot below shows a clearly linear pattern or form. It is easy to imagine drawing a line-of-best-fit through these points. The second plot shows a clearly curved form. A line would not make any sense, so this is non-linear. The third plot shows a great deal of scatter among the points with no clear pattern. It has no form whatsoever.
In example #2, the scatterplot for paper and glass recycling rates shows a very weak linear form. The relationship is very weak but we can still see in general, that as paper recycling rates increase, glass recycling rates also increase. If the two outliers were removed, the scatterplot would have a stronger linear pattern.

**Direction**

The direction of the graph is also important to mention. A graph that goes down to the right has a **negative association**. As the explanatory variable increases, the response variable decreases. The first plot below or on next page has a negative relationship between the variables. A graph that goes up to the right has a **positive association**. As the explanatory variable increases, the response variable also increases. The second plot shows a positive relationship between the variables. The third plot is an example of a graph that has neither a positive, nor a negative direction. If the relationship is linear and a line-of-best-fit is added to the graph, the slope of the line will be positive if the association is positive. Likewise, the line will have a negative slope if there is a negative linear association between the two variables.
S.C.O.F.D

When you describe the relationship between bivariate data there are several characteristics to include. The acronym S.C.O.F.D. will help you remember to describe the strength of the relationship, keep your description is in context, mention any outliers, describe the form, and state the direction of the graph.

Example 3

Consider the scatterplot to the right that shows the weights (pounds) and gas mileages (miles per gallon) for several cars.

a) Identify the explanatory and response variables.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).

Solution

a) The explanatory variable is the weight of the cars in pounds.

The response variable is the gas mileage of the cars in miles per gallon.

b) The relationship between these vehicles’ weights in pounds and gas mileage (mpg) is strong, negative, and linear. There are no clear outliers visible in the graph. As the weights of the vehicles increase, the gas mileages of the vehicles decrease.

Example 4

The scatterplot to the right shows the data collected by the professor who wanted to see whether or not there is an association between her students’ heights and their IQ scores. She gave each of her students an IQ test and had her TA measure each student’s height to the nearest inch. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).

Solution

There appears to be no relationship between height and IQ scores for these students. The graph has no form and no clear direction. Therefore, there are no outliers. The relationship has no strength. There is no pattern or trend between IQ scores and students’ heights.
Problem Set 6.1

Exercises

1) State whether or not you suspect that there will be an explanatory-response relationship between each of the following pairs of data. If yes, identify the explanatory and response variables.

   a) The number of semesters that students have been enrolled in college and the number of credits that they have earned.
   b) Student grades on a statistics test and their weights.
   c) Employees’ annual salaries and the number of years that the employee has been employed by the company.
   d) The number of applications downloaded on a student’s cell phone and the number of months that they have owned the cell phone.

2) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam (out of 100 points possible). At the end of the exam each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made the scatterplot shown to the right to examine the data. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).

3) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full (and her mother started to freak out). Her findings are listed in the table to the right.

   a) Identify the explanatory and response variables for this situation.
   b) Construct a scatterplot to show the results.
   c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).
4) Several brands of peanut butter were rated for quality. The following graph compares the price per ounce (in cents) and the quality rating (scale of 0 = lowest to 100 = highest) for each of these brands of peanut butter.

a) Identify the explanatory and response variables for this situation.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

5) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table:

<table>
<thead>
<tr>
<th># months owned machine</th>
<th># hours exercise per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

a) Identify the explanatory and response variables for this situation.

b) Construct a scatterplot to show the results.

c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

6) The following scatterplot shows the elevation and mean temperature for various locations in Nevada.

a) Identify the explanatory and response variables for this situation.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).
Review Exercises

7) If two cards are drawn from a standard deck of playing cards and laid face up on a table, what is the probability of getting two Queens?

8) A card is drawn from a standard deck. The card is put back, the deck is reshuffled, and another card is drawn. What is the probability of drawing two clubs?

9) A gum ball machine contains 14 pink, 7 blue, 9 white, and 11 green gumballs. A child buys two gumballs, one after the other. Find the following probabilities:
   a) $P(\text{blue, then green})$
   b) $P(\text{neither is pink})$
6.2 Correlation

Learning Objectives

- Understand the properties of the linear correlation coefficient
- Estimate and interpret linear correlation coefficients
- Understand the difference between correlation and causation
- Identify possible lurking variables in bivariate data
- Understand the effects outliers and influential points can have on correlation

The Correlation Coefficient

The correlation coefficient is a statistic that measures the strength and direction of a linear relationship between two numeric variables. The symbol for correlation is r, and r can take any value from -1 to +1. The correlation coefficient, r, tells us two things about the linear relationship between two variables, its strength and its direction. The direction of the relationship, positive or negative, is given by the sign of the r value. A positive value for r indicates that the relationship is positive (increasing to the right), and a negative r value indicates a negative relationship between the two variables (decreasing to the right). Bivariate data with a positive correlation tells us that as the explanatory variable increases, so does the response variable. Bivariate data with a negative correlation tells us that as the explanatory variable increases, the response variable decreases. A correlation of zero indicates neither of these trends.

The correlation coefficient also tells us about the strength of the linear relationship. It tells us how close the points are to forming a perfect line. A correlation (r-value) of exactly 1 or -1 has a perfect correlation. In other words the relationship will produce a scatterplot whose points lie in a perfectly straight line. An r-value of exactly +1 means that the relationship forms a perfect line with a positive slope and an r-value of exactly -1 means that the scatterplot will show a perfect line with a negative slope. The closer the correlation value is to either +1 or -1, the stronger the linear relationship is. As r gets closer to zero (either positive or negative), the weaker the linear relationship is. It is important to note that this is only measuring the linear relationship between the two variables. If the relationship shows a clear curved pattern for example, the correlation will tell us nothing about the strength of the relationship.
Here are some sample scatterplots with their correlation coefficients given:

\[ r = -1 \]
\[ r = -0.900 \]
\[ r = -0.536 \]
\[ r = -0.160 \]
\[ r = +0.803 \]
\[ r = +1 \]

We will be using either our calculator or a computer to calculate the correlation coefficient. The formula to calculate the correlation coefficient is quite tedious. It involves calculating the mean and standard deviation of all of the x-values and the mean and standard deviation of all of the y-values. It then compares the x-value of each ordered pair to the mean of x and every y-value to the mean of y (by subtracting and then dividing by the standard deviation), multiplies these newly calculated values, adds all of them, and divides by one less than the sample size. The correlation formula is shown on the next page, but we will be using technology rather than calculating by hand. See Appendix C for calculator instructions.
Example 1

Estimate the correlation coefficient for each of the following scatterplots.

Solution

Nevada: The correlation will be negative and fairly strong, so my estimate is $r \approx -0.85$.
Height & IQ: There seems to be no pattern to the graph, so my estimate is $r \approx 0$.

Properties of Correlation

When considering using correlation as a measure of the strength between two variables, you should construct and examine a scatterplot first. It is important to check for outliers, be sure that the relationship appears to be linear, be sure that your sample size is sufficient, and consider whether the individuals being examined were too much alike in some way to begin with. Thus, when examining correlation, there are four things that could affect our results: outliers, linearity, size of the sample, and homogeneity of the group.
An outlier, or a data point that lies outside of our overall pattern, can have a significant impact on the correlation. How great of an impact is determined by the sample size of the data set and by how much the outlier lies outside of the pattern of the data. The three plots below show scatterplots with their correlation coefficients \( r \). The first plot shows a positive and reasonably linear graph. Its correlation is \( r = 0.897 \), which is positive and fairly strong. The second plot shows the same data as plot one, with one outlier added in the upper left. Its correlation has dropped to \( r = 0.374 \), which is still positive, but much weaker. This demonstrates how outliers can bring the correlation closer to zero. However, some outliers can actually strengthen the correlation. This is demonstrated in the third plot, which shows the same data as the first with one outlier added in the upper right. With this outlier, the linear relationship becomes even stronger than the first plot with \( r = 0.973 \).

If the relationship is not linear, calculating the correlation coefficient is meaningless. It is only testing the linear relationship between the two variables. Imagine a scatterplot that shows a perfect parabolic relationship. We would know that there is a strong relationship between these two variables, but if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient might not always be the best statistic to use to understand the relationship between variables.

As we discussed in experimental design, a small sample size can be misleading. It can either appear to have a stronger or weaker relationship than is really accurate. The larger the sample, the more accurate of a predictor the correlation coefficient will be on the linearity of two variables.

When a group is too much alike in regard to some characteristics (homogeneous), the range of scores on either or both variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQ’s over 140) are sampled, we will most likely find a very low correlation between IQ and salary since most members will have a consistently high IQ, but their salaries will vary. This does not mean that there is not a relationship between IQ and salary. It simply means that the restriction of the sample limited the magnitude of the correlation coefficient.

Correlation is just a number and it has no units. A change in units of measurement will not impact the correlation. For example, suppose you calculated the correlation coefficient between height in inches and weight in pounds for a group of teenagers. If you later decided to convert the heights to centimeters or the weights to kilograms (or both), and then calculated the correlation coefficient again, you would have found that the value for ‘\( r \)’ did not change.
Lurking Variables

It is very important to know that a high correlation does not mean causation! Oftentimes, studies that show a high correlation between two variables will influence readers into thinking that one variable is the cause of the observed relationship. This is not always true! While in some situations we would agree that one variable is in fact causing the response in the other, it is important to remember that a high correlation simply does not prove that one variable is causing the other. The best way to prove such a direct cause-and-effect relationship is by carrying out a well-designed experiment. For example, smoking is strongly correlated with lung disease. Based upon much scientific evidence, we can now say that cigarette smoking is one cause of lung disease. However, this topic was highly debated for many years before the surgeon general announced that it was accepted that cigarette smoking causes lung cancer and emphysema. Many people refused to accept this for many years. People who stood to lose money if smoking was proven to be unsafe, suggested many other possible explanations. They suggested that it was simply a coincidence, or that all people who choose to smoke might have something else in common that was actually the cause of the lung disease, not the cigarettes. Because it was not ethical to experiment on humans in order to prove the direct cause-and-effect relationship, the debates went on for a long time.

Sometimes the relationship between variables is cause-and-effect, but many times it can be simply a coincidence that the two variables are highly correlated. It is also possible that some other outside factor, a lurking variable, is causing both variables to change. A situation where we have two variables that are both being impacted by some other outside lurking variable is called common response. For example, we can show a high correlation between the number of TV’s per household and the life expectancy per person among many countries. However, it makes no sense that TV’s cause people to live longer. Some lurking variable is playing a role here. It is likely that the economic status of the countries is causing both variables to change; more money means more TV’s and more money also means better health care. If a country is wealthy, it is much more likely to have citizens who own TV’s. Also, if a country is wealthy, it is much more likely to have good hospitals, roads, health education, and access to food and clean water. These all contribute to a longer life.
In some situations we will have two variables that are highly correlated, but we are unsure of the exact nature of the relationship. We may be unclear as to whether or not one is causing the other, if there is an outside factor impacting the response variable, or if there is some unknown lurking variable that is related in some other unknown way. Remember, lurking variables are not always obvious to the researchers. Such a situation is called confounding, because it is confusing to determine how the variables are related (if at all), and whether there may be some lurking variable and if it is related to the variables in question. The variables seem all mixed up and the relationship is unclear, even if highly correlated. An example of confounding is global warming. This is a highly debated topic in social media and web-blogs. Some people argue that human pollution is a major cause of the increase in CO2 and other greenhouse gases in the atmosphere. Others will argue that it is a part of a natural cycle that has normally occurred in our Earth’s history. Still some may think both explanations are at work. This is an example of confounding because there is confusion about the cause of global warming.

Finally, don’t forget that some relationships are occurring completely by chance, and their high correlation is then just a coincidence. For example, if you researched divorce rates and gas prices over the past 50 years you may note that both have gone up. A scatterplot comparing divorce rates and gas prices would show a strong positive relationship. The correlation would likely be a high, positive value. However, it makes no sense that divorce rates are causing high gas prices. It also is unlikely that there exists a common response or some form of confounding. So in this case, we would say that this is a relationship that is best explained by sheer coincidence.

Example 2
Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. Consider whether common response, confounding, or coincidence may be involved.

a) It has been shown that cities with more police officers also have higher numbers of violent crimes. Does this mean that more police officers are causing more violent crimes to occur?

b) Over the past 25 years, the percent of parents using car-seats has increased significantly. During this same time period, the rate of DUI arrests has also increased significantly. These two variables, when graphed, show a very high, positive correlation. Does this mean that car-seat use is causing DUI’s to increase?

c) A recently published study claimed that, “Teens who use social media a lot [are] more likely to try sex, drugs, and alcohol.” Does this mean that social media use causes teens to try sex, drugs and alcohol? Could we then limit teen behaviors such as these by eliminating social media?

Solutions

a) It makes no sense that the number of police officers would be causing the violent crime to occur. It is much more likely that this situation is reversed. Communities with high numbers of violent crimes need higher numbers of police officers. It is also probable that both variables increase in cities with higher populations. Due to the fact that we can think of more than one possible lurking variable and it is difficult to know how all of these variables actually relate, we would say that this is an example of confounding. The variables in question and the lurking variables are all mixed up.
b) It is clearly ridiculous to think that car-seat use is causing an increase in the rate of DUI’s. It also makes no sense that DUI’s cause car-seats to be used. It could be argued that this is simply a coincidence that these variables are both increasing.

Another possible argument is that there has been an increase in enforcement of laws for both over this time period. The awareness of the dangers of both have increased over the past 25 years, so maybe this is an example of common response.

A third explanation may be that many factors that have contributed to the increase of both, so perhaps this is an example of confounding. Our only certainty is that this is not cause-and-effect. Analysis of these sorts of situations can be very tricky!

c) It is unlikely that social media is actually the cause of these behaviors. There are most likely some lurking variables that are contributing factors. One probable lurking variable, when it comes to teenagers, is the parents. Perhaps this is an example of a common response to parents who are not very involved in their teens’ lives. Parents who are not very involved would not be aware that their teen is using social media too much and would also not be aware of what choices their teen is making during his or her free time. Perhaps teens that spend a lot of time unsupervised would be more likely to use social media and would also be more likely to try sex, drugs, and alcohol. All of these behaviors might be a common response to not having parents who prohibit or limit teens from doing these things. Eliminating social media would likely have little to no impact on other teen behaviors.

**Multimedia Links:**

**Calculating Correlation on the Internet,**

There are several websites where you can enter in data points and find correlations. You might find the two links below helpful in your understanding of correlation.


http://bcs.whfreeman.com/tps5e/default.asp#923932__929340__

Another, more lighthearted example of Correlation ≠ Causation can be found at http://www.exrx.net/ExInfo/Pickles.html which discusses the evil of the pickle.

**For a better understanding of correlation try these fun links below,**

http://www.istics.net/stat/Correlations

http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html
Problem Set 6.2

Exercises

1) What are the two things that the correlation coefficient measures?

2) The program used to create this scatterplot found the line-of-best-fit and reported the $r^2$ value as $r^2 = 0.805$ for the relationship between arm-span and height for several individuals. What is the correlation coefficient? Is it positive or negative? Explain how you know.

3) During the summer, Ms. Statsteacher lets her two daughters stay up later than during the school year. Their bedtimes during the summer range from 8:30 p.m. to 12:30 a.m. She has discovered that her older daughter Reily will wake up between 8:00 a.m. and 9:00 a.m. no matter what time she goes to bed. However, her younger daughter Neila tends to wake up later after she gets to stay up later, and earlier when she goes to bed earlier. Neila has been known to wake up anytime between 8:00 a.m. and 11:45 a.m.

   a) Create a separate scatterplot for each daughter that compares time going to sleep and time waking up. You will have to approximate your own points. Which variable will be explanatory and which will be response?

   b) Which of these do you think will best approximate the correlation for Reily?

      [A] close to $r = +1$      [B] close to $r = +.75$      [C] close to $r = 0$
      [D] close to $r = -.75$      [E] close to $r = -1$

   c) Which of these do you think will best approximate the correlation for Neila?

      [A] close to $r = +1$      [B] close to $r = +.75$      [C] close to $r = 0$
      [D] close to $r = -.75$      [E] close to $r = -1$
4) Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. Consider whether common response, confounding, or coincidence may be involved.

a) As ice cream sales increase, the rate of drowning deaths increases sharply. Does this mean that ice cream causes drowning?

b) With a decrease in the number of pirates, there has been an increase in global warming over the same time period. Does this mean global warming is caused by a lack of pirates?

c) The higher the number of fire-fighters fighting a fire, the more damage done by the fire. Does this mean that we can limit damage by sending fewer fire-fighters to fires?

d) Suppose that each of the hockey players on the high school team supplies his or her own hockey stick, with varying degrees of flex. The assistant coach has been keeping a record of the degree of flex for each player’s stick and their respective point totals (goals and assists). He has noted that there is a strong, negative correlation between these two variables. In other words, the players with less flex in their sticks are scoring more points and those with more flex are scoring fewer points. Does this prove that the amount of flex in a stick will causes the point totals for the players? Would we be able to give players less flexible sticks and expect to increase scoring?

5) In a recent study in Resource Manual, it was noted that divorced men were twice as likely to abuse alcohol as married men. The authors concluded that getting divorced caused alcohol abuse. Do you agree? Explain your reasoning.

6) A commercial for a new diet pill claims “You will lose weight while you sleep! No exercise needed!” They then show several before-and-after photos of people who have lost weight. People who were obese are now very buff. They then give the information for you to order the pills (“for three payments of just $19.95 each, plus shipping and handling”). Is this proof that these diet pills caused these people to lose weight? Suggest possible lurking variables. Explain your reasoning.

7) Use the “Beach Visitors” scatterplot to answer the questions that follow.

a) Identify the explanatory and response variables.

b) Estimate the correlation coefficient for the graph.

c) Describe what the scatterplot shows. (remember S.C.O.F.D.)
8) Match each graph with its correlation coefficient:

GRAPH #1
GRAPH #2
GRAPH #3
GRAPH #4
GRAPH #5

9) A correlation of $r = 0$ indicates no linear relationship between the two given variables. But, this does not mean that there is no relationship between the two variables. Sketch a scatterplot in which there is a strong relationship between the variables, but the correlation would be near $r = 0$.

**Review Exercises**

10) Zeke flips a coin 93 times and tails shows up 34 of those times. Based on these results, what is the experimental probability of getting tails?

11) If Stephanie’s batting average is 0.258, how many hits would you expect her to get out of her next 20 times at bat?

12) You have been playing the game Yahtzee with some friends and you have been keeping track of how often someone gets a Yahtzee (5 of the same dice) when they roll all 5 dice at once. The results today have been 3 Yahtzee’s on a single roll, out of 79 trials. Based on these results, what is the experimental probability of getting a Yahtzee in one roll?

13) What is the theoretical probability of getting a Yahtzee in one roll?
6.3 Least-Squares Regression

Learning Objectives

- Construct scatterplots using technology
- Calculate and graph the least-squares regression line using technology
- Calculate the correlation coefficient using technology
- Use the LSRL to make predictions
- Understand interpolation and extrapolation
- Interpret the slope and the y-intercept of the LSRL

Least-Squares Regression

In the last section we learned about the concept of correlation, which we defined as the measure of the strength of a linear relationship between two numerical variables. We saw that when the points of a scatterplot formed a clear linear pattern, we expected a high correlation. Scatterplots can have a strong correlation in either a positive (increasing to the right) or a negative (decreasing to the right) direction. We have also discussed the idea of drawing a line-of-best-fit through the data. In some scatterplots this is easy to do and all of us would end up with our lines in nearly the same place. However, if everyone were to simply draw a line where they thought it fits best, our lines and equations would almost certainly vary from person to person. To maintain consistency, we will use a specific formula to calculate the equation for the line-of-best-fit.

Linear regression involves using data to calculate a line that best fits the data and then using that line to predict scores. We will use the Least-Squares Regression Line (LSRL). The LSRL is the line that makes the sum of the squares of the vertical distance of each data point from the line the least possible value. This is the standard regression equation that is used most often. It is the equation that most calculators and spreadsheets will calculate for you. The formula and process to calculate this is quite tedious, so we will use technology to find the LSRL equation. The regression equation will be in the form $y = a + bx$, where $a$ is the y-intercept and $b$ is the slope of the equation. Your calculator will calculate the correlation coefficient ($r$) at the same time as it calculates the LSRL equation. Many will also report a value for $r^2$ (which is exactly what it says; $r$-squared). The $r^2$ value is called the coefficient of determination. It reports the percent of variation in our data that is explained by our LSRL equation. We will not be addressing its importance in this course.

To calculate the LSRL equation and correlation coefficient, use a graphing calculator, certain scientific calculators, or a computer program. See the Appendix C for instructions on finding the LSRL and correlation coefficient for certain graphing and scientific calculators.
Interpreting the slope and y-intercept

As with all of our statistics, these data, graphs, and equations are not meaningless. They represent the relationship between two numerical values measured on several specific individuals. Thus the slope and the y-intercept of our newly calculated regression equation mean something as well. We will be interpreting both the slope and y-intercept in context. Our interpretation of the slope of the regression equation will be the average rate of change in the response variable ($\hat{y}$) for each increase of one unit of the explanatory variable (x). You will say something like: *For each increase of one (explanatory variable), there will be average (an increase or decrease) of (slope value) in the (response variable).*

Our interpretation of the y-intercept of the regression equation will be the predicted value of the response variable (y) when the explanatory variable (x) is zero. You will say something like: *When (explanatory variable) is zero, the (response variable) is predicted to be (y-intercept value).* You will discover that the interpretation of the y-intercept often makes absolutely no sense when put into context. This is because actual data often does not involves x-values of zero.

**Least-Squares Regression Equation**

$$\hat{y} = a + bx$$

- $x$ = the explanatory variable
- $\hat{y}$ = the predicted response variable

- $a$ = the y-intercept (or the value of y, when $x = 0$)
- $b$ = the slope (or the rate of change in y for each increase of one unit in the x direction)
Example 1
To the right is data given by a canine expert. It relates a dog’s age in years to what they believe the equivalent age in human years to be.

The scatterplot showing this data, using dog age as the explanatory variable, is shown below.

<table>
<thead>
<tr>
<th>Dog Age (in Years)</th>
<th>Equivalent Human Age (in Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>5.5</td>
</tr>
<tr>
<td>1</td>
<td>10.5</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>57</td>
</tr>
</tbody>
</table>

a) Calculate the Least-Squares regression line for the Dog Year Data. Report your equation. Be sure to identify your variables.

b) Calculate the correlation coefficient (r). What two things does r tell us about this relationship?

c) Identify and interpret the slope in the context of the problem.

d) Identify and interpret the y-intercept in the context of the problem.

Solution

a) \( \hat{y} = 7.7947 + 4.6418x \) where x is the age of the dog in years and \( \hat{y} \) is the equivalent predicted age in human years.

b) \( r = +0.9907 \). This r-value tells us that the graph is increasing left to right and that this data produces a strong linear relationship.

c) The slope is 4.642. This means that for every increase of one year in dog age, there is an average increase of 4.642 years in the equivalent human age.

d) The y-intercept is 7.795. It means that if a dog were 0 years old, it would be predicted to be 7.795 years in human years. (Note that this clearly makes no sense in this case. It is reasonable that both values would start at zero.)
Making Predictions

The main use of the regression line is to predict values. After calculating this line, we are able to predict values by simply substituting a value for the explanatory variable (x) and solving the equation for the predicted response value ($\hat{y}$). In our example above, we can predict that the human year equivalence for a dog that is 6 years old is approximately 35.6 human years.

$\hat{y} = 7.795 + 4.642(6) = 35.647$. This prediction is reasonable and it matches our graph. However this is not always the case.

As you look at the LSRL drawn on the above scatterplot, you can see that the points to the far left do not appear to be very linear. So, using the line to the left of about 1 year will not make much sense. Also, we do not have any idea what will happen to the data beyond the 11 years that we have recorded. An LSRL is very useful in making predictions, but only within the range of the actual data that we have collected and can see. This is called interpolation. We can see that this line is a reasonably good fit between 1 and 11 dog years, but we simply do not know what happens beyond 11 years (and we cannot use negative years for obvious reasons). The prediction line that we have calculated will go forever in both directions, but it will not be appropriate to use it to predict for all values of x. Using a regression line to predict values that are outside the range of our actual data is called extrapolation. Extrapolation will sometimes yield ridiculous answers! However, even if the result seems reasonable, we should avoid extrapolating because we simply do not know what happens beyond our actual observations. Making decisions based on extrapolating can be dangerous as we are coming to conclusions that are not backed up by data.
Example 2

The table to the right lists the GPA and Verbal SAT Score for seven students. Analyze how well Verbal SAT Scores can be used to predict a student’s GPA based on this data.

a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes.

b) Calculate the Least-Squares Regression Line (LSRL) using your calculator and include it on the graph from part a). Be sure to identify your variables.

c) Calculate the correlation coefficient (r). What are the two things that this number tells us about the graph?

d) Identify and interpret the slope in the context of the problem.

e) Using your equation, what is the predicted GPA of a student who has a verbal SAT score of 500?

f) Using your equation, what is the predicted verbal SAT score of a student who has a GPA of 3.1?

Solution

\[ y = 0.097 + 0.0055x \] where \( x \) stands for the student’s verbal SAT score and \( y \) stands for the student’s GPA.

Here are the LSRL, correlation, and the scatterplot with the line added to the graph, from a screen shot of a TI-84 plus:
c) The correlation is $r = +0.9467$. This tells us that the relationship is positive and strong.

d) The slope is 0.0055. This tells us that for each increase of 1 point on the verbal SAT score, there will be an average increase of 0.0055 points in a student’s GPA.

e) $\hat{y} = 0.097 + 0.0055(500) = 2.847$ The predicted GPA for a student who scores 500 on the verbal portion of the SAT exam is 2.847.

f) $3.1 = 0.097 + 0.0055x$

$3.003 = 0.0055x$

$546 = x$

The predicted SAT verbal score for a student with a GPA of 3.1 is 546.

**Outliers and Influential Points**

An outlier is an extreme observation that does not fit the general pattern of the data (see the example data set to the right). Because an outlier is an extreme observation, the inclusion of it may impact both the correlation and the equation for the least-squares regression line. When examining a scatterplot and calculating the regression equation, it is worth considering whether extreme observations should be included or not.

Let’s use our GPA example to illustrate the effect of a single outlier. Suppose that we have a student who has scored very high on the SAT Verbal exam, but has a lower GPA. We will change Corbin’s GPA from 3.9 to 2.2 and see what happens to the LSRL and correlation.

Here is a screen shot of a TI-84 with the adjusted score for Corbin. Both the LSRL and the correlation changed.
As you can see, this one change turned Corbin into an outlier. This caused the correlation to drop from \( r = +0.947 \), all the way down to \( r = +0.317 \). This one data point caused a huge change. It made the relationship between the two variables extremely weak rather than very strong. In addition, the adjusted data point changed both the slope and the \( y \)-intercept of the LSRL equation dramatically. This means that predictions based on this LSRL will have different results than those based on the LSRL with Corbin’s old GPA.

There is no set rule when trying to decide how to deal with outliers in regression analysis, but you can now see how an outlier can dramatically change everything when it comes to scatterplots, correlation, and least-squares regression equations. Be sure to mention any potential outliers that you observe in any scatterplot.

**Multimedia Links**

For an introduction to what a least squares regression line represents,

See Bionic Turtle at http://www.youtube.com/watch?v=ocGEhiLwDVc (5:15).

For an applet that will calculate correlation and the least squares regression line,

Problem Set 6.3

Exercises

1) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full. Her findings are listed in the following table. In section 6.1 we constructed a scatterplot and described the plot, we are now going to analyze this data further.

   a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes. Use a ruler to draw in a best-fit line.
   b) Calculate the Least-Squares Regression Line (LSRL) using your calculator. Report your equation. Be sure to identify your variables.
   c) Calculate the correlation coefficient (r). What are the two things that this number tells us about this graph?
   d) Identify and interpret the slope in the context of the problem.
   e) Using your equation, what is the predicted depth of the water after 17 minutes? After one hour? Are these answers reasonable? Why or why not?

2) The following table shows the progression of the Federal Minimum Wage in the United States since 1938 (source:http://www.laborlawcenter.com). We are going to analyze the relationship between year and minimum wage to see if there is a predictable relationship between the variables.

   a) Using year only as the explanatory variable (ignore month & day), construct a scatterplot. Sketch the graph that the calculator shows. Be sure to label your axes.
   b) Describe the relationship between the two variables. (S.C.O.F.D.)
   c) Calculate the Least-Squares Regression Line (LSRL). Add the line to your graph and report your equation. Be sure to identify your variables.
   d) Calculate the correlation (r). Even though r is very high, do you feel that a line is the best model for this data? Why or why not?
   e) Based on the linear model from part d), what would you predict the Federal Minimum Wage to be in 2016? Is this an accurate prediction? Why or why not?
   f) Based on your model from part d), what would you predict the minimum wage to have been in 1968? How close is this to the actual minimum wage that year?
3) Suppose that some researchers analyzed the relationship between fathers’ and sons’ IQ scores for a group of men. Suppose they discovered that the relationship was reasonably linear and they calculated a regression line of $\hat{y} = 12 + 0.9x$ where $x$ = father’s IQ and $\hat{y}$ = son’s predicted IQ.

a) Identify the explanatory and response variables.
b) Identify and interpret the slope in the context of the problem.
c) Identify and interpret the $y$-intercept in the context of the problem.
d) Do your answers to (b) and (c) seem reasonable? Why or why not?
e) What would you predict a son’s IQ to be if his father has an IQ of 120? What if the father had an IQ of 140?
f) If you knew that the original data included fathers with IQs from 108 to 145, explain why it would be inappropriate to use your model to predict a son’s IQ if his father’s IQ were 170.

4) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table. We began to look at this data in section 6.1. We are now going to analyze it further.

<table>
<thead>
<tr>
<th># months owned machine</th>
<th># hours exercise per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6.5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

a) Construct a scatterplot. Calculate the LSRL and add it to your graph. Sketch your graph and report your equation. Be sure to identify your variables.
b) Identify and interpret the slope in the context of the problem.
c) Identify and interpret the $y$-intercept in the context of the problem.
d) What is the correlation coefficient? What are the two things that this statistic tells about the relationship between these two variables?
e) Based on your model, how many hours would you predict a person who has owned their machine for 12 months to exercise?
f) Based on your model, if a person claims to exercise 9 hours per week, how long would you suspect that they had owned the machine?
5) A college professor was becoming annoyed by how many of his students were absent during his 8:00 a.m. section of Philosophy 103. He decided to analyze whether these absences were impacting student scores. He assigned his TA the task of keeping track of attendance. At the end of the semester he compared each student’s grade on the final exam (100 points possible) with the number of times he or she had been absent. His findings are displayed in the graph to the right.

a) Identify the explanatory and response variables.

b) Describe the relationship between these two variables (S.C.O.F.D).

c) Jeremy was absent 25 times. What would you predict his score on the final exam to be?

d) Lucy often overslept and missed 43 class sessions. What would you predict for her score on the final?

e) Calculate the correlation coefficient (r). What two things does this statistic tell you about the association between these two variables? (Hint: You were given the value for $R^2$.)

f) Interpret the meaning of -1.654 in the context of this problem.

6) The table to the right shows the grade level and reading level for 5 students. Treat grade level as the explanatory variable as you answer the questions below.

a) Create a scatterplot and then calculate the LSRL and the correlation coefficient for this data set.

Suppose it was determined that student E was actually in grade 8. Let’s examine how this change would this impact the LSRL and the correlation.

b) Create a scatterplot for the new data. Then calculate the LSRL and the correlation coefficient for the changed data. The new data is shown in the table to the right.

c) What changes do you notice between your answers to (a) and (b)? Explain why these changes occurred.
7) The table below gives the nutritional information for Taco Bell Burritos as reported on the website: http://www.tacobell.com. Choose two of the variables to analyze. (Do not use trans fat.)

a) What will you be using as your explanatory and response variables?

b) Construct a scatterplot. Label your axes.

c) Describe the association (S.C.O.F.D).

d) Calculate the LSRL and the correlation. Report them. Be sure to define your variables. Add the line to your graph in part (b).

e) Use your model to make a prediction that involves interpolation.

f) Use your model to make a prediction that involves extrapolation.

<table>
<thead>
<tr>
<th>item</th>
<th>serving size (g)</th>
<th>calories</th>
<th>calories from fat</th>
<th>saturated fat (g)</th>
<th>total fat (g)</th>
<th>trans fat (g)</th>
<th>cholesterol (mg)</th>
<th>sodium (mg)</th>
<th>carbohydrates (g)</th>
<th>dietary fiber (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2 lb. * Cheesy Potato Burrito</td>
<td>248</td>
<td>540</td>
<td>230</td>
<td>7</td>
<td>26</td>
<td>0.5</td>
<td>45</td>
<td>1350</td>
<td>59</td>
<td>7</td>
</tr>
<tr>
<td>1/2 lb. * Combo Burrito</td>
<td>241</td>
<td>480</td>
<td>160</td>
<td>7</td>
<td>18</td>
<td>0.5</td>
<td>45</td>
<td>1330</td>
<td>53</td>
<td>9</td>
</tr>
<tr>
<td>7-Layer Burrito</td>
<td>283</td>
<td>500</td>
<td>160</td>
<td>6</td>
<td>18</td>
<td>0</td>
<td>20</td>
<td>1090</td>
<td>69</td>
<td>12</td>
</tr>
<tr>
<td>Bean Burrito</td>
<td>198</td>
<td>370</td>
<td>90</td>
<td>3.5</td>
<td>10</td>
<td>0</td>
<td>5</td>
<td>980</td>
<td>56</td>
<td>10</td>
</tr>
<tr>
<td>Beefy 5-Layer Burrito</td>
<td>245</td>
<td>540</td>
<td>190</td>
<td>8</td>
<td>22</td>
<td>0</td>
<td>35</td>
<td>1280</td>
<td>68</td>
<td>9</td>
</tr>
<tr>
<td>Beefy Nacho Burrito</td>
<td>186</td>
<td>470</td>
<td>180</td>
<td>6</td>
<td>20</td>
<td>0</td>
<td>30</td>
<td>990</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td>Burrito Supreme® - Chicken</td>
<td>248</td>
<td>400</td>
<td>110</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>40</td>
<td>1090</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td>Burrito Supreme® - Steak</td>
<td>248</td>
<td>390</td>
<td>110</td>
<td>5</td>
<td>13</td>
<td>0</td>
<td>30</td>
<td>1100</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td>Burrito Supreme® – Beef</td>
<td>248</td>
<td>420</td>
<td>140</td>
<td>6</td>
<td>16</td>
<td>0</td>
<td>35</td>
<td>1100</td>
<td>53</td>
<td>9</td>
</tr>
<tr>
<td>Chili Cheese Burrito</td>
<td>156</td>
<td>380</td>
<td>150</td>
<td>8</td>
<td>17</td>
<td>0.5</td>
<td>35</td>
<td>930</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>Fresco Bean Burrito</td>
<td>213</td>
<td>350</td>
<td>70</td>
<td>2.5</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>990</td>
<td>57</td>
<td>11</td>
</tr>
<tr>
<td>Grilled Chicken Burrito</td>
<td>177</td>
<td>430</td>
<td>170</td>
<td>5</td>
<td>16</td>
<td>0</td>
<td>35</td>
<td>870</td>
<td>48</td>
<td>3</td>
</tr>
<tr>
<td>XXL Grilled Stuf Burrito - Beef</td>
<td>445</td>
<td>880</td>
<td>370</td>
<td>14</td>
<td>42</td>
<td>1</td>
<td>75</td>
<td>2050</td>
<td>95</td>
<td>14</td>
</tr>
<tr>
<td>XXL Grilled Stuf Burrito - Chicken</td>
<td>445</td>
<td>840</td>
<td>310</td>
<td>11</td>
<td>35</td>
<td>0</td>
<td>85</td>
<td>1970</td>
<td>92</td>
<td>11</td>
</tr>
<tr>
<td>XXL Grilled Stuf Burrito - Steak</td>
<td>445</td>
<td>820</td>
<td>320</td>
<td>12</td>
<td>36</td>
<td>0.5</td>
<td>70</td>
<td>2050</td>
<td>92</td>
<td>11</td>
</tr>
</tbody>
</table>
8) Use the calculator output from a screenshot of a TI-84 to answer the questions. A lifeguard at Swimtastic Pool & Water-Slides decided to keep track of how many people came to the pool each day and compare this to the high temperature for that day. The temperatures ranged from 82° to 96° during his data collection time period. He used the number of people as the response variable.
   a) Write the regression equation. Define your variables.
   b) Identify and interpret the slope in the context of the problem.
   c) Report the correlation. What two things does the correlation tells us in this situation?
   d) Based on this model, how many people would you predict on a 91° day? How many would you predict on a 45° day? Are both of these predictions reasonable? Why or why not?

Review Exercises

Use the information below to answer review exercises 9 – 12.

Suppose that Marco, the star of the basketball team, makes 79% of the free-throws that he attempts. Assume that each free-throw is independent of any other free-throw.

9) What is the probability that Marco will make three free-throws in a row?
10) What is the probability that Marco will make exactly two out of three free-throws?
11) What is the probability that Marco will miss at least one of his next four free-throws?
12) If you were going to set up a simulation to estimate this scenario, which of the following would not be an appropriate way to assign the digits?
   [A] 01-79 represents makes, 80-99 & 00 represents misses
   [B] 01-21 represents misses, 22-99 & 00 represents makes
   [C] 00-79 represents makes, 80-99 represents misses
   [D] 00-20 represents misses, 21-99 represents makes
   [E] 00-78 represents makes, 79-99 represents misses

13) Here are the hourly salaries for the employees at Greezy’s Burger Boy:
    $9.35, $9.85, $9.25, $10.90, $10.25, $9.25, $12.05, $9.70, $18.90, $10.30, $9.75, and $9.55
    Use this salary data to answer the questions below.
   a) Calculate the mean and standard deviation for the salaries.
   b) Calculate the five number summary for the salaries.
   c) Construct an accurate box plot.
   d) Which numerical measures of center and spread (mean & standard deviation or median & IQR) would be more appropriate in this situation? Explain why.
   e) Describe the distribution. Include Shape, Outliers, Context, Center, & Spread (S.O.C.C.S.)
6.4 Chapter 6 Review

In this chapter, we have learned that when working with bivariate data, it is important to first identify whether there is an explanatory and response relationship between the two variables. Often one of the variables, the explanatory (independent) variable, can be identified as having an impact on the value of the other variable, the response (dependent) variable. The explanatory variable should be placed on the horizontal axis, and the response variable should be placed on the vertical axis. Next we learned how to construct a visual representation, in the form of a scatterplot, so that we can see what the association looks like. A scatterplot helps us see what, if any, association there is between the two variables. If there is an association between the two variables, it can be identified as being strong if the points form a very distinct form or pattern, or weak if the points appear to be somewhat randomly scattered. If the values of the response variable generally increase as the values of the explanatory variable increase, the data has a positive association. If the response variable generally decreases as the explanatory variable increases, the data has a negative association. We also are able to see the form of the pattern, if any, in the graph.

When the data looked reasonably linear, we learned how to use technology to calculate the least-squares regression line and the correlation coefficient. The least-squares regression line is often useful for making predictions for linear data. However, we now know to beware of extrapolating beyond the range of our actual data. Correlation is a measure of the linear relationship between two variables – it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other factors may be causing the two correlated variables to relate as they do. We learned how to interpret the linear correlation coefficient and that it can be greatly affected by outliers and influential points. Also, just because two variables have a high correlation, does not mean that they have a cause-and-effect relationship. Correlation does not necessarily imply causation!

Beyond constructing graphs and calculating statistics, we learned how to describe the relationship between the two variables in context. The acronym we learned to help us remember what to include in our descriptions is S.C.O.F.D. This tells us to describe the strength of the association, to be sure that our description is in context, to mention any outliers or influential points that we observe, and to describe the form and the direction of the relationship. We also learned how to interpret the slope and y-intercept of the least-squares regression line in context. Even though we are doing easy calculations, statistics is never about meaningless arithmetic and we should always be thinking about what a particular statistical measure means in the real context of the data.
Review Exercises

Answer the following as TRUE or FALSE.

1) A negative relationship between two variables means that for the most part, as the x variable increases, the y variable increases.

2) A correlation of -1 implies a perfect linear relationship between the variables.

3) The equation of the regression line used in statistics is \( \hat{y} = a + bx \)

4) When the correlation is high, one can assume that x causes y.

Complete the following statements with the best answer.

5) The variable used for the correlation coefficient is ____________.

6) A statistical graph of two variables is called a(n) ________________.

7) The ________________ variable is plotted along the x-axis.

8) The range of r is from ______________ to ________________.

9) The sign of r and __________________ will always be the same.

10) LSRL stands for ________________________.

11) If all the points fall on a straight line, the value of r will be ____________ or ____________.

12) If r = -0.86, then \( r^2 = ____________ \).

13) If \( r^2 = 0.77 \), then \( r = ____________ \) or ____________.

14) Using an LSRL to make predictions outside the range of our original data is called ____________.

15) Using an LSRL to make predictions within the range of our original data is called ____________.

16) When describing the relationship visible in a scatterplot, the acronym S.C.O.F.D. stands for ____________________________.

17) Suppose that a scatterplot shows a strong, linear, positive relationship, and the correlation coefficient is very high. However, both of the variables are actually increasing due to some outside lurking variable. This relationship is an example of ____________.

18) Suggest possible lurking variables to explain the high correlations between the following variables. Consider whether common response, confounding, or coincidence may be involved.

   a) The number of cell phones being made has been increasing over the past 15 years. So has the number of starving children. Do cell phones cause starvation?

   b) The stress level of all of the employees at a certain company has been going up consistently over the past year. During this time, they have received three pay bumps. Does this mean that higher pay is causing the stress?

   c) Suppose that a study shows that the number of hours of sleep a person gets is negatively correlated with the number of cigarettes a person smokes. In other words, as the number of hours of sleep goes down, the number of cigarettes smoked goes up. Does this mean that not sleeping causes a person to smoke more cigarettes?
19) Some researchers wanted to determine how well the number of beers consumed can predict what a person’s blood alcohol content will be after a given length of time. They set up an experiment in which several volunteers each drank a randomly selected number of beers during a given time period. The volunteers were between 21 and 25 years of age, but all ranged in gender and in weight. Exactly three hours after they began to drink the beers, their BAC level was measured three times. The three measurements were averaged and the results are given in the following table. (This is fictitious data, but it is based upon calculations from the BAC calculator at http://www.dot.wisconsin.gov.)

<table>
<thead>
<tr>
<th>Number of Beers Consumed (3 hours)</th>
<th>10</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>3</th>
<th>3</th>
<th>7</th>
<th>8</th>
<th>5</th>
<th>9</th>
<th>4</th>
<th>6</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC Level</td>
<td>0.29</td>
<td>0.034</td>
<td>0.094</td>
<td>0.1</td>
<td>0.135</td>
<td>0.025</td>
<td>0.062</td>
<td>0.23</td>
<td>0.225</td>
<td>0.127</td>
<td>0.137</td>
<td>0.13</td>
<td>0.06</td>
<td>0.012</td>
<td>0.139</td>
</tr>
</tbody>
</table>

a) Identify the explanatory and response variables and construct a scatter-plot (be neat & label your axes).

b) Calculate the LSRL and correlation. Report the equation and add it to your scatter-plot. Identify what your variables represent.

c) Identify and interpret the slope in context.

d) Identify and interpret the y-intercept in context.

e) If a person drinks 6 beers during this time period, on average what do you predict the person’s BAC will be?

f) If a person drinks 15 beers during this time period, on average what do you predict the person’s BAC will be?

g) Are you confident in both of the previous answers? Why or why not?

h) How many beers would predict had been consumed if the BAC was measured at 0.122?
20) When investigating car crashes, it is often necessary to try to determine the speed at which a vehicle was traveling at the time of the accident. Investigators are able to do this by measuring the length of the skid mark left by the vehicle in question. The table below lists several speeds (mph) based on the skid length (feet), according to the Forensic Dynamics website: http://forensicdynamics.com.

a) Identify the explanatory and response variables and construct a scatterplot. Be sure to properly scale and label your graph.

b) Calculate the LSRL and add it to your scatterplot. Report your equation and identify your variables.

c) Describe the relationship you see in the scatter-plot (S.C.O.F.D.). Be thorough & use complete sentences! Be sure that you explain the relationship in the context of the problem.

d) What is the correlation coefficient? Based on your scatterplot and the value of r, how well do you feel that your model fits this data? Explain.

e) What is the predicted speed if the skid mark is 157 feet?

f) What is the predicted speed for a skid mark of 36 feet?

g) Would you expect predictions beyond 250 feet to generally over-estimate or under-estimate the actual speed of the vehicle? Why?

### SPEED BASED ON SKID LENGTH

<table>
<thead>
<tr>
<th>Skid Length (feet)</th>
<th>Estimated Speed (mph)</th>
</tr>
</thead>
<tbody>
<tr>
<td>45</td>
<td>30.68</td>
</tr>
<tr>
<td>20</td>
<td>20.45</td>
</tr>
<tr>
<td>56</td>
<td>34.23</td>
</tr>
<tr>
<td>8</td>
<td>12.93</td>
</tr>
<tr>
<td>78</td>
<td>40.4</td>
</tr>
<tr>
<td>93</td>
<td>44.11</td>
</tr>
<tr>
<td>165</td>
<td>58.75</td>
</tr>
<tr>
<td>115</td>
<td>49.05</td>
</tr>
<tr>
<td>142</td>
<td>54.51</td>
</tr>
<tr>
<td>184</td>
<td>62.05</td>
</tr>
<tr>
<td>215</td>
<td>67.07</td>
</tr>
<tr>
<td>247</td>
<td>71.89</td>
</tr>
</tbody>
</table>

### Image References:
- Beach visitors & temperature: http://technomaths.edublogs.org
- Study Time & Test Scores: http://www.icoachmath.com
- Car weight & mpg: http://www.statcrunch.com
- Elevation & Temperature: http://staff.argyll.epsb.ca
- Peanut Butter & Quality Rating: http://intermath.coe.uga.edu
- Arm Span & Height: http://3.bp.blogspot.com
- Surgeon General’s Warning Labels: http://abibrands.com
- Outlier Example: http://mathworld.wolfram.com
- Recycling Rates: http://www.earth-policy.org
Chapter 7 - The Normal Distribution

7.1 Introduction to the Normal Curve

Learning Objectives

- Understand how a density curve can be used to approximate the data in a histogram
- Understand how to visually identify the mean and standard deviation of a normal distribution
- Be able to tie the concepts of percentages in the 68-95-99.7 Empirical Rule to normal distributions

In previous chapters we have seen how data can be represented by histograms. A density curve is a curve that gives an approximate description of a distribution. The curve is smooth, so any small irregularities in the data are ignored. An approximate density curve for one particular histogram is shown to the left. Perhaps the most important thought to remember about a density curve is that it represents 100% of the data. In other words, the area under any density curve is equal to 1. This is important because it allows us to ask probability questions about a population. For example, we might ask how likely is it that a teenager has a shoe size of 8 or larger.

In this chapter, we will focus on a special density curve called the normal curve. Have you ever wondered if you are ‘normal’? You probably are normal in most ways, but there may be some things about you that might not be considered normal by the mathematical definition. If you are on the high school baseball team, do you throw the baseball at a ‘normal’ speed? Is your hair a ‘normal’ length? Do you drive at a ‘normal’ speed on the freeway? Our goal this chapter is to gain an understanding of what ‘normal’ really is and how to properly calculate within the Normal Distribution. We have seen skewed distributions before. The density curves in the figure to the left show one density curve that is skewed left and one that is skewed right.
A normal curve is neither skewed left nor right and is often referred to as ‘the bell curve’ because of its shape. It is symmetrical. In addition, as you get closer and closer to the middle of the curve, there is a higher frequency of results. The mean (along with the median and mode) always lands at the center of a normal distribution. When dealing with the mean in previous chapters, we have used the symbol $\bar{x}$ because that calculation came from sample data. Normal distributions deal with an entire population instead of just a sample and we will use the symbol $\mu$ (Greek letter mu) to mark the mean of a normal distribution for an entire population. The mean is one of two key values needed to make a proper sketch and analysis of a normal distribution. The curve shown below represents a normal distribution and is a good representation of what a normal curve looks like.

Note that the amount of data to the right of the mean is the same as the amount of data to the left of the mean. Thinking about the definition of the median, this suggests that the mean and median are located at the same point. The other key component used to construct and analyze a normal distribution is the standard deviation. The standard deviation is a measure of spread and can be loosely thought of as a ‘typical’ distance from the mean. You may have calculated the standard deviation before for data sets either by hand or by using your calculator and looked for the $s_x$ in the statistical calculations summary screen. The symbol $s_x$ is used for the standard deviation whenever data is collected through the use of a sample from a population. When dealing with the normal distribution, we will use the symbol $\sigma$ (Greek letter sigma) to represent the standard deviation. The $\sigma$ symbol indicates that the standard deviation of the entire population is known. Visually, the standard deviation can be seen as the distance from the mean to an inflection point. An inflection point is located on a curve at the point where the curve changes from concave up (bent up) to concave down (bent down) or vice versa. On the normal curve in Figure 7.4, the mean is 23 and the standard deviation is 3.
The Empirical Rule (68-95-99.7 Rule)

It is now time to make use of some of the special characteristics of the normal curve. By definition, 100% of all results of a normal distribution, fall somewhere under the normal curve. It turns out that approximately 68% of all results are within one standard deviation of the mean, 95% of all results are within two standard deviations of the mean, and 99.7% of all results land within three standard deviations of the mean. These percentages are illustrated in Figure 7.5 to the right.

The numbers on the bottom represent the number of standard deviations from the mean. For example, the $\mu - 1\sigma$ marks the point one standard deviation below the mean. Some simple addition and subtraction allows us to be very specific in the percentages of the data that land in the sections of the normal curve as shown below.

Because 99.7% of all results lie within three standard deviations of the mean, both the area above 3 standard deviations and below -3 standard deviations would each contain 0.15%.

Can you see the 68-95-99.7 rule here?

Example 1

Suppose the mathematics portion of the SAT exam is normally distributed with a mean of 500 points and a standard deviation of 100 points.

a) Sketch a normal curve for this situation marking the mean and the values 1, 2, and 3 standard deviations above and below the mean.

b) Using the 68-95-99.7 Rule, approximately what percent of students scored at least 600 points on this test?

c) Between approximately which two scores did the middle 95% of students score?

d) Suppose that 4600 students take the exam this month. Approximately how many of those students should we expect to obtain a score of at least 700 points?
Solution

a)

b) We know that 50% of all results are below the 500 marker and that 34% of all results land between 500 and 600. We have used up 50% + 34% = 84% of all results. This tells us that 100% - 84% = 16% of all students scored above 600 on the mathematics portion of the SAT.

c) The middle 95% of all students scored within 2 standard deviations of the mean or between 300 and 700 points.

d) A score of 700 points marks the boundary two standard deviations above the mean such that only 2.5% of all test takers will score at least 700 points. Thus, 2.5% of 4600 is 115 students.

Example 2

The normal curve below represents the number of races that a typical racehorse will run in one calendar year.

a) Approximately what percent of racehorses will run between 5 and 11 races during a calendar year?

b) What are the values of the mean and standard deviation for the distribution shown?

Solution

a) Add 13.5% + 34% + 34% to get 81.5% so 81.5% of racehorses run between 5 and 11 races per year (see Figure 7.6).

b) The mean racehorse will run 9 races per year with a standard deviation of 2 races.
What is Normal?

Let’s now go back and try to think about our original question “What is normal?” In mathematics, the middle 95% is often (but not always) considered our ‘normal’ group. For example, suppose the ACT exam is normally distributed with a mean of 18 and a standard deviation of 6. Our ‘normal’ group would be comprised of those students who scored anywhere within two standard deviations from the mean or from 6 to 30 on the exam. A student who scored 31 or higher on the exam would have achieved an exceptional score. We might say that this student was not normal with regards to their ACT score.

Normal distributions are not as common as you might think. What if we measured the lengths of shoes of teenagers? Many students think that this would be normal when in fact; there are a couple of contributing factors that might tip us off that the situation may not be normal. First of all, teenagers encompass a large population. Most of those who are in their upper teen years have finished growing into their adult shoe size length whereas many of the younger teens are still growing. This would tend to give us a slightly larger percentage of smaller shoe lengths than we might expect from a normal distribution. In addition, teenagers include males and females. This could lead to a situation which might be bimodal (having two modes). We might expect to see a peak at the most common male shoe lengths and at the most common female shoe lengths.

Example 3

Which situation below is most likely to produce a normal distribution?

[A] The heights of all adults.


[C] The number of teeth that Americans adults have.

Solution

The correct answer is [B]. Three year-old American eagles have an average wingspan and we would expect that there are quite a few eagles at that wingspan or very close to it. As we move further and further up and down from that average, we would expect to see fewer and fewer eagles with those wingspans. Answer [A] could be ruled out quickly in that the heights here do not specify a particular group. For example, this data would include males and females. Answer [C] is out because the vast majority of American adults have 32 teeth. As we move away from 32, there are some people with fewer teeth due to a variety of reasons but there are virtually no people with more than 32 teeth. This distribution would be skewed left and therefore not a normal distribution.
Problem Set 7.1

Exercises

1) Consider the histogram shown to the right.
   a) Make a sketch of the histogram and overlay a sketch of a density curve for the histogram.
   b) What is the area under your density curve?
   c) What is the shape of the density curve?

2) A roadside bait salesman digs up worms to sell to fishermen. It turns out that the worms have a mean length of $\mu = 112$ mm and a standard deviation of $\sigma = 12$ mm.
   a) Draw and label a normal curve for this distribution. Include lines for the mean and for 1, 2, and 3 standard deviations above and below the mean.
   b) What percentage of the worms will have lengths longer than 112 mm?
   c) What percentage of the worms will have lengths between 100 and 124 mm long?
   d) What percentage of the worms will have lengths between 100 and 112 mm long?
   e) What percent of the worms are longer than 124 mm?
   f) What percent of the worms are shorter than 88 mm?

3) Sketch and label a normal curve which has a mean of 13 lbs. and a standard deviation of 3 lbs. Include lines for the mean and for 1, 2, and 3 standard deviations above and below the mean.

4) Not all 12-ounce cans of soda are the same. It turns out that the average 12-ounce can of soda does contain twelve ounces of soda, but the amount of soda is normally distributed with a standard deviation of 0.15 ounces. Fill in the blanks for each statement below.
   a) The middle 68% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.
   b) The middle 95% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.
   c) The middle 99.7% of all 12-ounce soda cans contain between _____ & _____ ounces of soda.
5) The graph to the right shows an approximate distribution of the number of fish caught by the competitors during a one hour pan-fishing contest. Give the approximate values of the mean and the standard deviation for the distribution.

6) Suppose the weights of adult males of a particular species of whale are distributed normally with a mean of 11,600 pounds and a standard deviation of 640 pounds.
   a) Draw and label a normal curve for this situation. Use vertical lines to mark and label the mean and 1, 2, and 3 standard deviations above and below the mean.
   b) Approximately what percent of these whales weigh less than 10,320 pounds?
   c) Between what two weights do the middle 99.7% of these whales weigh?
   d) About what percent of these whales weigh between 10,320 pounds and 12,240 pounds?

7) Which situation is most likely to be normally distributed? Explain your reasoning.
   a) The hair lengths for all the Statistics and Probability students who have Mr. Johnson as a teacher.
   b) The prices of all latest Samsung cell phones that are sold in Minnesota this week.
   c) The running times for all 4th grade males at Andover Elementary in the 50 yard dash.

8) Suppose a standard light bulb will run an average of 400 hours before burning out. Of course, some bulbs burn out sooner and some last longer. Suppose the lives of these bulbs are normally distributed with a standard deviation of 35 hours.
   a) Sketch and label a normal curve to illustrate this situation.
   b) What percent of these bulbs would we expect to burn out in 400 hours or less?
   c) Some bulbs will last longer than advertised. What percent of bulbs would we expect to last 435 hours or more? What percent of bulbs will last 470 hours or more?
   d) If there were 5000 bulbs needed for use in a large office building, how many would be expected to last at least 365 hours?
9) Suppose that the time that it takes for popcorn kernels to pop produces a normal distribution with a mean of 145 seconds and a standard deviation of 13 seconds for a standard microwave oven.

   a) It is usually not a good idea to let the microwave oven run until all the kernels are popped because some of the popcorn will start to burn. Suppose the ideal time to shut off the microwave oven is after about 97.5% of the kernels have popped. When will 97.5% of the kernels be popped?
   
   b) Between what two times will we see the middle 68% of kernels popped?

10) After a great deal of surveying, it is determined that the average wait times in the cafeteria line are normally distributed with a mean of 7 minutes and a standard deviation of 2 minutes. Suppose that 400 students are released to the cafeteria for 2nd lunch.

   a) Approximately how many students will have to wait more than 5 minutes for their food?
   
   b) Approximately how many students will have to wait more than 11 minutes for their food?

11) Sudoku is a popular logic game of number combinations. It originated in the late 1800s in the French press, Le Siècle. The mean time it takes the average 11th grader to complete a particular Sudoku puzzle was found to be 19.2 minutes, with a standard deviation of 3.1 minutes.

   a) Draw and label a normal distribution curve to represent this data.
   
   b) Suppose Andover High School is going to put together a Sudoku team. The coach has decided that she will only consider players who score in the fastest 2.5% of the junior class as she puts together the team. How fast must a student solve a puzzle to be in the top 2.5% of puzzle solvers?
   
   c) If there are 400 kids in the Andover junior class, how many of them will be able to solve the Sudoku puzzle below in 16.1 minutes or less?

12) In order to qualify for undercover detective training, a police officer must take a stress tolerance test. Scores on this test are normally distributed with a mean of 60 and a standard deviation of 10. Only the top 16% of police officers score high enough on the test to qualify for the detective training. What is the cutoff score that marks the top 16% of all scores?
Review Exercises

13) The ages of the kids at the YMCA summer day camp on Tuesday ranged from 3 to 8 years old. Use your calculator to find the mean and standard deviation for the ages in the data set below.
   3, 4, 4, 5, 5, 6, 6, 7, 7, 8

14) A pet store must select 2 dogs and 2 cats for display in their front window. In how many ways can this be done if there are 16 dogs and 12 cats available to choose from?

15) Several students were asked how many missing assignments they had in their math class. The results are reported below. Find the five number summary for the number of missing assignments for these students.
   3, 5, 5, 6, 8, 9, 10, 10, 12, 13, 13, 13, 14, 15, 17, 19, 19, 20

16) A student conducts a survey in which 100 tenth-graders are asked “What is your favorite item on the lunch menu at school today?” The student decides to conduct this survey by handing each tenth-grader a survey sheet while they are eating and asking them to fill it out and turn it in to room P202 by the end of the day. Why will this survey method have a problem with bias?
7.2 Z-Scores, Percentiles, and Normalcdf

Learning Objectives

• Be able to calculate and understand z-scores
• Understand the concept of a percentile and be able to calculate it for a particular result
• Be able to calculate percentages of data above, below, or in between any specific values in a normal distribution
• Be able to use z-scores to compare results for two different but related situations
• Be able to make all of these calculations by hand and with technology

In section 7.1, we analyzed normal distributions and specific situations in which analysis was done for data which followed the 68-95-99.7 Rule exactly. The truth of the matter is that most situations require us to answer questions that do not reference exact whole numbers of standard deviations above or below the mean. What if we asked a student what their actual score would be if they were in the top 10% of ACT test takers? We need a tool to help us deal with these types of situations.

Our first tool will be the z-score formula. The z-score is a measure of how many standard deviations above or below the mean a particular value is. If a z-score is negative, the result is below the mean and if it is positive, the result is above the mean. For example, if the ACT mathematics exam scores are normally distributed with a mean of 18 and a standard deviation of 6, then an ACT score of 30 would be equivalent to a z-score of 2 because 30 would be 2 standard deviations above the mean.

The formula below gives a quick way to calculate z-scores. In the formula, x is the observation, μ is the mean of the distribution, and σ is the standard deviation for the distribution.

\[
Z = \frac{x - \mu}{\sigma}
\]

\[
z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}}
\]
Example 1

Suppose the mean length of the hair of 10th grade girls is 10 inches with a standard deviation of 4 inches. What would be the z-score for hair length for a 10th grade girl whose hair is 16 inches long and what does it mean in terms of the normal curve?

Solution

It is often a good idea to draw a sketch for these sorts of situations so we can visualize what is happening.

Because 16 is located between 1 and 2 standard deviations above the mean, we expect a z-score between 1 and 2. Use the formula $z = \frac{x - \mu}{\sigma}$ to calculate the z-score. Our observation, $x$, is 16 inches while the mean is $\mu = 10$ inches and the standard deviation is $\sigma = 4$ inches.

$z = \frac{16 - 10}{4}$ or $z = 1.5$. This tells us that a hair length of 16 inches will be 1.5 standard deviations above the mean.

Example 2

Suppose that the z-score for a particular 10th grade girl’s hair length is $z = -1.25$. What is the length of the girl’s hair?

Solution

We will use the z-score formula to find our answer again. Note that this time it is the observation that is unknown.

$-1.25 = \frac{x - 10}{4}$

$-5 = x - 10$

$5 = x$ The length of the hair for this girl would be 5 inches.

Example 3

Suppose a student can either submit only their SAT score or their ACT score to a particular college. Suppose their SAT score was 620 points and that the SAT has a mean of 500 points and a standard deviation of 100 points. Suppose also that the same student scored a 25 on their ACT exam and that the ACT exam has a mean of 18 points and a standard deviation of 6 points. Which score should the student submit?
Solution

Looking at the diagram below, it is not exactly clear which score is better. They appear to be quite similar and we will need to do some calculations to make a distinction.

![SAT and ACT graphs]

Calculate the z-score for each exam. For the SAT, \( z = \frac{620 - 500}{100} = 1.2 \). For the ACT, \( z = \frac{25 - 18}{6} \approx 1.17 \). Since the z-score is higher on the SAT, the student should submit their SAT exam score.

Percentiles

In order to understand how to apply z-scores beyond what we have already done, we must first understand percentiles. A percentile is a marker on a normal curve such that the marker is greater than or equal to that percentage of results. For example, suppose you are at the 30th percentile for how fast you type. This means that you can type faster than 30% of all people. The percentile can also be thought of as the percent of area to the left of its marker. The graphic below shows where the 30th percentile is located. The shaded area to the left of the marker represents 30% of the normal curve.

![Graph showing 30th percentile]

It is very common for colleges and universities to use percentiles for entrance criteria. For example, a rather elite university might require that you score at the 90th percentile or higher on your ACT exam to be considered for admission. Doctors often use percentiles to track the growth of babies. For example, can you picture what a baby would look like that is at the 70th percentile for weight and the 25th percentile for length?
Now we must ask what percentiles have to do with z-scores. Find the Normal Distribution Table in Appendix A, Part 2, in the back of your book. Let’s examine the z-score of -1.25 from Example 2 to see how to use the table. Find the z-value of -1.2 and then go over until you are under the 0.05 column. A partial table is given in the graphic below. The value in the cell we are looking for is bold and underlined. The value of 0.1056 can be interpreted as a percentile. This means that the girl in Example 2 has hair that is longer than 10.56% of all girls. In other words, she is at about the 10th or 11th percentile for hair length for 10th grade girls.

<table>
<thead>
<tr>
<th>Z</th>
<th>0.09</th>
<th>0.08</th>
<th>0.07</th>
<th>0.06</th>
<th>0.05</th>
<th>0.04</th>
<th>0.03</th>
<th>0.02</th>
<th>0.01</th>
<th>0.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.3</td>
<td>0.0823</td>
<td>0.0838</td>
<td>0.0853</td>
<td>0.0869</td>
<td>0.0885</td>
<td>0.0901</td>
<td>0.0918</td>
<td>0.0934</td>
<td>0.0951</td>
<td>0.0968</td>
</tr>
<tr>
<td>-1.2</td>
<td>0.0985</td>
<td>0.1003</td>
<td>0.1020</td>
<td>0.1038</td>
<td>0.1056</td>
<td>0.1075</td>
<td>0.1093</td>
<td>0.1112</td>
<td>0.1131</td>
<td>0.1151</td>
</tr>
<tr>
<td>-1.1</td>
<td>0.1170</td>
<td>0.1190</td>
<td>0.1210</td>
<td>0.1230</td>
<td>0.1271</td>
<td>0.1271</td>
<td>0.1292</td>
<td>0.1314</td>
<td>0.1335</td>
<td>0.1357</td>
</tr>
</tbody>
</table>

Example 4

At what percentile for hair length is a 10th grade girl if her hair is 17 inches long? Recall that the mean is 10 inches and the standard deviation is 4 inches.

Solution

Start by determining her z-score which would be \( z = \frac{17 - 10}{4} = \frac{7}{4} = 1.75 \). We now go to the Normal Distribution Table in Appendix A, Part 2 of the book. We go across the row with \( z = 1.7 \) until we are under 0.05. This gives a value of 0.9599. This tells us the girl is at about the 96th percentile for hair length. In other words, this girl’s hair is longer than 96% of all 10th grade girls.

‘Between’ and ‘Above’ Problems

While it is nice to find percentiles for certain situations, we are often asked for the percentage of results that are between two given parameters or above a given parameter. For example, we might be asked to find the percentage of all 10th grade girls that have hair lengths between 8 inches and 15 inches long. To find these types of results, we often must do multiple z-score calculations and some addition or subtraction.

Example 5

Suppose the weights of adult males of a particular species of whale are distributed normally with a mean of 11,600 pounds and a standard deviation of 640 pounds.

a) What percent of these adult male whales will weigh between 11,000 and 12,000 pounds?

b) What percent of these adult male whales will weigh more than 12,000 pounds?
Solution

a) Begin by finding the z-scores for both of the weights given and get
\[ z = \frac{11,000 - 11,600}{640} = \frac{-600}{640} = -0.9375 \] and
\[ z = \frac{12,000 - 11,600}{640} = \frac{400}{640} = 0.625. \]
For \( z = -0.9375 \), our Normal Distribution Table from Appendix A, Part 2, gives us a value between 0.1736 and 0.1762. Since -0.9375 is closer to -0.94 than -0.93, we will use a value of 0.174. Likewise, we get a value between 0.7324 and 0.7357 for \( z = 0.625 \). We will split the difference on this and use 0.734. All that is left to do now is subtract 0.734 and 0.174 to get 0.56 or about 56% of all adult male whales of this species are between 11,000 and 12,000 pounds. The shaded region in Figure 7.4 represents about 56% of the normal curve.

b) Use \( z = 0.625 \) from part a) to get a value from the table of 0.734. This means that 73.4% of all whales weigh 12,000 pounds or less. Therefore, 100% - 73.4% = 26.6% of all whales weigh more than 12,000 pounds.

Technology

It is also important to note that graphing calculators can be used to quickly solve the types of problems discussed in this section by using the normalcdf command. Typically, this command requires that four values be entered, the lower bound, the upper bound, the mean, and the standard deviation. In Example 5, we can solve the problem in part a) simply by typing in the command string normalcdf(11000,12000,11600,640) and obtain a result of 0.5598 or 56%.

Be sure you know how to access this command if you have a graphing calculator. Appendix C on page 280 has some notes for common graphing calculators. An online calculator that is very similar to a graphing calculator and gives us the same information can be found at http://wolframalpha.com.

You might also be wondering how to solve a problem using the normalcdf command if only one parameter is given. Let’s revisit Example 4 to see how this works.

Example 6

At what percentile for hair length is a 10th grade girl if her hair is 17 inches long? Recall that the mean hair length is 10 inches with a standard deviation of 4 inches.

Solution

There is only one boundary given in this problem. It is your job to come up with a second boundary. In this case, the percentile we want to calculate is found by finding the percentage of all girls whose hair is 17 inches or less. We will use a lower bound of -100 and an upper bound of 17. We use -100 simply because we are confident that we will not find any results any further left than this. Typically, choose your missing parameter as being so extreme that it will not be in the realm of possible results. Normalcdf(-100,17,10,4)=0.9599 so the length of the girl’s hair is at about the 96th percentile.
Problem Set 7.2

Exercises

For problems 1) through 14) use the following information: On a particular stretch of road, the number of cars per hour produces a normal distribution with a mean of 125 cars per hour and a standard deviation of 40 cars per hour.

1) Sketch and label a normal curve for this situation. Be sure to label and mark the mean and 1, 2, and 3 standard deviations above and below the mean.
2) What is the z-score for an observation of 165 cars in one hour?
3) What is the z-score for an observation of 85 cars in one hour?
4) Calculate the z-score associated with an observation of 171 cars in one hour.
5) Suppose 135 cars are observed in one hour. At what percentile would this observation occur?
6) Suppose 70 cars are observed in one hour. At what percentile would this observation occur?
7) At what percentile would an observation of 125 cars occur?
8) What is the probability of observing at least 145 cars on the road in an hour?
9) What is the probability of observing between 100 and 150 cars on the road in an hour?
10) Determine the percentile for an observation of 140 cars on the road in one hour.
11) Determine the percentile for an observation of 65 cars on the road in one hour.
12) Determine the probability of observing between 90 and 130 cars on the road in one hour.
13) Determine the probability of observing at least 160 cars on the road in one hour.
14) Determine the probability of observing no more than 110 cars on the road in one hour.

For problems 15) through 20) use the following information: The number of ants found in a typical mature colony of leafcutter ants is normally distributed with a mean of 136 ants and a standard deviation of 14 ants.

15) One ant colony has 165 ants. At what percentile for size is this ant colony?
16) An ant colony has a z-score of \( z = -1.35 \) for size. Approximately how many ants would we expect to find in this colony?
17) Another ant colony has 131 ants. What is the z-score for this ant colony?
18) What is the probability of finding an ant colony with 160 ants or less?
19) What is the probability of finding an ant colony with 150 ants or more?
20) What is the probability of finding an ant colony that has between 120 and 155 ants in it?
21) Twin brothers Ricky and Robbie each took a college entrance exam. Ricky took the SAT which had a mean of 1000 with a Standard Deviation of 200 while Robbie took the ACT which had a mean of 18 with a standard deviation of 6. Which brother did better if Ricky scored a 1140 and Robbie scored a 22?

22) Suppose the average height of an adult American male is 69.5 inches with a standard deviation of 2.5 inches and the average height of an adult American female is 64.5 inches with a standard deviation of 2.3 inches. Who would be considered taller when compared to their gender, an adult American male who is 74 inches tall or an adult American female who is 68.5 inches tall? Explain your answer.

23) Professional golfer John Daly is one of the longest hitting golfers in history. Suppose his drives average 315 yards with a standard deviation of 12 yards. Will a drive of 345 yards be in his top 1% of his longest drives? Explain your answer.

**Review Exercises**

24) What is the area under any density curve?

25) In a standard deck of 52 cards, what is the probability of being dealt two queens if you are dealt two cards from the deck without replacement?

26) In a class competition, each grade (9-12) enters 10 students to run in a 500 meter race. Times for 9th grade males and 12th grade males are given in the table below in seconds. Build a back-to-back stem plot to compare data for the two groups of students.

| 9th Grade Times = 115, 118, 118, 121, 126, 127, 131, 134, 140 |
| 12th Grade Times = 106, 106, 109, 112, 114, 116, 116, 121, 122, 133 |

27) It turns out that countries that have higher percentages of people with computers also tend to have people who live longer. Is it logical to assume that shipping many computers to countries whose people have lower life-expectancies will help the people in those countries live longer? Answer the question including justification that references Cause and Effect, Common Response, Confounding, or Coincidence.

28) A sample survey at a local college campus asked 250 students how many textbooks they were currently carrying. The table below shows a summary of the findings. Use the table to determine the expected number of textbooks that an average college student at this campus would be carrying.

<table>
<thead>
<tr>
<th>Textbooks Carried by Students</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Books</td>
</tr>
<tr>
<td>Probability</td>
</tr>
</tbody>
</table>
7.3 Inverse Normal Calculations

Learning Objectives

- Understand how to use the Normal Distribution Table and the z-score formula to find values for a particular normal distribution given a percentile
- Be able to use the Inverse Normal command on a graphing calculator to find values for a particular normal distribution given a percentile
- Be able to find values for a particular normal distribution given a ‘middle’ percentage range

We can now comfortably calculate percentages, percentiles, and probabilities given key information about a normal distribution. It is possible to go the other direction. If you are told a certain result is at a specific percentile, you can figure out what the actual value is associated with that percentile. The process can be done using the Normal Distribution Table on page 265. Begin by identifying the percentile you are interested in and find its decimal equivalent in the table. From there, work backward to find the associated z-score. Finally, use that z-score in the z-score formula and solve it for the observation in question.

Example 1

Suppose that 10th grade girls have hair lengths that are normally distributed with a mean of 10 inches and a standard deviation of 4 inches. How long would a 10th grade girl’s hair have to be in order to be at the 80th percentile for length?

Solution

The figure below shows the distribution of hair lengths and also marks where the 80th percentile is located.

Begin by finding the value closest to 0.8000 in the Normal Distribution Table. We find our closest value to be 0.7995 which corresponds to a z-score of 0.84. Put this value into the

\[
0.84 = \frac{x - 10}{4}
\]

\[
0.84 = \frac{x - 10}{4}
\]

\[
3.36 = x - 10
\]

z-score formula to get

\[
x = 13.36
\]

A 10th grade girl would have to have a hair length of about 13.4 inches to be at the 80th percentile. This looks to be right based upon comparison to the figure above.
Technology
Once again, it is important to note that technology can be used to solve these types of problems without having to reference the Normal Distribution Table. The command that is commonly used for these types of problems is the Inverse Normal command or invNorm. The Inverse Normal command requires users to enter the percentile in question, the mean, and the standard deviation. To solve the problem in Example 1, we could have typed in invNorm(0.80,10,4) and we would be given an answer of 13.366 or about 13.4 inches of hair.

Be sure you know how to access this command if you have a graphing calculator. Appendix C on page 280 has some notes for users of graphing calculators. An online calculator that can produce the same information can be found at http://wolframalpha.com.

‘Middle’ and ‘Top’ Problems
Sometimes we are in situations where we want to know what range of results are found in a middle percentage interval or what value one would have to be at in order to be in a specific top percentage. For example, a car salesman might wish to know what sales prices comprise the middle 50% of his sales to help him learn more about the type of customer that buys at his dealership. A student might wish to know what they need to score on a test in order to be in the top 10% of the class. Once again, this process can be done with either the Normal Distributions Table or by using technology.

Example 2
Professional golfer John Daly is known for his long drives off the tee. Suppose his drives have a mean distance of 315 yards with a standard deviation of 12 yards. What lengths of drives will constitute the middle 60% of all of his drives?

Solution
The sketch to the right is helpful in understanding what is happening here.

It is easy to calculate that marker line ‘a’ is at the 20th percentile and marker line ‘b’ is at the 80th percentile simply by noting their relationship to the 50th percentile marker. In addition, note that ‘a’ and ‘b’ clearly enclose the middle 60 percent of all data. From the Normal Distributions Table, we can see that the z-score associated with the 20th percentile is -0.84 and the z-score associated with the 80th percentile is 0.84. We now calculate $-0.84 = \frac{x-315}{12}$ or $x = 304.9$ yards. A similar calculation at the 80th percentile gives us $x = 325.1$ yards. We conclude that the middle 60% of John Daly’s drives will travel between 304.9 yards and 325.1 yards.

We also could have used the Inverse Normal command once we knew the percentiles; invNorm(0.20,315,12) = 304.9 yards and invNorm(0.80,335,14) = 325.1 yards.
Example 3
In a weightlifting competition, the amount that the competitors can lift is normally distributed with μ = 196 kg and σ = 11 kg. Only the top 20% of all competitors will be able to advance to the next phase of the competition. What amount must a competitor lift in order to move into the next phase of the competition?

Solution
The key to this problem is noticing that to be in the top 20%, a competitor would actually have to be at the 80th percentile. The z-score at the 80th percentile is z = 0.84.

\[
0.84 = \frac{x - 196}{11}
\]

\[
9.24 = x - 196
\]

\[
205.24 = x
\]

The competitor would have to lift about 205 or 206 kg. Using a calculator, we get invNorm(0.8,196,11) = 205.26 kg.
Problem Set 7.3

Exercises

1) The Standard Normal Curve is defined as having a mean of 0 and a standard deviation of 1.
   a) What is the z-score associated with a result at the 84th percentile?
   b) What is the z-score associated with a result at the 16th percentile?
   c) Find a z-score such that only 5% of the Standard Normal Curve is to the right of that z-score.
   d) Find a z-score such that only 35% of the Standard Normal Curve is the left of that z-score.
   e) Find the two z-scores such that the middle 50% of the Standard Normal Curve is between the two z-scores.

2) Doctors often monitor their patients’ blood-glucose levels. Suppose that blood-glucose levels are known to be normally distributed with, \( \mu = 85 \) mg/dL and \( \sigma = 25 \) mg/dL.
   a) Draw and label sketch of the normal distribution for this situation marking the mean and 1, 2, and 3 standard deviations above and below the mean.
   b) It turns out that doctors consider the blood-glucose level of a patient to be normal if the level is in the middle 94% of all results. What range of blood-glucose levels constitute the middle 94% of all results?
   c) Patients are considered to be at high risk for diabetes if their blood-glucose test comes back in the top 1% of all results. What blood-glucose level marks the start of the top 1% of blood-glucose levels?
   d) Doctors also show concern if there is too little blood-glucose in a patient’s system. They will prescribe treatments to patients if their blood-glucose is in the lowest 2% of all patients. What is the blood-glucose level that marks this boundary?

3) For a given population of high school juniors and seniors, the SAT math scores are normally distributed with a mean of 500 and a standard deviation of 100. For that same population, the ACT math exam has a mean of 18 with a standard deviation of 6.
   a) One school requires that students score in the top 10% on their SAT math exam for admission. What is the minimum score that a student must achieve to be considered for this school?
   b) Another school requires that students score in the top 40% on their ACT math exam for admission. What is the minimum score that a student must achieve to be considered for this school?
   c) One particular school likes to focus on mid-level students and so they only accept students who are in the middle 50% of all ACT math test takers. Between what two scores must a student achieve in order to be considered for acceptance into this school?
   d) One student boasts that they scored at the 85th percentile on their ACT math exam. Another student brags that they scored a 620 on the SAT math exam. Who did better?
4) Many athletes train to try to be selected for the U.S. Olympic team. Suppose for the men’s 100 meters, the athletes being considered for the team have a mean time of 10.06 seconds with a standard deviation of 0.07 seconds. In the final qualifying event for the team, only the top 20% of runners will be selected. What time must a runner get to be in the top 20%?

5) A high school basketball coach notices that taller players tend to have more success on his team. As a result, the coach decides that only the tallest 25% of the males in the 11th and 12th grades will be allowed to try out for the team this year. Suppose that the mean height of 11th and 12th grade males is 5 feet 9 inches (69 inches) with a standard deviation of 2.5 inches. How tall must a player be in order to be able to try out for the team?

6) A student comes home to his parents and excitedly claims that he is in the top 90% of his class. Explain why this might not be worth getting excited about.

7) At a certain fast-food restaurant, automatic soft drink filling machines have been installed. For 20-ounce cups, the machine is set to fill up the cups with 19 ounces of soda. Unfortunately, the machine is not perfectly consistent and does not always dispense 19 ounces of soda. Suppose the amount it dispenses produces a normal distribution with a mean of 19 ounces and a standard deviation of 0.6 ounces. It turns out that the 20 ounce cup will actually hold a bit more than 20 ounces. A mathematically inclined worker notices this and starts to record what happens when the machine fills the cups. It turns out that the cups overfill 2% of the time. How much soda will the 20-ounce cup actually hold?

Review Exercises

8) Adult male American bald eagles have a mean wingspan of 79 inches with a standard deviation of 3.5 inches. What percent of these eagles have wingspans longer than 7 feet?
9) Consider the data in the table below where the number of pages is the explanatory variable.

The table lists the weights of ten books and the number of pages in each one.

<table>
<thead>
<tr>
<th>Number of pages</th>
<th>85</th>
<th>150</th>
<th>100</th>
<th>120</th>
<th>90</th>
<th>140</th>
<th>137</th>
<th>105</th>
<th>115</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight (g)</td>
<td>165</td>
<td>325</td>
<td>200</td>
<td>250</td>
<td>180</td>
<td>285</td>
<td>250</td>
<td>170</td>
<td>230</td>
<td>340</td>
</tr>
</tbody>
</table>

a) Create a scatterplot for the data set. Label your axes.

b) Determine the correlation coefficient, r, for the scatterplot.

c) Give the least-squares linear regression equation. Be sure to define your variables.

d) Using your answer from part c), predict the weight of a book that has 130 pages.

e) Using your answer from part c), predict the number of pages for a book that weighs 295 grams.

10) Consider a standard set of 15 pool balls. Pool balls #1-8 are solid and pool balls #9-15 are striped.

a) If you randomly select one pool ball, what is the probability that it is both solid and odd?

b) If you randomly select one pool ball, what is the probability that it is either solid or odd?

c) If you randomly select two pool balls without replacement, what is the probability that they are either both solid or both striped?
Chapter 7 Review

In this chapter we have discussed what a density curve is and specifically focused on a special density curve called the normal distribution. The two critical pieces of information that are necessary for analysis of a density curve are the mean and standard deviation. The mean is the center of the distribution while the standard deviation is a measure of spread. We have focused on several key concepts including the 68-95-99.7 Rule and z-scores. We then introduced the Normal Distribution Table and the normalcdf and invNorm commands on our calculators to help us move back and forth between probabilities and percentiles and specific values in our distributions.

Review Exercises

1) Suppose a teacher gives a test in which the scores on the test are normally distributed with a mean of 10 points and a standard deviation of 2 points.

   a) Draw and label a normal curve to represent this situation. Clearly mark the mean and 1, 2, and 3 standard deviations above and below the mean.

   b) Using the 68-95-99.7 Rule, approximately what percent of students will get a score between 6 and 14?

   c) Using the 68-95-99.7 Rule, approximately what percent of students will get a score between 8 and 16?

   d) Find the percent of students that will get a score between 8 points and 13 points on this test.

   e) What percent of students will score at least 11 points on this test?

   f) What percent of students will score between 5 points and 12 points on this test?

   g) How many points would a student have to score in order to be at the 90th percentile on this test?

   h) What is the z-score associated with a test score of 13 points?

   i) How many points did a student score if their z-score was z = -1.5?

2) Which situation below is most likely to be normally distributed?

   a) The heights of all the trees in a forest.

   b) The distances that all the kids at Blaine High School can hit a golf ball.

   c) The number of siblings that each student at Anoka High School has.

   d) The length of time that 6th grade boys at Roosevelt Middle School can hold their breath.
3) The weights of adult male African elephants are normally distributed with a mean weight of 11,000 pounds and standard deviation of 900 pounds.

   a) Between what two weights do the middle 50% of all adult male African elephants weigh?

   b) Suppose one of these elephants weighs 13,400 pounds. At what percentile is this weight?

   c) At what weight would we find the 70th percentile of weights for these elephants?

4) Suppose that IQ test scores are normally distributed with a mean of 100 points and a standard deviation of 15 points.

   a) What z-score is associated with an IQ score of 125 points?

   b) The intelligence organization MENSA requires that members score in the top 2.5% of all IQ test takers to gain membership in the organization. What IQ score must a person score to qualify for MENSA?

   c) What percentage of IQ scores are greater than 125 points?

   d) What percentage of IQ scores are less than 70 points? Use the 68-95-99.7 Rule to approximate your answer.

   e) Who did better, a person with an IQ score of 143 points or someone who was at the 99th percentile on the IQ test? Justify your answer.

5) In a certain city, the number of pounds of newspaper recycled each month by a household produces a normal distribution with a mean of 8.5 pounds and a standard deviation of 2.7 pounds.

   a) Draw and label a sketch for this normal distribution and shade in the region that represents the households that recycle between 6 and 12 pounds of newspaper each month.

   b) What percent of households recycle between 6 and 12 pounds of newspaper each month?

   c) A local newspaper wants to do a story on newspaper recycling in the city. They decide that they would like to base their story on a typical household. After some thought, they decide that ‘typical’ means that they are in the middle 60% of all households in terms of newspaper recycling. Between what two weights are the ‘typical’ households?
6) Snowfall each winter in the Twin Cities is normally distributed with a mean of 56 inches and a standard deviation of 11 inches.

   a) In what percentage of years does the Twin Cities get less than 3 feet of snow?

   b) In what percentage of years does the Twin Cities get more than 6 feet of snow?

   c) The winter of 2010-2011 was the fifth snowiest on record for the Twin Cities with a total snowfall of 85 inches. What percentage of years will have snowfalls of more than 85 inches?

   d) A winter is considered to be dry if it is in the lowest 10% of snowfall totals. What is the maximum amount of snow the Twin Cities could receive to still be called a dry winter?

7) You just got your history test back and found out you scored 37 points. The scores were normally distributed with a mean of 31 points and a standard deviation of 4 points. When you tell your parents how you did, your little brother pipes in that he got a 56 on his math test which was normally distributed with a mean of 40 points and a standard deviation of 11 points. How could you use z-scores to explain to your parents that your score was more impressive than your little brother’s score?

8) In 1941, Ted Williams batted 0.406 for the baseball season. He is the last player to hit over 0.400 for an entire major league baseball season. In 2009, Joe Mauer hit 0.365 for the baseball season. In 1941, the batting averages were normally distributed with a mean of 0.260 and a standard deviation of 0.041. In 2009, the batting averages were normally distributed with a mean of 0.262 and a standard deviation of 0.035. Decide which player had a better season compared to the rest of the league during their respective year by comparing z-scores.

9) Suppose that medals will be given out to any student at Champlain Park High School that scores at least 200 points on an aptitude test. The mean score on the aptitude test is 150 points with a standard deviation of 22 points. Approximately how many medals should be ordered if there are 456 students who sign up for the test?
Image References

Density Curve www.madscientist.blogspot.com
Earthworms http://www.flowers.vg
Pet Store Window http://perezhilton.com/teddyhilton/
Traffic Jam https://www.rnw.org/
Leafcutter Ant http://www.orkin.com/ants/
Pair of Queens http://www.123rf.com
American Diabetes Association https://americandiabetesassn.wordpress.com/
Track Race http://www.tierraunica.com
Eagle http://www.esa.org
Elephant http://animals.nationalgeographic.com/animals/
Blizzard http://www.csc.cs.colorado.edu
Joe Mauer http://www.mauersquickswing.com
Appendices

Appendix A – Tables

Appendix A, Part 1 - Random Digit Table

<p>| Line 101 | 19223 | 95034 | 05756 | 28713 | 96409 | 12531 | 42544 | 82853 |
| Line 102 | 73676 | 47150 | 99400 | 01927 | 27754 | 42648 | 82425 | 36290 |
| Line 103 | 45467 | 71709 | 77558 | 00095 | 32863 | 29485 | 82226 | 90056 |
| Line 104 | 52711 | 38889 | 93074 | 60227 | 40011 | 85848 | 48767 | 52573 |
| Line 105 | 95592 | 94007 | 69971 | 91481 | 60779 | 53791 | 17297 | 59335 |
| Line 106 | 68417 | 35013 | 15529 | 72765 | 85089 | 57067 | 50211 | 47487 |
| Line 107 | 82739 | 57890 | 20807 | 47511 | 81676 | 55300 | 94383 | 14893 |
| Line 108 | 60940 | 72024 | 17868 | 24943 | 61790 | 90656 | 87964 | 18883 |
| Line 109 | 36009 | 19365 | 15412 | 39638 | 85453 | 46816 | 83485 | 41979 |
| Line 110 | 38448 | 48789 | 18338 | 24697 | 39364 | 42006 | 76688 | 08708 |
| Line 111 | 81486 | 69487 | 60513 | 09297 | 00412 | 71238 | 27649 | 39950 |
| Line 112 | 59636 | 88804 | 04634 | 71197 | 19352 | 73089 | 84898 | 45785 |
| Line 113 | 62568 | 70206 | 40325 | 03699 | 71080 | 22553 | 11486 | 11776 |
| Line 114 | 45149 | 32992 | 75730 | 66280 | 03819 | 56202 | 02938 | 70915 |
| Line 115 | 61041 | 77684 | 94322 | 24709 | 73698 | 14526 | 31893 | 32592 |
| Line 116 | 14459 | 26056 | 31424 | 80371 | 65103 | 62253 | 50490 | 61181 |
| Line 117 | 38167 | 98532 | 62183 | 70632 | 23417 | 26185 | 41448 | 75532 |
| Line 118 | 73190 | 32533 | 04470 | 29669 | 84407 | 90785 | 65956 | 86382 |
| Line 119 | 95857 | 07118 | 87664 | 92099 | 58806 | 66979 | 98624 | 84826 |
| Line 120 | 35476 | 55972 | 39421 | 65850 | 04266 | 35435 | 43742 | 11937 |</p>
<table>
<thead>
<tr>
<th>Line</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>71487 09984 29077 14863 61683 47052 62224 51025</td>
</tr>
<tr>
<td>122</td>
<td>13873 81598 95052 90908 73592 75186 87136 95761</td>
</tr>
<tr>
<td>123</td>
<td>54580 81507 27102 56027 55892 33063 41842 81868</td>
</tr>
<tr>
<td>124</td>
<td>71035 09001 43367 49497 72719 96758 27611 91596</td>
</tr>
<tr>
<td>125</td>
<td>96746 12149 37823 71868 18442 35119 62103 39244</td>
</tr>
<tr>
<td>126</td>
<td>96927 19931 36089 74192 77567 88741 48409 41903</td>
</tr>
<tr>
<td>127</td>
<td>43909 99477 25330 64359 40085 16925 85117 36071</td>
</tr>
<tr>
<td>128</td>
<td>15689 14227 06565 14374 13352 49367 81982 87209</td>
</tr>
<tr>
<td>129</td>
<td>36759 58984 68288 22913 18638 54303 00795 08727</td>
</tr>
<tr>
<td>130</td>
<td>69051 64817 87174 09517 84534 06489 87201 97245</td>
</tr>
<tr>
<td>131</td>
<td>05007 16632 81194 14873 04197 85576 45195 96565</td>
</tr>
<tr>
<td>132</td>
<td>68732 55259 84292 08796 43165 93739 31685 97150</td>
</tr>
<tr>
<td>133</td>
<td>45740 41807 65561 33302 07051 93623 18132 09547</td>
</tr>
<tr>
<td>134</td>
<td>27816 78416 18329 21337 35213 37741 04312 68508</td>
</tr>
<tr>
<td>135</td>
<td>66925 55658 39100 78458 11206 19876 87151 31260</td>
</tr>
<tr>
<td>136</td>
<td>08421 44753 77377 28744 75592 08563 79140 92454</td>
</tr>
<tr>
<td>137</td>
<td>53645 66812 61421 47836 12609 15373 98481 14592</td>
</tr>
<tr>
<td>138</td>
<td>66831 68908 40772 21558 47781 33586 79177 06928</td>
</tr>
<tr>
<td>139</td>
<td>55588 99404 70708 41098 43563 56934 48394 51719</td>
</tr>
<tr>
<td>140</td>
<td>12975 13258 13048 45144 72321 81940 00360 02428</td>
</tr>
<tr>
<td>141</td>
<td>96767 35964 23822 96012 94591 65194 50842 53372</td>
</tr>
<tr>
<td>142</td>
<td>72829 50232 97892 63408 77919 44575 24870 04178</td>
</tr>
<tr>
<td>143</td>
<td>88565 42628 17797 49376 61762 16953 88604 12724</td>
</tr>
<tr>
<td>144</td>
<td>62964 88145 83083 69453 46109 59505 69680 00900</td>
</tr>
<tr>
<td>145</td>
<td>19687 12633 57857 95806 09931 02150 43163 58636</td>
</tr>
<tr>
<td>146</td>
<td>37609 59057 66967 83401 60705 02384 90597 93600</td>
</tr>
<tr>
<td>147</td>
<td>54973 86278 88737 74351 47500 84552 19909 67181</td>
</tr>
<tr>
<td>148</td>
<td>00694 05977 19664 65441 20903 62371 22725 53340</td>
</tr>
<tr>
<td>149</td>
<td>71546 05233 53946 68743 72460 27601 45403 88692</td>
</tr>
<tr>
<td>150</td>
<td>07511 88915 41267 16853 84569 79367 32337 03316</td>
</tr>
</tbody>
</table>
### Appendix A, Part 2 – The Normal Distribution Table

For $z$-scores with $z$ less than or equal to zero

<table>
<thead>
<tr>
<th>$z$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.0</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0013</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0011</td>
<td>0.0010</td>
<td>0.0010</td>
</tr>
<tr>
<td>-2.9</td>
<td>0.0019</td>
<td>0.0018</td>
<td>0.0018</td>
<td>0.0017</td>
<td>0.0016</td>
<td>0.0016</td>
<td>0.0015</td>
<td>0.0015</td>
<td>0.0014</td>
<td>0.0014</td>
</tr>
<tr>
<td>-2.8</td>
<td>0.0026</td>
<td>0.0025</td>
<td>0.0024</td>
<td>0.0023</td>
<td>0.0023</td>
<td>0.0022</td>
<td>0.0021</td>
<td>0.0021</td>
<td>0.0020</td>
<td>0.0019</td>
</tr>
<tr>
<td>-2.7</td>
<td>0.0035</td>
<td>0.0034</td>
<td>0.0033</td>
<td>0.0032</td>
<td>0.0031</td>
<td>0.0030</td>
<td>0.0029</td>
<td>0.0028</td>
<td>0.0027</td>
<td>0.0026</td>
</tr>
<tr>
<td>-2.6</td>
<td>0.0047</td>
<td>0.0045</td>
<td>0.0044</td>
<td>0.0043</td>
<td>0.0042</td>
<td>0.0041</td>
<td>0.0040</td>
<td>0.0039</td>
<td>0.0038</td>
<td>0.0037</td>
</tr>
<tr>
<td>-2.5</td>
<td>0.0062</td>
<td>0.0060</td>
<td>0.0059</td>
<td>0.0057</td>
<td>0.0056</td>
<td>0.0055</td>
<td>0.0054</td>
<td>0.0052</td>
<td>0.0051</td>
<td>0.0049</td>
</tr>
<tr>
<td>-2.4</td>
<td>0.0082</td>
<td>0.0080</td>
<td>0.0078</td>
<td>0.0075</td>
<td>0.0073</td>
<td>0.0071</td>
<td>0.0069</td>
<td>0.0068</td>
<td>0.0066</td>
<td>0.0064</td>
</tr>
<tr>
<td>-2.3</td>
<td>0.0107</td>
<td>0.0104</td>
<td>0.0102</td>
<td>0.0099</td>
<td>0.0096</td>
<td>0.0094</td>
<td>0.0091</td>
<td>0.0089</td>
<td>0.0087</td>
<td>0.0084</td>
</tr>
<tr>
<td>-2.2</td>
<td>0.0139</td>
<td>0.0136</td>
<td>0.0132</td>
<td>0.0129</td>
<td>0.0125</td>
<td>0.0122</td>
<td>0.0119</td>
<td>0.0116</td>
<td>0.0113</td>
<td>0.0110</td>
</tr>
<tr>
<td>-2.1</td>
<td>0.0179</td>
<td>0.0174</td>
<td>0.0170</td>
<td>0.0166</td>
<td>0.0162</td>
<td>0.0158</td>
<td>0.0154</td>
<td>0.0150</td>
<td>0.0146</td>
<td>0.0143</td>
</tr>
<tr>
<td>-2.0</td>
<td>0.0228</td>
<td>0.0222</td>
<td>0.0217</td>
<td>0.0212</td>
<td>0.0207</td>
<td>0.0202</td>
<td>0.0197</td>
<td>0.0192</td>
<td>0.0188</td>
<td>0.0183</td>
</tr>
<tr>
<td>-1.9</td>
<td>0.0287</td>
<td>0.0281</td>
<td>0.0274</td>
<td>0.0268</td>
<td>0.0262</td>
<td>0.0256</td>
<td>0.0250</td>
<td>0.0244</td>
<td>0.0239</td>
<td>0.0233</td>
</tr>
<tr>
<td>-1.8</td>
<td>0.0359</td>
<td>0.0351</td>
<td>0.0344</td>
<td>0.0336</td>
<td>0.0329</td>
<td>0.0322</td>
<td>0.0314</td>
<td>0.0307</td>
<td>0.0301</td>
<td>0.0294</td>
</tr>
<tr>
<td>-1.7</td>
<td>0.0446</td>
<td>0.0436</td>
<td>0.0427</td>
<td>0.0418</td>
<td>0.0409</td>
<td>0.0401</td>
<td>0.0392</td>
<td>0.0384</td>
<td>0.0375</td>
<td>0.0367</td>
</tr>
<tr>
<td>-1.6</td>
<td>0.0548</td>
<td>0.0537</td>
<td>0.0526</td>
<td>0.0516</td>
<td>0.0505</td>
<td>0.0495</td>
<td>0.0485</td>
<td>0.0475</td>
<td>0.0465</td>
<td>0.0455</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.0668</td>
<td>0.0655</td>
<td>0.0643</td>
<td>0.0630</td>
<td>0.0618</td>
<td>0.0606</td>
<td>0.0594</td>
<td>0.0582</td>
<td>0.0571</td>
<td>0.0559</td>
</tr>
<tr>
<td>-1.4</td>
<td>0.0808</td>
<td>0.0793</td>
<td>0.0778</td>
<td>0.0764</td>
<td>0.0749</td>
<td>0.0735</td>
<td>0.0721</td>
<td>0.0708</td>
<td>0.0694</td>
<td>0.0681</td>
</tr>
<tr>
<td>-1.3</td>
<td>0.0968</td>
<td>0.0951</td>
<td>0.0934</td>
<td>0.0918</td>
<td>0.0901</td>
<td>0.0885</td>
<td>0.0869</td>
<td>0.0853</td>
<td>0.0838</td>
<td>0.0823</td>
</tr>
<tr>
<td>-1.2</td>
<td>0.1151</td>
<td>0.1131</td>
<td>0.1112</td>
<td>0.1093</td>
<td>0.1075</td>
<td>0.1056</td>
<td>0.1038</td>
<td>0.1020</td>
<td>0.1003</td>
<td>0.0985</td>
</tr>
<tr>
<td>-1.1</td>
<td>0.1357</td>
<td>0.1335</td>
<td>0.1314</td>
<td>0.1292</td>
<td>0.1271</td>
<td>0.1251</td>
<td>0.1230</td>
<td>0.1210</td>
<td>0.1190</td>
<td>0.1170</td>
</tr>
<tr>
<td>-1.0</td>
<td>0.1587</td>
<td>0.1562</td>
<td>0.1539</td>
<td>0.1515</td>
<td>0.1492</td>
<td>0.1469</td>
<td>0.1446</td>
<td>0.1423</td>
<td>0.1401</td>
<td>0.1379</td>
</tr>
<tr>
<td>-0.9</td>
<td>0.1841</td>
<td>0.1814</td>
<td>0.1788</td>
<td>0.1762</td>
<td>0.1736</td>
<td>0.1711</td>
<td>0.1685</td>
<td>0.1660</td>
<td>0.1635</td>
<td>0.1611</td>
</tr>
<tr>
<td>-0.8</td>
<td>0.2119</td>
<td>0.2090</td>
<td>0.2061</td>
<td>0.2033</td>
<td>0.2005</td>
<td>0.1977</td>
<td>0.1949</td>
<td>0.1922</td>
<td>0.1894</td>
<td>0.1867</td>
</tr>
<tr>
<td>-0.7</td>
<td>0.2420</td>
<td>0.2389</td>
<td>0.2358</td>
<td>0.2327</td>
<td>0.2296</td>
<td>0.2266</td>
<td>0.2236</td>
<td>0.2206</td>
<td>0.2177</td>
<td>0.2148</td>
</tr>
<tr>
<td>-0.6</td>
<td>0.2743</td>
<td>0.2709</td>
<td>0.2676</td>
<td>0.2643</td>
<td>0.2611</td>
<td>0.2578</td>
<td>0.2546</td>
<td>0.2514</td>
<td>0.2483</td>
<td>0.2451</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.3085</td>
<td>0.3050</td>
<td>0.3015</td>
<td>0.2981</td>
<td>0.2946</td>
<td>0.2912</td>
<td>0.2877</td>
<td>0.2843</td>
<td>0.2810</td>
<td>0.2776</td>
</tr>
<tr>
<td>-0.4</td>
<td>0.3446</td>
<td>0.3409</td>
<td>0.3372</td>
<td>0.3336</td>
<td>0.3300</td>
<td>0.3264</td>
<td>0.3228</td>
<td>0.3192</td>
<td>0.3156</td>
<td>0.3121</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.3821</td>
<td>0.3783</td>
<td>0.3745</td>
<td>0.3707</td>
<td>0.3669</td>
<td>0.3632</td>
<td>0.3594</td>
<td>0.3557</td>
<td>0.3520</td>
<td>0.3483</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.4207</td>
<td>0.4168</td>
<td>0.4129</td>
<td>0.4090</td>
<td>0.4052</td>
<td>0.4013</td>
<td>0.3974</td>
<td>0.3936</td>
<td>0.3897</td>
<td>0.3829</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.4602</td>
<td>0.4562</td>
<td>0.4522</td>
<td>0.4483</td>
<td>0.4443</td>
<td>0.4404</td>
<td>0.4364</td>
<td>0.4325</td>
<td>0.4286</td>
<td>0.4247</td>
</tr>
<tr>
<td>-0.0</td>
<td>0.5000</td>
<td>0.4960</td>
<td>0.4920</td>
<td>0.4880</td>
<td>0.4840</td>
<td>0.4801</td>
<td>0.4761</td>
<td>0.4721</td>
<td>0.4681</td>
<td>0.4641</td>
</tr>
</tbody>
</table>
For $z$-scores with $z$ greater than or equal to 0

<table>
<thead>
<tr>
<th>$z$</th>
<th>0.0</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5000</td>
<td>0.5040</td>
<td>0.5080</td>
<td>0.5120</td>
<td>0.5160</td>
<td>0.5199</td>
<td>0.5239</td>
<td>0.5279</td>
<td>0.5319</td>
<td>0.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5398</td>
<td>0.5438</td>
<td>0.5478</td>
<td>0.5517</td>
<td>0.5557</td>
<td>0.5596</td>
<td>0.5636</td>
<td>0.5675</td>
<td>0.5714</td>
<td>0.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5793</td>
<td>0.5832</td>
<td>0.5871</td>
<td>0.5910</td>
<td>0.5948</td>
<td>0.5987</td>
<td>0.6026</td>
<td>0.6064</td>
<td>0.6103</td>
<td>0.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6179</td>
<td>0.6217</td>
<td>0.6255</td>
<td>0.6293</td>
<td>0.6331</td>
<td>0.6368</td>
<td>0.6406</td>
<td>0.6443</td>
<td>0.6480</td>
<td>0.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6554</td>
<td>0.6591</td>
<td>0.6628</td>
<td>0.6664</td>
<td>0.6700</td>
<td>0.6736</td>
<td>0.6772</td>
<td>0.6808</td>
<td>0.6844</td>
<td>0.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6915</td>
<td>0.6950</td>
<td>0.6985</td>
<td>0.7019</td>
<td>0.7054</td>
<td>0.7088</td>
<td>0.7123</td>
<td>0.7157</td>
<td>0.7190</td>
<td>0.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7257</td>
<td>0.7291</td>
<td>0.7324</td>
<td>0.7357</td>
<td>0.7389</td>
<td>0.7422</td>
<td>0.7454</td>
<td>0.7486</td>
<td>0.7517</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7580</td>
<td>0.7611</td>
<td>0.7642</td>
<td>0.7673</td>
<td>0.7704</td>
<td>0.7734</td>
<td>0.7764</td>
<td>0.7794</td>
<td>0.7823</td>
<td>0.7852</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7881</td>
<td>0.7910</td>
<td>0.7939</td>
<td>0.7967</td>
<td>0.7995</td>
<td>0.8023</td>
<td>0.8051</td>
<td>0.8078</td>
<td>0.8106</td>
<td>0.8133</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8159</td>
<td>0.8186</td>
<td>0.8212</td>
<td>0.8238</td>
<td>0.8264</td>
<td>0.8289</td>
<td>0.8315</td>
<td>0.8340</td>
<td>0.8365</td>
<td>0.8389</td>
</tr>
<tr>
<td>1.0</td>
<td>0.8413</td>
<td>0.8438</td>
<td>0.8461</td>
<td>0.8485</td>
<td>0.8508</td>
<td>0.8531</td>
<td>0.8554</td>
<td>0.8577</td>
<td>0.8599</td>
<td>0.8621</td>
</tr>
<tr>
<td>1.1</td>
<td>0.8643</td>
<td>0.8665</td>
<td>0.8686</td>
<td>0.8708</td>
<td>0.8729</td>
<td>0.8749</td>
<td>0.8770</td>
<td>0.8790</td>
<td>0.8810</td>
<td>0.8830</td>
</tr>
<tr>
<td>1.2</td>
<td>0.8849</td>
<td>0.8869</td>
<td>0.8888</td>
<td>0.8907</td>
<td>0.8925</td>
<td>0.8944</td>
<td>0.8962</td>
<td>0.8980</td>
<td>0.8997</td>
<td>0.9015</td>
</tr>
<tr>
<td>1.3</td>
<td>0.9032</td>
<td>0.9049</td>
<td>0.9066</td>
<td>0.9082</td>
<td>0.9099</td>
<td>0.9115</td>
<td>0.9131</td>
<td>0.9147</td>
<td>0.9162</td>
<td>0.9177</td>
</tr>
<tr>
<td>1.4</td>
<td>0.9192</td>
<td>0.9207</td>
<td>0.9222</td>
<td>0.9236</td>
<td>0.9251</td>
<td>0.9265</td>
<td>0.9279</td>
<td>0.9292</td>
<td>0.9306</td>
<td>0.9319</td>
</tr>
<tr>
<td>1.5</td>
<td>0.9332</td>
<td>0.9345</td>
<td>0.9357</td>
<td>0.9370</td>
<td>0.9382</td>
<td>0.9394</td>
<td>0.9406</td>
<td>0.9418</td>
<td>0.9429</td>
<td>0.9441</td>
</tr>
<tr>
<td>1.6</td>
<td>0.9452</td>
<td>0.9463</td>
<td>0.9474</td>
<td>0.9484</td>
<td>0.9495</td>
<td>0.9505</td>
<td>0.9515</td>
<td>0.9525</td>
<td>0.9535</td>
<td>0.9545</td>
</tr>
<tr>
<td>1.7</td>
<td>0.9554</td>
<td>0.9564</td>
<td>0.9573</td>
<td>0.9582</td>
<td>0.9591</td>
<td>0.9599</td>
<td>0.9608</td>
<td>0.9616</td>
<td>0.9625</td>
<td>0.9633</td>
</tr>
<tr>
<td>1.8</td>
<td>0.9641</td>
<td>0.9649</td>
<td>0.9656</td>
<td>0.9664</td>
<td>0.9671</td>
<td>0.9678</td>
<td>0.9686</td>
<td>0.9693</td>
<td>0.9699</td>
<td>0.9706</td>
</tr>
<tr>
<td>1.9</td>
<td>0.9713</td>
<td>0.9719</td>
<td>0.9726</td>
<td>0.9732</td>
<td>0.9738</td>
<td>0.9744</td>
<td>0.9750</td>
<td>0.9756</td>
<td>0.9761</td>
<td>0.9767</td>
</tr>
<tr>
<td>2.0</td>
<td>0.9772</td>
<td>0.9778</td>
<td>0.9783</td>
<td>0.9788</td>
<td>0.9793</td>
<td>0.9798</td>
<td>0.9803</td>
<td>0.9808</td>
<td>0.9812</td>
<td>0.9817</td>
</tr>
<tr>
<td>2.1</td>
<td>0.9821</td>
<td>0.9826</td>
<td>0.9830</td>
<td>0.9834</td>
<td>0.9838</td>
<td>0.9842</td>
<td>0.9846</td>
<td>0.9850</td>
<td>0.9854</td>
<td>0.9857</td>
</tr>
<tr>
<td>2.2</td>
<td>0.9861</td>
<td>0.9864</td>
<td>0.9868</td>
<td>0.9871</td>
<td>0.9875</td>
<td>0.9878</td>
<td>0.9881</td>
<td>0.9884</td>
<td>0.9887</td>
<td>0.9890</td>
</tr>
<tr>
<td>2.3</td>
<td>0.9893</td>
<td>0.9896</td>
<td>0.9898</td>
<td>0.9901</td>
<td>0.9904</td>
<td>0.9906</td>
<td>0.9909</td>
<td>0.9911</td>
<td>0.9913</td>
<td>0.9916</td>
</tr>
<tr>
<td>2.4</td>
<td>0.9918</td>
<td>0.9920</td>
<td>0.9922</td>
<td>0.9925</td>
<td>0.9927</td>
<td>0.9929</td>
<td>0.9931</td>
<td>0.9932</td>
<td>0.9934</td>
<td>0.9936</td>
</tr>
<tr>
<td>2.5</td>
<td>0.9938</td>
<td>0.9940</td>
<td>0.9941</td>
<td>0.9943</td>
<td>0.9945</td>
<td>0.9946</td>
<td>0.9948</td>
<td>0.9949</td>
<td>0.9951</td>
<td>0.9952</td>
</tr>
<tr>
<td>2.6</td>
<td>0.9953</td>
<td>0.9955</td>
<td>0.9956</td>
<td>0.9957</td>
<td>0.9959</td>
<td>0.9960</td>
<td>0.9961</td>
<td>0.9962</td>
<td>0.9963</td>
<td>0.9964</td>
</tr>
<tr>
<td>2.7</td>
<td>0.9965</td>
<td>0.9966</td>
<td>0.9967</td>
<td>0.9968</td>
<td>0.9969</td>
<td>0.9970</td>
<td>0.9971</td>
<td>0.9972</td>
<td>0.9973</td>
<td>0.9974</td>
</tr>
<tr>
<td>2.8</td>
<td>0.9974</td>
<td>0.9975</td>
<td>0.9976</td>
<td>0.9977</td>
<td>0.9977</td>
<td>0.9978</td>
<td>0.9979</td>
<td>0.9979</td>
<td>0.9980</td>
<td>0.9981</td>
</tr>
<tr>
<td>2.9</td>
<td>0.9981</td>
<td>0.9982</td>
<td>0.9982</td>
<td>0.9983</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9985</td>
<td>0.9985</td>
<td>0.9986</td>
<td>0.9986</td>
</tr>
<tr>
<td>3.0</td>
<td>0.9987</td>
<td>0.9987</td>
<td>0.9987</td>
<td>0.9988</td>
<td>0.9988</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9989</td>
<td>0.9990</td>
<td>0.9990</td>
</tr>
</tbody>
</table>
Appendix A, Part 3 – A standard deck of 52 cards

<table>
<thead>
<tr>
<th>Black cards</th>
<th>Red cards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clubs</td>
<td>Spades</td>
</tr>
<tr>
<td>A ♠</td>
<td>A ♠</td>
</tr>
<tr>
<td>2 ♠</td>
<td>2 ♠</td>
</tr>
<tr>
<td>3 ♠</td>
<td>3 ♠</td>
</tr>
<tr>
<td>4 ♠</td>
<td>4 ♠</td>
</tr>
<tr>
<td>5 ♠</td>
<td>5 ♠</td>
</tr>
<tr>
<td>6 ♠</td>
<td>6 ♠</td>
</tr>
<tr>
<td>7 ♠</td>
<td>7 ♠</td>
</tr>
<tr>
<td>8 ♠</td>
<td>8 ♠</td>
</tr>
<tr>
<td>9 ♠</td>
<td>9 ♠</td>
</tr>
<tr>
<td>10 ♠</td>
<td>10 ♠</td>
</tr>
<tr>
<td>Jack ♣</td>
<td>Jack ♣</td>
</tr>
<tr>
<td>Queen ♣</td>
<td>Queen ♣</td>
</tr>
<tr>
<td>King ♣</td>
<td>King ♣</td>
</tr>
</tbody>
</table>

Appendix A, Part 4 - Results for the total of two 6-sided dice

<table>
<thead>
<tr>
<th>+</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1,1)</td>
<td>(1,2)</td>
<td>(1,3)</td>
<td>(1,4)</td>
<td>(1,5)</td>
<td>(1,6)</td>
</tr>
<tr>
<td>2</td>
<td>(2,1)</td>
<td>(2,2)</td>
<td>(2,3)</td>
<td>(2,4)</td>
<td>(2,5)</td>
<td>(2,6)</td>
</tr>
<tr>
<td>3</td>
<td>(3,1)</td>
<td>(3,2)</td>
<td>(3,3)</td>
<td>(3,4)</td>
<td>(3,5)</td>
<td>(3,6)</td>
</tr>
<tr>
<td>4</td>
<td>(4,1)</td>
<td>(4,2)</td>
<td>(4,3)</td>
<td>(4,4)</td>
<td>(4,5)</td>
<td>(4,6)</td>
</tr>
<tr>
<td>5</td>
<td>(5,1)</td>
<td>(5,2)</td>
<td>(5,3)</td>
<td>(5,4)</td>
<td>(5,5)</td>
<td>(5,6)</td>
</tr>
<tr>
<td>6</td>
<td>(6,1)</td>
<td>(6,2)</td>
<td>(6,3)</td>
<td>(6,4)</td>
<td>(6,5)</td>
<td>(6,6)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>+</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>
Appendix B – Glossary and Index

95% confidence statement – Page 120, Section 4.4
A confidence statement is a summary statement of the findings of a study. All confidence statements have the form ‘We are 95% confident that the true proportion of (parameter of interest) will be between (low value of confidence interval) and (high value of confidence interval).’

Back to Back Stem Plots – Page 184, Section 5.6
A stem plot in which two sets of numerical data share the stems in the middle, with one set having its leaves going to the right and the other set having its leaves going to the left.

Bar Graph – Page 137, Section 5.1
A graph in which each bar shows how frequently a given category occurs. The bars can go either horizontally or vertically. Bars should be of consistent width and need to be equally spaced apart. The categories may be placed in any order along the axis.

Bias - Page 95, 103, Section 4.1, 4.2
Bias occurs when a measurement repeatedly reports values that are either too high or too low.

Bin Width
See Class Size

Bivariate Data - Page 200, Section 6.1
Numerical data that measures two variables.

Blind Study - Page 126, Section 4.5
A study in which the subject does not know exactly what treatment they are getting.

Block Design - Page 128, Section 4.5
A study in which subjects are divided into distinct categories with certain characteristics (for example, males and females) before being randomly assigned treatments in an experiment.
**Box Plot (Box and Whisker Plot)** - Page 171, Section 5.5
A display in which a numerical data set is divided into quarters. The ‘box’ marks the middle 50% of the data and the ‘whiskers’ mark the upper 25% and lower 25% of the data.

**Categorical Variable** - Page 93, 136, Section 4.1, 5.1
Variables that can be put into categories, like favorite color, type of car you own, your sports jersey number, etc...

**Census** - Page 97, 101, Section 4.1, 4.2
A special type of study in which data is gathered from every single member of the population.

**Center** - Page 147, 156, Section 5.2, 5.3
Typically, it is the mean, median, or the mode of a data set. In a normal distribution curve the mean, median, and mode all mark the center. If a data set is skewed or has outliers, it is standard practice to use the median as the center.

**Chance Behavior** - Page 26, Section 2.1
Events whose outcomes are not predictable in the short term, but have long term predictability.

**Class Size (Bin Width)** - Page 164, Section 5.4
A consistent width that all bars on a histogram have. A quick estimation of a reasonable class size is to roughly divide the range by a value from about 7 to 10.

**Coincidence** - Page 215, Section 6.2
A relationship between two variables that simply occurs by chance.

**Combination** - Page 15, Section 1.4
An arrangement of a set of objects in which the order does not matter.

\[ \binom{n}{r} = \frac{n!}{r!(n-r)!} \]

**Common Response** - Page 214, Section 6.2
A situation in which two variables have a strong correlation but are actually responding to an additional lurking variable.

**Complement of an Event** - Page 26, Section 2.1
The probability of an event, ‘A’, NOT occurring. It can be thought of the opposite of an event and can be notated as Ac or A’. P(A’) = 1 – P(A)
Compound Event - Page 33, Section 2.2
An event with two or more steps such as drawing a card and then rolling a die.

Conditional Probability - Page 54, Section 2.5
The probability of a particular outcome happening assuming a certain prerequisite condition has already been met. A clue that a conditional probability is being considered is the word ‘given’ or the vertical bar symbol, |.

Confidence Interval - Page 119, Section 4.4
The range of answers included within the margin of error. Typically, we use a 95% confidence interval meaning it is very likely (95% chance) that the parameter lies within this range.

Confounding - Page 215, Section 6.2
Occurs when two variables are related, but it is not a clear cause/effect relationship because there may be other variables that are influencing the observed effect.

Context - Page 156, 204 Section 5.3
The specific realities of the situation we are considering. We often consider the labels and units when defining the context.

Contingency Table
See Two-Way Table

Control - Page 125, Section 4.5, 6.1
A researcher in an experiment establishes control when one of the treatment groups receives either a placebo or the currently accepted treatment.

Control Group - Page 125, Section 4.5
A group in an experiment that does not receive the actual treatment, but rather receives a placebo or a known treatment.

Convenience Sample - Page 106, Section 4.2
A biased sampling method in which data is only gathered from those individuals who are easy to access or are conveniently located.

Correlation (r) - Page 210-213, Section 6.2
A statistic that is used to measure the strength and direction of a linear correlation whose values range from -1 to 1. The sign of the correlation (+/-) matches the sign of the slope of the regression equation. A correlation value of 0 indicates no linear relationship whatsoever.
Data - Page 93, Section 4.1
A collection of facts, measurements, or observations about a set of individuals.

Density Curve - Page 236, Section 7.1
A curve that gives a rough description of a distribution. The curve is smooth and always has an area equal to 1 or 100%.

Direct Cause and Effect - Page 214, Section 6.2
A situation in which one variable causes a specific effect to occur with no lurking variables.

Direction - Page 210, Section 6.2
One of three general results reported for a linear regression. It will be reported as either being positive, negative, or 0.

Disjoint
See Mutually Exclusive Events

Dot Plot - Page 154, Section 5.3
A simple display that places a dot above each marked value on the x-axis. There is a dot for each result, so results that occur more than once will be shown by stacked dots.

Double Blind - Page 126, Section 4.5
A study in which neither the person administering the treatments nor the subject knows which treatment is being given.

Empirical Rule (68-95-99.7 Rule) - Page 238, Section 7.1
A rule stating that in a normal distribution, 68% of the data is located within one standard deviation of the mean, 95% of the data is located within two standard deviations of the mean, and 99.7% of the data is located within three standard deviations of the mean.

Event - Page 1, Section 1.1
Any action from which a result will be recorded or measured.

Expected Value - Page 67, Section 3.1
The average result over the long run for an event if repeated a large number of times.

Experiment - Page 97, 124, Section 4.1, 4.5
A study in which the researchers impose a treatment on the subjects.
**Explanatory Variable** - Page 125, 200, Section 4.5, 6.1
The x-axis variable. It can often be viewed as the ‘cause’ variable or the independent variable.

**Factorial** - Page 7, Section 1.2
A number followed by an exclamation point indicated repeated multiplication down to 1. For example, $4! = 4 \times 3 \times 2 \times 1$.

**Fair Game** - Page 76, Section 3.2
A game in which neither the player nor the house has an advantage. An average player over the long run will neither gain nor lose money. In other words, the expected value of the game is the same as the cost to play the game.

**Five-Number Summary** - Page 171, Section 5.5
A description of data that includes the minimum, first quartile, median, third quartile, and maximum numbers which can be used to create a box plot.

**Form** - Page 204, Section 6.1
A general description of the pattern in a scatterplot. Typical descriptions include linear, curved, or random (no specific form).

**Frequency Table** - Page 137, Section 5.1
A table that shows the number of occurrences in each category.

**Fundamental Counting Principle** - Page 4, Section 1.2
A rule that states that in order to find the number of outcomes for a multi-step event, simply multiply the number of possibilities from each step of the event.

**Histogram** - Page 164, Section 5.4
A special bar graph for a numerical data set. In a histogram, each bar has the same bin width and there is no space between consecutive bars. Each bar tracks the number or frequency of results in its given range.

**Independent Events** - Page 33, Section 2.2
Two events are independent if the outcome of one event does not change the probability for the outcome for the other event.

**Individual** - Page 93, Section 4.1
This is the person, animal, or object being studied.

**Interquartile Range (IQR)** - Page 174, Section 5.5
The distance between the lower and upper quartiles. $IQR = Q_3 - Q_1$
**Instrument of Measurement** - Page 94, Section 4.1
This is the tool used to make measurements. Some examples of instruments include rulers, scales, thermometers, or speedometers.

**Intersection of Events** - Page 42, Section 2.3
In a Venn Diagram, it includes the results that are members of more than one group simultaneously. We use the symbol, ∩, to indicate the intersection and think of the intersection as those parts of the diagram that include both A and B.

**Law of Large Numbers** - Page 26, 84, Section 2.1, 3.3
A rule that states that we will eventually get closer to the theoretical probability as we greatly increase the number of times an event is repeated.

**Line Graph**
See Time Plot

**Lurking Variable** - Page 124, 214, Section 4.5, 6.2
An additional variable that was not taken into account in a particular situation.

**Margin of Error** - Page 119, Section 4.4
It is the distance we move above and below the mean to help establish a 95% confidence interval in which we believe the true parameter is located. An approximation for the margin of error for a 95% confidence interval is M.O.E = ± \( \frac{1}{\sqrt{n}} \) where n represents the sample size.

**Mean (Average)** - Page 147, 237, Section 5.2, 7.1
The sum of all the numbers divided by the number of values in a data set. It is also located at the center of a normal distribution and is a good measure of center for symmetric data sets.

**Median** - Page 147, Section 5.2
The data result in the middle of a data list that has been organized from smallest to largest. If there are two middle data values, then the median is located halfway between those two values. Visually, it marks the spot where half of the area of a graph is below the median and half of the area is above the median. It is common to use the median as your measure of center for skewed data sets or data sets that contain outliers.
Mode - Page 147, Section 5.2
The result that appears most frequently in a data set. It also occurs at the highest point of a density curve.

Multistage Random Sample - Page 104, Section 4.2
A sampling technique that uses randomly selected sub-groups of a population before random selection of individuals occurs.

Mutually Exclusive Events (Disjoint) - Page 41, Section 2.3
Outcomes that cannot occur at the same time. For example, if a single card is drawn from a standard deck, the outcomes of a diamond and a black card are mutually exclusive.

Negative Linear Association - Page 205, Section 6.1
A situation such that as one numerical variable increases, another numerical variable decreases.

Non-Response - Page 108, Section 4.2
A non-sampling error in which individuals selected for a study do not participate or do not answer questions in a survey.

Normal Distribution Curve - Page 237, Section 7.1
A bell-shaped curve that describes a symmetrical data set such that the most frequent results occur near the mean and results become less frequent as you move further from the mean.

Numerical Variable - Page 93, Section 4.1
A variable that can be assigned a numerical value, such as height, distance, or temperature.

Observational Study - Page 97, 124, Section 4.1, 4.5
A study in which researchers do not impose a treatment on the individuals being studied. Data is collected by observing the individuals, surveying the individuals, or collecting data from the individuals from information that is already available. (Observe but do not disturb)

Outcome - Page 1, Section 1.1
A possible result of an event.

Outlier - Page 155, 178, 204, Section 5.3, 5.5, 6.1
A value that is unusual when compared to the rest of a data set. High outliers will be greater than $Q_3 + 1.5 \text{ IQR}$. Low outliers will be below $Q_1 - 1.5 \text{ IQR}$. 
**Parallel Box Plots** - Page 183, Section 5.6

Multiple box plots graphed on the same axes to compare multiple data sets.

**Parameter** - Page 111, Section 4.2

A value that describes the truth about a population. The value is frequently unknown so a parameter is often given as a description of truth.

**Permutation** - Page 10, Section 1.3

A specific order or arrangement of a set of objects or items. In a permutation, the order in which the items are selected matters.

**Pictograph** - Page 141, Section 5.1

A bar graph that uses pictures instead of bars. These graphs can be misleading because pictures measure height and width, where bar graphs measure only height. To be effective, all the pictures used must be the same size.

**Pie chart** - Page 139, Section 5.1

A graph which shows each category as a part of the whole in a circle graph. Pie charts can be used if exactly 100% of the results from a particular situation are known.

**Placebo** - Page 126, Section 4.5

A fake treatment that is similar in appearance to the real treatment.

**Placebo Effect** - Page 126, Section 4.5

The placebo effect occurs when a subject starts to experience changes simply because they believe they are receiving a treatment.

**Population** - Page 101, Section 4.2

The entire group of individuals we are interested in. A population is often described using the word ‘all’.

**Positive Linear Association** - Page 205, Section 6.1

A situation in which as one numerical variable increases, the other numerical variable also increases.

**Prime Number** - Page 42, Section 2.3

A number that has exactly 2 factors. Remember, 1 is not a prime number!

**Probability** - Page 26, Section 2.1

The likelihood of a particular outcome occurring.
**Probability Model** - Page 49, Section 2.4
A table that lists all the values for the outcomes of an event and their respective probabilities. The sum of all the probabilities in a probability model must equal 1.

**Processing Errors** - Page 109, Section 4.2
An error commonly made due to issues like poor calculations or inaccurate recording of results.

**Prospective Studies** - Page 124, Section 4.5
A study which follows up with study subjects in the future in an effort to see if there were any long-term effects.

**Quartile 1** - Page 172, Section 5.5
The median of all the values to the left of the median. Do not include the median itself in this calculation if the median is one of the data points.

**Quartile 3** - Page 172, Section 5.5
The median of all the values to the right of the median. Do not include the median itself in this calculation if the median is one of the data points.

**Random Digit Table** - Pages 82, 114, Section 3.3, 4.3, Appendix A
A long list of randomly chosen digits from 0 to 9, usually generated by computer software or calculators. A table of random digits can be found in Appendix A, Part 1.

**Random Event** - Page 26, Section 2.1
An event is random if it does not have short-term predictability but it has long-term predictability. For example, a coin flip is a random event because we do not know what will happen on the next flip, but we can be reasonably sure that about 50% of a long series of flips will land on heads.

**Random Sampling Error** - Page 107, Section 4.2
Even though a sample is randomly selected, it is entirely possible that a particular result within the population will be over-represented causing us to be significantly different from the parameter. Larger sample sizes reduce random sampling error. The margin of error is stated with most studies to account for random sampling error.

**Range** - Page 148, 174, Section 5.2, 5.5
A basic description of how spread out a data set is. It is calculated by subtracting the smallest number from the largest number in a data set.
Reliability - Page 95, Section 4.1
How consistently a particular measurement technique gives the same, or nearly the same measurement.

Response Bias - Page 109, Section 4.2
Occurs when an individual responds to a survey with an incorrect or untruthful answer. This type of bias can frequently happen when questions are potentially sensitive or embarrassing.

Response Variable - Page 125, 200, Section 4.5, 6.1
This is the y-axis variable. It can often be thought of as the ‘effect’ variable or dependent variable.

Retrospective Study - Page 124, Section 4.5
A study in which information about a subject’s past is used in the study.

Sample - Page 102, Section 4.2
A representative subset of a population.

Sample Space - Page 1, Section 1.1
A list of all the possible outcomes that may occur.

Sample Survey - Page 97, Section 4.1
A survey that uses a subset of the population in order to try to make predictions about the entire population.

Sampling Frame - Page 103, Section 4.2
A list of all members of a population.

Scatterplot - Page 200, Section 6.1
Graphs that represent a relationship between two numerical variables where each data point is shown as a coordinate point on a scaled grid.

SCOFD - Page 203-206, Section 6.1
This is an acronym used for the description of a scatterplot and stands for Strength, Context, Outliers, Form, and Direction.

Simple Random Sample (SRS) - Page 103, Section 4.2
A sample where all possible groups of a particular size are equally possible. It can be thought of as putting names of all members of a population in a hat and randomly drawing until the desired sample size is reached.
**Simulation** - Page 82, Section 3.3
A model of a real situation that can be used to make predictions about what might really happen. Often, tables of random digits are used to carry out simulations.

**Skewed Distribution** - Page 155, 236, Section 5.3, 7.1
A distribution in which the majority of the data is concentrated on one end of the distribution. Visually, there is a ‘tail’ on the side with less data and this is the direction of the skew.

**SOCCS** - Page 154-156, Section 5.3
An acronym used to remember the key information to discuss for a distribution: Shape, Outliers, Center, Context, and Spread.

**Spread** - Page 156, Section 5.3
A way to measure variability of a data set. Common measures of spread are the range, standard deviation, and IQR.

**Standard Deviation** - Page 174, 237, Section 5.5, 7.1
A measure of spread relative to the mean of a data set. Use this measurement for any data set which is approximately normally distributed.

**Statistic** - Page 111, Section 4.2
A number that describes results from sample. This number is often a percentage and is used to make an approximation of the parameter.

**Stem Plot** - Page 157, Section 5.3
A method of organizing data that sorts the data in a visual fashion. The stem is made up of all the leading digits of a piece of data and the leaf is the final digit. No commas or decimal points should be used in a stem plot.

**Stratified Random Sample** - Page 104, Section 4.2
A sample in which the population is divided into distinct groups called strata before a random sample is chosen from each strata.

**Strength** - Page 203, 210, Section 6.1, 6.2
One of three measurements reported for a best-fit line that describes how close the data is to being perfectly linear.

**Subjects** - Page 125, Section 4.5
The individuals that are being studied in an experiment.
**Symmetrical Distribution** - Page 155, Section 5.3
A distribution in which the left side of the distribution looks like a mirror image of the right side of the distribution.

**Systematic Random Sample** - Page 104, Section 4.2
A sampling method in which the first selection is made randomly and then a 'system' is used to make the remaining selections. For example, randomly select one person from a list and then select every 14th person after that.

**Theoretical Model** - Page 26, Section 2.1, 3.3
A model that gives a picture of exactly the frequencies of what should happen in a situation involving probability.

**Theoretical Probability** - Page 26, Section 2.1
A mathematical calculation of the likelihood that a given outcome will occur.

**Time Plot (Line Graph)** - Page 145, Section 5.2
A graph that shows how a numerical variable changes over time.

**Tree Diagram** - Page 2, 4, 48 Section 1.1, 1.2, 2.4
A visual representation of a multi-step event where each successive step branches off from the previous step.

**Two-Way Table (Contingency Table)** - Page 55, Section 2.5
A table which tracks two characteristics from a set of individuals. For example, we might track gender and grade of all the students in your high school.

**Undercoverage** - Page 107, Section 4.2
A sampling error in which an entire group or groups of subjects are left out or underrepresented in a study.

**Union of Events** - Page 41, Section 2.3
A union includes all results that are in either one category, another category, or both categories in a Venn diagram. We use the symbol $\cup$ and can think of a union as anything belonging to either A, B, or both A and B.

**Validity** - Page 95, Section 4.1
A measurement technique is valid if it is a reasonable way to collect data.

**Variables** - Page 93, Section 4.1
Characteristics about the individuals in a study in which researchers might have interest.
**Venn Diagrams** - Page 29, 42, Section 2.1, 2.3
Diagrams that represent outcomes or categories using intersecting circles.

**Voluntary Response Survey** - Page 105, Section 4.2
A biased sampling method in which participants get to choose whether or not to participate in the survey. The bias occurs because those who are most passionate about an issue will be more likely to respond.

**Wording of a Question** - Page 108, Section 4.2
The wording of a question can be used to manipulate individuals in a survey such that they are more likely to respond a certain way in the survey which causes bias.

**Z-Score** - Page 245, Section 7.2
A measure of the number of standard deviations a particular data point is away from the mean in a normal distribution. If a z-score is positive, the value is larger than the mean and if it is negative, it is less than the mean.
Appendix C – Calculator Help

This appendix is not meant to be a full guide for calculators common to students who take this course. Rather, it is intended to highlight some of the locations to access a variety of commands commonly used on a TI-30XS Multiview Scientific Calculator and a TI-84 Plus Graphing Calculator. One online source that can be helpful for those of you with graphing calculator issues can be found on the Prentice Hall website at http://www.prenhall.com/divisions/esm/app/calc_v2/.

**Topic 1 - Combinations, Permutations, and Factorials**

**TI-30 XS Multiview**

Access located in the **prb** menu. Enter the value for n, select nCr or nPr, and then enter the value for r.

**TI-84 Plus**

Access located in the **Math, PRB** menu. Enter the value for n, select nCr or nPr, and then enter the value for r.

**Topic 2 – Random Number Generators**

**TI-30 XS Multiview**

Access located in the **prb** menu. Select rand, enter lowest value, enter highest value.

**TI-84 Plus**

Access located in the **Math, PRB** menu. Select RandInt, enter lowest value, enter highest value, enter number of random values desired.
Topic 3 – Means and Standard Deviations

TI-30 XS Multiview
Enter data into L1 in the data menu. Press 2nd data (stat) and select 1-Var Stats. Arrow down to find the mean, $\bar{x}$, and the standard deviation, $s_x$.

TI-84 Plus
Enter data in L1 by selecting STAT and EDIT. Press STAT and CALC and then select 1-Var Stats. Arrow down to find the mean, $\bar{x}$, and the standard deviation, $s_x$.

Topic 4 – Correlations, Slopes, and Y-Intercepts

TI-30 XS Multiview
Enter data into L1 and L2 in the data menu. Press 2nd data (stat) and select 2-Var Stats for L1 and L2. Arrow down to find the slope (a), the y-intercept (b) and the correlation coefficient (r).

TI-84 Plus
Enter data in L1 and L2 by selecting STAT and EDIT. Press STAT and CALC and then select LinReg(ax+b). Be sure the Xlist and Ylist are L1 and L2. If you wish to store you equation into the Y= menu, press VARS, Y-VARS, Function, and Y1. If the correlation (r) does not show up, go to 2nd CATALOG and select DiagnosticOn.

Topic 5 – Normal Distributions

TI-30 XS Multiview
This calculator cannot perform normal distribution calculations.

TI-84 Plus
To find the percent of area in a normal curve, select 2nd DISTR and select normalcdf(). Enter the lower bound, upper bound, mean, and standard deviation. To find a value from a percentile in a normal distribution, select 2nd DISTR and select invNorm( . Enter the %tile, mean, and standard deviation.

Image References
Random Digit Table   http://uwsp.edu/math
Normal Distribution Table http://www.regentsprep.org
TI-30XS Multiview Calculator   http://education.ti.com
TI-84 Plus Graphing Calculator   http://education.ti.com
Appendix D – Selected Answers

Problem Set 1.1
1a) S={HH, HT, TH, TT}
1b) 4
2a) S={HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
2b) 8
3a) S={HHHH, HHHT, HHTH, HHTT, HTTH, HTHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}
3b) 16
4) S={1, 2, 3, 4}
5a) Grid
5b) S={2, 3, 4, 5, 6, 7, 8}
6a) Grid
6b) S={1, 2, 3, 4, 6, 8, 9, 12, 16}
6c) 9
6d) 4

Problem Set 1.2
1) 60
2) 100,000
3) 30,240
4) 362,880
5) 720
6) 144
7) 40,320
8) 27,600
9) 35,152
10) 10,000
11) 504
12) 5,040
13) 10
14) 256
15) 360
16) 1,296
17) 256
18) 64
19) HH, HT, TH, TT
20a) Tree Diagram
20b) Answers will vary.
20c) 12
21) Tree Diagram
22) Tree Diagram
23) Tree Diagram
24a) S = { P1C1, P1C2, P2C1, P2C2, P3C1, P3C2, P4C1, P4C2 }
24b) 8
25a) Grid
25b) 9

Problem Set 1.3
1a) 336
1b) 24
1c) 60
1d) 1
2) 840
3) 336
4) 151,200
5) 5,040
6) 12
7) 210
8) 5,527,200
9) 120
10) 60
11) 120
12) 120
13) 300
14a) No
14b) Yes
14c) No
14d) Yes
15) 48
16) Tree Diagram, 12
17) 9

Problem Set 1.4
1a) 1
1b) 15
1c) 1
1d) 35
2) 22,100
3) 120
4) 126
5) 35
6) 462
7) 35
8) 4,102,565,544
9a) 351
9b) 171
9c) 1035
9d) 513
10) 5,040
11a) 625
11b) 120
12) S={HH, HT, TH, TT}
13a) No
13b) Yes
13c) No

Problem Set 1.5
1) Answers will vary.
2) 252
3) 30,240
4) 5,100,480
5) 42,504
6a) 2,598,960
6b) 311,875,200
7) 1,260
8a) 495
8b) 210
8c) 420
9) 18,480
10) 166,320
11) 2,025
12) 646,646
13) 194,040
14) 3,024
15) 120
16) 90
17) 35
18) 120
19) 7
20) 35,904

Chapter 1 Review
1a) Grid
1b) S={2, 3, 4, 5, 6, 7, 8, 9, 10}
1c) Lose
2) Tree Diagram
Problem Set 2.1
1a) 0.5
1b) 0.23
1c) 0.08
1d) 0.08
1e) 0.25
1f) 0.02
1g) 0.04
1h) 0.75
2a) 0
2b) 0.25
2c) 0.7
2d) 0.7
3a) 0.17
3b) 0
3c) 0.5
3d) 0.5
3e) 0.33
4a) 0.2
4b) 0.8
4c) 0
4d) 1
4e) 0.2
5a) 0.32
5b) 0.68
5c) 0.03
5d) 0
5e) 0.97
5f) 0.52
6) 0.47

Problem Set 2.2
1) Answers will vary.
2a) 0.25
2b) 0.06
2c) 0.01
2d) 0.05
3a) Independent
3b) Not Independent
3c) Not Independent
3d) Not Independent
3e) Independent
3f) Not Independent
3g) Not Independent
4) 0.17
5a) 0.06
5b) 0.24
5c) 1/221
5d) 1/5225
6a) Venn Diagram
6b) 25%
7a) Venn Diagram
7b) Venn Diagram
7c) Venn Diagram
8) 40%
9a) 0.73
9b) 0.73
9c) 1
9d) 0.2
10a) 0.28
10b) 0.07
10c) 0.28
11) 12
12) 0.05
13) 0.66
14) 0.13
15) 56
16) 665

Problem Set 2.3
1) Answers will vary.
2) Answers will vary.
3) Answers will vary.
4a) Not Mutually Exclusive
4b) Mutually Exclusive
4c) Mutually Exclusive
4d) Not Mutually Exclusive
4e) Not Mutually Exclusive
4f) Mutually Exclusive
4g) Mutually Exclusive
5a) Venn Diagram
5b) 8
5c) 32
5d) 50
5e) 110
6a) Venn Diagram
6b) 25%
7a) Venn Diagram
7b) Venn Diagram
7c) Venn Diagram
8) 40%
9a) 0.73
9b) 0.73
9c) 1
9d) 0.2
10a) 0.28
10b) 0.07
11) 12
12) 0.05
13) 0.66
14) 0.13
15) 56
16) 665
Problem Set 2.4
1) P(RR) = 0.22, P(RG) = 0.5, P(GG) = 0.28
2) 0.53
3) P($40) = 0.44, P($60) = 0.22, P($200) = 0.22, P($300) = 0.11
4) P(0) = 0.02, P(1) = 0.14, P(2) = 0.42, P(3) = 0.42
5) P(HD) = 0.08, P(HD') = 0.42, P(TD) = 0.08, P(TD') = 0.42
6a) 8
6b) Each probability = 0.125
7) P($40) = 0.44, P($60) = 0.22, P($200) = 0.22, P($300) = 0.11
8) P(H2) = 0.22, P(H3) = 0.2, P(D2) = 0.27, P(D3) = 0.31
9) P(SS) = 0.765, P(SF) = 0.135, P(FS) = 0.085, P(FF) = 0.015
10) 0.62
11a) 0.49
11b) 0
12) 15,600
13) 17,576
14) 5%

Problem Set 2.5
1a) 0.52
1b) 0.48
1c) 0.44
1d) 0.66
2) 0.002
3a) 0.53
3b) 0.69
3c) 0.57
4) 0.8
5) 0.08
6) 0.17
7a) 0.86
7b) 0.57
7c) 1
7d) 0.6
8a) 0.55
8b) 0.8
8c) 0.6
8d) 0.44
9a) 0.12
9b) 0.85
9c) 0.47
9d) 0.35
10a) 0.5
10b) 0.43
11a) 0.48
11b) 0.77
11c) 0.3
11d) 0.4
12) 0.04
13) P(High, Sport) = 0.15, P(High, Sport') = 0.10, P(High', Sport) = 0.45, P(High', Sport') = 0.30
14a) 0.25
14b) 0.125
14c) 0.55
15) 0.84
16) 18,928,000

Chapter 2 Review
1a) 0.14
1b) 0.11
1c) 0.76
2a) 0.17
2b) 0.17
2c) 0.67
2d) 0.25
2e) 0.42
2f) 0.56
2g) 0.64
3a) Venn Diagram
3b) 20
3c) 0.75
3d) 0.44
3e) 0.27
4a) Tree Diagram
4b) 0.72
5a) 0.33
5b) 0.21
5c) 0.06
5d) 0.13
5e) 0.16
6a) 0.54
6b) 0.37
6c) 0.6
6d) 0.23
6e) 0.4
7a) 0.27
7b) 0.33
7c) 0.31
7d) 0.58
7e) 0.32
8) 0.59
9) 0.14
10a) Tree Diagram
10b) 0.17
10c) 0.39
11a) 0.04
11b) 0.86
11c) 0.14
11d) 0.43
12a) 0.43
12b) 0.27
12c) 0.52
12d) 0.05
13a) Tree Diagram
13b) Probability Model
13c) 0.13
14a) 0.10
14b) 0.90
14c) 0.38
15a) Venn Diagram
15b) 0.47
15c) 0.43
16a) 240
16b) 0.09
16c) 0.25
17a) 7,140
17b) 0.14
17c) 0.16
18a) No
18b) Yes

Problem Set 3.1
1) 2.25
2a) Probability Model
2b) $7.60
3) $6.00
4) $8.40
5) 2.25
6a) $8
6b) $3
6c) $3
7) 3.5
8) $25
9a) 0.06
9b) 0.33
9c) $9.65
10a) 0.1 = 10%
10b) $11
11a) 2x + 3y = 1.07
11b) x + y = 0.5
11c) x = 0.43, y = 0.07
12a) 0.4462
12b) 0.5385
Problem Set 3.2
1) $2
2) $4.50
3) Yes
4) No
5) 58
6) -$0.50
7) Lose $1.17
8) $260
9) Yes, $39,000
10) $124.25
11a) Tree Diagram
11b) Probability Model
11c) Lose $3.02
12) $3.33
13) No
14a) 0.55
14b) $3.70
14c) $5.70
15) 1,960
16) 0.2353
17a) 9
17b) $S = \{aa, ab, ac, ba, bb, bc, ca, cb, cc\}$

Problem Set 3.3
1) Answers will vary.
2a) 0.1
2b) Answers will vary.
2c) Answers will vary.
3a) Answers will vary.
3b) Answers will vary.
3c) Answers will vary.
4a) Answers will vary.
4b) Answers will vary.
4c) Answers will vary.
5a) Answers will vary.
5b) Answers will vary.
6a) Yes
6b) Approximately 10%
6c) Go to next line, continue
7a) Answers will vary.
7b) Answers will vary.
8) 0.0121
9a) 0.5306
9b) 0.4286
10a) 55
10b) 0.4330
10c) 0.2340

Chapter 3 Review
1a) Probability table
1b) Lose $1.75
1c) Don't play.
2a) Probability table
2b) Lose $3.85
2c) $0
3a) Answers will vary.
3b) Answers will vary.
4a) Answers will vary.
4b) Answers will vary.
5a) Tree diagram
5b) Probability table
5c) $70
6) $10.67
7a) 0.16
7b) Answers will vary.
7c) Answers will vary.
8a) 17
8b) 17.53
9) $3.67

Problem Set 4.1
1) Ind. = Players
   Cat. = Pos.
   Team Num. = BA, #AB’s, #hits, #SB
2) Ind. = Employees
   Cat. = Deg., Job Title
   Num. = Sal., Yrs, 401K, #Chil.
   Neither = Add., Phone#
3a) Numerical
3b) Categorical
3c) Categorical
3d) Numerical
3e) Neither
3f) Categorical
4a) Var. = Temp., Inst. = Therm., Units = Deg.
4b) Answers will vary
5a) Inst. = Foot Ruler, Units = # of 'feet'
5b) Answers will vary
6a) Reliability

Problem Set 4.2
1a) Parameter
1b) Parameter or Statistic
1c) Statistic
1d) Parameter
1e) Parameter
1f) Statistic
2a) All MBHS Seniors
2b) True # of seats
2c) 148 seniors
2d) 6.2 seats
2e) Voluntary Response
2f) 27.3%
2g) Voluntary Response and Non-Response Bias
2h) Likely too high, too big of a venue
3a) Stratified Rand. Samp.
3b) Systematic Rand. Samp.
3c) Voluntary Response
3d) SRS  
3e) Multi-Stage Rand. Samp.  
3f) Convenience Sample  
4a) Response Bias  
4b) Non-Response Bias  
4c) Response Bias  
4d) Random Sampling Error  
4e) Processing Error  
4f) Voluntary Response  
4g) Under-Coverage  
4h) Wording of Question  
5) 66.7%  
6) 16.7%  
7) 27.8%  
8) 2.78%  

Problem Set 4.3  
1a) Answers will vary  
1b) Common Results = 07, 20, 24, 17, 49, 43, 09, 06, 56, 41, 36, and 15  
2) Answers will vary  
3) Answers will vary  
4) Venn Diagram  
5a) Venn Diagram  
5b) 59%  
5c) 91%  

Problem Set 4.4  
1a) All MBHS Students  
1b) True % that text  
1c) 270 students  
1d) 0.6593  
1e) ± 6.09%  
1f) 0.5984 to 0.7202  
1g) Confidence Statement  
1h) Answers will vary  
2a) All eligible voters  
2b) True % leaning Dem.  
2c) 814 voters  
2d) 38.2%  
2e) ± 3.50%  
2f) 34.7% to 41.7%  
2g) Confidence Statement  
3a) 38.8% to 45.8%  
3b) No  
4a) All cans of Spaz  
4b) True % too little sugar  
4c) 480 cans  
4d) 8.96%  
4e) 4.56%  
4f) 4.40% to 13.52%  
4g) Confidence Statement  
4h) Yes  
5) History, Higher %  
6) Probability, Higher %  
7) 78.91%  
8) Sketches  

Problem Set 4.5  
1a) 450 pets  
1b) Treatment given  
1c) Change in Allergies  
1d) Yes  
1e) Experimental Design  
2a) 70 Prob. Students  
2b) The music  
2c) Probability Retention  
2d) Answers will vary  
2e) Answers will vary  
2f) Experimental Design  
3a) 982 contact wearers  
3b) Type of drop taken  
3c) Dry eye symptoms  
3d) Experimental Design  
3e) Answers will vary, Common result = 367, 468, 288, 229, 131  
4a) 744 UW students and 3,057 YPC attendees  
4b) The commercials  
4c) Desire to purchase phone  
4d) Experimental Design  
4e) Answers will vary  
5a) 39.7%  
5b) All MN teens  
5c) True % who like SBSP  
5d) 516 MN teens  
5e) 84.1%  
5f) 4.4%  
5g) 88.5%  
5h) Confidence Statement  
5i) Answers will vary  
6) 1,188,137,600  
7) 0.33  

Chapter 4 Review  
1) Flashcards  
2a) Bias  
2b) Validity  
2c) Reliability  
3a) Answers will vary  
3b) Common result = 1261, 4214, 1260, 4592, 1689  
4a) All women at UC  
4b) True % ‘Not Safe’  
4c) 350 women  
4d) 21.14%  
4e) ± 5.35%  
4f) 15.79% to 26.49%  
4g) Confidence Statement  
5a) 69 students  
5b) Type of test taken  
5c) Student scores on test  
5d) Answers will vary  
5e) Experimental Design  
6a) Observational Study  
6b) Sample Survey  
6c) Census  
6d) Experiment  
7a) Answers will vary  
7b) Answers will vary  
8) Experiment  
9a) $2.30  
9b) 158.6%  
10a) 200 arthritis sufferers  
10b) Treatment taken  
10c) Swelling of knuckles  
10d) Answers will vary  
10e) Experimental Design  
11) 23.3%  
12a) Eisenhower  
12b) McArthur = 83.7%  
12c) Meade = 88.0%  
12d) Eisenhower = 96.7%  
13a) Voluntary Response, Undercoverage  
13b) Non-Response  
13c) Response Bias  
13d) Wording of Question  
13e) Processing Error  
13f) Random Sampling Error  
13g) Voluntary Response, Undercoverage  
14a) Systematic Rand. Samp.  
14c) Stratified Rand. Samp.  
14e) Voluntary Resp. Samp.  
14f) Convenience Sample  
14g) Stratified Rand. Samp.  
14h) MultiStage Rand. Samp.
14j) SRS
15) Answers will vary

**Problem Set 5.1**
1a) Bar Graph
1b) Possibly
1c) 12.3%, 44.3°; 14.5%, 52.1°; 27.0%, 97.4°; 12.0%, 43.3°; 18.0%, 64.9°; 6.0%, 21.6°; 10.1%, 36.4°
1d) Pie Chart
2) Answers will vary
3a) – c) Answers will vary
4a) – c) Answers will vary
5a) Answers will vary
5b) Bar Graph
6a) – c) Answers will vary
7a) Players
7b) # = Number, POS = Position, GP = Games Played, G = Goals, A = Assists, P = Points, +/- = Plus/Minus Rating, PIM = Penalty Minutes, PP = Power Play Goals, SH = Shorthanded Goals, GW = Game Winning Goals, S = Shots, S% = Shooting Percentage
7c) Cat.=POS, #=Neither, Num.=All Other Variables
8) 32,768
9) 0.0000305

**Problem Set 5.2**
1a) $\bar{x} = 34.33$, Med = 29, No Mode, Range = 50
1b) $\bar{x} = 44.57$, Med = 48, Mode = 22, Range = 54
1c) $\bar{x} = 62.80$, Med = 62, No Mode, Range = 100
2) 171.6 lbs.
3a) 61.2 feet
3b) 38.4 feet
3c) 4.98 feet
4) Median = 32 or $\bar{x} = 31$
5a) $\bar{x} = 63.4$, Med = 70.5, No Mode, Range = 72, D
5b) $\bar{x} = 70.9$, Med = 70.5, Mode = 70, Range = 24, Mean Changed, C-
5c) $\bar{x} = 70.3$
5d) 74.2
6a) $5,750; 5,150; 2,760; 3,100; 2,500; 5,450; 1,280; 3,130; 2,350; 3,675$
6b) 30.7%; 24.2%; 13.9%; 14.0%; 17.6%; 24.6%; 9.8%; 12.3%; 8.6%; 12.8%
6c) $\bar{x} = 16.9\%$, Med = 14.0%, No Mode, Range = 22.1%
6d) Answers will vary
7) Answers will vary
8a) Time Plot
8b) Answers will vary
9) 30,240
10) 210
11) 220
12) 168
13) 990

**Problem Set 5.3**
1a) $\bar{x} = 65$, Med = 70, Mode = 70, Range = 64
1b) Answers will vary
2a) Stem Plot
2b) Answers will vary
3a) Split-Stem Plot
3b) Med = 72.5%
3c) Lower
3d) Answers will vary
4a) Split-Stem Plot
4b) Answers will vary
4c) Answers will vary
5a) $\bar{x} = 75.62$, Med = 77, Mode = 92
5b) Median
6a) i) Skewed Left
   ii) Roughly Symmetrical
   iii) Roughly Symmetrical
   iv) Skewed Right
6b) i) Median ii) Median
   iii) Similar iv) Mean
6c) Answers will vary
7a) Stem Plot
7b) $\bar{x} = 63.4$, Med = 71, Mode = 82, Range = 66
7c) Answers will vary
8a) Plus/Minus
8b) Dot Plot
8c) Answers will vary
9a) All Springfield Res.
9b) True % who like Simp.
9c) 1.245 Springfield Res.
9d) 80.5%
9e) ±2.8%
9f) 77.7% to 83.3%
9g) Confidence Statement

**Problem Set 5.4**
1a) 450
1b) $11,000
1c) Answers will vary
2a) Answers will vary
2b) 2 or 137 to 139 lbs.
2c) 76.7%
3a) Over 75 bin wrong
3b) Answers will vary
4a) Histogram
4b) Answers will vary
5a) – e) Sketches, Answers will vary
6a) Histogram
6b) Median
6c) $\bar{x} = 71.52$, Med = 74, Mode = 82, Range = 35
6d) Median
7a) Probability Model
7b) $-6$
7c) No
8a) 0.191
8b) 0.375
8c) 0.078
9a) 0.083
9b) 0.525
9c) 0.55

**Problem Set 5.5**
1a) Med = 210, IQR = 130
1b) Med = 85, IQR = 55
2a) {240, 340, 440, 600, 750}...
2b) No Outliers
2c) Answers will vary
3a) Box Plot
3b) Med = 1030, IQR = 500
3c) $\bar{x} = 992$, SD = 291.43
3d) Yes
4a) {2, 11, 31.5, 56, 206}...
4b) One High Outlier = 206
4c) $\bar{x} = 48.8$, SD = 54.7, Mean is larger than Median
4d) {2, 11, 20, 54.5, 106} $\bar{x} = 36.7$, SD = 31.9. Median, Mean, Q3, Max, and SD all changed.

5a) Box Plot

5b) Answers will vary.

5c) Answers will vary.

6a) {1142, 1215.5, 1371.5, 1601, 2717} Box Plot

6b) Outlier is Burj Khalifa

6c) Skewed Right

6d) Range = 109, IQR = 27

6f) Outlier is Burj Khalifa

7a) {48, 66.5, 77, 86, 97} Box Plot

7b) Answers will vary.

7c) Answers will vary.

8a) Bias

8b) Validity

8c) Reliability

9) Tree Diagram

10a) 0.125

10b) 0.375

10c) 0.875

**Problem Set 5.6**

1a) % Fat = {12, 22, 36, 45, 66}, % Sat. Fat = {15, 29, 49, 63, 96}.

1b) Box Plots

1c) Answers will vary.

2a) Boys: SD = 3.8, Range = 12, IQR = 7; Girls: SD = 5.2, Range = 11, IQR = 5

2b) Answers will vary.

2c) Boys: $\bar{x} = 69.1, \text{ Med } = 69, \text{ Mode } = 67 & 73;$ Girls: $\bar{x} = 63.2, \text{ Med } = 63.5, \text{ Mode } = 61, 64, & 66$

2d) Answers will vary

2e) Answers will vary.

3a) Back to Back Stem Plot

3b) Class 3 = [14, 24, 35, 38.5, 51]; Class 4 = [20, 27, 31, 38, 46]

3c) Class 3: $\bar{x} = 32.0, \text{ SD } = 9.8, \text{ Mode } = 35 & 37, \text{ Range } = 37, \text{ IQR } = 14.5;$ Class 4: $\bar{x} = 32.9, \text{ SD } = 6.5, \text{ Many Modes, Range } = 26, \text{ IQR } = 11$

3d) Answers will vary.

4a) BR: $\bar{x} = 43.9, \text{ SD } = 11.3, \text{ IQR } = 19, \{22, 35, 46, 54, 60\}; \text{ MM: } \bar{x} = 38.7, \text{ SD } = 18.1, \text{ IQR } = 23, \{9, 29, 39, 52, 70\}; \text{ BB: } \bar{x} = 36.1, \text{ SD } = 13.5, \text{ IQR } = 19, \{16, 25, 34, 44, 73\}; \text{ RM: } \bar{x} = 26.1, \text{ SD } = 15.6, \text{ IQR } = 19, \{8, 14, 24.5, 33, 61\}$

4b) Box Plots

4c) One outlier only for BB at 73 home runs.

4d) Answers will vary.

5a) Answers will vary.

**Chapter 5 Review**

1) C

2a) $\bar{x} = 16.592, \text{ SD } = 0.165, \text{ Mode } = 16.6, \text{ Range } = 0.8$

2b) {16.1, 16.5, 16.6, 16.7, 16.9}; Box Plot

2c) Dot Plot

2d) 11.5%

3a) Pie Chart

3b) Answers will vary

3c) Answers will vary.

4) B

5) D

6) A

7) C

8) A

9) B

10) A

11) D

12) E

13) B

14) C

15) A

16) E

17) B

18a) Stem Plot

18b) Tornadoes: {48, 58, 62, 71, 90}, $\bar{x} = 63.133, \text{ SD } = 10.315$; Bengals: {38, 66, 73, 84, 95}, $\bar{x} = 73, \text{ SD } = 14.147$

18c) Parallel Box Plots

18d) Answers will vary.

18e) Answers will vary.

19a) Time Plots

19b) Answers will vary.

19c) Answers will vary.

19d) 1997 = 0.2; 2005 = 0.75; 2018 = 1.15

19e) 2000 = 0.7; 2015 = 0.25

20a) Histogram

20b) {79.6, 81.9, 87.7, 89.4, 91.3}

20c) No Outliers

20d) Box Plot

20e) Range = 11.7, IQR = 7.5, Mode = 73 & 85

20f) $\bar{x} = 86.27, \text{ SD } = 14.147$

20g) Med is 1.4 larger than $\bar{x}$

20h) Answers will vary.

20i) Answers will vary.

20j) 1st

21a) Split Stem Plot

21b) {8, 18, 20, 22, 50}

21c) 8, 9, 10, 10, 29, 30, 35, 40, 40, 50

21d) Box Plot

21e) Range = 42, IQR = 4, Mode = 20

21f) $\bar{x} = 20.90, \text{ SD } = 7.65$

21g) $\bar{x}$ is 0.9 larger than Med

21h) 5-number summary

21i) Answers will vary.

**Problem Set 6.1**

1a) Yes, Exp. = Semesters, Res. = Credits

1b) No

1c) Yes, Exp. = Years, Res. = Salary

1d) Yes, Exp. = Months, Res. = Apps Downloaded

2) +, Moderately Strong, Linear, No Outliers

3a) Exp. = Minutes, Res. = Depth

3b) Scatterplot
Problem Set 6.2
1) Strength, Direction
2) +0.8972
3a) Scatterplot, Exp. = Bed Time, Res. = Wake Time
3b) C
3c) B
4a) No, Common Response
4b) No, Coincidence
4c) No, Common Response
4d) No, Confounding
5) Answers will vary
6) No, Answers will vary
7a) Exp. = Temp, Res. = Visitors
7b) Answers will vary, 0.9
7c) +, Strong, Linear, 1 Outlier
8) 1=E, 2=C, 3=B, 4=A, 5=D
9) Scatterplot, Answers will vary
10) 0.3656
11) 5 or 6
12) 0.03797
13) 0.00077

Problem Set 6.3
1a) Scatterplot
1b) $\hat{y} = 2.3818 + 1.9795x$
1c) +0.9950
1d) Slope = 1.9795
1e) 36 cm, 121 cm
2a) Scatterplot
2b) +, Strong, Curved, No Outliers
2c) $\hat{y} = -181.5867 + 0.0933x$
2d) +0.9624
2e) $\$6.51$, Not Accurate
2f) $\$2.03$, Too High by $\$0.43$
3a) Exp. = Father’s IQ, Res. = Son’s IQ
3b) Slope = 0.9
3c) Y-Intercept = 12
3d) Slope is Reasonable, Y-Intercept is Not
3e) 120, 138
3f) Answers will vary
4a) Scatterplot, $\hat{y} = 7.757 - 0.478x$
4b) Slope = -0.478
4c) Y-Intercept = 7.757
4d) -0.8564
4e) 2.021 hours
4f) -2.6 months
5a) Exp. = Absences, Res. = Grade
5b) -, Strong, Linear, No Outliers
5c) 50.354
5d) 20.582
5e) -0.9345
5f) Answers will vary
6a) $\hat{y} = 2.419 + 1.884x$, $r = +0.9858$
6b) $\hat{y} = 10.2 - 0.2x$, $r = -0.1129$
6c) Answers will vary
7a) – 7f) Answers will vary
8a) $\hat{y} = -2714.859 + 35.078x$
8b) 35.078
8c) 0.7823
8d) 477, -1136, Answers vary
9) 0.4930
10) 0.3932
11) 0.6105
12) C
13a) mean = $\$10.76$, Standard Deviation = $\$2.69$
13b) $\$9.25$, $\$9.45$, $\$9.80$, $\$10.60$, $\$18.90$
13c) Box Plot
13d) Median, 5# Summary
13e) Skewed Right, 1 Outlier, Median = $\$7.80$, IQR = $\$1.15$

Chapter 6 Review
1) False
2) True
3) True
4) False
5) r
6) Scatterplot
7) Explanatory
8) -1 to 1
9) The Slope
10) Least Squares Regression Line
11) -1 or 1
12) 0.7396
13) +0.8775 or -0.8775
14) Extrapolation
15) Interpolation
16) Strength, Context, Outliers, Form, and Direction
17) Common Response
18a) No. Coincidence
18b) No. Answers will vary
18c) No. Answers will vary
19a) Scatterplot, Exp. = Beers, Res. = BAC
19b) $\hat{y} = -0.0215 + 0.0259x$, $r = +0.8209$
19c) Slope = 0.0259
19d) Y-Intercept = -0.0215
19e) 0.1339
19f) 0.367
19g) Answers will vary
19h) 5 to 6
20a) Scatterplot, Exp. = Skid Length, Res. = Speed
20b) $\hat{y} = 18.825 + 0.2341x$
20c) +, Strong, Curved, No Outliers
20d) +0.9805, Answers vary
20e) 55.6 mph
20f) 27.3 mph
20g) Overestimate

Problem Set 7.1
1a) Sketch
1b) 1
1c) Skewed Right
2a) Sketch
2b) 50%
2c) 68%
2d) 34%
2e) 16%
2f) 2.5%
3) Sketch
4a) 11.85, 12.15
4b) 11.70, 12.30
4c) 11.55, 12.45
5) $\mu = 15, \sigma \approx 3$
6a) Sketch
6b) 2.5%
6c) 9,680 & 13,320
6d) 81.5%
7) c)
8a) Sketch
8b) 50%
8c) 16%, 2.5%
8d) 4,200
9a) 171 seconds
9b) 132 & 158 seconds
10a) 336
10b) 10
11a) Sketch
11b) 13 minutes
11c) 64
12) 70
13) $\bar{x} \approx 5.45, S_x \approx 1.51$
14) 7,920
15) 3, 8, 12.5, 15, 20
16) Voluntary Response

Problem Set 7.2
1) Sketch
2) 1
3) -1
4) 1.15
5) 60th
6) 8th
7) 50th
8) 31%
9) 47%
10) 65th
11) 7th
12) 36%
13) 19%
14) 35%
15) 98th
16) 117
17) -0.36
18) 96%
19) 16%
20) 79%
21) Ricky

22) The male
23) Yes
24) 1
25) 0.005
26) Stem Plot
27) No, Common Response
28) 1.31

Problem Set 7.3
1a) 0.99
1b) -0.99
1c) 1.65
1d) -0.39
1e) -0.67 & 0.67
2a) Sketch
2b) 38 & 132
2c) 143
2d) 34
3a) 628
3b) 20
3c) 14 & 22
3d) The SAT student
4) 10 seconds or less
5) 5ft, 10.5 in.
6) Answers will vary.
7) 20.23 oz.
8) 8%
9a) Scatterplot
9b) 0.96
9c) $\hat{y} = 2.34x - 41.59$
9d) 263 grams
9e) 144 pages
10a) 0.27
10b) 0.8
10c) 0.47

Chapter 7 Review
1a) Sketch
1b) 95%
1c) 83.85%
1d) 77.5%
1e) 31%
1f) 83.5%
1g) 12.6
1h) 1.5
1i) 7
1j) 2
2) d
3a) 10,393 & 11,607
3b) 99th
3c) 11,472
4a) 1.67