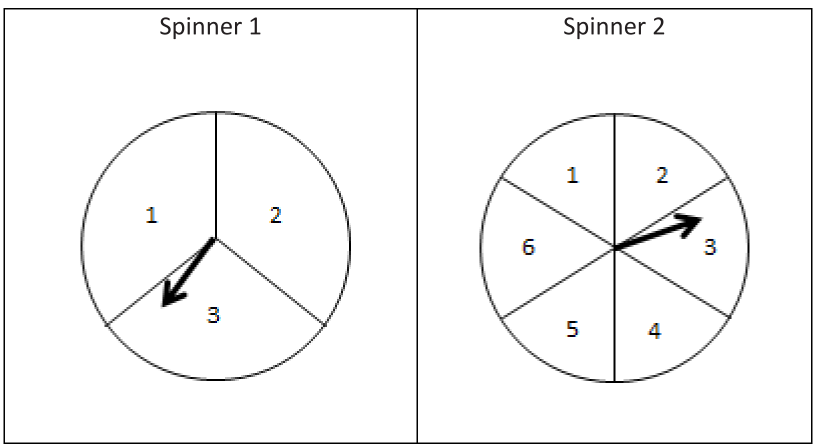
Lesson 1: Chance Experiments, Sample Spaces, and Events

Classwork

Alan is designing a probability game. He plans to present the game to people who will consider financing his idea. Here is a description of the game:

- The game includes the following materials:
  - A fair coin with a "heads" and a "tails".
  - Spinner 1 with three equal area sectors identified as 1, 2, and 3.
  - Spinner 2 with six equal area sectors identified as 1, 2, 3, 4, 5, and 6.
  - A card bag contains six cards. Four cards are blue with the letter "A" written on one card, "B" on another card, "C" on a third card, and "D" on the fourth card. Two cards are red with the letter "E" written on one card and the letter "F" written on the other. (Although actually using colored paper is preferable, slips of paper with the words "blue" or "red" written will also work.)
  - A set of scenario cards, each describing a chance experiment and a set of five possible events based on the chance experiment.



Card bag:

Blue A	Blue B	Blue C	Blue D	Red E	Red F

- The game is played by two players (or two small groups of players) identified as Player 1 and Player 2.
- Rules of the game:
  - The scenario cards are shuffled and one is selected.
  - Each player reads the description of the chance experiment and the description of the five possible outcomes.
  - Players independently assign the numbers 1–5 (no repeats) to the five events described on the Scenario Card based on how likely they think the event is to occur, with 5 being most likely and 1 being least likely.
  - Once players have made their assignments, the chance experiment described on the scenario card is performed. Points are then awarded based on the outcome of the chance experiment. If the event described on the scenario card has occurred, the player earns the number of points corresponding to the number that player assigned to that event (1–5 points). If an event occurs that is not described on the scenario card, then no points are awarded for that event.
  - If an outcome is described by two or more events on the scenario card, the player selects the higher point value.
  - The chance experiment is repeated four more times with points being awarded each time the chance experiment is performed.
  - The player with the largest number of points at the end of the game is the winner.

Alan developed two Scenario Cards for his demonstration to the finance people. A table in which the players can make their assignments and keep track of their scores accompanies each scenario card. Consider the first scenario card Alan developed:

Scenario Card 1

Game Tools: **Spinner 1** (three equal sectors with the number 1 in one sector, the number 2 in the second sector, and the number 3 in the third sector)
Card Bag (Blue-A, Blue-B, Blue-C, Blue-D, Red-E, Red-F)

Directions (chance experiment): Spin Spinner 1 and randomly select a card from the card bag (four blue cards and two red cards). Record the number from your spin and the color of the card selected.

Five events of interest:

Outcome is an odd number on Spinner 1 and a red card from the card bag.	Outcome is an odd number on Spinner 1.	Outcome is an odd number on Spinner 1 and a blue card from the card bag.	Outcome is an even number from Spinner 1 or a red card from the card bag.	Outcome is not a blue card from the card bag.

Player:

Scoring Card for Scenario 1:

Turn	Outcome from Spinner 1	Outcome from the card bag	Points
1			
2			
3			
4			
5			

Here is an example of Alan demonstrating the first scenario card: The chance experiment for Scenario Card 1 is: "Spin Spinner 1 and record the number. Randomly select a card from the card bag (four blue cards and two red cards). Record the color of the card selected."

Alan assigned the numbers 1–5 to the following descriptions as shown below. Once a number is assigned, it cannot be used again.

Five events of interest:

Outcome is an odd number on Spinner 1 and a red card from the card bag.	Outcome is an odd number on Spinner 1.	Outcome is an odd number on Spinner 1 and a blue card from the card bag.	Outcome is an even number from Spinner 1 or a red card from the card bag.	Outcome is not a blue card from the card bag.
3	**1**	**4**	**2**	**5**

Alan is now ready to take his five turns. The results were recorded from the spinner and the card bag. Based on the results, Alan earned the points indicated for each turn.

Player: Player 1

Scoring Card for Scenario 1:

Turn	Outcome from Spinner 1	Outcome from the card bag	Points based on Alan's assignment of the numbers to the five events
1	**2**	Blue	**2**
2	**1**	Red	**5**
3	**1**	Red	**5**
4	**3**	Blue	**4**
5	**2**	Blue	**2**

Alan earned a total of 18 points. The game now turns to Player 2. Player 2 assigns the numbers 1–5 to the same description of outcomes. Player 2 does not have to agree with the numbers Alan assigned. After five turns, the player with the most number of points is the winner.

Exploratory Challenge/Exercises 1–13

1. Would you change any of the assignments of 1–5 that Alan made? Explain your answer. Assign the numbers 1–5 to the event descriptions based on what you think is the best strategy to win the game.

Outcome is an odd number on Spinner 1 and a Red card from the card bag.	Outcome is an odd number on Spinner 1.	Outcome is an odd number on Spinner 1 and a Blue card from the card bag.	Outcome is an even number from Spinner 1 or a Red card from the card bag.	Outcome is not a Blue card from the card bag.

2. Carry out a turn by observing an outcome from spinning Spinner 1 and picking a card. How many points did you earn from this first turn?

3. Complete four more turns (for a total of five) and determine your final score.

   Player: Your Turn

   Scoring Card for Scenario 1:

Trial	Outcome from Spinner 1	Outcome from the card bag	Points based on your assignment of numbers to the events
1			
2			
3			
4			
5			

4. If you changed the numbers assigned to the descriptions, was your score better than Alan's score? Did you expect your score to be better? Explain. If you did not change the numbers from those that Alan assigned, explain why you did not change them.

5. Spinning Spinner 1 and drawing a card from the card bag is a **chance experiment**. One possible outcome of this experiment is (1, Blue-A). Recall that the **sample space** for a chance experiment is the set of all possible outcomes. What is the sample space for the chance experiment of Scenario Card 1?

6. Are the outcomes in the sample space equally likely? Explain your answer.

7. Recall that an **event** is a collection of outcomes from the sample space. One event of interest for someone with Scenario Card 1 is "Odd number on Spinner 1 and a Red card." What are the outcomes that make up this event? List the outcomes of this event in the first row of the Table 1 (see Exercise 9).

8. What is the probability of getting an odd number on Spinner 1 and picking a Red card from the card bag? Also enter this probability in Table 1 (see Exercise 9).

9. Complete the Table 1 by listing the outcomes for the other events and their probabilities based on the chance experiment for this scenario card.

Table 1

Event	Outcomes	Probability
Odd number on Spinner 1 and a red card from the card bag		
Odd number on Spinner 1		
Odd number on Spinner 1 and a blue card from the card bag		
Even number on Spinner 1 or a Red card from the card bag		
Not picking a blue card from the card bag		

10. Based on the above probabilities, how would you assign the numbers 1 to 5 to each of the game descriptions? Explain.

11. If you changed any of the points assigned to the game descriptions, play the game again at least three times and record your final scores for each game. Do you think you have the best possible assignment of numbers to the events for this scenario card? If you did not change the game descriptions, also play the game so that you have at least three final scores. Compare your scores with other members of your class. Do you think you have the best assignment of numbers to the events for this scenario card?

Turn	Outcome from Spinner 1	Outcome from the card bag	Points based on the assignment of points in Exercise 10
1			
2			
3			
4			
5			

12. Why might you not be able to answer the question of whether or not you have the best assignment of numbers to the game descriptions with at least three final scores?

13. Write your answers to the following questions independently, and then share your responses with a neighbor.
    a. How did you make decisions about what to bet on?

    b. How do the ideas of probability help you make decisions?

Lesson Summary

- The sample space of a chance experiment is the collection of all possible outcomes for the experiment.
- An event is a collection of outcomes of a chance experiment.
- For a chance experiment in which outcomes of the sample space are equally likely, the probability of an event is the number of outcomes in the event divided by the number of outcomes in the sample space.
- Some events are described in terms of "or," "and," or "not."

Problem Set

Consider a second scenario card that Alan created for his game:

Scenario Card 2

Tools: **Spinner 1**

Spinner 2: A spinner with six equal sectors. (Place the number 1 in a sector, the number 2 in a second sector, the number 3 in a third sector, the number 4 in a fourth sector, the number 5 in a fifth sector, and the number 6 in a the last sector.)

Directions (chance experiment): Spin Spinner 1, and spin Spinner 2. Record the number from Spinner 1, and record the number from Spinner 2.

Five events of interest:

Outcome is an odd number on Spinner 2.	Outcome is an odd number on Spinner 1 and an even number on Spinner 2.	Outcome is the sum of 7 from the numbers received from Spinner 1 and Spinner 2.	Outcome is an even number on Spinner 2.	Outcome is the sum of 2 from the numbers received from Spinner 1 and Spinner 2.

Player:

Scoring Card for Scenario 2:

Turn	Outcome from Spinner 1	Outcome from Spinner 2	Points
1			
2			
3			
4			
5			

1. Prepare Spinner 1 and Spinner 2 for the chance experiment described on this second scenario card. (Recall that Spinner 2 has six equal sectors.)

2. What is the sample space for the chance experiment described on this scenario card?

3. Based on the sample space, determine the outcomes and the probabilities for each of the events on this scenario card. Complete the table below.

Event	Outcomes	Probability
Outcome is an odd number on Spinner 2.		
Outcome is an odd number on Spinner 1 and an even number on Spinner 2.		
Outcome is the sum of 7 from the numbers received from Spinner 1 and Spinner 2.		
Outcome is an even number on Spinner 2.		
Outcome is the sum of 2 from the numbers received from Spinner 1 and Spinner 2.		

4. Assign the numbers 1–5 to the events described on the scenario card.

Five events of interest: Scenario Card 2

Outcome is an odd number on Spinner 2.	Outcome is an odd number on Spinner 1 and an even number on Spinner 2.	Outcome is the sum of 7 from the numbers received from Spinner 1 and Spinner 2.	Outcome is an even number on Spinner 2.	Outcome is the sum of 2 from the numbers received from Spinner 1 and Spinner 2.

5. Determine at least three final scores based on the numbers you assigned to the events.

Player: Scott

Trial	Outcome from Spinner 1	Outcome from Spinner 2	Points (see Problem 4)
1			
2			
3			
4			
5			

Player: Scott

Trial	Outcome from Spinner 1	Outcome from Spinner 2	Points (see Problem 4)
1			
2			
3			
4			
5			

Player: Scott

Trial	Outcome from Spinner 1	Outcome from Spinner 2	Points (see Problem 4)
1			
2			
3			
4			
5			

6. Alan also included a fair coin as one of the scenario tools. Develop a scenario card (Scenario Card 3) that uses the coin and one of the spinners. Include a description of the chance experiment and descriptions of five events relevant to the chance experiment.

Scenario Card 3

Tools: **Fair coin (Head or Tail)**
Spinner

Directions (chance experiment):

Five events of interest:

7. Determine the sample space for your chance experiment. Then, complete the table below for the five events on your scenario card. Assign the numbers 1–5 to the descriptions you created.

Event	Outcomes	Probability

8. Determine a final score for your game based on five turns.

Turn			Points
1			
2			
3			
4			
5			

Lesson 2: Calculating Probabilities of Events Using Two-Way Tables

Classwork

Example 1: Building a New High School

The School Board of Waldo, a rural town in the Midwest, is considering building a new high school primarily funded by local taxes. They decided to interview eligible voters to determine if the school board should build a new high school facility to replace the current high school building. There is only one high school in the town. Every registered voter in Waldo was interviewed. In addition to asking about support for a new high school, data on gender and age group were also recorded. The data from these interviews are summarized below.

	Should our town build a new high school?					
	Yes		No		No answer	
Age (in years)	Male	Female	Male	Female	Male	Female
18 – 25	29	32	8	6	0	0
26 – 40	53	60	40	44	2	4
41 – 65	30	36	44	35	2	2
66 and older	7	26	24	29	2	0

Exercises 1–8

1. Based on this survey, do you think the school board should recommend building a new high school? Explain your answer.

2. An eligible voter is picked at random. If this person is 21 years old, do you think he or she would indicate that the town should build a high school? Why or why not?

3. An eligible voter is picked at random. If this person is 55 years old, do you think he or she would indicate that the town should build a high school? Why or why not?

4. The School Board wondered if the probability of recommending a new high school was different for different age categories. Why do you think the survey classified voters using the age categories 18–25 years old, 26–40 years old, 41–65 years old, and 66 years old and older?

5. It might be helpful to organize the data in a two-way frequency table. Use the given data to complete the following two-way frequency table. Note that the age categories are represented as rows, and the possible responses are represented as columns.

	Yes	No	No answer	Total
$18-25$ years old				
$26-40$ years old				
$41-65$ years old				
66 years old and older				
Total				

6. A local news service plans to write an article summarizing the survey results. Three possible headlines for this article are provided below. Is each headline accurate or inaccurate? Support your answer using probabilities calculated using the table above.

Headline 1: Waldo Voters Likely to Support Building a New High School

Headline 2: Older Voters Less Likely to Support Building a New High School

Headline 3: Younger Voters Not Interested in Building a New High School

7. The School Board decided to put the decision on whether or not to build the high school up for a referendum in the next election. At the last referendum regarding this issue, only 25 of the eligible voters ages 18–25 voted, 110 of the eligible voters ages 26–40 voted, 130 of the eligible voters ages 41–65 voted, and 80 of the eligible voters ages 66 and older voted. If the voters in the next election turnout in similar numbers, do you think this referendum will pass? Justify your answer.

8. Is it possible that your prediction of the election outcome might be incorrect? Explain.

Example 2: Smoking and Asthma

Health officials in Milwaukee, Wisconsin were concerned about teenagers with asthma. People with asthma often have difficulty with normal breathing. In a local research study, researchers collected data on the incidence of asthma among students enrolled in a Milwaukee Public High School.

Students in the high school completed a survey that was used to begin this research. Based on this survey, the probability of a randomly selected student at this high school having asthma was found to be 0.193. Students were also asked if they had at least one family member living in their house who smoked. The probability of a randomly selected student having at least one member in their household who smoked was reported to be 0.421.

Exercises 9–14

It would be easy to calculate probabilities if the data for the students had been organized into a two-way table like the one used in Exercise 5. But there is no table here, only probability information. One way around this is to think about what the table might have been if there had been $1{,}000$ students at the school when the survey was given. This table is called a *hypothetical 1000* two-way table.

What if the population of students at this high school was $1{,}000$? The population was probably not exactly $1{,}000$ students, but using an estimate of $1{,}000$ students provides an easier way to understand the given probabilities. Connecting these estimates to the actual population is completed in a later exercise. Place the value of $1{,}000$ in the cell representing the total population. Based on a hypothetical 1000 population, consider the following table.

	No household member smokes	At least one household member smokes	Total
Student indicates he or she has asthma	Cell 1	Cell 2	Cell 3
Student indicates he or she does not have asthma	Cell 4	Cell 5	Cell 6
Total	Cell 7	Cell 8	$\mathbf{1{,}000}$

9. The probability that a randomly selected student at this high school has asthma is 0.193. This probability can be used to calculate the value of one of the cells in the table above. Which cell is connected to this probability? Use this probability to calculate the value of that cell.

10. The probability that a randomly selected student has at least 1 household member who smokes is 0.421. Which cell is connected to this probability? Use this probability to calculate the value of that cell.

11. In addition to the previously given probabilities, the probability that a randomly selected student has at least one household member who smokes and has asthma is 0.120. Which cell is connected to this probability? Use this probability to calculate the value of that cell.

12. Complete the two-way frequency table above by calculating the values of the other cells in the table.

13. Based on your completed two-way table, estimate the following probabilities as a fraction and also as a decimal (rounded to three decimal places):
    a. A randomly selected student has asthma. What is the probability this student has at least 1 household member who smokes?

    b. A randomly selected student does not have asthma. What is the probability he or she has at least one household member who smokes?

    c. A randomly selected student has at least one household member who smokes. What is the probability this student has asthma?

14. Do you think that whether or not a student has asthma is related to whether or not this student has at least one family member who smokes? Explain your answer.

Lesson Summary

Data organized in a two-way frequency table can be used to calculate probabilities.

In certain problems, probabilities that are known can be used to create a hypothetical 1000 two-way table. The hypothetical population of 1,000 can then be used to calculate probabilities.

Probabilities are always interpreted in context.

Problem Set

1. The Waldo School Board asked eligible votes to evaluate the town's library service. Data are summarized in the following table.

	How would you rate our town's library services?							
	Good		Average		Poor		Do not use library	
Age (in years)	Male	Female	Male	Female	Male	Female	Male	Female
18–25	10	8	5	7	5	5	17	18
26–40	30	28	25	30	20	30	20	20
41 –65	30	32	26	21	15	10	5	10
66 and older	21	25	8	15	2	10	2	5

a. What is the probability that a randomly selected person who completed the survey rated the library as "good?"

b. Imagine talking to a randomly selected male who had completed the survey. How do you think this person rated the library services? Explain your answer.

c. Use the given data to construct a two-way table that summarizes the responses on gender and rating of the library services. Use the following template as your guide:

	Good	Average	Poor	Do Not Use	Total
Male					
Female					
Total					

d. Based on your table, answer the following.

i. A randomly selected person who completed the survey is male. What is the probability he rates the library services as "good?"

ii. A randomly selected person who completed the survey is female. What is the probability she rates the library services as "good?"

e. Also based on your table, answer the following.

i. A randomly selected person who completed the survey rated the library services as "good." What is the probability this person is a male?

ii. A randomly selected person who completed the survey rated the library services as "good." What is the probability this person is a female?

f. Do you think there is a difference in how males and females rated library services? Explain your answer.

2. *Obedience School for Dogs* is a small franchise that offers obedience classes for dogs. Some people think that larger dogs are easier to train and, therefore, should not be charged as much for the classes. To investigate this claim, dogs enrolled in the classes were classified as large (30 pounds or more) or small (under 30 pounds). The dogs were also classified by whether or not they passed the obedience class offered by the franchise. 45% of the dogs involved in the classes were large. 60% of the dogs passed the class. Records indicate that 40% of the dogs in the classes were small and passed the course.

   a. Complete the following hypothetical 1000 two-way table.

	Passed the course	Did not pass the course	Total
Large Dogs			
Small Dogs			
Total			

   b. Estimate the probability that a dog selected at random from those enrolled in the classes passed the course.

   c. A dog was randomly selected from the dogs that completed the class. If the selected dog was a large dog, what is the probability this dog passed the course?

   d. A dog was randomly selected from the dogs that completed the class. If the selected dog is a small dog, what is the probability this dog passed the course?

   e. Do you think dog size and whether or not a dog passes the course are related?

   f. Do you think large dogs should get a discount? Explain your answer.

Lesson 3: Calculating Conditional Probabilities and Evaluating Independence Using Two-Way Tables

Classwork

Example 1

Students at Rufus King High School were discussing some of the challenges of finding space for athletic teams to practice after school. Part of the problem, according to Kristin, is that the females are more likely to be involved in after-school athletic programs than males. However, the athletic director assigns the available facilities as if males are more likely to be involved. Before suggesting changes to the assignments, the students decided to investigate.

Suppose the following information is known about Rufus King High School: 40% of students are involved in one or more of the after-school athletic programs offered at the school. It is also known that 58% of the school's students are female. The students decide to construct a **hypothetical 1000 two-way table**, like Table 1, to organize the data.

Table 1

Participation in after-school athletic program (Yes or No) by gender

	Yes - Participate in After-School Athletic Program	No - Do Not Participate in After-School Athletic Program	Total
Females	Cell 1	Cell 2	Cell 3
Males	Cell 4	Cell 5	Cell 6
Total	Cell 7	Cell 8	Cell 9

Exercises 1–6

1. What cell in Table 1 represents a hypothetical group of $1,000$ students at Rufus King High School?

2. What cells in Table 1 can be filled based on the information given about the student population? Place these values in the appropriate cells of the table based on this information.

3. Based only on the cells you completed in Exercise 2, which of the following probabilities can be calculated, and which cannot be calculated? Calculate the probability if it can be calculated. If it cannot be calculated, indicate why.
   a. The probability that a randomly selected student is female.
   b. The probability that a randomly selected student participates in after school athletics programs.
   c. The probability that a randomly selected student who does not participate in the after school athletics program is male.
   d. The probability that a randomly selected male student participates in the after school athletics program.

4. The athletic director indicated that 23.2% of the students at Rufus King are females and participate in after school athletics programs. Based on this information, complete Table 1.

5. Consider the cells 1, 2, 4, and 5 of Table 1. Identify which of these cells represent students who are female or who participate in after-school athletic programs.

6. What cells of the two-way table represent students who are males who do not participate in after-school athletic programs?

Example 2

The completed hypothetical 1000 table organizes information in a way that makes it possible to answer various questions. For example, you can investigate whether females at the school are more likely to be involved in the after-school athletic programs.

Consider the following events:

- Let "A" represent the event "a randomly selected student is female."
- Let "not A" represent the "complement of A." The complement of A represents the event "a randomly selected student is not female," which is equivalent to the event "a randomly selected student is male."
- Let "B" represent the event "a randomly selected student participates in the after-school athletic program."
- Let "not B" represent the "complement of B." The complement of B represents the event "a randomly selected student does not participate in the after-school athletic program."
- Let "A or B" (described as A union B) represent the event "a randomly selected student is female or participates in the after-school athletic program."
- Let "A and B" (described as A intersect B) represent the event "a randomly selected student is female and participates in the after-school athletic program."

Exercises 7–9

7. Based on the descriptions above, describe the following events in words:
   a. Not A or Not B.

   b. A and Not B.

8. Based on the above descriptions and Table 1, determine the probability of each of the following events:
   a. A

   b. B

   c. Not A

d. Not B

e. A or B

f. A and B

9. Determine the following values:
   a. The probability of A plus the probability of Not A.

   b. The probability of B plus the probability of Not B.

   c. What do you notice about the results of parts (a) and (b)? Explain.

Example 3: Conditional Probability

Another type of probability is called a conditional probability. Pulling apart the two-way table helps to define a conditional probability.

	Yes - Participate in After-School Athletic Program	No - Do Not Participate in After-School Athletic Program	Total
Females	Cell 1	Cell 2	Cell 3

Suppose that a randomly selected student is female. What is the probability that the selected student participates in the after-school athletic program? This probability is an example of what is called a conditional probability. This probability is calculated as the number of students who are female students and participate in the after-school athletic program (or the students in cell 1) divided by the total number of female students (or the students in cell 3).

Exercises 10–15

10. The following are also examples of conditional probabilities. Answer each question.
    a. What is the probability that if a randomly selected student is female, she participates in the after-school athletic program?

    b. What is the probability that if a randomly selected student is female, she does not participate in after-school athletics?

11. Describe two conditional probabilities that can be determined from the following row in Table 1.

	Yes - Participate in After-School Athletic Program	No - Do Not Participate in After-School Athletic Program	Total
Males	Cell 4	Cell 5	Cell 6

12. Describe two conditional probabilities that can be determined from the following column in Table 1.

	Yes - Participate in After-School Athletic Program
Females	Cell 1
Males	Cell 4
Total	Cell 7

13. Determine the following conditional probabilities.

    a. A randomly selected student is female. What is the probability she participates in the after-school athletic program? Explain how you determined your answer.

    b. A randomly selected student is male. What is the probability he participates in the after-school athletic program?

    c. A student is selected at random. What is the probability this student participates in the after-school athletic program?

14. Based on the answers to Exercise 13, do you think that female students are more likely to be involved in after-school athletics? Explain your answer.

15. What might explain the concern females expressed in the beginning of this lesson about the problem of assigning space?

Lesson Summary

Data organized in a two-way frequency table can be used to calculate probabilities. The two-way frequency tables can also be used to calculate conditional probabilities.

In certain problems, probabilities that are known can be used to create a hypothetical 1000 two-way table. This hypothetical population of $1{,}000$ can be used to calculate conditional probabilities.

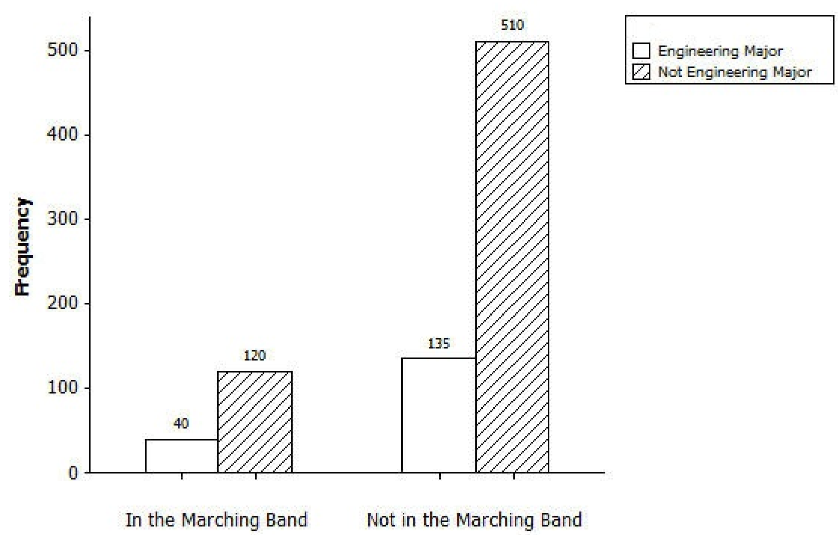
Probabilities are always interpreted by the context of the data.

Problem Set

Oostburg College has a rather large marching band. Engineering majors were heard bragging that students majoring in engineering are more likely to be involved in the marching band than students from other majors.

1. If the above claim is accurate, does that mean that most of the band is engineering students? Explain your answer.

2. The following graph was prepared to investigate the above claim.



Based on the graph, complete the following two-way frequency table:

	In the Marching Band	Not in the Marching Band	Total
Engineering major			
Not an engineering major			
Total			

3. Let M represent the event that a randomly selected student is in the marching band. Let E represent the event that a randomly selected student is an engineering major.

   a. Describe the event represented by the complement of M.

   b. Describe the event represented by the complement of E.

   c. Describe the event A and B (A intersect B).

   d. Describe the event A or B (A union B).

4. Based on the completed two-way frequency table, determine the following and explain how you got your answer.

   a. The probability that a randomly selected student is in the marching band.

   b. The probability that a randomly selected student is an engineering major.

   c. The probability that a randomly selected student is in the marching band and an engineering major.

   d. The probability that a randomly selected student is in the marching band and not an engineering major.

5. Indicate if the following conditional probabilities would be calculated using the rows or the columns of the two-way frequency table.

   a. A randomly selected student is majoring in engineering. What is the probability this student is in the marching band?

   b. A randomly selected student is not in the marching band. What is the probability that this student is majoring in engineering?

6. Based on the two-way frequency table, determine the following conditional probabilities.

   a. A randomly selected student is majoring in engineering. What is the probability that this student is in the marching band?

   b. A randomly selected student is not majoring in engineering. What is the probability that this student is in the marching band?

7. The claim that started this investigation was that students majoring in engineering are more likely to be in the marching band than students from other majors. Describe the conditional probabilities that would be used to determine if this claim is accurate.

8. Based on the two-way frequency table, calculate the conditional probabilities identified in Problem 7.

9. Do you think the claim that students majoring in engineering are more likely to be in the marching band than students for other majors is accurate? Explain your answer.

10. There are 40 students at Oostburg College majoring in computer science. Computer science is not considered an engineering major. Calculate an estimate of the number of computer science majors you think are in the marching band. Explain how you calculated your estimate.

Lesson 4: Calculating Conditional Probabilities and Evaluating Independence Using Two-Way Tables

Classwork

Exercises

In previous lessons, conditional probabilities were used to investigate whether or not there is a connection between two events. This lesson formalizes this idea and introduces the concept of independence.

1. Several questions are posed below. Each question is about a possible connection between two events. For each question, identify the two events and indicate whether or not you think that there would be a connection. Explain your reasoning.

   a. Are high school students whose parents or guardians set a midnight curfew less likely to have a traffic violation than students whose parents or guardians have not set such a curfew?

   b. Are left-handed people more likely than right-handed people to be interested in the arts?

   c. Are students who regularly listen to classical music more likely to be interested in mathematics than students who do not regularly listen to classical music?

   d. Are people who play video games more than 10 hours per week more likely to select football as their favorite sport than people who do not play video games more than 10 hours per week?

Two events are independent when knowing that one event has occurred does not change the likelihood that the second event has occurred. How can conditional probabilities be used to tell if two events are independent or not independent?

Recall the hypothetical 1000 two-way frequency table that was used to classify students at Rufus King High School according to gender and whether or not they participated in the after-school athletic program.

Table 1

Participation in the after-school athletic program (Yes or No) of males and females

	Participate in the after-school athletic program	Do not participate in the after-school athletic program	Total
Females			
Males			
Total			

2. For each of the following, indicate whether the probability described is one that can be calculated using the values in Table 1. Also indicate whether or not it is a conditional probability.
   a. The probability that a randomly selected student participates in the after-school athletic program.
   b. The probability that a randomly selected student who is female participates in the after-school athletic program.
   c. The probability that a randomly selected student who is male participates in the after-school athletic program.

3. Use Table 1 to calculate each of the probabilities described in Exercise 2.
   a. The probability that a randomly selected student participates in the after-school athletic program.
   b. The probability that a randomly selected student who is female participates in the after-school athletic program.

c. The probability that a randomly selected student who is male participates in the after-school athletic program.

4. Would your prediction of whether or not a student participates in the after-school athletic program change if you knew the gender of the student? Explain your answer.

Two events are *independent* if knowing that one event has occurred does not change the probability that the other event has occurred. For example, consider the following two events:

F: the event that a randomly selected student is female

S: the event that a randomly selected student participates in the after-school athletic program.

F and S would be independent if the probability that a randomly selected student participates in the after-school athletic program is equal to the probability that a randomly selected student who is female participates in the after-school athletic program. If this were the case, knowing that a randomly selected student is female does not change the probability that the selected student participates in the after-school athletic program. Then F and S would be independent.

5. Based on the definition of independence, are the events *randomly selected student is female* and *randomly selected student participates in the after-school athletic program* independent? Explain.

6. A randomly selected student participates in the after-school athletic program.

a. What is the probability this student is a female?

b. Using only your answer from part (a), what is the probability that this student is a male? Explain how you arrived at your answer.

Consider data below.

	No household member smokes	At least one household member smokes	Total
Student indicates he or she has asthma	69	113	182
Student indicates he or she does not have asthma	473	282	755
Total	542	395	937

7. You are asked to determine if the two events *a randomly selected student has asthma* and *a randomly selected student has a household member who smokes* are independent. What probabilities could you calculate to answer this question?

8. Calculate the probabilities you described in Exercise 7.

9. Based on the probabilities you calculated in Exercise 8, are these two events independent or not independent? Explain.

10. Is the probability that a randomly selected student who has asthma and who has a household member who smokes the same as or different than the probability that a randomly selected student who does not have asthma but does have a household member who smokes? Explain your answer.

11. A student is selected at random. The selected student indicates that he or she has a household member who smokes. What is the probability that the selected student has asthma?

Lesson Summary

Data organized in a two-way frequency table can be used to calculate conditional probabilities.

Two events are independent if knowing that one event has occurred does not change the probability that the second event has occurred.

Probabilities calculated from two-way frequency tables can be used to determine if two events are independent or not independent.

Problem Set

1. Consider the following questions.
   a. A survey of the students at a Midwest high school asked the following questions:
      "Do you use a computer at least 3 times a week to complete your school work?"
      "Are you taking a mathematics class?"

      Do you think the events *a randomly selected student is taking a mathematics class* and *a randomly selected student uses a computer at least* 3 *times a week* are independent or not independent? Explain your reasoning.

   b. The same survey also asked students the following:
      "Do you participate in any extracurricular activities at your school?"
      "Do you know what you want to do after high school?"

      Do you think the events *a randomly selected student participates in extracurricular activities* and *a randomly selected student knows what he or she wants to do after completing high school* are independent or not independent? Explain your reasoning.

   c. People attending a professional football game in 2013 completed a survey that included the following questions:
      "Do you think football is too violent?"
      "Is this the first time you have attended a professional football game?"

      Do you think the events *a randomly selected person who completed the survey is attending a professional football game for the first time* and *a randomly selected person who completed the survey thinks football is too violent* are independent or not independent? Explain your reasoning.

2. Complete the table below in a way that would indicate the two events *uses a computer* and *is taking a mathematics class* are independent.

	Uses a computer at least 3 times a week for school work	Does not use a computer at least 3 times a week for school work	Total
In a mathematics class			700
Not in a mathematics class			
Total	600		$1,000$

3. Complete the following hypothetical 1000 table. Are the events *participates in extracurricular activities* and *know what I want to do after high school* independent or not independent? Justify your answer.

	Participate in extracurricular activities	Do not participate in extracurricular activities	Total
Know what I want to do after high school			800
Do not know what I want to do after high school	50		
Total	600		$1,000$

4. The following hypothetical 1000 table is from Lesson 2.

	No household member smokes	At least one household member smokes	Total
Student indicates he or she has asthma	73	120	193
Student indicates he or she does not have asthma	506	301	807
Total	579	421	$1,000$

The actual data from the entire population is given in the table below.

	No household member smokes	At least one household member smokes	Total
Student indicates he or she has asthma	69	113	182
Student indicates he or she does not have asthma	473	282	755
Total	542	395	937

a. Based on the hypothetical 1000 table, what is the probability that a randomly selected student who has asthma has at least one household member who smokes?

b. Based on the actual data, what is the probability that a randomly selected student who has asthma has at least one household member who smokes (round your answer to 3 decimal places)?

c. Based on the hypothetical 1000 table, what is the probability that a randomly selected student who has no household member who smokes has asthma?

d. Based on the actual data, what is the probability that a randomly selected student who has no household member who smokes has asthma?

e. What do you notice about the probabilities calculated from the actual data and the probabilities calculated from the hypothetical 1000 table?

5. As part of the asthma research, the investigators wondered if students who have asthma are less likely to have a pet at home than students who do not have asthma. They asked the following two questions:

   "Do you have asthma?"
   "Do you have a pet at home?"

   Based on the responses to these questions, you would like to set up a two-way table that you could use to determine if the following two events are independent or not independent:

   Event 1: a randomly selected student has asthma
   Event 2: a randomly selected student has a pet at home.

   a. What would you use to label the rows of the two-way table?

   b. What would you use to label the columns of the two-way table?

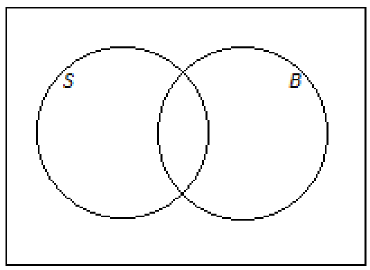
   c. What probabilities would you calculate to determine if Event 1 and Event 2 are independent?

Lesson 5: Events and Venn Diagrams

Classwork

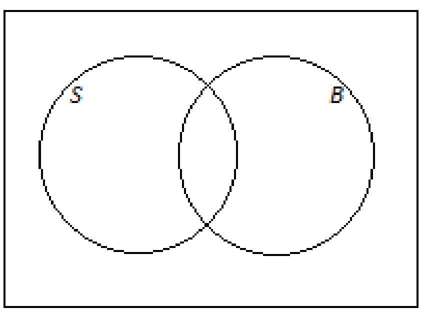
Example 1: Shading Regions of a Venn Diagram

At a high school, some students play soccer and some do not. Also, some students play basketball and some do not. This scenario can be represented by a Venn diagram, as shown below. The circle labeled S represents the students who play soccer, the circle labeled B represents the students who play basketball, and the rectangle represents all the students at the school.

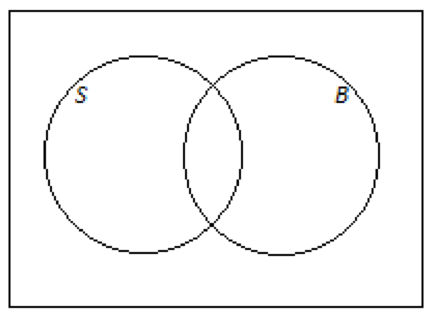


On the Venn diagrams provided, shade the region representing the students who

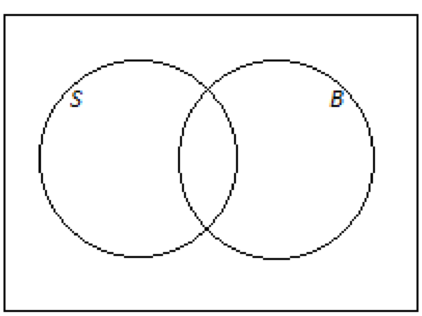
a. play soccer.



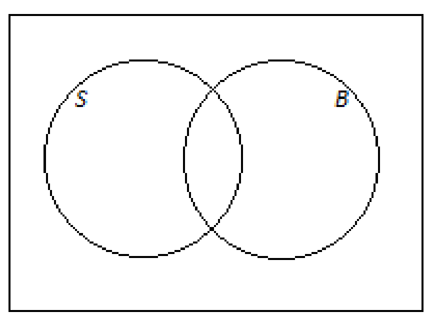
b. do not play soccer.



c. play soccer and basketball.



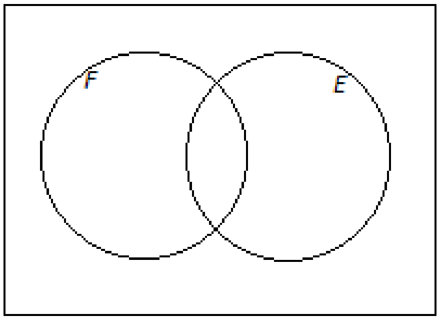
d. play soccer or basketball.



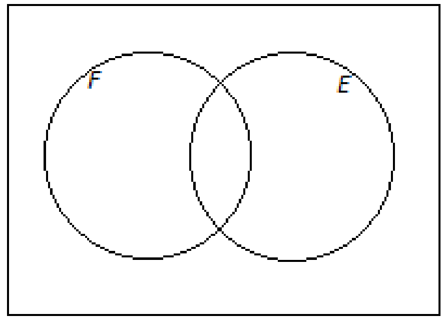
Exercise 1

1. An online bookstore offers a large selection of books. Some of the books are works of fiction, and some are not. Also, some of the books are available as e-books, and some are not. Let F be the set of books that are works of fiction, and let E be the set of books that are available as e-books. On the Venn diagrams provided, shade the regions representing books that are

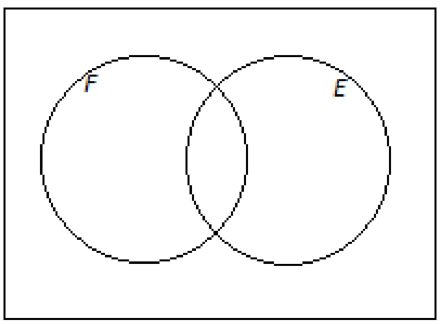
   a. available as e-books.



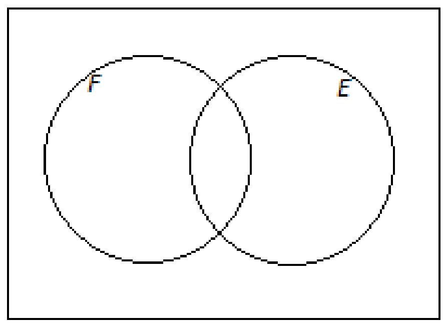
   b. not works of fiction.



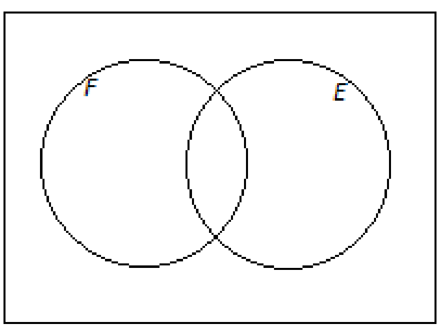
   c. works of fiction and available as e-books.



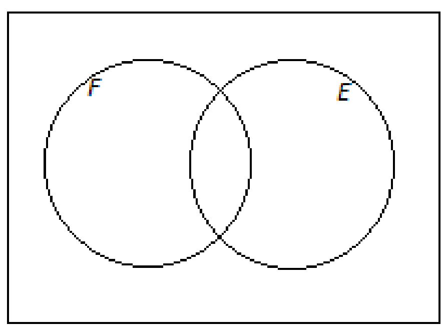
   d. works of fiction or available as e-books.



   e. neither works of fiction nor available as e-books.



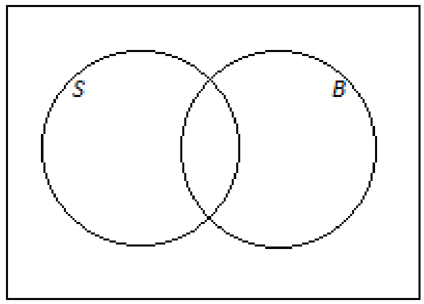
   f. works of fiction that are not available as e-books.



Example 2: Showing Numbers of Possible Outcomes (and Probabilities) in a Venn Diagram

Think again about the school introduced in Example 1. Suppose that 230 students play soccer, 190 students play basketball, and 60 students play both sports. There are a total of 500 students at the school.

a. Complete the Venn diagram below by writing the numbers of students in the various regions of the diagram.

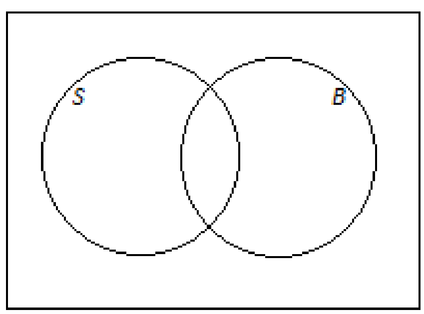


b. How many students play basketball but not soccer?

c. Suppose that a student will be selected at random from the school.

i. What is the probability that the selected student plays both sports?

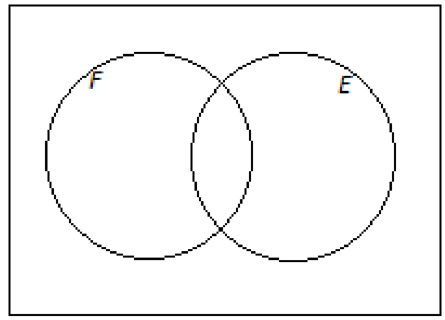
ii. Complete the Venn diagram below by writing the probabilities associated with the various regions of the diagram.



Example 3: Adding and Subtracting Probabilities

Think again about the online bookstore introduced in Exercise 1, and suppose that 62% of the books are works of fiction, 47% are available as e-books, and 14% are available as e-books but are not works of fiction. A book will be selected at random.

a. Using a Venn diagram, find the probability that the book will be



i. a work of fiction and available as an e-book.

ii. neither a work of fiction nor available as an e-book.

b. Return to the information given at the beginning of the question: 62% of the books are works of fiction, 47% are available as e-books, and 14% are available as e-books but are not works of fiction.

i. How would this information be shown in a hypothetical 1000 table? (Show your answers in the table provided below.)

	Fiction	Not Fiction	Total
Available as E-Book			
Not Available as E-Book			
Total			$\mathbf{1,000}$

ii. Complete the hypothetical 1000 table given above.

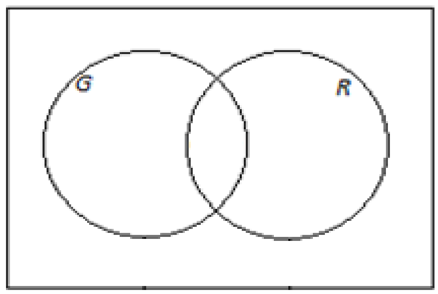
iii. Complete the table below showing the probabilities of the events represented by the cells in the table.

	Fiction	Not Fiction	Total
Available as E-Book			
Not available as E-Book			
Total			

iv. How do the probabilities in your table relate to the probabilities you calculated in part (a)?

Exercises 2–3

2. When a fish is selected at random from a tank, the probability that it has a green tail is 0.64, the probability that it has red fins is 0.25, and the probability that it has both a green tail and red fins is 0.19.
   a. Draw a Venn diagram to represent this information.

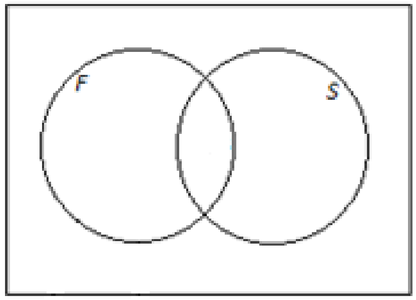


   b. Find the probability that the fish has
      i. red fins but does not have a green tail.
      ii. a green tail but not red fins.
      iii. neither a green tail nor red fins.
   c. Complete the table below showing the probabilities of the events corresponding to the cells of the table.

	Green tail	Not green tail	Total
Red Fins			
Not Red Fins			
Total			

3. In a company, 43% of the employees have access to a fax machine, 38% have access to a fax machine and a scanner, and 24% have access to neither a fax machine nor a scanner. Suppose that an employee will be selected at

random. Using a Venn diagram, calculate the probability that the randomly selected employee will not have access to a scanner. (Note that Venn diagrams and probabilities use decimals or fractions, not percentages.) Explain how you used the Venn diagram to determine your answer.



Lesson Summary

In a probability experiment, the events can be represented by circles in a Venn diagram.

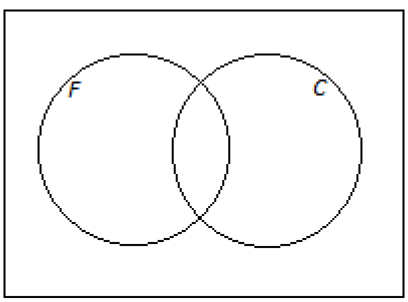
Combinations of events using "and," "or," and "not" can be shown by shading the appropriate regions of the Venn diagram.

The number of possible outcomes can be shown in each region of the Venn diagram; alternatively, probabilities may be shown. The number of outcomes in a given region (or the probability associated with it) can be calculated by adding or subtracting the known numbers of possible outcomes (or probabilities).

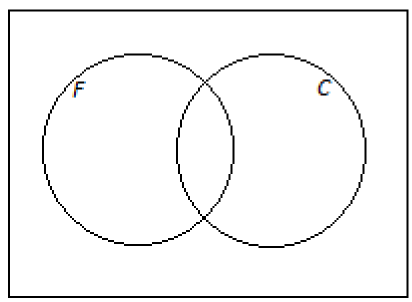
Problem Set

1. On a flight, some of the passengers have frequent flier status and some do not. Also, some of the passengers have checked baggage and some do not. Let the set of passengers who have frequent flier status be F and the set of passengers who have checked baggage be C. On the Venn diagrams provided, shade the regions representing passengers who

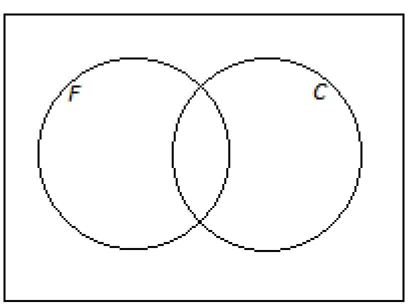
   a. have frequent flier status and have checked baggage.

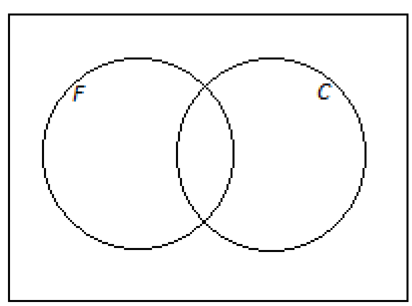
   b. have frequent flier status or have checked baggage.

   c. do not have both frequent flier status and checked baggage.

   d. have frequent flier status or do not have checked baggage.

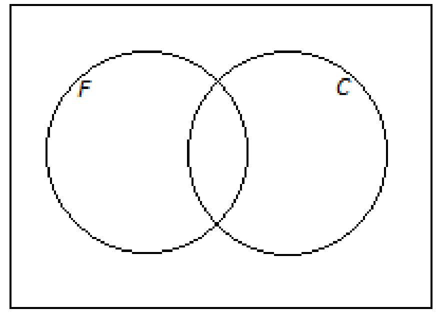
   

2. For the scenario introduced in Problem 1, suppose that, of the 400 people on the flight, 368 have checked baggage, 228 have checked baggage but do not have frequent flier status, and 8 have neither frequent flier status nor checked baggage.

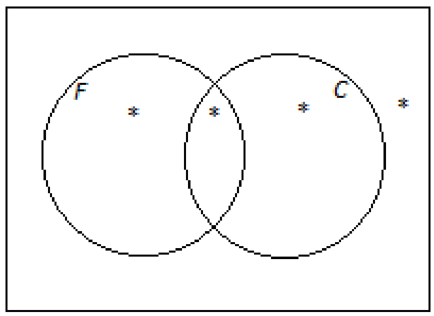
   a. Using a Venn diagram, calculate the number of people on the flight who

      i. have frequent flier status and have checked baggage.

      ii. have frequent flier status.

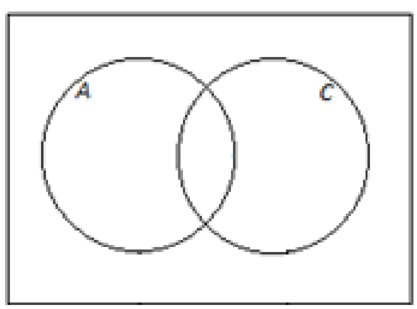


   b. In the Venn diagram provided below, write the probabilities of the events associated with the regions marked with a star (*).



3. When an animal is selected at random from those at a zoo, the probability that it is North American (meaning that its natural habitat is in the North American continent) is 0.65, the probability that it is both North American and a carnivore is 0.16, and the probability that it is neither American nor a carnivore is 0.17.

   a. Using a Venn diagram, calculate the probability that a randomly selected animal is a carnivore.



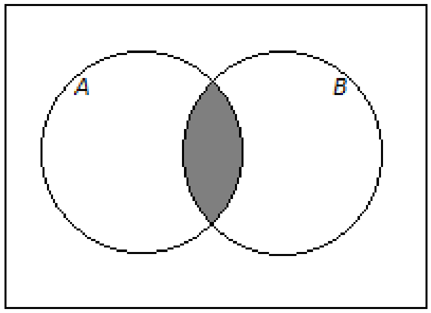
   b. Complete the table below showing the probabilities of the events corresponding to the cells of the table.

	American	Not American	Total
Carnivore			
Not carnivore			
Total			

4. This question introduces the mathematical symbols for "and," "or," and "not."

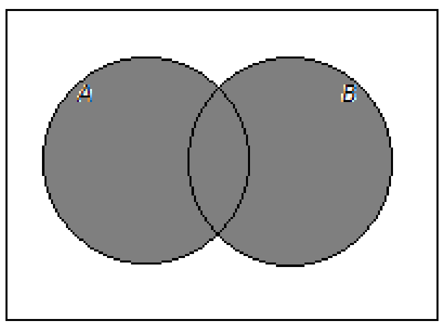
   Considering all the people in the world, let A be the set of Americans (citizens of the United States), and let B be the set of people who have brothers.

   - The set of people who are Americans and have brothers is represented by the shaded region in the Venn diagram below.



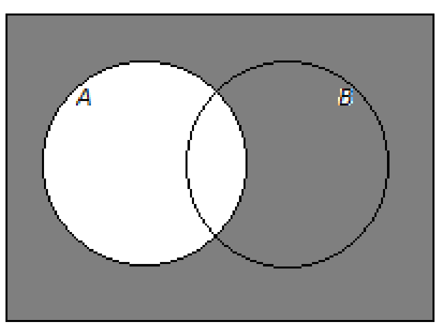
     This set is written $A \cap B$ (read "A intersect B"), and the probability that a randomly selected person is American and has a brother is written $P(A \cap B)$.

   - The set of people who are Americans or have brothers is represented by the shaded region in the Venn diagram below.



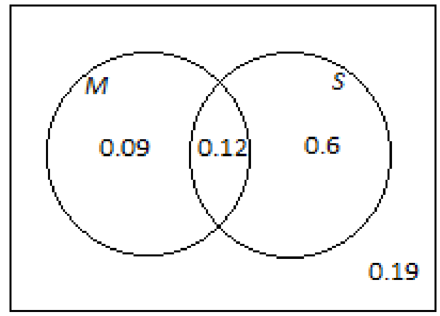
     This set is written $A \cup B$ (read "A union B"), and the probability that a randomly selected person is American or has a brother is written $P(A \cup B)$.

   - The set of people who are not Americans is represented by the shaded region in the Venn diagram below.



     This set is written A^C (read "A complement"), and the probability that a randomly selected person is not American is written $P(A^C)$.

Now think about the cars available at a dealership. Suppose a car is selected at random from the cars at this dealership. Let the event that the car has manual transmission be denoted by M, and let the event that the car is a sedan be denoted by S. The Venn diagram below shows the probabilities associated with four of the regions of the diagram.



a. What is the value of $P(M \cap S)$?

b. Complete this sentence using and/or:

   $\boldsymbol{P(M \cap S)}$ is the probability that a randomly selected car has a manual transmission __________ is a sedan.

c. What is the value of $P(M \cup S)$?

d. Complete this sentence using and/or:

   $P(M \cup S)$ is the probability that a randomly selected car has a manual transmission __________ is a sedan.

e. What is the value of $P(S^C)$?

f. Explain the meaning of $P(S^C)$.

Lesson 6: Probability Rules

Classwork

Example 1: The Complement Rule

In previous lessons you have seen that to calculate the probability that an event *does not happen* you can subtract the probability of the event from 1. If the event is denoted by A, then this rule can be written:

$$\boldsymbol{P}(\text{not } \boldsymbol{A}) = \mathbf{1} - \boldsymbol{P}(\boldsymbol{A}).$$

For example, suppose that the probability that a particular flight is on time is 0.78. What is the probability that the flight is not on time?

Example 2: Formula for Conditional Probability

When a room is randomly selected in a downtown hotel, the probability that the room has a king-sized bed is 0.62, the probability that the room has a view of the town square is 0.43, and the probability that it has a king-sized bed *and* a view of the town square is 0.38. Let A be the event that the room has a king-sized bed, and let B be the event that the room has a view of the town square.

a. What is the meaning of "$P(A \text{ given } B)$" in this context?

b. Use a hypothetical 1000 table to calculate $\mathrm{P}(\mathrm{A} \text{ given } \mathrm{B})$.

	$\boldsymbol{A}$ (room has a king-sized bed)	Not $\boldsymbol{A}$ (room does not have king-sized bed)	Total
$\boldsymbol{B}$ (room has a view of the town square)			
Not $\boldsymbol{B}$ (room does not have view of the town square)			
Total			

c. There is also a formula for calculating a conditional probability. The formula for conditional probability is

$$\boldsymbol{P}(\boldsymbol{A}\text{ given }\boldsymbol{B}) = \frac{\boldsymbol{P}(\boldsymbol{A}\text{ and }\boldsymbol{B})}{\boldsymbol{P}(\boldsymbol{B})}.$$

Use this formula to calculate $P(A\text{ given }B)$, where the events A and B are as defined in this example.

d. How does the probability you calculated using the formula compare to the probability you calculated using the hypothetical 1000 table?

Exercise 1

A credit card company states that 42% of its customers are classified as long-term cardholders, 35% pay their bills in full each month, and 23% are long-term cardholders who also pay their bills in full each month. Let the event that a randomly selected customer is a long-term cardholder be L, and the event that a randomly selected customer pays his or her bill in full each month be F.

a. What are the values of $P(L)$, $P(F)$, and $P(L\text{ and }F)$?

b. Draw a Venn diagram, and label it with the probabilities from part (a).

c. Use the conditional probability formula to calculate $P(L\text{ given }F)$. (Round your answer to the nearest thousandth.)

d. Use the conditional probability formula to calculate $P(F \text{ given } L)$. (Round your answer to the nearest thousandth.)

e. Which is greater, $P(F \text{ given } L)$ or $P(F)$? Explain why this is relevant.

f. Remember that two events A and B are said to be independent if $P(A \text{ given } B) = P(A)$. Are the events F and L independent? Explain.

Example 3: Using the Multiplication Rule for Independent Events

A number cube has faces numbered 1 through 6, and a coin has two sides, *heads* and *tails*.

The number cube will be rolled and the coin will be flipped. Find the probability that the cube shows a 4 and the coin lands *heads*. Because the events are independent, we can use the multiplication rules we just learned.

If you toss the coin five times, what is the probability you will see a *head* on all five tosses?

If you tossed the coin five times and got five *heads*, would you think that this coin is a fair coin? Why or why not?

If you roll the number cube three times, what is the probability that it will show 4 on all three throws?

If you rolled the number cube three times and got a 4 on all three rolls, would you think that this number cube is fair? Why or why not?

Suppose that the credit card company introduced in Exercise 1 states that when a customer is selected at random, the probability that the customer pays his or her bill in full each month is 0.35, the probability that the customer makes regular online purchases is 0.83, and these two events are independent. What is the probability that a randomly selected customer pays his or her bill in full each month *and* makes regular online purchases?

Exercise 2

A spinner has a pointer, and when the pointer is spun, the probability that it stops in the red section of the spinner is 0.25.

a. If the pointer is spun twice, what is the probability that it will stop in the red section on both occasions?

b. If the pointer is spun four times, what is the probability that it will stop in the red section on all four occasions? (Round your answer to the nearest thousandth.)

c. If the pointer is spun five times, what is the probability that it never stops on red? (Round your answer to the nearest thousandth.)

Lesson Summary

For any event A, $P(\text{not } A) = 1 - P(A)$.

For any two events A and B, $P(A \text{ given } B) = \dfrac{P(A \text{ and } B)}{P(A)}$.

Events A and B are independent if and only if $P(A \text{ and } B) = P(A)P(B)$.

Problem Set

1. When an avocado is selected at random from those delivered to a food store, the probability that it is ripe is 0.12, the probability that it is bruised is 0.054, and the probability that it is ripe and bruised is 0.019.
   a. Rounding your answers to the nearest thousandth where necessary, find the probability that an avocado randomly selected from those delivered to the store is
      i. not bruised.
      ii. ripe given that it is bruised.
      iii. bruised given that it is ripe.
   b. Which is larger, the probability that a randomly selected avocado is bruised given that it is ripe or the probability that a randomly selected avocado is bruised? Explain in words what this tells you.
   c. Are the events *ripe* and *bruised* independent? Explain.

2. Return to the probability information given in Problem 1. Complete the hypothetical 1000 table given below, and use it to find the probability that a randomly selected avocado is bruised given that it is not ripe. (Round your answer to the nearest thousandth.)

	Ripe	Not Ripe	Total
Bruised			
Not bruised			
Total			

3. According to the website www.census.gov, based on the US population in 2010, the probability that a randomly selected male is 65 or older is 0.114, and the probability that a randomly selected female is 65 or older is 0.146. (In the questions that follow, round your answers to the nearest thousandth.)
   a. If a male is selected at random and a female is selected at random, what is the probability that both people selected are 65 or older? (Hint: Use the multiplication rule for independent events.)
   b. If two males are selected at random, what is the probability that both of them are 65 or older?
   c. If two females are selected at random, what is the probability that neither of them is 65 or older?

4. In a large community, 72% of the people are adults, 78% of the people have traveled outside the state, and 11% are adults who have not traveled outside the state.

   a. Using a Venn diagram or a hypothetical 1000 table, calculate the probability that a randomly selected person from the community is an adult and has traveled outside the state.

   b. Use the multiplication rule for independent events to decide whether the events *is an adult* and *has traveled outside the state* are independent.

5. In a particular calendar year, 10% of the registered voters in a small city are called for jury duty. In this city, people are selected for jury duty at random from all registered voters in the city, and the same individual cannot be called more than once during the calendar year.

   a. What is the probability that a registered voter is not called for jury duty during a particular year?

   b. What is the probability that a registered voter is called for jury duty two years in a row?

6. A survey of registered voters in a city in New York was carried out to assess support for a new school tax. 51% of the respondents supported the school tax. Of those with school-age children, 56% supported the school tax, while only 45% of those who did not have school-age children supported the school tax.

   a. If a person who responded to this survey is selected at random, what is the probability that

      i. the person selected supports the school tax?

      ii. the person supports the school tax given that he or she does not have school-age children?

   b. Are the two events *has school-age children* and *supports the school tax* independent? Explain how you know this.

   c. Suppose that 35% of those responding to the survey were over the age of 65 and that 10% of those responding to the survey were both over age 65 and supported the school tax. What is the probability that a randomly selected person who responded to this survey supported the school tax given that he or she was over age 65?

Lesson 7: Probability Rules

Classwork

Exercise 1

When a car is brought to a repair shop for a service, the probability that it will need the transmission fluid replaced is 0.38, the probability that it will need the brake pads replaced is 0.28, and the probability that it will need both the transmission fluid and the brake pads replaced is 0.16. Let the event that a car needs the transmission fluid replaced be T and the event that a car needs the brake pads replaced be B.

a. What are the values of

i. $P(T)$

ii. $P(B)$

iii. $P(T \text{ and } B)$

b. Use the addition rule to find the probability that a randomly selected car needs the transmission fluid or the brake pads replaced.

Exercise 2

Josie will soon be taking exams in math and Spanish. She estimates that the probability she passes the math exam is 0.9 and the probability that she passes the Spanish exam is 0.8. She is also willing to assume that the results of the two exams are independent of each other.

a. Using Josie's assumption of independence, calculate the probability that she passes both exams.

b. Find the probability that Josie passes at least one of the exams. (Hint: Passing at least one of the exams is passing math or passing Spanish.)

Example 1: Use of the Addition Rule for Disjoint Events

A set of 40 cards consists of

- 10 black cards showing squares.
- 10 black cards showing circles.
- 10 red cards showing Xs.
- 10 red cards showing diamonds.

A card will be selected at random from the set. Find the probability that the card is black or shows a diamond.

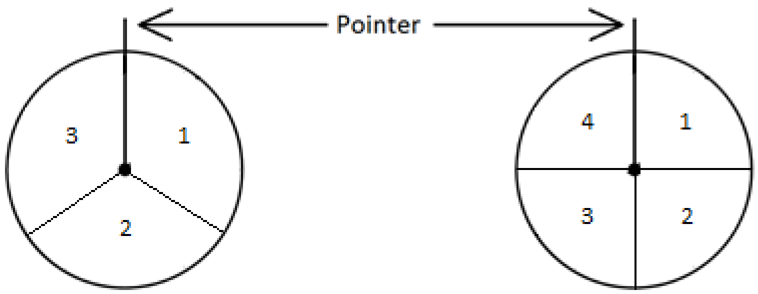
Example 2: Combining Use of the Multiplication and Addition Rules

A red cube has faces labeled 1 through 6, and a blue cube has faces labeled in the same way. The two cubes are rolled. Find the probability that

a. both cubes show 6s.

b. the total score is at least 11.

Exercise 3



The diagram above shows two spinners. For the first spinner, the scores 1, 2, and 3 are equally likely, and for the second spinner, the scores 1, 2, 3, and 4 are equally likely. Both pointers will be spun. Writing your answers as fractions in lowest terms, find the probability that

a. the total of the scores on the two spinners is 2.

b. the total of the scores on the two spinners is 3.

c. the total of the scores on the two spinners is 5.

d. the total of the scores on the two spinners is not 5.

Lesson Summary

The addition rule states that for any two events A and B, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

The addition rule can be used in conjunction with the multiplication rule for independent events: Events A and B are independent if and only if $P(A \text{ and } B) = P(A)P(B)$.

Two events are said to be disjoint if they have no outcomes in common. If A and B are disjoint events, then $P(A \text{ or } B) = P(A) + P(B)$.

The addition rule for disjoint events can be used in conjunction with the multiplication rule for independent events.

Problem Set

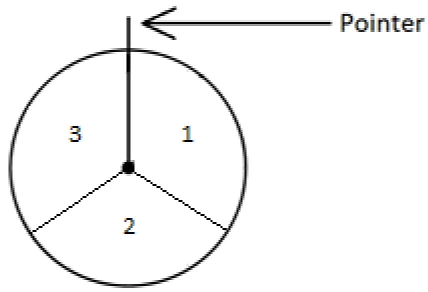
1. Of the works of art at a large gallery, 59% are paintings and 83% are for sale. When a work of art is selected at random, let the event that it is a painting be A and the event that it is for sale be B.
   a. What are the values of $P(A)$ and $P(B)$?
   b. Suppose you are told that $P(A \text{ and } B) = 0.51$. Find $P(A \text{ or } B)$.
   c. Suppose now that you are not given the information in part (b), but you are told that the events A and B are independent. Find $P(A \text{ or } B)$.

2. A traveler estimates that, for an upcoming trip, the probability of catching malaria is 0.18, the probability of catching typhoid is 0.13, and the probability of catching neither of the two diseases is 0.75.
   a. Draw a Venn diagram to represent this information.
   b. Calculate the probability of catching both of the diseases.
   c. Are the events catches malaria and catches typhoid independent? Explain your answer.

3. A deck of 40 cards consists of
   - 10 black cards showing squares, numbered 1–10,
   - 10 black cards showing circles, numbered 1–10,
   - 10 red cards showing Xs, numbered 1–10,
   - 10 red cards showing diamonds, numbered 1–10.

   A card will be selected at random from the deck.
   a. i. Are the events *the card shows a square* and *the card is red* disjoint? Explain.
      ii. Calculate the probability that the card will show a square or will be red.
   b. i. Are the events the card shows a 5 and the card is red disjoint? Explain.
      ii. Calculate the probability that the card will show a 5 or will be red.

4. The diagram below shows a spinner. When the pointer is spun, it is equally likely to stop on 1, 2, or 3. The pointer will be spun three times. Expressing your answers as fractions in lowest terms, find the probability and explain how the answer was determined that the total of the values from all thee spins is



   a. 9.
   b. 8.
   c. 7.

5. A number cube has faces numbered 1 through 6, and a coin has two sides, "heads" and "tails". The number cube will be rolled once, and the coin will be flipped once. Find the probabilities of the following events. (Express your answers as fractions in lowest terms.)
   a. The number cube shows a 6.
   b. The coin shows "heads."
   c. The number cube shows a 6, and the coin shows "heads."
   d. The number cube shows a 6, or the coin shows "heads."

6. Kevin will soon be taking exams in math, physics, and French. He estimates the probabilities of his passing these exams to be as follows:
   - Math: 0.9,
   - Physics: 0.8,
   - French: 0.7.

   Kevin is willing to assume that the results of the three exams are independent of each other. Find the probability that Kevin will
   a. pass all three exams.
   b. pass math but fail the other two exams.
   c. pass exactly one of the three exams.

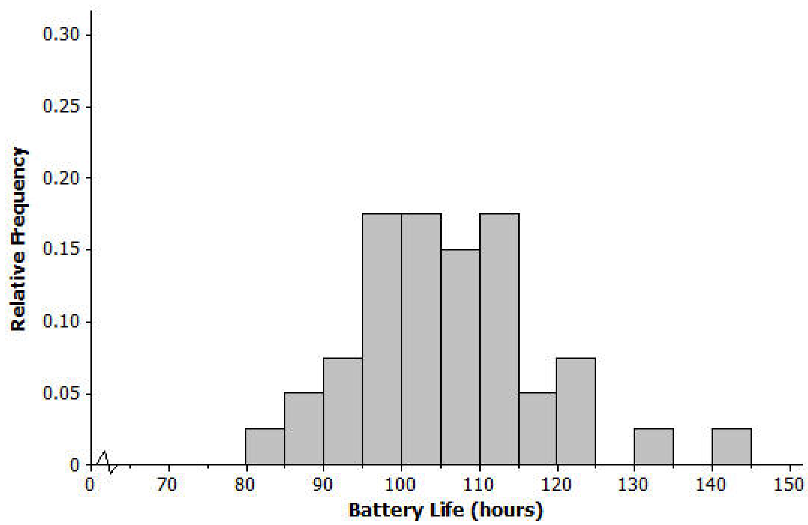
Lesson 8: Distributions—Center, Shape, and Spread

Classwork

Example 1: Center, Shape and Spread

Have you ever noticed how sometimes batteries seem to last a long time, and other times the batteries seem to last only a short time?

The histogram below shows the distribution of battery life (hours) for a sample of 40 batteries of the same brand. When studying a distribution, it is important to think about the shape, center, and spread of the data.



Exercises

1. Would you describe the distribution of battery life as approximately symmetric or as skewed? Explain your answer.

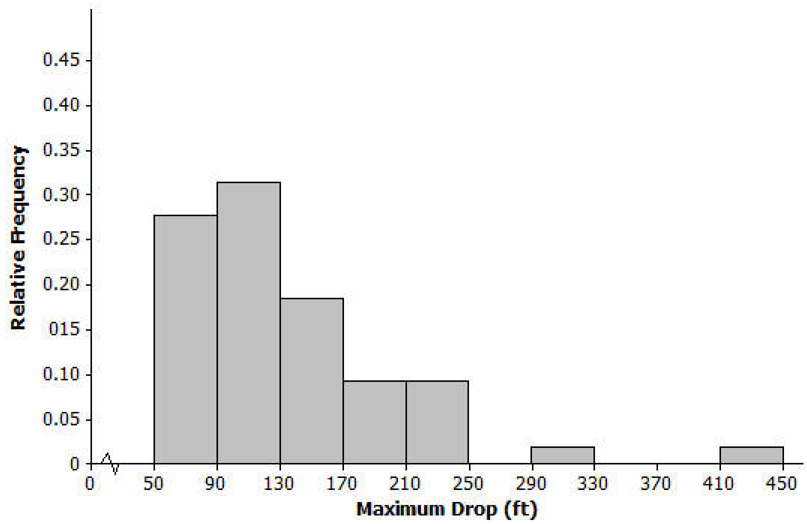
2. Is the mean of the battery life distribution closer to 95, 105, or 115 hours? Explain your answer.

3. Consider 5, 10, or 25 hours as an estimate of the standard deviation for the battery life distribution.
   a. Consider 5 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

   b. Consider 10 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

   c. Consider 25 hours as an estimate of the standard deviation. Is it a reasonable description of a typical distance from the mean? Explain your answer.

The histogram below shows the distribution of the greatest drop (in feet) for 55 major roller coasters in the U.S.



4. Would you describe this distribution of roller coaster maximum drop as approximately symmetric or as skewed? Explain your answer.

5. Is the mean of the maximum drop distribution closer to 90, 135, or 240 feet? Explain your answer.

6. Is the standard deviation of the maximum drop distribution closer to 40, 70, or 100 hours? Explain your answer.

7. Consider the following histograms: Histogram 1, Histogram 2, Histogram 3, and Histogram 4. Descriptions of four distributions are also given. Match the description of a distribution with the appropriate histogram.

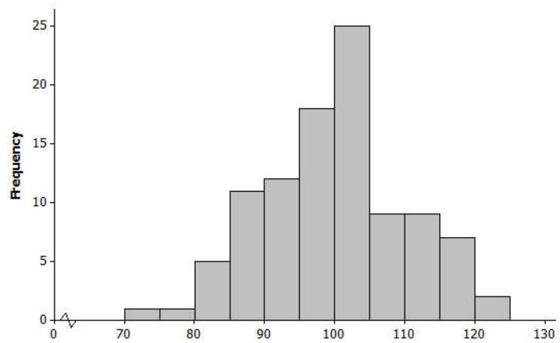
Histogram	Distribution
1	
2	
3	
4	

Description of distributions:

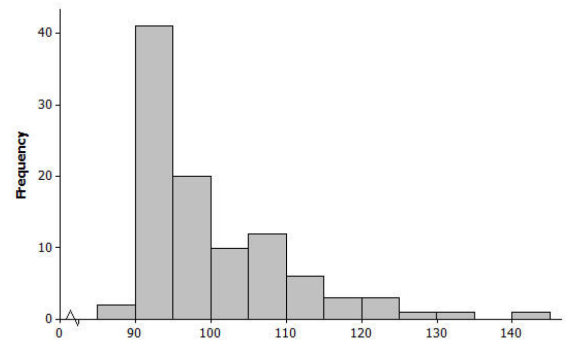
Distribution	Shape	Mean	Standard Deviation
A	Skewed to the right	100	10
B	Approximately symmetric, mound shaped	100	10
C	Approximately symmetric, mound shaped	100	40
D	Skewed to the right	100	40

Histograms:

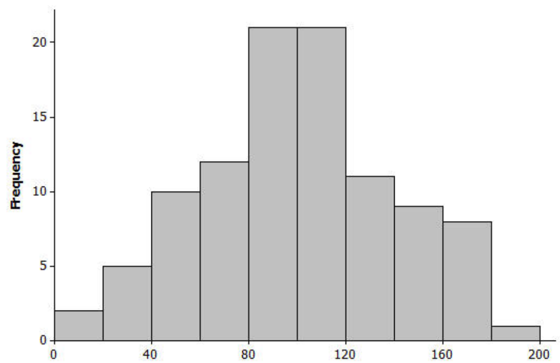
Histogram 1



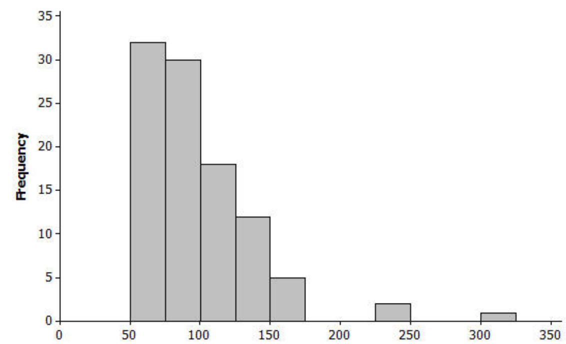
Histogram 2



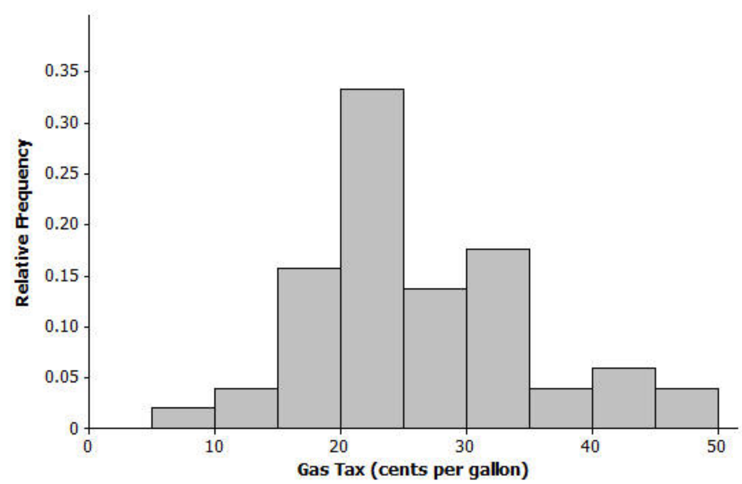
Histogram 3



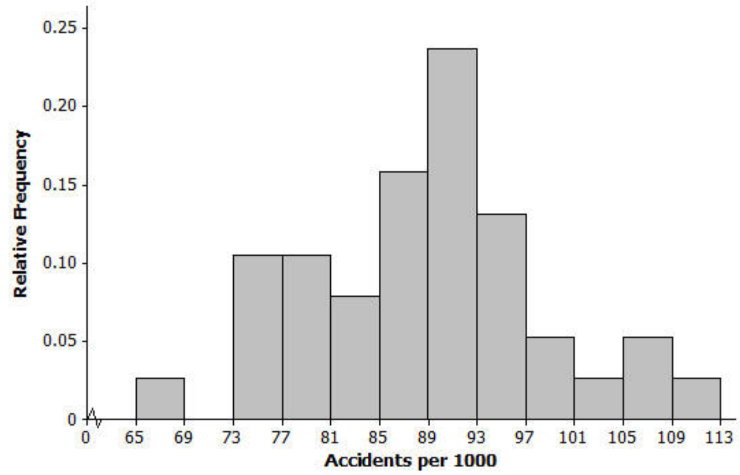
Histogram 4



8. The histogram below shows the distribution of gasoline tax per gallon for the 50 states and the District of Columbia in 2010. Describe the shape, center, and spread of this distribution.



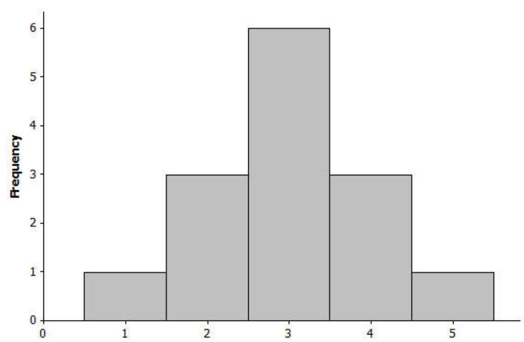
9. The histogram below shows the distribution of the number of automobile accidents per year for every 1,000 people in different occupations. Describe the shape, center, and spread of this distribution.

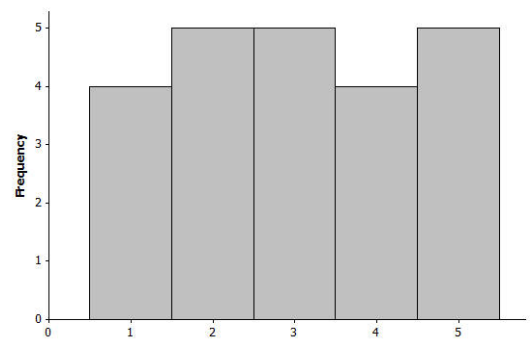


Lesson Summary

Distributions are described by the shape (symmetric or skewed), the center, and the spread (variability) of the distribution.

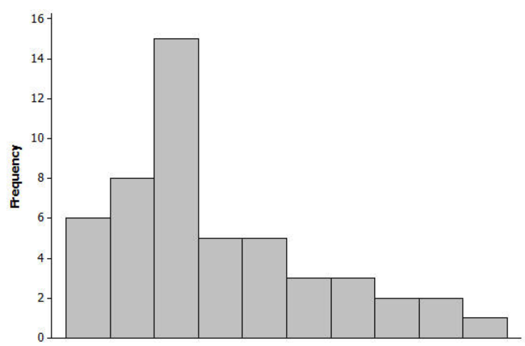
A distribution that is approximately symmetric can take different forms.



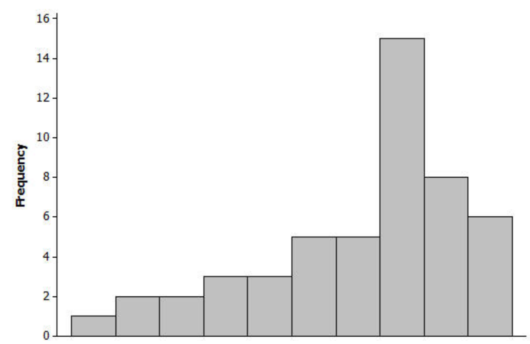


A distribution is described as mound shaped if it is approximately symmetric and has a single peak.

A distribution is skewed to the right or skewed to the left if one of its tails is longer than the other.



Skewed to the right



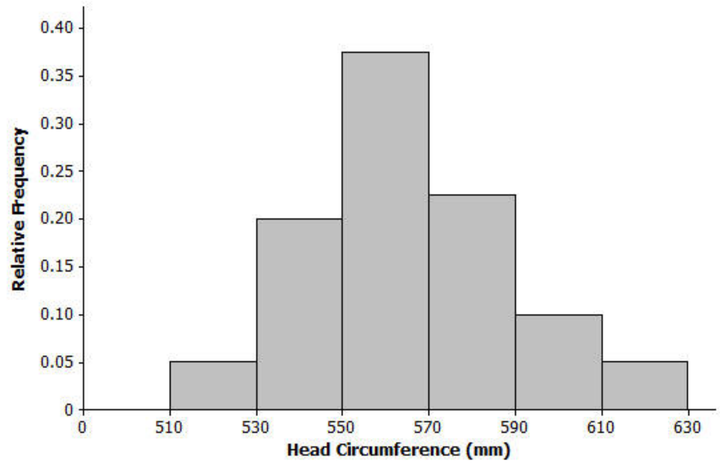
Skewed to the left

The mean of a distribution is interpreted as a typical value and is the average of the data values that make up the distribution.

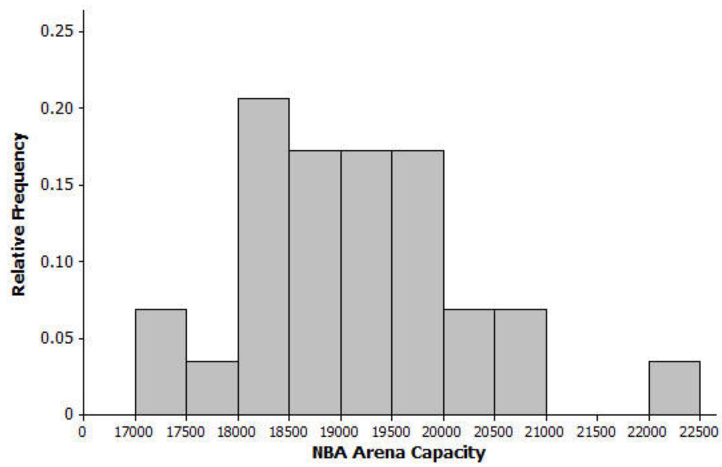
The standard deviation is a value that describes a typical distance from the mean.

Problem Set

1. For each of the following histograms, describe the shape, and give estimates of the mean and standard deviation of the distributions.
   a. Distribution of head circumferences (mm)



   b. Distribution of NBA arena seating capacity



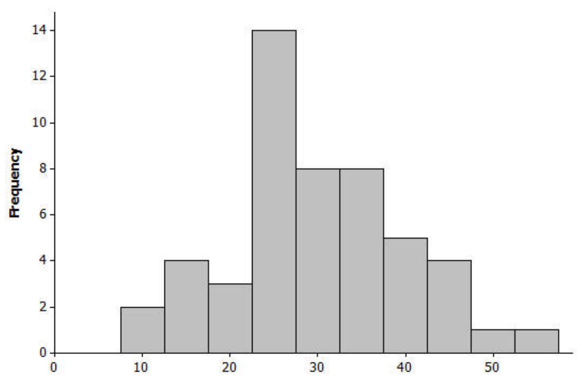
2. For the each of the following, match the description of each distribution with the appropriate histogram.

Histogram	Distribution
1	
2	
3	
4	

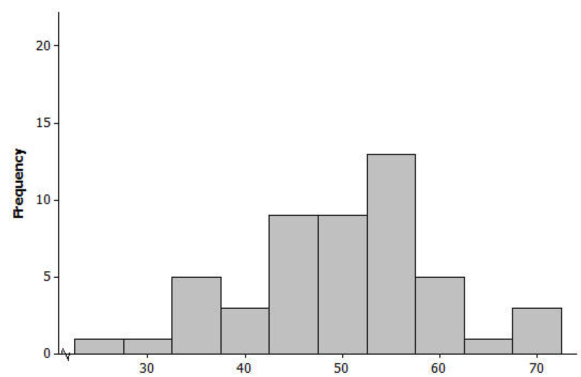
Description of distributions:

Distribution	Shape	Mean	Standard Deviation
A	Approximately symmetric, mound shaped	50	5
B	Approximately symmetric, mound shaped	50	10
C	Approximately symmetric, mound shaped	30	10
D	Approximately symmetric, mound shaped	30	5

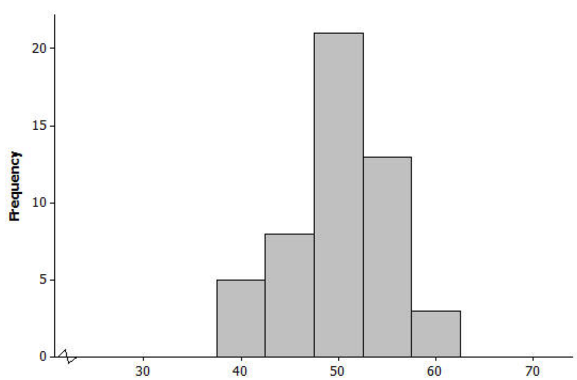
Histogram 1



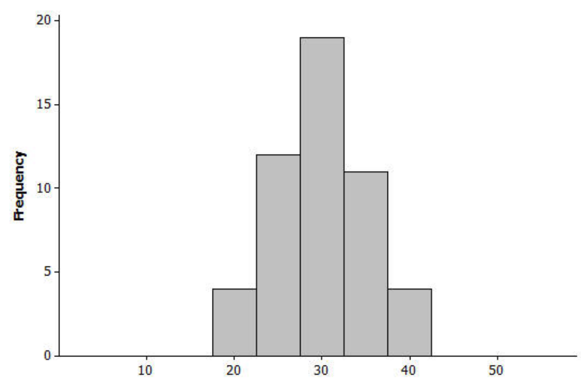
Histogram 2



Histogram 3



Histogram 4



3. Following are the number of calories in a basic hamburger (one meat patty with no cheese) at various fast food restaurants around the country.

   $380,\ 790,\ 680,\ 460,\ 725,\ 1130,\ 240,\ 260,\ 930,\ 331,\ 710,\ 680,\ 1080,\ 612,\ 1180,\ 400, 866, 700, 1060,$
   $270,\ 550,\ 380,\ 940,\ 280,\ 940,\ 550,\ 549,\ 937,\ 820,\ 870,\ 250,\ 740$

   a. Draw a dot plot on the scale below.

   b. Describe the shape of the calorie distribution.
   c. Using technology, find the mean and standard deviation of the calorie data.
   d. Why do you think there is a lot of variability in the calorie data?

Lesson 9: Using a Curve to Model a Data Distribution

Classwork

Example 1: Heights of Dinosaurs and the Normal Curve

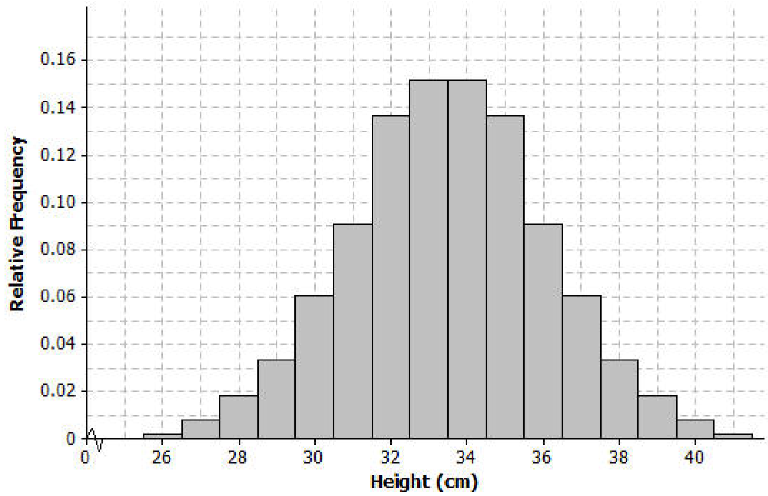
A paleontologist studies prehistoric life and sometimes works with dinosaur fossils. The table below shows the distribution of heights (rounded to the nearest inch) of 660 procompsognathids or "compys."

The heights were determined by studying the fossil remains of the compys.

Height (cm)	Number of Compys	Relative Frequency
26	1	0.002
27	5	0.008
28	12	0.018
29	22	0.033
30	40	0.061
31	60	0.091
32	90	0.136
33	100	0.152
34	100	0.152
35	90	0.136
36	60	0.091
37	40	0.061
38	22	0.033
39	12	0.018
40	5	0.008
41	1	0.002
Total	660	1.00

Exercises 1–8

The following is a relative frequency histogram of the compy heights.



1. What does the relative frequency of 0.136 mean for the height of 32 cm?

2. What is the width of each bar? What does the height of the bar represent?

3. What is the area of the bar that represents the relative frequency for compys with a height of 32 cm?

4. The mean of the distribution of compy heights is 33.5 cm, and the standard deviation is 2.56 cm. Interpret the mean and standard deviation in this context.

5. Mark the mean on the graph and mark <u>one</u> deviation above and below the mean.
   a. Approximately what percent of the values in this data set are within one standard deviation of the mean? (i.e., between $33.5 - 2.56 = 30.94$ cm and $33.5 + 2.56 = 36.06$ cm.)

   b. Approximately what percent of the values in this data set are within <u>two</u> standard deviations of the mean?

6. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Describe the shape of the distribution.

7. Shade the area under the curve that represents the proportion of heights that are within one standard deviation of the mean.

8. Based on our analysis, how would you answer the question, "How tall was a compy?"

Example 2: Gas Mileage and the Normal Distribution

A normal curve is a smooth curve that is symmetric and bell shaped. Data distributions that are mound shaped are often modeled using a normal curve, and we say that such a distribution is approximately normal. One example of a distribution that is approximately normal is the distribution of compy heights from Example 1. Distributions that are approximately normal occur in many different settings. For example, a salesman kept track of the gas mileage for his car over a 25-week span.

The mileages (miles per gallon rounded to the nearest whole number) were

$23,\ 27,\ 27,\ 28,\ 25,\ 26,\ 25,\ 29,\ 26,\ 27,\ 24,\ 26,\ 26,\ 24,\ 27,\ 25,\ 28,\ 25,\ 26,\ 25,\ 29,\ 26,\ 27,\ 24,\ 26.$

Exercise

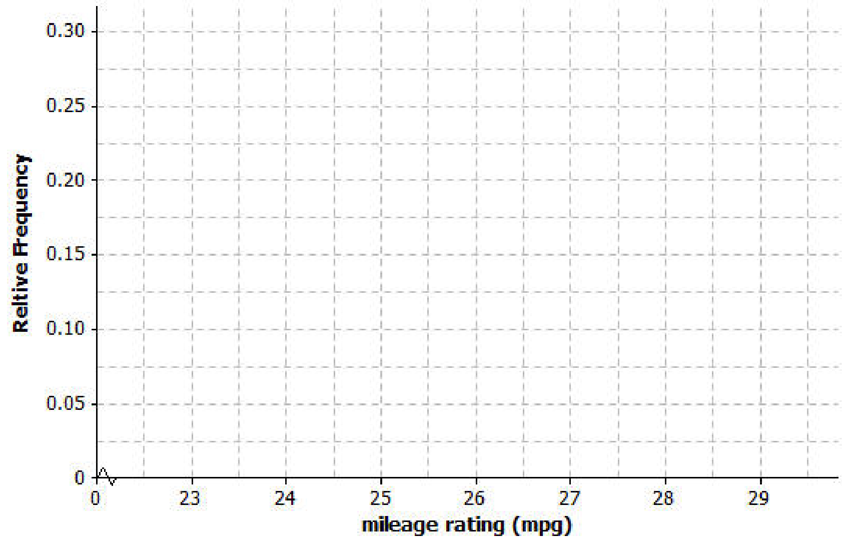
9. Consider the following:

   a. Use technology to find the mean and standard deviation of the mileage data. How did you use technology to assist you?

   b. Calculate the relative frequency of each of the mileage values. For example, the mileage of 26 mpg has a frequency of 7. To find the relative frequency, divide 7 by 25, the total number of mileages recorded. Complete the following table.

Mileage	Frequency	Relative Frequency
23		
24		
25		
26	7	
27		
28		
29		
Total	25	

c. Construct a relative frequency histogram using the scale below.

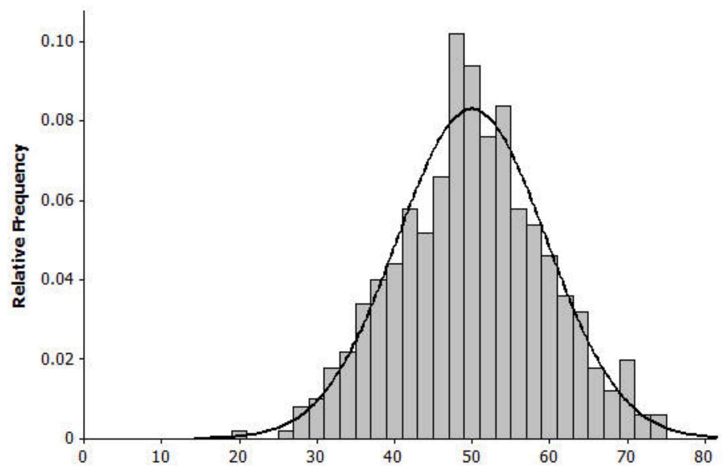


d. Describe the shape of the mileage distribution. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Is this approximately a normal curve?

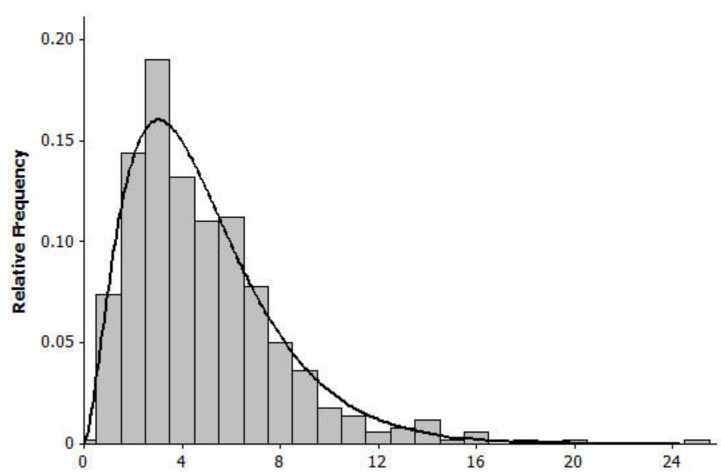
e. Mark the mean on the histogram. Mark one standard deviation to the left and right of the mean. Shade the area under the curve that represents the proportion of data within one standard deviation of the mean. Find the proportion of the data within one standard deviation of the mean.

Lesson Summary

- A normal curve is symmetric and bell shaped. The mean of a normal distribution is located in the center of the distribution. Areas under a normal curve can be used to estimate the proportion of the data values that fall within a given interval.

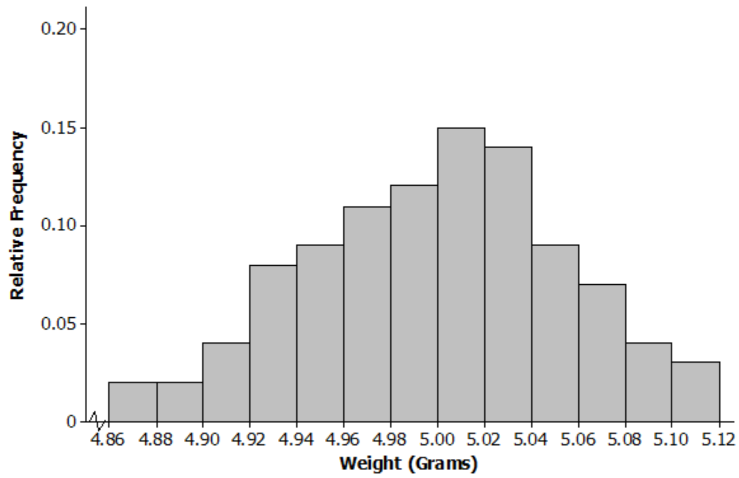


- When a distribution is skewed, it is not appropriate to model the data distribution with a normal curve.



Problem Set

1. Periodically the U.S. Mint checks the weight of newly minted nickels. Below is a histogram of the weights (in grams) of a random sample of 100 new nickels.

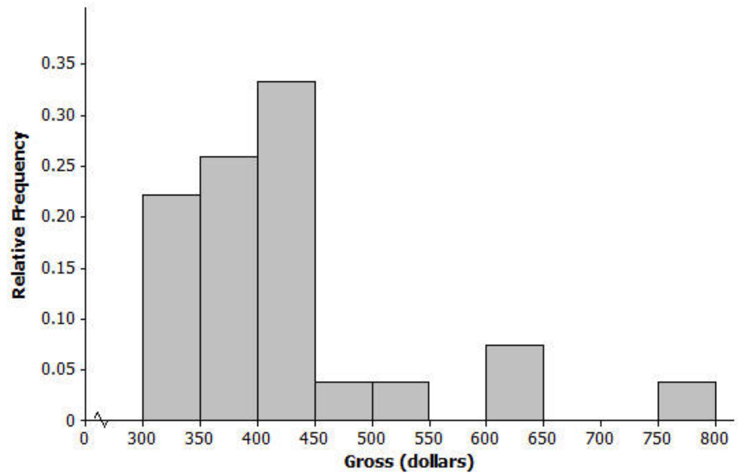


   a. The mean and standard deviation of the distribution of nickel weights are 5.00 grams and 0.06 grams, respectively. Mark the mean on the histogram. Mark one standard deviation above the mean and one standard deviation below the mean.

   b. Describe the shape of the distribution. Draw a smooth curve that comes reasonably close to passing through the midpoints of the tops of the bars in the histogram. Is this approximately a normal curve?

   c. Shade the area under the curve that represents the proportion of data within one standard deviation above and below the mean. Find the proportion of the data within one standard deviation above and below the mean.

2. Below is a relative frequency histogram of the gross (in millions of dollars) for the all-time top-grossing American movies (as of the end of 2012). Gross is the total amount of money made before subtracting out expenses, like advertising costs and actors' salaries.



a. Describe the shape of the distribution of all-time top-grossing movies. Would a normal curve be the best curve to model this distribution? Explain your answer.

b. Which of the following is a reasonable estimate for the mean of the distribution? Explain your choice.

   i. 325 million

   ii. 375 million

   iii. 425 million

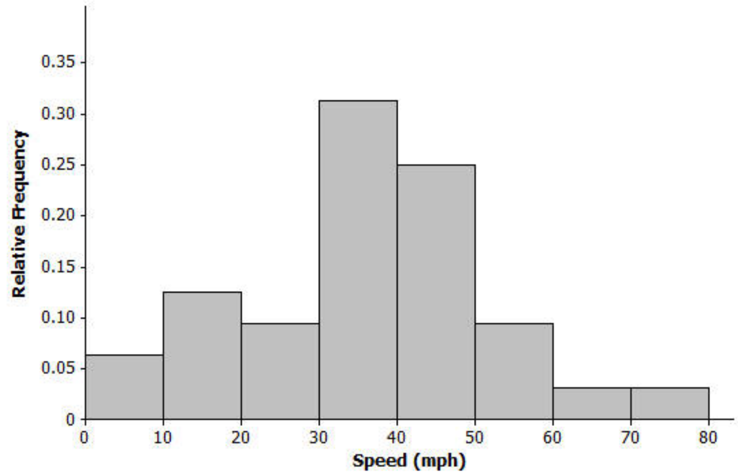
c. Which of the following is a reasonable estimate for the sample standard deviation? Explain your choice.

   i. 50 million

   ii. 100 million

   iii. 200 million

3. Below is a histogram of the top speed of different types of animals.



a. Describe the shape of the top speed distribution.

b. Estimate the mean and standard deviation of this distribution. Describe how you made your estimate.

c. Draw a smooth curve that is approximately a normal curve. The actual mean and standard deviation of this data set are 34.1 and 15.3. Shade the area under the curve that represents the proportion of the data values that are within one standard deviation of the mean.

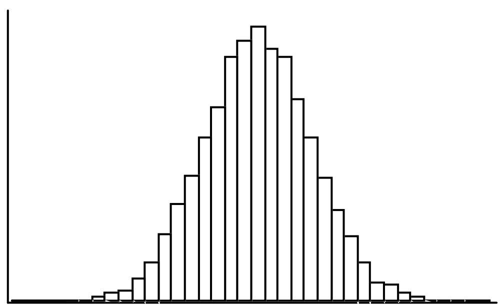
Lesson 10: Normal Distributions

Classwork

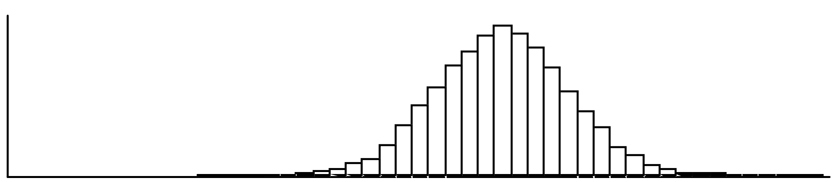
Exercise 1

Consider the following data distributions. In the previous lesson, you distinguished between distributions that were approximately normal and those that were not. For each of the following distributions, indicate if it is approximately normal, skewed, or neither, and explain your choice.

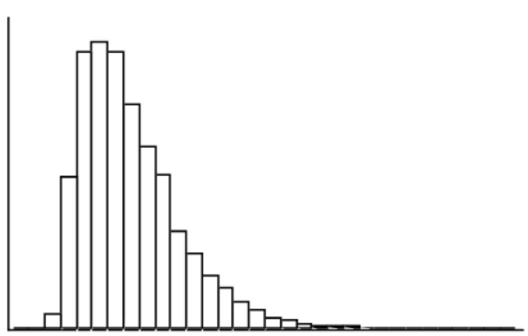
a.



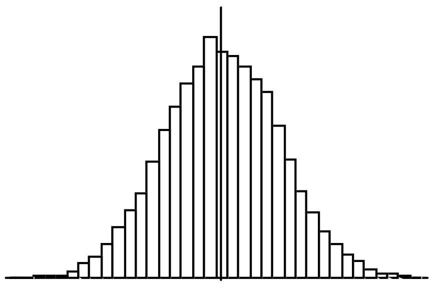
b.



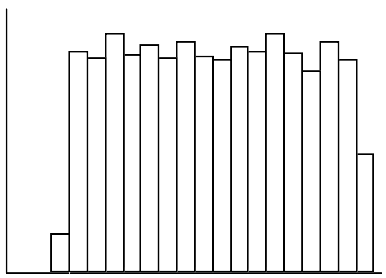
c.



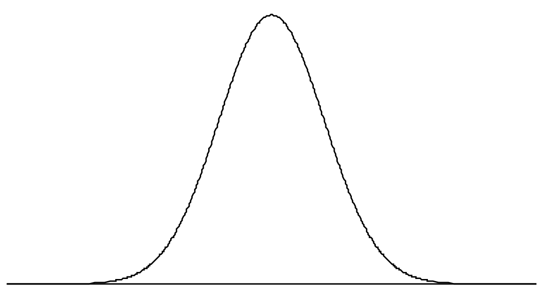
d.



e.



A normal distribution is a distribution that has a particular symmetric mound shape, as shown below.



Exercise 2

When calculating probabilities associated with normal distributions, z scores are used. A z score for a particular value measures the number of standard deviations away from the mean. A positive z score corresponds to a value that is above the mean, and a negative z score corresponds to a value that is below the mean. The letter z was used to represent a variable that has a standard normal distribution where the mean is 0 and standard deviation is 1. This distribution was used to define a z score. A z score is calculated by

$$z = \frac{value - mean}{standard\ deviation}.$$

a. The prices of the printers in a store have a mean of $\$240$ and a standard deviation of $\$50$. The printer that you eventually choose costs $\$340$.

   i. What is the z score for the price of your printer?

   ii. How many standard deviations above the mean was the price of your printer?

b. Ashish's height is 63 inches. The mean height for boys at his school is 68.1 inches, and the standard deviation of the boys' heights is 2.8 inches.

   i. What is the z score for Ashish's height? (Round your answer to the nearest hundredth.)

   ii. What is the meaning of this value?

c. Explain how a z score is useful in describing data.

Example 1: Use of z Scores and a Graphing Calculator to find Normal Probabilities

A swimmer named Amy specializes in the 50 meter backstroke. In competition her mean time for the event is 39.7 seconds, and the standard deviation of her times is 2.3 seconds. Assume that Amy's times are approximately normally distributed.

a. Estimate the probability that Amy's time is between 37 and 44 seconds.

b. Using z scores and a graphing calculator and rounding your answers to the nearest thousandth, find the probability that Amy's time in her next race is between 37 and 44 seconds.

c. Estimate the probability that Amy's time is more than 45 seconds.

d. Using z scores and a graphing calculator and rounding your answers to the nearest thousandth, find the probability that Amy's time in her next race is more than 45 seconds.

e. What is the probability that Amy's time would be at least 45 seconds?

f. Using z scores and a graphing calculator and rounding your answers to the nearest thousandth, find the probability that Amy's time in her next race is less than 36 seconds.

Exercise 3

The distribution of lifetimes of a particular brand of car tires has a mean of $51{,}200$ miles and a standard deviation of $8{,}200$ miles.

a. Assuming that the distribution of lifetimes is approximately normally distributed and rounding your answers to the nearest thousandth, find the probability that a randomly selected tire lasts

   i. between $55{,}000$ and $65{,}000$ miles.

   ii. less than $48{,}000$ miles.

   iii. at least $41{,}000$ miles.

b. Explain the meaning of the probability that you found in part (a–iii).

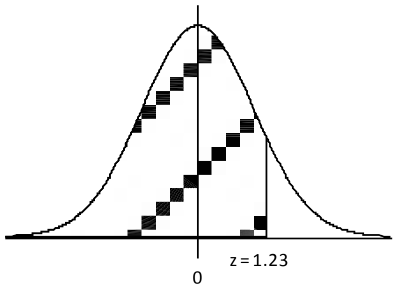
Exercise 4

Think again about the brand of tires described in Exercise 3. What is the probability that the lifetime of a randomly selected tire is within $10{,}000$ miles of the mean lifetime for tires of this brand?

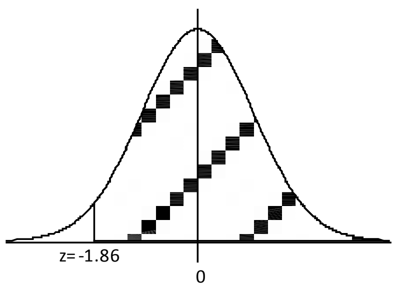
Example 2: Using Table of Standard Normal Curve Areas

The standard normal distribution is the normal distribution with a mean of 0 and a standard deviation of 1. The diagrams below show standard normal distribution curves. Use a table of standard normal curve areas to determine the shaded areas.

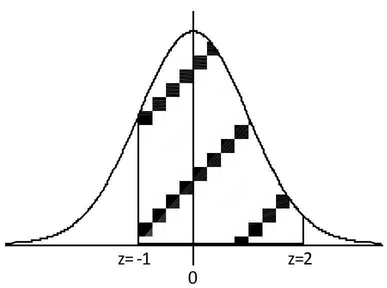
a.



b.



c.



Lesson Summary

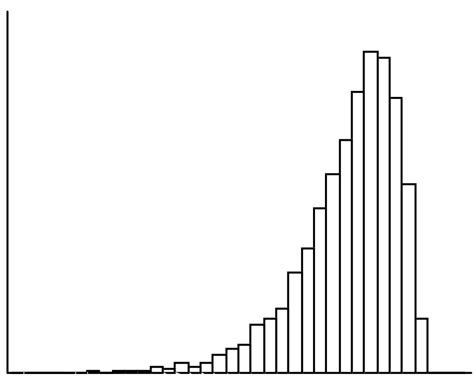
A normal distribution is a continuous distribution that has the particular symmetric mound-shaped curve that is shown at the beginning of the lesson.

Probabilities associated with normal distributions are determined using scores and can be found using a graphing calculator or tables of standard normal curve areas.

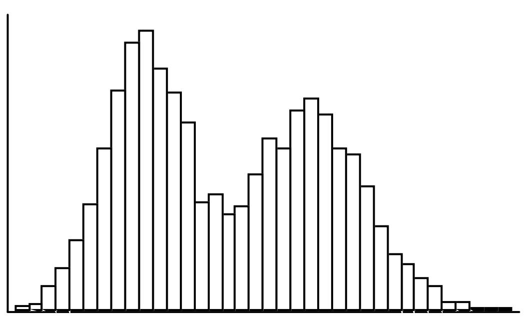
Problem Set

1. Which of the following histograms show distributions that are approximately normal?

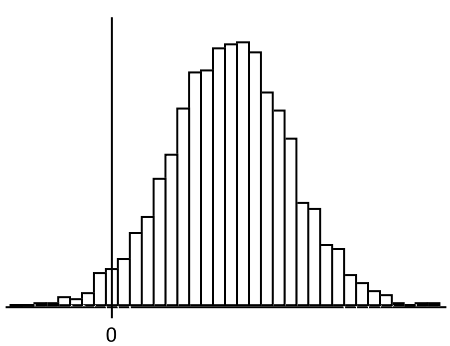
   a.



   b.



   c.



2. Suppose that a particular medical procedure has a cost that is approximately normally distributed with a mean of $\$19{,}800$ and a standard deviation of $\$2{,}900$. For a randomly selected patient, find the probabilities of the following events. (Round your answers to the nearest thousandth.)
   a. The procedure costs between $\$18{,}000$ and $\$22{,}000$.
   b. The procedure costs less than $\$15{,}000$.
   c. The procedure costs more than $\$17{,}250$.

3. Consider the medical procedure described in the previous question, and suppose a patient is charged $\$24{,}900$ for the procedure. The patient is reported as saying, "I've been charged an outrageous amount!" How justified is this comment? Use probability to support your answer.

4. Think again about the medical procedure described in Question 2.
   a. Rounding your answers to the nearest thousandth, find that probability that, for a randomly selected patient, the cost of the procedure is
      i. within two standard deviations of the mean cost.
      ii. more than one standard deviation from the mean cost.
   b. If the mean or the standard deviation were to be changed, would your answers to part (a) be affected? Explain.

5. Use a table of standard normal curve areas to find
   a. the area to the left of $z = 0.56$.
   b. the area to the right of $z = 1.20$.
   c. the area to the left of $z = -1.47$.
   d. the area to the right of $z = -0.35$.
   e. the area between $z = -1.39$ and $z = 0.80$.
   f. Choose a response from parts (a) through (f), and explain how you determined your answer.

Lesson 11: Normal Distributions

Classwork

Example 1: Calculation of Normal Probabilities Using z scores and Tables of Standard Normal Areas

The U.S. Department of Agriculture, in its Official Food Plans (www.cnpp.usda.gov), states that the average cost of food for a 14–18-year-old male (on the "Moderate-cost" plan) is $\$261.50$ per month. Assume that the monthly food cost for a 14–18-year-old male is approximately normally distributed with a mean of $\$261.50$ and a standard deviation of $\$16.25$.

a. Use a table of standard normal curve areas to find the probability that the monthly food cost for a randomly selected $14-18$ year old male is

   i. less than $\$280$.

   ii. more than $\$270$.

   iii. more than $\$250$.

iv. between $\$240$ and $\$275$.

b. Explain the meaning of the probability that you found in part (a–iv).

Exercise 1

The USDA document described in Example 1 also states that the average cost of food for a 14–18 year old female (again, on the "Moderate-cost" plan) is $\$215.20$ per month. Assume that the monthly food cost for a 14–18 year old female is approximately normally distributed with mean $\$215.20$ and standard deviation $\$14.85$.

a. Use a table of standard normal curve areas to find the probability that the monthly food cost for a randomly selected 14–18 year old female is

i. less than $\$225$.

ii. less than $\$200$.

iii. more than $250.

iv. between $190 and $220.

b. Explain the meaning of the probability that you found in part (a–iv).

Example 2: Use of a Graphing Calculator to Find Normal Probabilities Directly

Return to the information given in Example 1. Using a graphing calculator, and *without* using z scores, find the probability (rounded to the nearest thousandth) that the monthly food cost for a randomly selected 14–18 year old male is

a. between $260 and $265.

b. at least $252.

c. at most $248.

Exercise 2

Return to the information given in Exercise 1.

a. In Exercise 1, you calculated the probability that that the monthly food cost for a randomly selected 14–18 year old female is between $190 and $220. Would the probability that the monthly food cost for a randomly selected 14–18 year old female is between $195 and $230 be greater than or smaller than the probability for between $190 and $220? Explain your thinking.

b. Do you think that the probability that the monthly food cost for a randomly selected 14–18 year old female is between $195 and $230 is closer to 0.50, 0.75, or 0.90? Explain your thinking.

c. Using a graphing calculator, and without using z scores, find the probability (rounded to the nearest thousandth) that the monthly food cost for a randomly selected 14–18 year old female is between $195 and $230. Is this probability consistent with your answer to part (b)?

d. How does the probability you calculated in part (c) compare to the probability that would have been obtained using the table of normal curve areas?

e. What is one advantage to using a graphing calculator to calculate this probability?

f. In Exercise 1, you calculated the probability that the monthly food cost for a randomly selected 14–18 year old female is at most \$200. Would the probability that the monthly food cost for a randomly selected 14–18 year old female is at most $\$210$ be greater than or less than the probability for at most $\$200$? Explain your thinking.

g. Do you think that the probability that the monthly food cost for a randomly selected 14−18 year old female is at most $\$210$ is closer to 0.10, 0.30, or 0.50? Explain your thinking.

h. Using a graphing calculator, and without using z scores, find the probability (rounded to the nearest thousandth) that the monthly food cost for a randomly selected 14–18 year old female is at most $\$210$.

i. Using a graphing calculator, and without using z scores, find the probability (rounded to the nearest thousandth) that the monthly food cost for a randomly selected 14–18 year old female is at least $\$235$.

Example 3: Using a Spreadsheet to Find Normal Probabilities

Return to the information given in Example 1. The U.S. Department of Agriculture, in its Official Food Plans (www.cnpp.usda.gov), states that the average cost of food for a 14–18-year-old male (on the "moderate-cost" plan) is $\$261.50$ per month. Assume that the monthly food cost for a 14–18-year-old male is approximately normally distributed with a mean of $\$261.50$ and a standard deviation of $\$16.25$. Round your answers to four decimal places.

Use a spreadsheet to find the probability that the monthly food cost for a randomly selected 14–18 year old male is

a. less than $\$280$.

b. more than $\$270$.

c. more than $\$250$.

d. between $\$240$ and $\$275$.

Exercise 3

The USDA document described in Example 1 also states that the average cost of food for a 14–18 year old female (again, on the "moderate-cost" plan) is $\$215.20$ per month. Assume that the monthly food cost for a 14–18 year old female is approximately normally distributed with a mean of $\$215.20$ and a standard deviation of $\$14.85$. Round your answers to 4 decimal places.

Use a spreadsheet to find the probability that the monthly food cost for a randomly selected 14–18 year old female is

a. less than $\$225$.

b. less than $\$200$.

c. more than $\$250$.

d. between $\$190$ and $\$220$.

Exercise 4

The reaction times of 490 people were measured. The results are shown in the frequency distribution below.

Reaction Time (seconds)	0.1 to < 0.15	0.15 to < 0.2	0.2 to < 0.25	0.25 to < 0.3	0.3 to < 0.35	0.35 to < 0.4
Frequency	9	82	220	138	37	4

a. Construct a histogram that displays these results.

b. Looking at the histogram, do you think a normal distribution would be an appropriate model for this distribution?

c. The mean of the reaction times for these 490 people is 0.2377, and the standard deviation of the reaction times is 0.0457. For a normal distribution with this mean and standard deviation, what is the probability that a randomly selected reaction time is at least 0.25?

d. The actual proportion of these 490 people who had a reaction time that was at least 0.25 is 0.365 (this can be calculated from the frequency distribution). How does this proportion compare to the probability that you calculated in part (c)? Does this confirm that the normal distribution is an appropriate model for the reaction time distribution?

Lesson Summary

Probabilities associated with normal distributions can be found using z scores and tables of standard normal curve areas.

Probabilities associated with normal distributions can be found directly (without using z scores) using a graphing calculator.

When a data distribution has a shape that is approximately normal, a normal distribution can be used as a model for the data distribution. The normal distribution with the same mean and the standard deviation as the data distribution is used.

Problem Set

1. Use a table of standard normal curve areas to find
   a. the area to the left of $z = 1.88$.
   b. the area to the right of $z = 1.42$.
   c. the area to the left of $z = -0.39$.
   d. the area to the right of $z = -0.46$.
   e. the area between $z = -1.22$ and $z = -0.5$.

2. Suppose that the durations of high school baseball games are approximately normally distributed with mean 105 minutes and standard deviation 11 minutes. Use a table of standard normal curve areas to find the probability that a randomly selected high school baseball game lasts
   a. less than 115 minutes.
   b. more than 100 minutes.
   c. between 90 and 110 minutes.

3. Using a graphing calculator, and *without* using z values, check your answers to Problem 2. (Round your answers to the nearest thousandth
   a.
   b.
   c.

4. In Problem 2, you were told that the durations of high school baseball games are approximately normally distributed with mean 105 minutes and standard deviation 11 minutes. Suppose also that the durations of high school softball games are approximately normally distributed with a mean of 95 minutes and the same standard deviation, 11 minutes. Is it more likely that a high school baseball game will last between 100 and 110 minutes or that a high school softball game will last between 100 and 110 minutes? Answer this question without doing any calculations!

5. A farmer has 625 female adult sheep. The sheep have recently been weighed, and the results are shown in the table below.

Weight (pounds)	140 to < 150	150 to < 160	160 to < 170	170 to < 180	180 to < 190	190 to < 200	200 to < 210
Frequency	8	36	173	221	149	33	5

   a. Construct a histogram that displays these results.
   b. Looking at the histogram, do you think a normal distribution would be an appropriate model for this distribution?
   c. The weights of the 625 sheep have mean 174.21 pounds and standard deviation 10.11 pounds. For a normal distribution with this mean and standard deviation, what is the probability that a randomly selected sheep has a weight of at least 190 pounds? (Round your answer to the nearest thousandth.)

Lesson 12: Types of Statistical Studies

Classwork

Opening Exercise

You want to know what proportion of the population likes rock music. You carefully consider three ways to conduct a study. What are the similarities and differences between the following three alternatives? Do any display clear advantages or disadvantages over the others?

a. You could pick a random sample of people and ask them the question, "Do you like rock music?" and record their answers.

b. You could pick a random sample of people and follow them for a period of time, noting their music purchases, both in stores and online.

c. You could pick a random sample of people, separate it into groups, and have each group listen to a different genre of music. You would collect data on the people who display an emotional response to the rock music.

A statistical study begins by asking a question that can be answered with data. The next steps are to collect appropriate data, organize and analyze it, and arrive at a conclusion in the context of the original question. This lesson focuses on the three main types of statistical studies: observational studies, surveys, and experiments. The objective of an observational study and a survey is to learn about characteristics of some population, so the data should be collected in a way that would result in a representative sample. This speaks to the importance of random selection of subjects for the study. The objective of an experiment is to answer such questions as, "What is the effect of treatments on a response variable?" Data in an experiment need to be collected in a way that does not favor one treatment over another. This demonstrates the importance of random assignment of subjects in the study to the treatments.

An observational study is one in which the values of one or more variables are observed with no attempt to affect the outcomes. One kind of observational study is a survey. A survey requires asking a group of people to respond to one or more questions. (A poll is one example of a survey.) An experiment differs from an observational study: In an experiment, subjects are assigned to treatments for the purpose of seeing what effect the treatment has on some response while an observational study makes no attempt to affect the outcomes, i.e., no treatment is given. Note that subjects could be people, animals, or any set of items that produce variability in their responses. Here is an example of an observational study: In a random sample of students, it was observed that those students who played a musical instrument had better grades than those who did not play a musical instrument. In an experiment, a group of students who do not currently play a musical instrument would be assigned at random to having to play a musical instrument or not having to play a musical instrument for a certain period of time. Then, at the end of the period of time we would compare academic performance.

Classify each of the three study methods about rock music as an observational study, a survey or an experiment.

Example 1: Survey

Item	I like the item.	I don't like the item.	I have never tried the item.
Salad			
Veg pizza			
Turkey sandwich			
Raspberry tea			

a. It is easy to determine if a study is a survey. A survey asks people to respond to questions. But surveys can be flawed in several ways. Questions may be confusing. For example, consider the following question:

What kind of computer do you own? (*Circle one*) Mac IBM-PC

How do you answer that question if you don't own a computer? How do you answer that question if you own a different brand? A better question would be:

Do you own a computer? (*Circle one*) Yes No

If you answered yes, what brand of computer is it? ________________________"

Now consider the question, "Do you like your school's cafeteria food?"

Re-write the question in a better form. Keep in mind that not all students may use the school's cafeteria, and even if they do, there may be some foods that they like and some that they don't like.

b. Something else to consider with surveys is how survey participants are chosen. If the purpose of the survey is to learn about some population, ideally participants would be randomly selected from the population of interest. If people are not randomly selected, misleading conclusions from the survey data may be drawn. There are many famous examples of this. Perhaps the most famous case was in 1936 when *The Literary Digest* magazine predicted that Alf Landon would beat incumbent President Franklin Delano Roosevelt by 370 electoral votes to 161. Roosevelt won 523 to 8.

Ten million questionnaires were sent to prospective voters (selected from the magazine's subscription list, automobile registration lists, phone lists, and club membership lists), and over two million questionnaires were returned. Surely such a large sample should represent the whole population. How could *The Literary Digest* prediction be so far off the mark?

c. Write or say to your neighbor two things that are important about surveys.

Example 2: Observational Study

a. An observational study records the values of variables for members of a sample but does not attempt to influence the responses. For example, researchers investigated the link between use of cell phones and brain cancer. There are two variables in this study: One is the extent of cell-phone usage, and the second is whether a person has brain cancer. Both variables were measured for a group of people. This is an observational study. There was no attempt to influence peoples' cell-phone usage to see if different levels of usage made any difference in whether or not a person developed brain cancer.

   Why would studying any relationship between asbestos exposure and lung cancer be an observational study and not an experiment?

b. In an observational study (just as in surveys), the people or objects to be observed would ideally be selected at random from the population of interest. This would eliminate bias and make it possible to generalize from a sample to a population. For example, to determine if the potato chips made in a factory contain the desired amount of salt, a sample of chips would be selected randomly so that the sample can be considered to be representative of the population of chips.

   Discuss how a random sample of 100 chips might be selected from a conveyor belt of chips.

c. Suppose that an observational study establishes a link between asbestos exposure and lung cancer. Based on that finding, can we conclude that asbestos exposure causes lung cancer? Why or why not?

d. Write or say to your neighbor two things that are important about observational studies.

Example 3: Experiment

a. An experiment imposes treatments to see the effect of the treatments on some response. Suppose that an observational study indicated that a certain type of tree did not have as much termite damage as other trees. Researchers wondered if resin from the tree was toxic to termites. They decided to do an experiment where they exposed some termites to the resin and others to plain water and recorded whether the termites survived. The explanatory variable (treatment variable) is the exposure type (resin, plain water), and the response variable is whether or not the termite survived. We know this is an experiment because the researchers imposed a treatment (exposure type) on the subjects (termites).

   Is the following an observational study or an experiment? Why? If it is an experiment, identify the treatment variable and the response variable. If it is an observational study, identify the population of interest.

   > A study was done to answer the question, "What is the effect of different durations of light and dark on the growth of radish seedlings?" Three similar growth chambers (plastic bags) were created in which $\mathbf{30}$ seeds randomly chosen from a package were placed in each chamber. One chamber was randomly selected and placed in $\mathbf{24}$ hours of light, another for $\mathbf{12}$ hours of light and $\mathbf{12}$ hours of darkness, and a third for $\mathbf{24}$ hours of darkness. After three days, researchers measured and recorded the lengths of radish seedlings for the germinating seeds.

b. In an experiment, random assignment of subjects to treatments is done to create comparable treatment groups. For example, a university biologist wants to compare the effects of two weed killers on pansies. She chooses 24 plants. If she applies weed killer A to the 12 healthiest plants and B to the remaining 12 plants, she won't know which plants died due to the type of weed killer used and which plants subjected to weed killer B were already on their "last legs." Randomly selecting twelve plants to receive weed killer A and then assigning the rest to B would help ensure that the plants in each group are fairly similar.

   How might the biologist go about randomly assigning 12 plants from the 24 candidates to receive weed killer A? Could she be sure to get exactly 12 plants assigned to weed killer A and 12 plants to weed killer B by tossing a fair coin for each plant and assigning "heads up" plants to weed killer A and "tails up" to weed killer B? If not, suggest a method that you would use.

c. Write or say to your neighbor two things that are important about experiments.

Exercises 1–3

1. For each of the following study descriptions, identify whether the study is a survey, observational study, or experiment, and give a reason for your answer. For observational studies, identify the population of interest. For experiments, identify the treatment and response variables.
   a. A study investigated whether boys are quicker at learning video games than girls. Twenty randomly selected boys and twenty randomly selected girls played a video game that they had never played before. The time it took them to reach a certain level of expertise was recorded.

   b. As your statistics project, you collect data by posting five questions on poster board around your classroom and recording how your classmates respond to them.

   c. A professional sports team traded its best player. The local television station wanted to find out what the fans thought of the trade. At the beginning of the evening news program, they asked viewers to call one number if they favored the trade and a different number if they were opposed to the trade. At the end of the news program they announced that 53.7% of callers favored the trade.

   d. The local Department of Transportation is responsible for maintaining lane and edge lines on its paved roads. There are two new paint products on the market. Twenty comparable stretches of road are identified. Paint A is randomly assigned to ten of the stretches of road and paint B to the other ten. The department finds that paint B lasts longer.

   e. The National Highway Traffic Safety Administration conducts annual studies on drivers' seatbelt use at a random selection of roadway sites in each state in the United States. To determine if seatbelt usage has increased, data are analyzed over two successive years.

   f. People should brush their teeth at least twice a day for at least two to three minutes with each brushing. For a statistics class project, you ask a random number of students at your school questions concerning their tooth-brushing activities.

g. A study determines whether taking aspirin regularly helps to prevent heart attacks. A large group of male physicians of comparable health were randomly assigned equally to taking an aspirin every second day or to taking a placebo. After several years, the proportion of the males who had suffered heart attacks in each group was compared.

2. For the following, is the stated conclusion reasonable? Why or why not?

   A study found a positive relationship between the happiness of elderly people and the number of pets they have. Therefore, having more pets causes elderly people to be happier.

3. A researcher wanted to find out whether higher levels of a certain drug given to experimental rats would decrease the time it took them to complete a given maze to find food.

   a. Why would the researcher have to carry out an experiment rather than an observational study?

   b. Describe an experiment that the researcher might carry out based on 30 comparable rats and three dosage levels: 0 mg, 1 mg, and 2 mg.

Lesson Summary

- There are three major types of statistical studies: observational studies, surveys, and experiments.
  - An *observational study* records the values of variables for members of a sample.
  - A *survey* is a type of observational study that gathers data by asking people a number of questions.
  - An *experiment* assigns subjects to treatments for the purpose of seeing what effect the treatments have on some response.
- To avoid bias in observational studies and surveys, it is important to select subjects randomly.
- Cause and effect conclusions cannot be made in observational studies or surveys.
- In an experiment, it is important to assign subjects to treatments randomly in order to make cause-and-effect conclusions.

Problem Set

1. State if the following is an observational study, survey, or experiment, and give a reason for your answer.

   Linda wanted to know if it is easier for students to memorize a list of common three-letter words (such as fly, pen, red, …) than a list of three-letter nonsense words (such as vir, zop, twq, …). She randomly selected 28 students from all tenth-graders in her district. She put 14 blue and 14 red chips in a jar, and without looking each student chose a chip. Those with red chips were given the list of common words; those with blue chips were given the list of nonsense words. She gave all students one minute to memorize their list. After the minute, she collected the lists and asked them to write down all the words that they could remember. She recorded the number of correct words recalled.

2. State if the following is an observational study, survey, or experiment, and give a reason for your answer.

   Ken wants to compare how many hours a week that sixth graders spend doing mathematics homework to how many hours a week that eleventh graders spend doing mathematics homework. He randomly selects ten sixth graders and ten eleventh graders and records how many hours each student spent on mathematics homework in a certain week.

3. Suppose that in your health class you read two studies on the relationship between eating breakfast and success in school for elementary school children. Both studies concluded that eating breakfast causes elementary school children to be successful in school.
   a. Suppose that one of the studies was an observational study. Describe how you would recognize that they had conducted an observational study. Were the researchers correct in their causal conclusion?
   b. Suppose that one of the studies was an experiment. Describe how you would recognize that they had conducted an experiment. Were the researchers correct in their causal conclusion?

4. Data from a random sample of 50 students in a school district showed a positive relationship between reading score on a standardized reading exam and shoe size. Can it be concluded that having bigger feet causes one to have a higher reading score? Explain your answer.

Use the following scenarios for Problems 5–7.

A. Researchers want to determine if there is a relationship between whether or not a woman smoked during pregnancy and the birth weight of her baby. Researchers examined records for the past five years at a large hospital.

B. A large high school wants to know the proportion of students who currently use illegal drugs. Uniformed police officers asked a random sample of 200 students about their drug use.

C. A company develops a new dog food. The company wants to know if dogs would prefer its new food over the competition's dog food. One hundred dogs, who were food-deprived overnight, were given equal amounts of the two dog foods; the new food vs. competitor's food. The proportion of dogs preferring the new food was recorded.

5. Which scenario above describes an experiment? Explain why.

6. Which scenario describes a survey? Will the results of the survey be accurate? Why or why not?

7. The remaining scenario is an observational study. Is it possible to perform an experiment to determine if a relationship exists? Why or why not?

Lesson 13: Using Sample Data to Estimate a Population Characteristic

Classwork

Example 1: Population and Sample

Answer the following questions, and then share your responses with a neighbor.

a. A team of scientists wants to determine the average length and weight of fish in Lake Lucerne. Name a sample that can be used to help answer their question.

b. Golf balls from different manufacturers are tested to determine which brand travels the farthest. What is the population being studied?

Exercise 1

For each of the following, does the group described constitute a population or a sample? Or, could it be considered to be either a population or a sample? Explain your answer.

a. The animals that live in Yellowstone National Park.

b. The first-run movies released last week that were shown at the local theater complex last weekend.

c. People who are asked how they voted in an exit poll.

d. Some cars on the lot of the local car dealer.

e. The words of the Gettysburg Address.

f. The colors of pencils available in a 36-count packet of Crayola colored pencils.

g. The students from your school who watched your school's soccer team play yesterday.

Example 2: Representative Sample

If a sample is taken for the purpose of generalizing to a population, the sample must be representative of the population. In other words, it must be similar to the population even though it is smaller than the population. For example, suppose you are the campaign manager for your friend who is running for Senior Class President. You would like to know what proportion of students would vote for her if the election were held today. The class is too big to ask everyone (314 students). What would you do?

Comment on whether or not each of the following sampling procedures should be used. Explain why or why not.

a. Poll everyone in your friend's math class.

b. Assign every student in the senior class a number from 1 to 314. Then use a random number generator to select 30 students to poll.

c. Ask each student that is going through the lunch line in the cafeteria who they will vote for.

Exercise 2

There is no procedure that guarantees a representative sample. But the best procedure to obtain a representative sample is one that gives every different possible sample an equal chance to be chosen. The sample resulting from such a procedure is called a **random sample**.

Suppose that you want to randomly select 60 employees from a group of 625 employees.

Explain how to use a random number table or a calculator with a random number generator to choose 60 different numbers at random and include the students with these numbers in the sample.

Example 3: Population Characteristics and Sample Statistics

A statistical study begins with a question of interest that can be answered by data. Depending on the study, data could be collected from all individuals in the population or from a random sample of individuals selected from the population. Read through the following and identify which of the summary measures represents a *population characteristic* and which represents a *sample statistic*. Explain your reasoning for each.

Suppose the population of interest is the words of the Gettysburg Address. There are 269 of them (depending on the version).

a. The proportion of nouns in all words of the Gettysburg Address.

b. The proportion of nouns or the proportion of words containing the letter "e" in a random sample of words taken from the Gettysburg Address.

c. The mean length of the words in a random sample of words taken from the Gettysburg Address.

d. The proportion of all words in the Gettysburg Address that contain the letter “e.”

e. The mean length of all words in the Gettysburg Address.

Exercise 3

For the following items of interest, describe an appropriate population, population characteristic, sample, and sample statistic. Explain your answer.

a. Time it takes students to run a quarter-mile.

b. National forests that contain bald eagle nests.

c. Curfew time of boys compared to girls.

d. Efficiency of electric cars.

Exercise 4

Consider the following questions:

- What proportion of eleventh graders at our high school are taking at least one advanced placement course?
- What proportion of eleventh graders at our high school have a part-time job?
- What is the typical number of hours an eleventh grader at our high school studies outside of school hours on a weekday (Monday, Tuesday, Wednesday, or Thursday)?
- What is the typical time (in minutes) that students at our high school spend getting to school?
- What is the proportion of students at our high school who plan to attend a college or technical school after graduation?
- What is the typical amount of time (in hours per week) that students at our high school are involved in community service?

Select one of these questions (or a different statistical question that has been approved by your teacher). Working with your group, write a paragraph that:

- states the statistical question of interest pertaining to the students in the population for the statistical question selected.
- identifies a population characteristic of interest.
- identifies the appropriate statistic based on a sample of 40 students.
- states what property your sample must have for you to be able to use its results to generalize to all students in your high school.
- includes the details on how you would select your sample.

Lesson Summary

We refer to summary measures calculated using data from an entire population as population characteristics. We refer to summary measures calculated using data from a sample as sample statistics. To generalize from a sample to the corresponding population, it is important that the sample be a random sample from the population. A random sample is one that is selected in a way that gives every different possible sample an equal chance of being chosen.

Problem Set

1. In the following, identify whether the subjects being measured are the sample or the population. In some cases, they could be considered a sample or a population. Explain each answer.

Subjects	What is being measured	Sample or Population? Explain
Some students in your class	Number of books in backpack	
AA batteries of a certain brand	Lifetime	
Birds in Glacier National Park	Number of species	
Students in your school	Number absent or present today	
Words in the Constitution of the U.S.	Whether a noun or not	
Americans of voting age	Opinion on an issue	

2. For the following items of interest, describe an appropriate population, a population characteristic, sample, and sample statistic.
   a. Whether or not a driver is speeding in your school zone during school hours in a day.
   b. Seatbelt usage of men compared to women.
   c. Impact of a new antidepressant on people with severe headaches.

3. What are the identification numbers for ten students chosen at random from a population of 78 students based on the following string of random digits? Start at the left.

27816 78416 01822 73521 37741 016312 68000 53645 56644 97892 63408 77919 44575

Lesson 14: Sampling Variability in the Sample Proportion

Classwork

Example 1: Polls

A recent poll stated that 40% of Americans pay "a great deal" or a "fair amount" of attention to the nutritional information that restaurants provide. This poll was based on a random sample of $2{,}027$ adults living in the U.S.

The 40% corresponds to a proportion of 0.40, and 0.40 is called a sample proportion. It is an estimate of the proportion of all adults who would say they pay "a great deal" or a "fair amount" of attention to the nutritional information that restaurants provide. If you were to take a random sample of 20 Americans, how many would you predict would say that they pay attention to nutritional information? In this lesson, you will investigate this question by generating distributions of sample proportions and investigating patterns in these distributions.

Your teacher will give your group a container of dried beans. Some of the beans in the container are black. With your classmates, you are going see what happens when you take a sample of beans from the container and use the proportion of black beans in the sample to estimate the proportion of black beans in the container (a population proportion).

Exploratory Challenge 1/Exercises 1-9

1. Each person in the group should randomly select a sample of 20 beans from the container by carefully mixing all the beans and then selecting one bean and recording its color. Replace the bean, mix the bag, and continue to select one bean at a time until 20 beans have been selected. Be sure to replace each bean and mix the bag before selecting the next bean. Count the number of black beans in your sample of 20.

2. What is the proportion of black beans in your sample of 20? (Round your answer to 2 decimal places.) This value is called the sample proportion of black beans.

3. Write your sample proportion on a post-it note, and place the note on the number line that your teacher has drawn on the board. Place your note above the value on the number line that corresponds to your sample proportion.

   The graph of all the students' sample proportions is called a distribution of the sampling distribution of sample proportions. This sampling distribution is an approximation of the actual sampling distribution of all possible samples of size 20.

4. Describe the shape of the distribution.

5. What was the smallest sample proportion observed?

6. What was the largest sample proportion observed?

7. What sample proportion occurred most often?

8. Using technology, find the mean and standard deviation of the sample proportions used to construct the sampling distribution created by the class.

9. How does the mean of the sampling distribution compare with the population proportion of 0.40?

Example 2: Sampling Variability

What do you think would happen to the sampling distribution if everyone in class took a random sample of 40 beans from the container? To help answer this question, you will repeat the process described in Example 1, but this time you will draw a random sample of 40 beans instead of 20.

Exploratory Challenge 2/Exercises 10-21

10. Take a random sample with replacement of 40 beans from the container. Count the number of black beans in your sample of 40 beans.

11. What is the proportion of black beans in your sample of 40? (Round your answer to 2 decimal places.)

12. Write your sample proportion on a post-it note, and place it on the number line that your teacher has drawn on the board. Place your note above the value on the number line that corresponds to your sample proportion.

13. Describe the shape of the distribution.

14. What was the smallest sample proportion observed?

15. What was the largest sample proportion observed?

16. What sample proportion occurred most often?

17. Using technology, find the mean and standard deviation of the sample proportions used to construct the sampling distribution created by the class.

18. How does the mean of the sampling distribution compare with the population proportion of 0.40?

19. How does the mean of the sampling distribution based on random samples of size 20 compare to the mean of the sampling distribution based on random samples of size 40?

20. As the sample size increased from 20 to 40 describe what happened to the sampling variability (standard deviation of the distribution of sample proportions)?

21. What do you think would happen to the variability (standard deviation) of the distribution of sample proportions if the sample size for each sample were 80 instead of 40? Explain.

Lesson Summary

The sampling distribution of the sample proportion can be approximated by a graph of the sample proportions for many different random samples. The mean of the sampling distribution of the sample proportions will be approximately equal to the value of the population proportion.

As the sample size increases, the sampling variability in the sample proportion decreases – the standard deviation of the sampling distribution of the sample proportions decreases.

Problem Set

1. A class of 28 eleventh graders wanted to estimate the proportion of all juniors and seniors at their high school with part-time jobs after school. Each eleventh grader took a random sample of 30 juniors and seniors and then calculated the proportion with part-time jobs. Following are the 28 sample proportions.

   $0.7, 0.8, 0.57, 0.63, 0.7, 0.47, 0.67, 0.67, 0.8, 0.77, 0.4, 0.73, 0.63, 0.67, 0.6, 0.77, 0.77, 0.77, 0.53, 0.57, 0.73, 0.7, 0.67, 0.7, 0.77, 0.57, 0.77, 0.67$

   a. Construct a dot plot of the sample proportions.

   b. Describe the shape of the distribution.

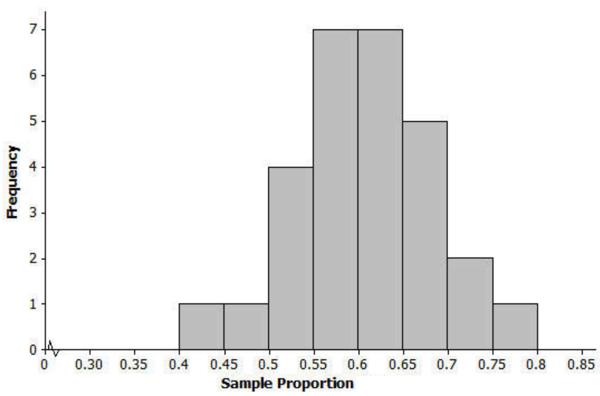
   c. Using technology, find the mean and standard deviation of the sample proportions.

   d. Do you think that the proportion of all juniors and seniors at the school with part-time jobs could be 0.7? Do you think it could be 0.5? Justify your answers based on your dot plot.

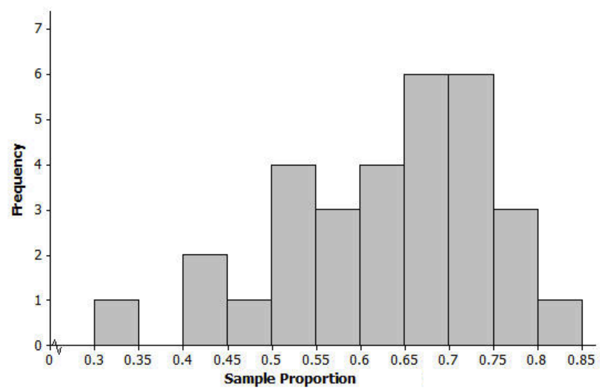
   e. Suppose the eleventh graders had taken random samples of size 60. How would the distribution of sample proportions based on samples of size 60 differ from the distribution for samples of size 30?

2. A group of eleventh graders wanted to estimate the proportion of all students at their high school who suffer from allergies. Each student in one group of eleventh graders took a random sample of 20 students, while another group of eleventh graders each took a random sample of 40 students. Below are the two sampling distributions (shown as histograms) of the sample proportions of high school students who said that they suffer from allergies. Which histogram is based on random samples of size 40? Explain.

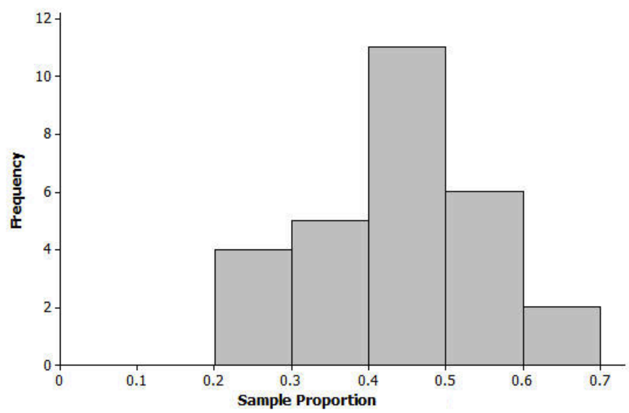
Histogram A



Histogram B



3. The nurse in your school district would like to study the proportion of all high school students in the district who usually get at least eight hours of sleep on school nights. Suppose each student in your class takes a random sample of 20 high school students in the district and each calculates their sample proportion of students who said that they usually get at least eight hours of sleep on school nights. Below is a histogram of the sampling distribution.



a. Do you think that the proportion of all high school students who usually get at least eight hours of sleep on school nights could have been 0.4? Do you think it could have been 0.55? Could it have been 0.75? Justify your answers based on the histogram.

b. Suppose students had taken random samples of size 60. How would the distribution of sample proportions based on samples of size 60 differ from those of size 20?

Lesson 15: Sampling Variability in the Sample Proportion

Classwork

Example 1

A high school principal claims that 50% of the school's students walk to school in the morning. A student attempts to verify the principal's claim by taking a random sample of 40 students and asking them if they walk to school in the morning. Sixteen of the sampled students say they usually walk to school in the morning, giving a sample proportion of $\frac{16}{40} = 0.40$, which seem to dispel the principal's claim of 50%. But could the principal be correct that the proportion of all students who walk to school is 50%?

a. Make a conjecture about the answer.

b. Develop a plan for how to respond.

Help the student make a decision on the principal's claim by investigating what kind of sample proportions you would expect to see if the principal's claim of $\mathbf{50}\%$ is true. You will do this by using technology to simulate the flipping of a coin $\mathbf{40}$ times.

Exploratory Challenge 1/Exercises 1-9

In Exercises 1–9, students should assume that the principal is correct that 50% of the population of students walk to school. Designate heads to represent a student who walks to school.

1. Simulate 40 flips of a fair coin. Record your observations in the space below.

2. What is the sample proportion of heads in your sample of 40? Report this value to your teacher.

3. Repeat Exercises 1 and 2 to obtain a second sample of 40 coin flips.

Your teacher will display a graph of all the students' sample proportions of heads.

4. Describe the shape of the distribution.

5. What was the smallest sample proportion observed?

6. What was the largest sample proportion observed?

7. Estimate the center of the distribution of sample proportions.

Your teacher will report the mean and standard deviation of the sampling distribution created by the class.

8. How does the mean of the sampling distribution compare with the population proportion of 0.50?

9. Recall that a student took a random sample of 40 students and found that the sample proportion of students who walk to school was 0.40. Would this have been a surprising result if the actual population proportion were 0.50 as the principal claims?

Example 2: Sampling Variability

What do you think would happen to the sampling distribution you constructed in the previous exercises had everyone in class taken a random sample of size 80 instead of 40? Justify your answer. This will be investigated in the following exercises.

Exploratory Challenge 2/Exercises 10-22

10. Use technology and simulate 80 coin flips. Calculate the proportion of heads. Record your results in the space below.

11. Repeat flipping a coin 80 times until you have recorded a total of 40 sample proportions.

.

12. Construct a dot plot of the 40 sample proportions.

13. Describe the shape of the distribution.

14. What was the smallest proportion of heads observed?

15. What was the largest proportion of heads observed?

16. Using technology, find the mean and standard deviation of the distribution of sample proportions.

17. Compare your results with the others in your group. Did you have similar means and standard deviations?

18. How does the mean of the sampling distribution based on 40 simulated flips of a coin (Exercise 1) compare to the mean of the sampling distribution based on 80 simulated coin flips?

19. Describe what happened to the sampling variability (standard deviation) of the distribution of sample proportions as the number of simulated coin flips increased from 40 to 80.

20. What do you think would happen to the variability (standard deviation) of the distribution of sample proportions if the sample size for each sample were 200 instead of 80? Explain.

21. Recall that a student took a random sample of 40 students and found that the sample proportion of students who walk to school was 0.40. If the student had taken a random sample of 80 students instead of 40, would this have been a surprising result if the actual population proportion was 0.50 as the principal claims?

22. What do you think would happen to the sampling distribution you constructed in the previous exercises if everyone in class took a random sample of size 80 instead of 40? Justify your answer.

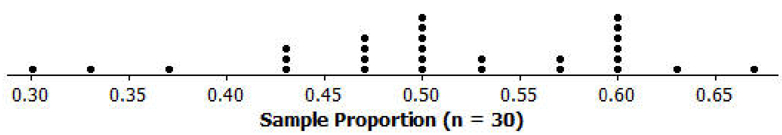
Lesson Summary

The sampling distribution of the sample proportion can be approximated by a graph of the sample proportions for many different random samples. The mean of the sample proportions will be approximately equal to the value of the population proportion.

As the sample size increases, the sampling variability in the sample proportion decreases – the standard deviation of the sample proportions decreases.

Problem Set

1. A student conducted a simulation of 30 coin flips. Below is a dot plot of the sampling distribution of the proportion of heads. This sampling distribution has a mean of 0.51 and a standard deviation of 0.09.

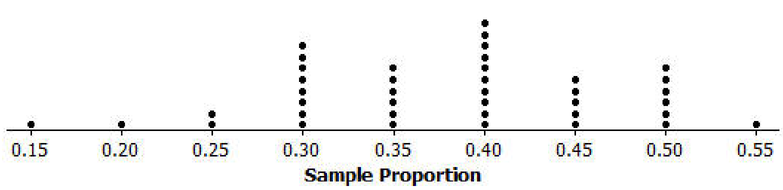


   a. Describe the shape of the distribution.
   b. Describe what would have happened to the mean and the standard deviation of the sampling distribution of the sample proportions if the student had flipped a coin 50 times, calculated the proportion of heads, and then repeated this process for a total of 30 times.

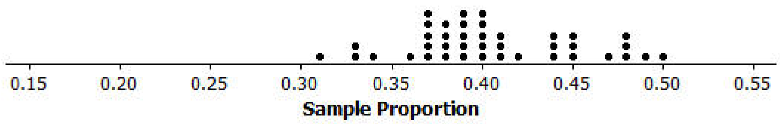
2. What effect does increasing the sample size have on the mean of the sampling distribution?

3. What effect does increasing the sample size have on the standard deviation of the sampling distribution?

4. A student wanted to decide whether or not a particular coin was fair (i.e., the probability of flipping a head is 0.5). She flipped the coin 20 times, calculated the proportion of heads, and repeated this process a total of 40 times. Below is the sampling distribution of sample proportions of heads. The mean and standard deviation of the sampling distribution is 0.379 and 0.091. Do you think this was a fair coin? Why or why not?



5. The same student flipped the coin 100 times, calculated the proportion of heads, and repeated this process a total of 40 times. Below is the sampling distribution of sample proportions of heads. The mean and standard deviation of the sampling distribution is 0.405 and 0.046. Do you think this was a fair coin? Why or why not?



Lesson 16: Margin of Error when Estimating a Population Proportion

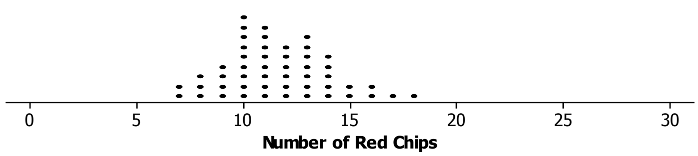
Classwork

Exercises 1-4

In this lesson you will use data from a random sample drawn from a mystery bag to estimate a population proportion and learn how to find and interpret a margin of error for your estimate.

1. Write down your estimate for the proportion of red chips in the mystery bag based on the random sample of 30 chips drawn in class.

2. Tanya and Raoul had a paper bag that contained red and black chips. The bag was marked 40% red chips. They drew random samples of 30 chips, with replacement, from the bag. (They were careful to shake the bag after they replaced a chip.) They had nine red chips in their sample. They drew another random sample of 30 chips from the bag, and this time they had 12 red chips. They repeated this sampling process 50 times and made a plot of the number of red chips in each sample. A plot of their sampling distribution is shown below.



   a. What was the most common number of red chips in the 50 samples? Does this seem reasonable? Why or why not?

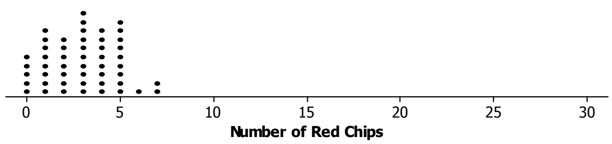
b. What number of red chips, if any, never occurred in any of the samples?

c. Give an interval that contains the "likely" number of red chips in samples of size 30 based on the simulated sampling distribution.

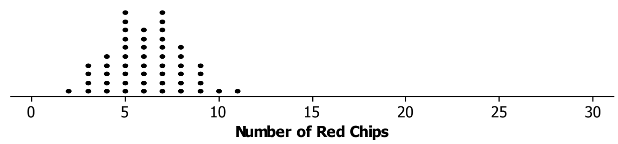
d. Do you think the number of red chips in the mystery bag could have come from a sample drawn from a bag that had 40% red chips? Why or why not?

Nine different bags of chips were distributed to small teams of students in the class. Each bag had a different proportion of red chips. Each team simulated drawing 50 different random samples of size 30 from their bag and recorded the number of red chips for each sample. The graphs of their simulated sampling distributions are shown below.

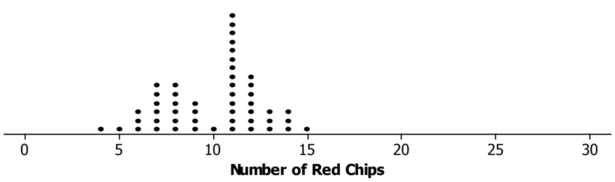
Population 10% red chips



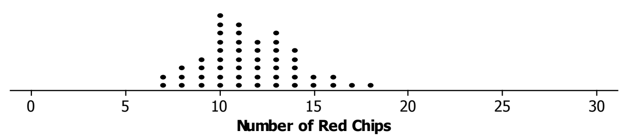
Population 20% red chips



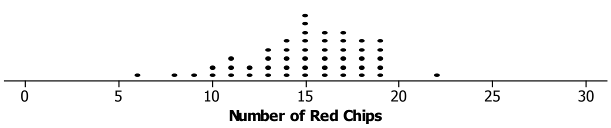
Population 30% red chips



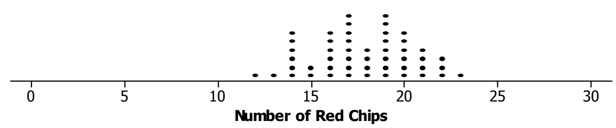
Population 40% red chips



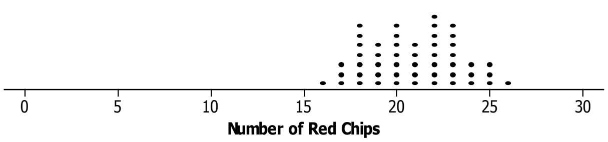
Population 50% red chips



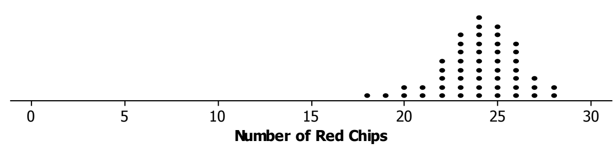
Population 60% red chips



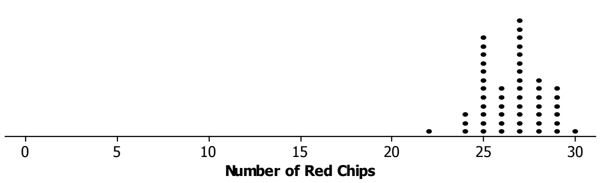
Population 70% red chips



Population 80% red chips



Population 90% red chips



3. Think about the number of red chips in the random sample of size 30 that was drawn from the mystery bag.
   a. Based on the simulated sampling distributions, do you think that the mystery bag might have had 10% red chips? Explain your reasoning.

   b. Based on the simulated sampling distributions, which of the percentages 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% might reasonably be the percentage of red chips in the mystery bag?

   c. Let p represent the proportion of red chips in the mystery bag. (For example, $p = 0.40$ if there are 40% red chips in the bag.) Based on your answer to part (b), write an inequality that describes plausible values for p. Interpret the inequality in terms of the mystery bag population.

4. If the inequality like the one you described in part (c) of Exercise 3 went from 0.30 to 0.60, it is sometimes written as 0.45 ± 0.15. The value 0.15 is called a "margin of error." The margin of error represents an interval from the expected proportion that would not contain any proportions or very few proportions based on the simulated sampling distribution. Proportions in this interval are not expected to occur when taking a sample from the mystery bag.
   a. Write the inequality you found in Exercise 3 part (c) using this notation. What is the margin of error?

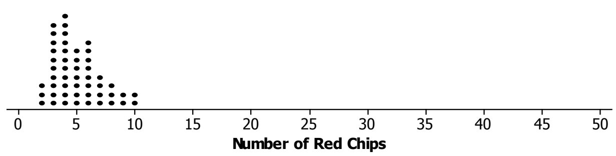
   b. Suppose Sol said, "So this means that the actual proportion of red chips in the mystery bag was 60%." Tonya argued that the actual proportion of red chips in the mystery bag was 20%. What would you say?

Exploratory Challenge 2: Samples of Size 50/Exercises 5–7

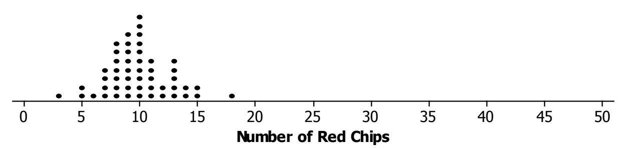
5. Do you think the "margin of error" would be different in Exercise 4 if you had sampled 50 chips instead of 30? Try to convince a partner that your conjecture is correct.

6. Below are simulated sampling distributions of the number of red chips for samples of size 50 from populations with various percentages of red chips.

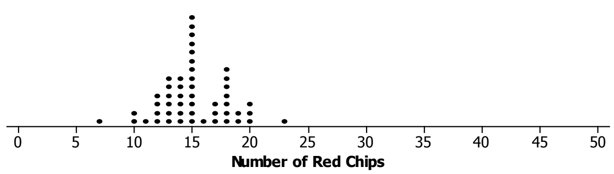
Population with 10% red chips



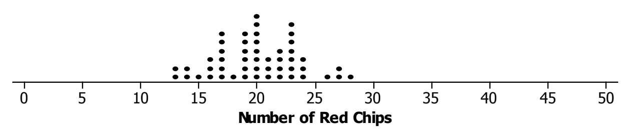
Population with 20% red chips



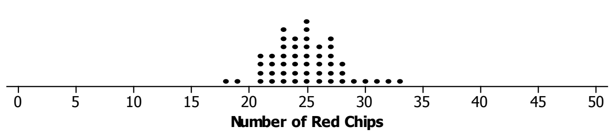
Population with 30% red chips



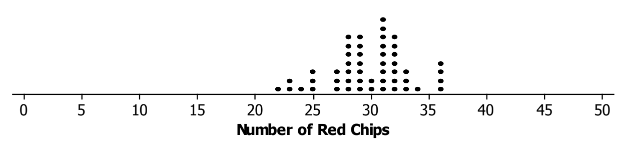
Population with 40% red chips



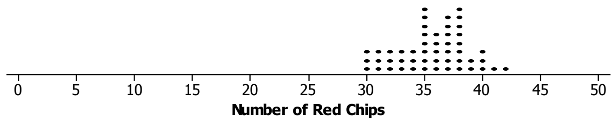
Population with 50% red chips



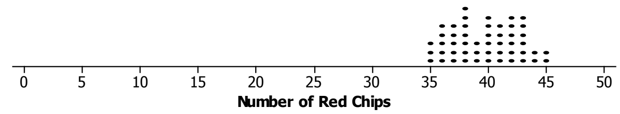
Population with 60% red chips



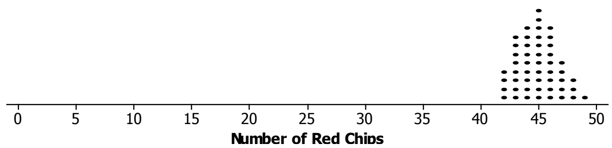
Population with 70% red chips



Population with 80% red chips



Population with 90% red chips



a. Suppose you drew 30 red chips in a random sample of 50 from the mystery bag. What are plausible values for the proportion of red chips in the mystery bag? Explain your reasoning.

b. Write an expression that contains the margin of error based on your answer to part (a).

7. Remember your conjecture from Exercise 5, and compare the margin of error you found for a sample of size 30 (from Exercise 3) to the margin of error you found for a sample of size 50.

   a. Was your reasoning in Exercise 5 correct? Why or why not?

   b. Explain why the change in the margin of error makes sense.

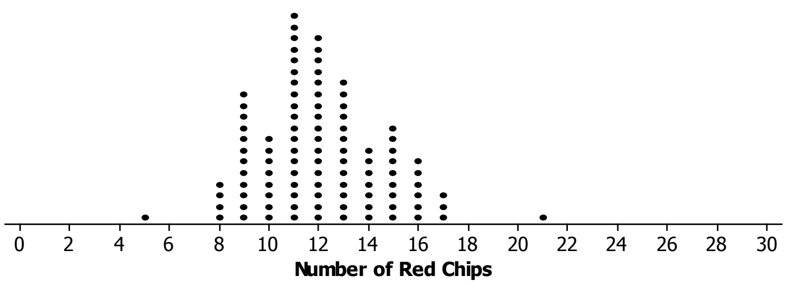
Lesson Summary

In this lesson, you investigated how to make an inference about an unknown population proportion based on a random sample from that population.

- You learned how random samples from populations with known proportions of successes behave by simulating sampling distributions for samples drawn from those populations.
- Comparing an observed proportion of successes from a random sample drawn from a population with an unknown proportion of successes to these sampling distributions gives you some information about what populations might produce a random sample like the one you observed.
- These plausible population proportions can be described as $p \pm M$. The value of M is called a margin of error.

Problem Set

1. Tanya simulated drawing a sample of size 30 from a population of chips and got the following simulated sampling distribution for the number of red chips:



Which of the following results seem like they might have come from this population? Explain your reasoning.

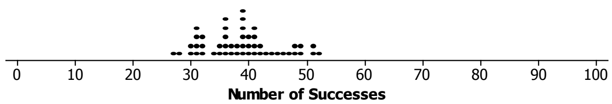
I. 8 red chips in a random sample of size 30.

II. 12 red chips in a random sample of size 30.

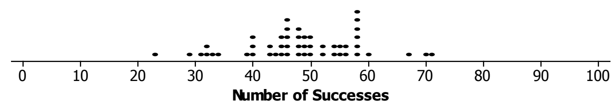
III. 24 red chips in a random sample of size 30.

2. 64% percent of the students in a random sample of 100 high school students intended to go onto college. The graphs below show the result of simulating random samples of size 100 from several different populations where the success percentage was known and recording the percentage of successes in the sample.

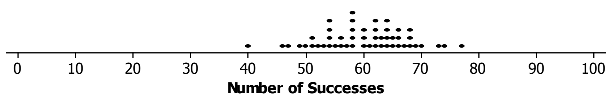
(i) Population with $\mathbf{40}\%$ successes



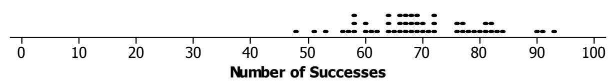
(ii) Population with $\mathbf{50}\%$ successes



(iii) Population with $\mathbf{60}\%$ successes



(iv) Population with $\mathbf{70}\%$ successes



a. Based on these graphs, which of the following are plausible values for the percentage of successes in the population from which the sample was selected: 40%, 50%, 60%, or 70%? Explain your thinking.

b. Would you need more information to determine plausible values for the actual proportion of the population of high school students who intend to go to some postsecondary school? Why or why not?

3. Suppose the mystery bag had resulted in the following number of red chips. Using the simulated sampling distributions found earlier in this lesson, find a margin of error in each case.

a. The number of red chips in a random sample of size 30 was 10.

b. The number of red chips in a random sample of size 30 was 21.

c. The number of red chips in a random sample of size 50 was 22.

4. The following intervals were plausible population proportions for a given sample. Find the margin of error in each case.

a. from 0.35 to 0.65

b. from 0.72 to 0.78

c. from 0.84 to 0.95

d. from 0.47 to 0.57

5. Decide if each of the following statements is true or false. Explain your reasoning in each case.
   a. The smaller the sample size, the smaller the margin of error.
   b. If the margin of error is 0.05 and the observed proportion of red chips is 0.35, then the true population proportion is likely to be between 0.40 and 0.50.

6. Extension: The margin of error for a sample of size 30 is 0.20; for a sample of 50, is 0.10. If you increase the sample size to 70, do you think the margin of error for the percent of successes will be 0.05? Why or why not?

Lesson 17: Margin of Error when Estimating a Population Proportion

Classwork

In this lesson, you will find and interpret the standard deviation of a simulated distribution for a sample proportion and use this information to calculate a margin of error for estimating the population proportion

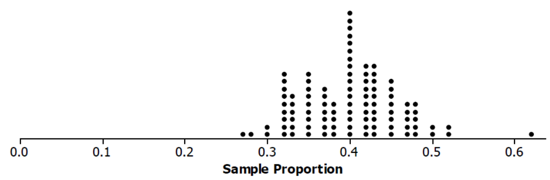
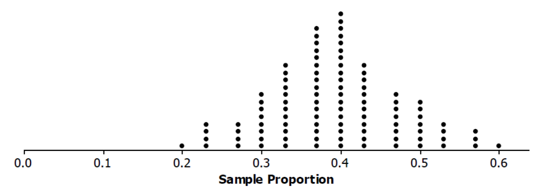
Exercises 1-6: Standard Deviations for Proportions

In the previous lesson, you used simulated sampling distributions to learn about sampling variability in the sample proportion and the margin of error when using a random sample to estimate a population proportion. However, finding a margin of error using simulation can be cumbersome and take a long time for each situation. Fortunately, given the consistent behavior of the sampling distribution of the sample proportion for random samples, statisticians have developed a formula that will allow you to find the margin of error quickly and without simulation.

1. 30% of students participating in sports at Union High School are female (a proportion of 0.30).
   a. If you took many random samples of 50 students play sports and made a dot plot of the proportion of females in each sample, where do you think this distribution will be centered? Explain your thinking.

   b. In general, for any sample size, where do you think the center of a simulated distribution of the sample proportion of females in sports at Union High School will be?

2. Below are two simulated sampling distributions for the sample proportion of females in random samples from all the students at Union High School.

Simulated Sampling Distribution 2	Simulated Sampling Distribution 2
	

a. Based on the two sampling distributions above, what do you think is the population proportion of females?

b. One of the sampling distributions above is based on random samples of size 30, and the other is based on random samples of size 60. Which sampling distribution corresponds to the sample size of 30? Explain your choice.

3. Remember from your earlier work in statistics that distributions were described using shape, center, and spread. How was spread measured?

4. In previous lessons, you saw a formula for the standard deviation of the sampling distribution of the sample mean. There is also a formula for the standard deviation of the sampling distribution of the sample proportion. For random samples of size n, the standard deviation can be calculated using the following formula:

   standard deviation $= \sqrt{\frac{p(1-p)}{n}}$, where p is the value of the population proportion and n is the sample size.

   a. If the proportion of females at Union High School is 0.4, what is the standard deviation of the distribution of the sample proportions of females for random samples of size 50? Round your answer to three decimal places.

   b. The proportion of males at Union High School is 0.6. What is the standard deviation of the distribution of the sample proportions of males for samples of size 50? Round your answer to three decimal places.

   c. Think about the graphs of the two distributions in parts (a) and (b). Explain the relationship between your answers using the center and spread of the distributions.

5. Think about the simulations that your class performed in the previous lesson and the simulations in Exercise 2 above.

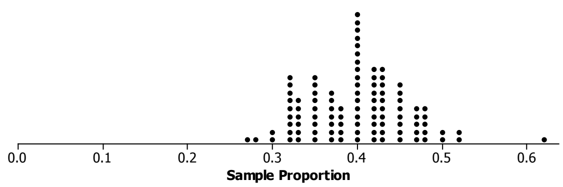
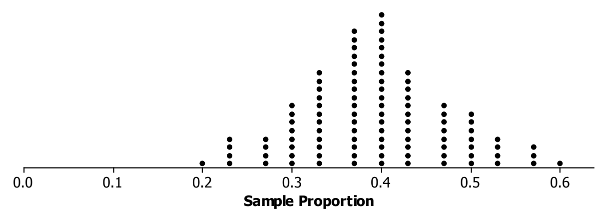
   a. Was the sampling variability in the sample proportion greater for samples of size 30 or for samples of size 50? In other words, does the sample proportion tend to vary more from one random sample to another when the sample size is 30 or 50?

   b. Explain how the observation that the variability in the sample proportions decreases as the sample size increases is supported by the formula for the standard deviation of the sample proportion.

6. Consider the two simulated sampling distributions of the proportion of females in Exercise 2 where the population proportion was 0.4. Recall that you found $n = 60$ for Distribution 1, and $n = 30$ for Distribution 2.

   a. Find the standard deviation for each distribution. Round your answer to three decimal places.

   b. Make a sketch and mark off the intervals one standard deviation from the mean for each of the two distributions. Interpret the intervals in terms of the proportion of females in a sample.

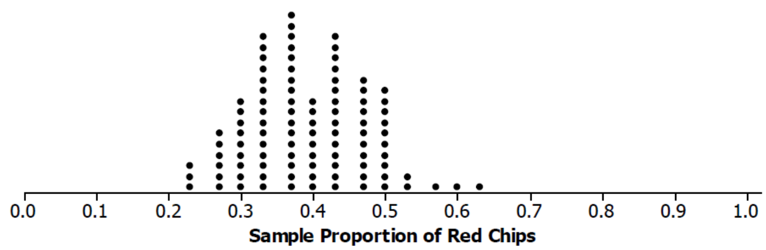
Simulated Sampling Distribution 1: $n = 60$	Simulated Sampling Distribution 2: $n = 30$
	

In general, three results about the sampling distribution of the sample proportion are known:

- The sampling distribution of the sample proportion is centered at the actual value of the population proportion, p.
- The sampling distribution of the sample proportion is less variable for larger samples than for smaller samples. The variability in the sampling distribution is described by the standard deviation of the distribution, and the standard deviation of the sampling distribution for random samples of size n is $\sqrt{\frac{p(1-p)}{n}}$, where p is the value of the population proportion. This standard deviation is usually estimated using the sample proportion, which is denoted by $\hat{p}$ (read as p-hat), to distinguish it from the population proportion. The formula for the estimated standard deviation of the distribution of sample proportions is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- As long as the sample size is large enough that the sample includes at least 10 successes and failures, the sampling distribution is approximately normal in shape. That is, a normal distribution would be a reasonable model for the sampling distribution.

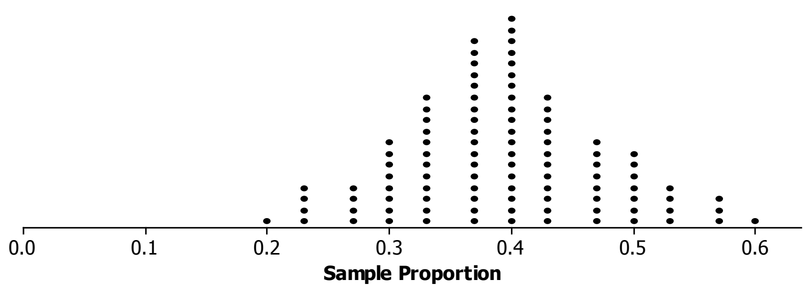
Exercises 7–12: Using the Standard Deviation with Margin of Error

7. In the work above, you investigated a simulated sampling distribution of the proportion of females in a sample of size 30 drawn from a population with a known proportion of 0.4 females. The simulated distribution of the proportion of red chips in a sample of size 30 drawn from a population with a known proportion of 0.4 is displayed below.

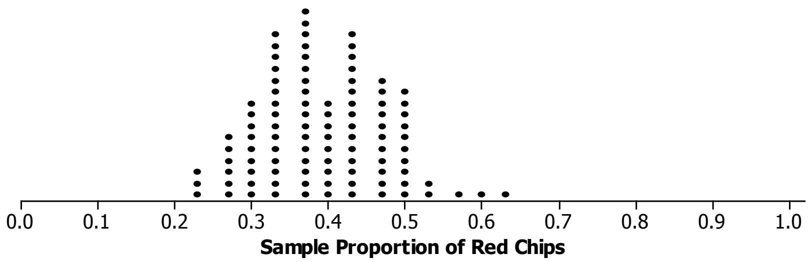


a. Use the formula for the standard deviation of the sample proportion to calculate the standard deviation of the sampling distribution. Round your answer to three decimal places.

b. The distribution from Exercise 2 for a sample of size 30 is below. How do the two distributions compare?



c. How many of the values of the sample proportions are within one standard deviation of 0.4? How many are within two standard deviations of 0.4?



In general, for a known population proportion, about 95% of the outcomes of a simulated sampling distribution of a sample proportion will fall within two standard deviations of the population proportion. One caution is that if the proportion is close to 1 or 0, this general rule may not hold unless the sample size is very large. You can build from this to estimate a proportion of successes for an unknown population proportion and calculate a margin of error without having to carry out a simulation.

If the sample is large enough to have at least 10 of each of the two possible outcomes in the sample, but small enough to be no more than 10% of the population, the following formula (based on an observed sample proportion $\hat{p}$) can be used to calculate the margin of error. The standard deviation involves the parameter p that is being estimated. Because p is often not known, statisticians replace p with its estimate $\hat{p}$ in the standard deviation formula. This estimated standard deviation is called the standard error of the sample proportion.

8.

a. Suppose you draw a random sample of 36 chips from a mystery bag and find 20 red chips. Find $\hat{p}$, the sample proportion of red chips, and the standard error.

b. Interpret the standard error.

When estimating a population proportion, **margin of error** can be defined as the **maximum expected difference** between the value of the population proportion and a sample estimate of that proportion (the farthest away from the actual population value that you think your estimate is likely to be).

If $\hat{p}$ is the sample proportion for a random sample of size n from some population, and if the sample size is large enough:

$$\text{estimated margin of error} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

9. Henri and Terence drew samples of size 50 from a mystery bag. Henri drew 42 red chips, and Terence drew 40 red chips. Find the margins of error for each student.

10. Divide the problems below among your group, and find the sample proportion of successes and the estimated margin of error in each situation:
    a. Sample of size 20, 5 red chips
    b. Sample of size 40, 10 red chips
    c. Sample of size 80, 20 red chips
    d. Sample of size 100, 25 red chips

11. Look at your answers to Exercise 2.
    a. What conjecture can you make about the relation between sample size and margin of error? Explain why your conjecture makes sense.

    b. Think about the formula for a margin of error. How does this support or refute your conjecture?

12. Suppose that a random sample of size 100 will be used to estimate a population proportion.
    a. Would the estimated margin of error be greater if $\hat{p} = 0.4$ or $\hat{p} = 0.5$? Support your answer with appropriate calculations.

    b. Would the estimated margin of error be greater if $\hat{p} = 0.5$ or $\hat{p} = 0.8$? Support your answer with appropriate calculations.

    c. For what value of $\hat{p}$ do you think the estimated margin of error will be greatest? (Hint: Draw a graph of $\hat{p}(1 - \hat{p})$ as $\hat{p}$ ranges from 0 to 1.)

Lesson Summary

- Because random samples behave in a consistent way, a large enough sample size allows you to find a formula for the standard deviation of the sampling distribution of a sample proportion. This can be used to calculate the margin of error: $M = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p}$ is the proportion of successes in a random sample of size n.
- The sample size is large enough to use this result for estimated margin of error if there are at least 10 of each of the two outcomes.
- The sample size should not exceed 10% of the population.
- As the sample size increases, the margin of error decreases.

Problem Set

1. Different students drew random samples of size 50 from the mystery bag. The number of red chips each drew is given below. In each case, find the margin of error for the proportions of the red chips in the mystery bag.
   a. 10 red chips
   b. 28 red chips
   c. 40 red chips

2. The school newspaper at a large high school reported that 120 out of 200 randomly selected students favor assigned parking spaces. Compute the margin of error. Interpret the resulting interval in context.

3. A newspaper in a large city asked 500 women the following: "Do you use organic food products (such as milk, meats, vegetables, etc.)?" 280 women answered "yes." Compute the margin of error. Interpret the resulting interval in context.

4. The results of testing a new drug on 1000 people with a certain disease found that 510 of them improved when they used the drug. Assume these 1000 people can be regarded as a random sample from the population of all people with this disease. Based on these results, would it be reasonable to think that more than half of the people with this disease would improve if they used the new drug? Why or why not?

5. A newspaper in New York took a random sample of 500 registered voters from New York City and found that 300 favored a certain candidate for governor of the state. A second newspaper polled 1000 registered voters in upstate New York and found that 550 people favored this candidate. Explain how you would interpret the results.

6. In a random sample of $1{,}500$ students in a large suburban school, $1{,}125$ reported having a pet, resulting in the interval 0.75 ± 0.022. While in a large urban school, 840 out of $1{,}200$ students reported having a pet, resulting in the interval 0.7 ± 0.026. Because these two intervals do not overlap, there appears to be a difference in the proportion of suburban students owning a pet and the proportion of urban students owning a pet. Suppose the sample size of the suburban school was only 500 but 75% still reported having a pet. Also, suppose the sample size of the urban school was 600 and 70% still reported having a pet. Is there still a difference in the proportion of students owning a pet in suburban schools and urban schools? Why does this occur?

7. Find an article in the media that uses a margin of error. Describe the situation (an experiment, an observational study), and interpret the margin of error for the context.

Lesson 18: Sampling Variability in the Sample Mean

Classwork

Exercises 1-7: Random Segments

The worksheet contains 100 segments of different lengths. The length of a segment is the number of rectangles spanned on the grid. For example, segment 2 has length 5.

1. Briefly review the sheet and estimate of the mean length of the segments. Will your estimate be close to the actual mean? Why or why not?

2. Look at the sheet. With which of the statements below would you agree? Explain your reasoning.

   The mean length of the segments is:

   a. close to 1

   b. close to 8

   c. around 5

   d. between 2 and 5

3. Follow your teacher's directions to select ten random numbers between 1 and 100. For each random number, start at the upper left cell with a segment value of 2, and count down and to the right the number of cells based on the random number selected. The number in the cell represents the length of a randomly selected segment.

   a. On a number line, graph the lengths of the corresponding segments on the worksheet.

b. Find the mean and standard deviation of the lengths of the segments in your sample. Mark the mean length on your graph from part (a).

4. Your sample provides some information about the mean length of the segments in one random sample of size 10, but that sample is only one among all the different possible random samples. Let's look at other random samples and see how the means from those samples compare to the mean segment length from your random sample.

   Record the mean segment length for your random sample on a post-it note, and post the note in the appropriate place on the number line your teacher set up.

   a. Jonah looked at the plot and said, "Wow, our means really varied." What do you think he meant?

   b. Describe the simulated sampling distribution of mean segment lengths for samples of size 10.

   c. How did your first estimate (from Exercise 1) compare to your sample mean from the random sample? How did it compare to the means in the simulated distribution of the sample means from the class?

5. Collect the values of the sample means from the class.

   a. Find the mean and standard deviation of the simulated distribution of the sample means.

   b. Interpret the standard deviation of the simulated sampling distribution in terms of the length of the segments.

   c. What do you observe about the values of the means in the simulated sampling distribution that are within two standard deviations from the mean of the sampling distribution?

6. Generate another set of ten random numbers, find the corresponding lengths on the sheet, and calculate the mean length for your sample. Put a post-it note with your sample mean on the second number line. Then answer the following questions:
   a. Find the mean and standard deviation of the simulated distribution of the sample means.

   b. Interpret the standard deviation of the simulated sampling distribution in terms of the length of the segments.

   c. What do you observe about the values of the means in the simulated sampling distribution that are within two standard deviations from the mean of the sampling distribution?

7. Suppose that we know the actual mean of all the segment lengths is 2.78 units.
   a. Describe how the population mean relates to the two simulated distributions of sample means.

   b. Tonya was concerned that neither of the simulated distributions of sample means had a value around 5, but some of the segments on the worksheet were 5 units long and some were as big as 8 units long. What would you say to Tonya?

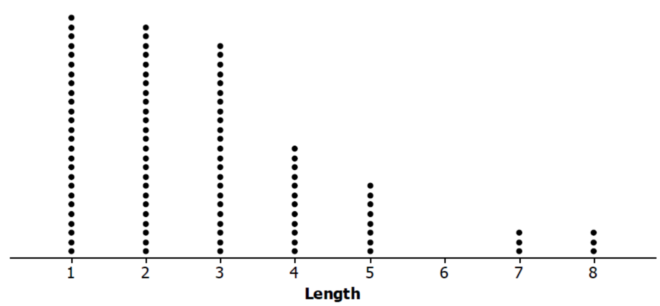
Lesson Summary

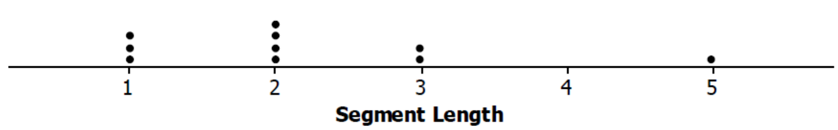
In this lesson you drew a sample from a population and found the mean of that sample.

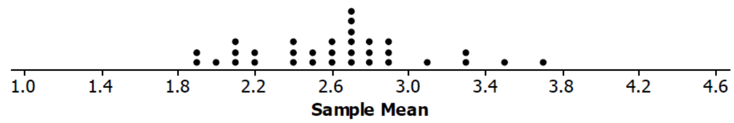
- Drawing many samples of the same size from the same population and finding the mean of each of those samples allows you to build a simulated sampling distribution of the sample means for the samples you generated.
- The mean of the simulated sampling distribution of sample means is close to the population mean.
- In the two examples of simulated distributions of sample means we generated, most of the sample means seemed to fall within two standard deviations of the mean of the simulated distribution of sample means.

Problem Set

1. The three distributions below relate to the population of all of the random segment lengths and to samples drawn from that population. The eight phrases below could be used to describe a whole graph or a value on the graph. Identify where on the appropriate graph the phrases could be placed. (For example, segment of length 2 could be placed by any of the values in the column for 2 on the plot labeled "Length.")







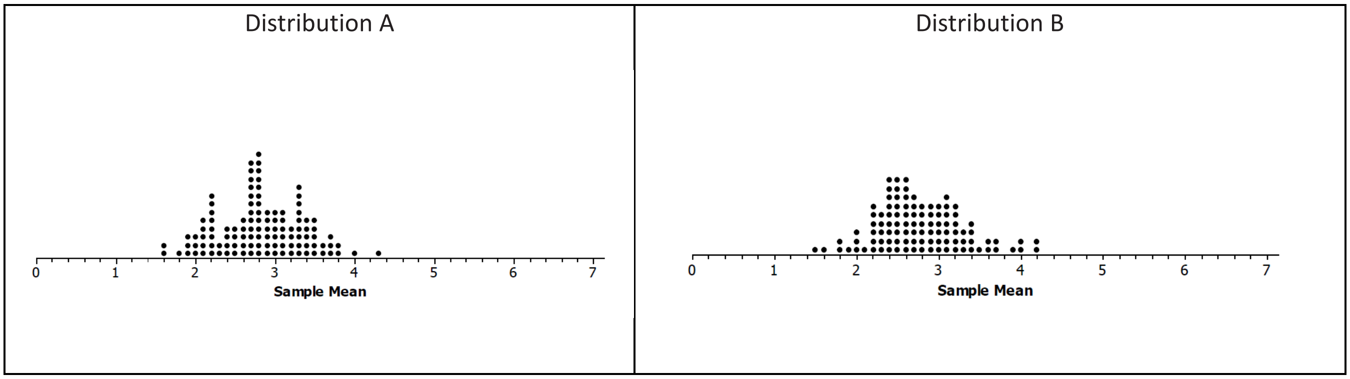
a. Random sample of size 10 of segment lengths
b. Segment of length 2
c. Sample mean segment length of 2
d. Mean of sampling distribution, 2.6
e. Simulated distribution of sample means
f. Sample segment lengths
g. Population of segment lengths
h. Mean of all segment lengths, 2.78

2. The following segment lengths were selected in four different random samples of size 10.

Lengths Sample A	Lengths Sample B	Lengths Sample C	Lengths Sample D
1	1	1	2
2	3	5	2
1	1	1	7
5	2	3	2
3	1	4	5
1	5	2	2
2	3	2	3
2	4	4	5
3	3	3	5
1	3	4	4

a. Find the mean segment length of each sample.
b. Find the mean and standard deviation of the four sample means.
c. Interpret your answer to part (b) in terms of the variability in the sampling process.

3. Two simulated sampling distributions of the mean segment lengths from random samples of size 10 are displayed below.



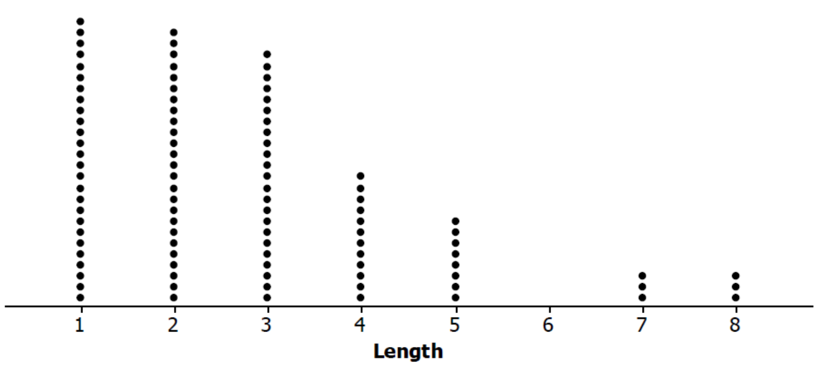
a. Compare the two distributions with respect to shape, center, and spread.

b. Distribution A has a mean of 2.82, and Distribution B has a mean of 2.77. How do these means compare to the population mean of 2.78?

c. Both Distribution A and Distribution B have a standard deviation 0.54. Make a statement about the distribution of sample means that makes use of this standard deviation.

4. The population distribution of all the segment lengths is shown in the dot plot below. How does the population distribution compare to the two simulated sampling distributions of the sample means in Problem 3?

Distribution of Lengths of 100 Segments



Lesson 19: Sampling Variability in the Sample Mean

Classwork

This lesson uses simulation to approximate the sampling distribution of the sample mean for random samples from a population, explores how the simulated sampling distribution provides insight into the anticipated estimation error when using a sample mean to estimate a population mean, and covers how sample size affects the distribution of the sample mean.

Exercises 1–6: SAT scores

1. SAT test scores vary a lot. The table displays the 506 scores for students in one New York school district for a given year.

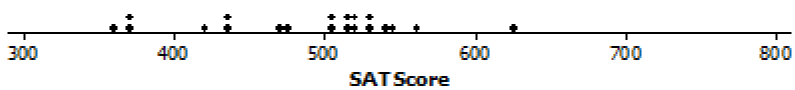
Table 1: SAT scores for district students

441	395	369	350	521	691	648	521	498	413	486
440	415	481	392	800	448	603	503	486	476	500
391	359	447	550	432	158	379	394	495	442	507
395	504	399	424	456	729	356	392	514	388	518
445	436	386	493	467	493	440	387	512	431	467
499	412	457	389	323	319	550	450	517	405	506
486	519	369	373	348	532	496	488	504	444	
396	473	319	367	679	472	613	561	522	408	
451	427	369	560	602	520	567	495	473	424	
362	391	371	407	436	366	582	528	533	463	
328	613	357	438	436	713	603	525	553	446	
414	466	382	362	777	259	557	508	495	466	
409	486	627	589	749	410	639	516	520	632	
526	334	608	374	634	443	556	506	506	526	
391	497	378	358	566	442	496	568	544	546	
529	392	387	373	198	555	499	476	525	529	
529	426	470	378	345	431	613	490	548	455	
574	379	380	561	712	197	556	547	543	431	
363	382	370	379	504	254	596	489	474	386	
486	434	365	530	685	372	580	506	529	434	
418	722	674	504	645	501	605	511	566	362	
527	437	388	525	509	662	445	489	487	426	
441	395	377	561	448	503	602	523	510	404	
467	463	427	519	491	448	638	530	518	493	
387	433	446	525	352	662	570	507	515	515	
503	371	394	569	779	158	558	504	516	407	
350	392	368	484	689	691	535	522	505	409	
583	416	406	416	513	729	623	503	536	422	
370	370	350	446	624	493	465	524	547	612	
499	422	344	420	465	319	460	523	528	486	
399	532	347	446	504	532	375	524	527	394	
374	545	377	462	390	472	540	501	523	424	
372	427	391	528	576	520	564	482	540	393	
559	371	339	533	756	366	547	502	480	420	
330	390	404	543	451	713	568	503	516	415	
567	529	377	460	505	259	588	439	501	394	
371	341	469	391	540	410	502	474	452	473	
503	356	417	623	436	443	510	477	507	531	
327	351	356	587	298	442	589	458	486	469	
528	377	370	528	449	555	537	494	500	453	
447	404	355	356	352	431	410	447	507	442	
572	369	364	523	574	197	330	517	518	509	
379	396	383	404	518	460	500	457	467	435	
456	396	400	505	682	623	531	471	506	427	
406	535	404	512	474	587	509	541	509	489	
420	388	375	514	629	528	571	513	597	480	
395	370	398	516	656	523	527	441	509	516	
355	417	376	498	539	505	457	489	567	501	
423	419	451	460	553	514	552	498	509	452	
438	348	369	541	400	629	561	538	597	507	

a. Looking at the table above, how would you describe the population of SAT scores?

b. Jason used technology to draw a random sample of size 20 from all of the scores and found a sample mean of 487. What does this value represent in terms of the graph below?

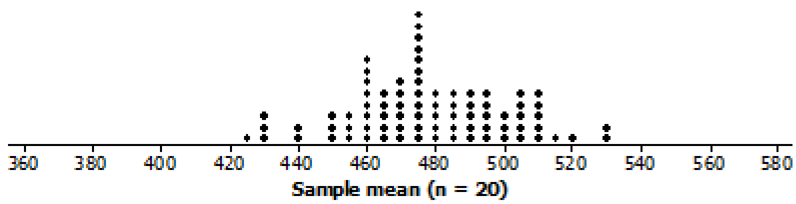
Random sample from District SAT scores



2. If you were to take many different random samples of 20 from this population, describe what you think the sampling distribution of these sample means would look like.

3. Everyone in Jason's class drew several random samples of size 20 and found the mean SAT score. The plot below displays the distribution of the mean SAT scores for their samples.

Random sample from District SAT scores



a. How does the simulated sampling distribution compare to your conjecture in Exercise 2? Explain any differences.

b. Use technology to generate many more samples of size 20, and plot the means of those samples. Describe the shape of the simulated distribution of sample mean SAT scores.

c. How did the simulated distribution using more samples compare to the one you generated in Exercise 3?

d. What are the mean and standard deviation of the simulated distribution of the sample mean SAT scores you found in part (b)? (Use technology and your simulated distribution of the sample means to find the values.)

e. Write a sentence describing the distribution of sample means that uses the mean and standard deviation you calculated in part (d).

4. Reflect on some of the simulated sampling distributions you have considered in previous lessons.
   a. Make a conjecture about how you think the size of the sample might affect the distribution of the sample SAT means.

   b. To test the conjecture, investigate the following sample sizes: $5, 10, 40$, and 50 as well as the simulated distribution of sample means from Exercise 3. Divide the sample sizes among your group members, and use technology to simulate sampling distributions of mean SAT scores for samples of the different sizes. Find the mean and standard deviation of each simulated sampling distribution.

   c. How does the sample size seem to affect the simulated distributions of the sample SAT mean scores? Include the simulated distribution from part (b) of Exercise 3 in your response. Why do you think this is true?

5.
   a. For each of the sample sizes, consider how the standard deviation seems to be related to the range of the sample means in the simulated distributions of the sample SAT means you found in Exercise 4.

   b. How do your answers to part (a) compare to the answers from other groups?

6.

a. Make a graph of the distribution of the population consisting of the SAT scores for all of the students.

b. Find the mean of the distribution of SAT scores. How does it compare to the mean of the sampling distributions you have been simulating?

Lesson Summary

For a given sample you can find the sample mean.

- There is variability in the sample mean. The value of the sample mean varies from one random sample to another.
- A graph of the distribution of sample means from many different random samples is a simulated sampling distribution.
- Sample means from random samples tend to cluster around the value of the population mean. That is, the simulated sampling distribution of the sample mean will be centered close to the value of the population mean.
- The variability in the sample mean decreases as the sample size increases.
- Most sample means are within two standard deviations of the mean of the simulated sampling distribution.

Problem Set

1. Which of the following will have the smallest standard deviation? Explain your reasoning.

   A sampling distribution of sample means for samples of size:

   a. 15 b. 25 c. 100

2. In light of the distributions of sample means you have investigated in the lesson, comment on the statements below for random samples of size 20 chosen from the District SAT scores.
   a. Josh claimed he took a random sample of size 20 and had a sample mean score of 320.
   b. Sarfina stated she took a random sample of size 20 and had a sample mean of 520.
   c. Ana announced that it would be pretty rare for the mean SAT score in a random sample to be more than three standard deviations from the mean SAT score of 475.

3. Refer to your answers for Exercise 4, and then comment on each of the following:
   a. A random sample of size 50 produced a mean SAT score of 400.
   b. A random sample of size 10 produced a mean SAT score of 400.
   c. For what sample sizes was a sample mean SAT score of 420 plausible? Explain your thinking.

4. Explain the difference between the sample mean and the mean of the sampling distribution.

Lesson 20: Margin of Error when Estimating a Population Mean

Classwork

Example 1: Describing a Population of Numerical Data

The course project in a computer science class was to create 100 computer games of various levels of difficulty that had ratings on a scale from 1 (easy) to 20 (difficult). We will examine a representation of the data resulting from this project. Working in pairs, your teacher will give you a page that contains 100 rectangles of various sizes.

a. What do you think the rectangles represent in the context of the 100 computer games?

b. What do you think the sizes of the rectangles represent in the context of the 100 computer games?

c. Why do you think the rectangles are numbered from 00 to 99 instead of from 1 to 100?

Exploratory Challenge 1/Exercises 1–3: Estimate the Population Mean Rating

1. Working with your partner, discuss how you would calculate the mean rating of all 100 computer games (the population mean).

2. Discuss how you might select a random sample to estimate the population mean rating of all 100 computer games.

3. Calculate an estimate of the population mean rating of all 100 computer games based on a random sample of size 10. Your estimate is called a sample mean, and it is denoted by $\bar{x}$. Use the following random numbers to select your sample.

$$34\ \ 86\ \ 80\ \ 58\ \ 04\ \ 43\ \ 96\ \ 29\ \ 44\ \ 51$$

Exploratory Challenge 2/Exercises 4–6: Build a Distribution of Sample Means

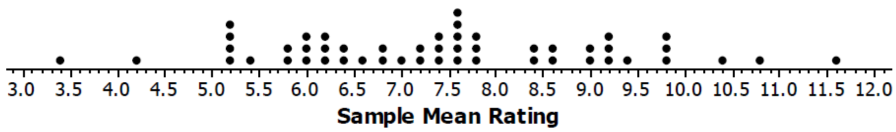
4. Work in pairs. Using a table of random digits or a calculator with a random-number generator, generate four sets of ten random numbers. Use these sets of random numbers to identify four random samples of size 10. Calculate the sample mean rating for each of your four random samples.

5. Write your sample means on separate sticky notes and post them on a number line that your teacher has prepared for your class.

6. The actual population mean rating of all 100 computer games is 7.5. Does your class distribution of sample means center at 7.5? Discuss why it does. Or, if it doesn't, discuss why it doesn't.

Example 2: Margin of Error

Suppose that 50 random samples each of size ten produced the sample means displayed in the following dot plot.



Note that almost all of the sample means are between 4 and 11. That is, almost all are roughly within 3.5 rating points of the population mean 7.5. The value 3.5 is a visual estimate of the margin of error. It is not really an "error" in the sense of "mistake." Rather, it is how far our estimate for the population mean is likely to be from the actual value of the population mean.

Based on the class distribution of sample means, is the visual estimate of margin of error close to 3.5?

Example 3: Standard Deviation as a Refinement of Margin of Error

Note that the margin of error is measuring how spread out the sample means are relative to the value of the actual population mean. From previous lessons, you know that the standard deviation is a good measure of spread. So, rather than producing a visual estimate for the margin of error from the distribution of sample means, another approach is to use the standard deviation of the sample means as a measure of spread. For example, the standard deviation of the 50 sample means in the example above is 1.7. Note that if you double 1.7, you get a value for margin of error close to the visual estimate of 3.5.

Another way to estimate margin of error is to use two times the standard deviation of a distribution of sample means. For the above example, the refined margin of error (based on the standard deviation of sample means) is $2(1.7) = 3.4$ rating points.

An interpretation of the margin of error is that plausible values for the population mean rating are from $7.5 - 3.4$ to $7.5 + 3.4$ (i.e., 4.1 to 10.9 rating points).

Exploratory Challenge 3/Exercise 7

Calculate and interpret the margin of error for your estimate of the population mean rating of 100 computer games based on the standard deviation of your class distribution of sample means.

Lesson Summary

This lesson revisited margin of error. Previously, you estimated a population proportion of successes and described the accuracy of the estimate by its margin of error. This lesson also focused on margin of error but in the context of estimating the mean of a population of numerical data.

Margin of error was estimated in two ways:

- The first was through a visual estimation in which you judged the amount of spread in the distribution of sample means.
- The second was more formalized by defining margin of error as twice the standard deviation of the distribution of sample means.

Problem Set

1. Suppose you are interested in knowing how many text messages eleventh graders send daily.

   Describe the steps that you would take to estimate the mean number of text messages per day sent by all eleventh graders at a school.

2. Suppose that 62 random samples based on ten student responses to the question, "How many text messages do you send per day?" resulted in the 62 sample means (rounded) shown below.

65	68	76	76	78	82	83	83	85	86	87	88	88
88	89	89	89	90	91	91	91	91	92	92	92	92
92	93	93	93	93	93	94	94	94	94	94	94	95
95	95	95	95	95	95	95	96	96	97	97	97	98
98	98	98	98	99	100	100	101	104	106			

   a. Draw a dot plot for the distribution of sample means.
   b. Based on your dot plot, would you be surprised if the actual mean number of text messages sent per day for all eleventh graders in the school is 91.7? Why or why not?

3. Determine a visual estimate of the margin of error when a random sample of size 10 is used to estimate the population mean number of text messages sent per day.

4. The standard deviation of the above distribution of sample mean number of text messages sent per day is 7.5. Use this to calculate and interpret the margin of error for an estimate of the population mean number of text messages sent daily by eleventh graders (based on a random sample of size 10 from this population).

Lesson 21: Margin of Error when Estimating a Population Mean

Classwork

This lesson continues to discuss using the sample mean as an estimate of the population mean and judging its accuracy based on the concept of margin of error. In the last lesson, the margin of error was defined as twice the standard deviation of the sampling distribution of the sample mean. In this lesson, a formula will be given for the margin of error that allows you to calculate the margin of error from a single random sample rather than having to create a sampling distribution of sample means.

Example 1: Estimating a Population Mean Using a Random Sample

Provide a one-sentence summary of our findings from the previous lesson.

What were drawbacks of the calculation method?

In practice, you do not have to use that process to find the margin of error. Fortunately, just as was the case with estimating a population proportion, there are some general results that lead to a formula that allows you to estimate the margin of error using a single sample. You can then gauge the accuracy of your estimate of the population mean by calculating the margin of error using the sample standard deviation.

> The standard deviation of the distribution of sample means is approximated by $\frac{s}{\sqrt{n}}$, where s is the standard deviation of the sample, and n is the size of the sample.

Exercises 1–5

1. Suppose a random sample of size ten produced the following ratings in the computer games rating example in the last lesson: $12, 5, 2, 4, 1, 4, 18, 10, 1, 16$. Estimate the population mean rating based on these ten sampled ratings.

2. Calculate the sample standard deviation. Round your answer to three decimal places.

3. Use the formula given above to calculate the approximate standard deviation of the distribution of sample means. Round your answer to three decimal places.

4. Recall that the margin of error is twice the standard deviation of the distribution of sample means. What is the value of the margin of error based on this sample? Write a sentence interpreting the value of the margin of error in the context of this problem on computer game ratings.

5. Based on the sample mean and the value of the margin of error, what is an interval of plausible values for the population mean?

Exercises 6-13

The Gettysburg Address is considered one of history's greatest speeches. Some students noticed that the speech was very short (about 268 words, depending on the version) and wondered if the words were also relatively short. To estimate the mean length of words in the population of words in the Gettysburg Address, work with a partner on the following steps. Your teacher will give you a copy of the Gettysburg Address with words numbered from 001 to 268.

6. Develop and describe a plan for collecting data from the Gettysburg Address and determining the typical length of a word. Then, implement your plan, and report your findings.

7. Use a random-number table or a calculator with a random-number generator to obtain ten different random numbers from 001 to 268.

8. Use the random numbers found in Exercise 7 as identification numbers for the words that will make up your random sample of words from the Gettysburg Address. Make a list of the ten words in your sample.

9. Count the number of letters in each of the ten words in your sample.

10. Calculate the sample mean number of letters for the ten words in your sample.

11. Calculate the sample standard deviation of the number of letters for the ten words in your sample. Round your answer to three decimal places.

12. Use the sample standard deviation from Exercise 11 to calculate the margin of error associated with using your sample mean as an estimate of the population mean. Round your answer to three decimal places.

13. Write a few sentences describing what you have learned about the mean length of the population of 268 words in the Gettysburg Address. Be sure to include an interpretation of the margin of error.

Lesson Summary

- When using the sample mean to estimate a population mean, it is important to know something about how accurate that estimate might be.
- Accuracy can be described by the margin of error.
- The margin of error can be estimated using data from a single random sample (without the need to create a simulated sampling distribution) by using the formula $2\left(\frac{s}{\sqrt{n}}\right)$, where s is the standard deviation of a single sample, and n is the sample size.

Problem Set

1. A new brand of hot dog claims to have a lower sodium content than the leading brand.
   a. A random sample of ten of these new hot dogs results in the following sodium measurements (mg):
   $$370\ \ 326\ \ 322\ \ 297\ \ 326\ \ 289\ \ 293\ \ 264\ \ 327\ \ 331.$$
   Estimate the population mean sodium content of this new brand of hot dog based on the ten sampled measurements.
   b. Calculate the margin of error associated with your estimate of the population mean from part (a). Round your answer to three decimal places.
   c. The mean sodium content of the leading brand of hot dogs is known to be 350 mg. Based on the sample mean and the value of the margin of error for the new brand, is a mean sodium content of 350 mg a plausible value for the mean sodium content of the new brand? Comment on whether you think the new brand of hot dog has a lower sodium content on average than the leading brand.
   d. Another random sample of 40 new brand hot dogs is taken. Should this larger sample of hot dogs produce a more accurate estimate of the population mean sodium content than the sample of size 10? Explain your answer by appealing to the formula for margin of error.

2. It is well known that astronauts increase their height in space missions because of the lack of gravity. A question is whether or not we increase height here on Earth when we are put into a situation where the effect of gravity is minimized. In particular, do people grow taller when confined to a bed? A study was done in which the heights of six men were taken before and after they were confined to bed for three full days.
   a. The before-after differences in height measurements (mm) for the six men were:
   $$12.6\ \ 14.4\ \ 14.7\ \ 14.5\ \ 15.2\ \ 13.5.$$
   Assuming that the men in this study are representative of the population of all men, what is an estimate of the population mean increase in height after three full days in bed?
   b. Calculate the margin of error associated with your estimate of the population mean from part (a). Round your answer to three decimal places.
   c. Based on your sample mean and the margin of error from parts (a) and (b), what are plausible values for the population mean height increase for all men who stay in bed for three full days?

Lesson 22: Evaluating Reports Based on Data from a Sample

Classwork

Exercises 1–5: Election Results

The following is part of an article that appeared in a newspaper:

> "With the election for governor still more than a year away, a new poll shows the race is already close. The Republican governor had 47%, and the Democratic challenger had 45% in a poll released Tuesday of 800 registered voters.
>
> 'That's within the poll's margin of error of 3.5 percentage points, making it essentially a tossup,' said the poll's director."

1. Why don't the two percentages add up to 100%?

2. What is meant by the margin of error of 3.5 percentage points?

3. Using the sample size of 800 and the proportion 0.47, calculate the margin of error associated with the estimate of the proportion of all registered voters who would vote for the Republican governor.

4. Why did the poll director say that the election is "essentially a tossup"?

5. If the sample size had been $2{,}500$ registered voters, and the results stated 47% would vote for the Republican Governor and 45% said they would vote for the Democratic challenger, what would the margin of error have been? Could the director still say that the election was a tossup?

Exercises 6–8: Chocolate Chip Claim

The Nabisco Company claims that there are at least $1,000$ chocolate chips in every 18-ounce bag of their Chips Ahoy! cookies. An article in a local newspaper reported the efforts of a group of students in their attempt to validate the Nabisco claim. The article reported that the students randomly selected 42 bags of cookies from local grocery stores and counted the number of chocolate chips in the cookies in each bag. The students found the sample mean was $1,261.6$ chips, and the sample standard deviation was 117.6 chips. The article stated that the students' data supported the Nabisco Company claim.

6. Using the students' results, calculate the margin of error associated with the estimate of the mean number of chocolate chips in an 18-ounce bag of Chips Ahoy! chocolate chip cookies. Write a sentence interpreting the margin of error.

7. Do you agree that the student data supported the Nabisco Company claim? Explain.

8. Comment on the procedure that the students used to collect their data.

Exercises 9–15: Understanding a Poll

George Gallup founded the American Institute of Public Opinion (Gallup Poll) in 1935. The company is famous for its public opinion polls, which are conducted in the United States and other countries.

Gallup published the following graph in May 2013.

Source: http://www.gallup.com/poll/162194/americans-exercise-habits-worsen-slightly-2013.aspx

9. What percent of those surveyed said that they exercise at least 30 minutes three or more days a week at the start of 2013?

10. Describe the patterns that you observe in the graph.

11. Give some reasons why you think the graph follows the pattern that you described.

Following are the survey methods that Gallup used to collect the data:

> "Results are based on telephone interviews conducted as part of the Gallup-Healthways Well-Being Index survey June 1-30, 2013, with a random sample of $15{,}235$ adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia.

For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ± 1 percentage point."

12. Using the value of 0.538 for the proportion of those surveyed who said they exercise at least 30 minutes three or more days a week in the most recent poll, calculate the margin of error. How does your margin of error compare to the value reported by Gallup?

13. Interpret the phrase "margin of sampling error is ± 1 percentage point."

14. Why is it important that Gallup selects a random sample of adults?

15. If Gallup had used a random sample of $1{,}500$, what would happen to the margin of error? Explain your answer.

Lesson Summary

- The estimated margin of error when a sample proportion from a random sample is used to estimate a population proportion is $ME = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is the sample proportion.
- The estimated margin of error when a sample mean from a random sample is used to estimate a population mean is $ME = 2\left(\frac{s}{\sqrt{n}}\right)$ where $\bar{x}$ is the sample mean.
- It is important to interpret margin of error in context.
- It is unlikely that the estimate of a population proportion or mean will be farther from the actual population value than the margin of error.

Problem Set

1. The British Medical Journal published a study whose objective was to investigate estimation of calorie content of meals from fast food restaurants. Below are the published results.

   Participants: 1,877 adults and 330 school age children visiting restaurants at dinnertime (evening meal) in 2010 and 2011; 1,178 adolescents visiting restaurants after school or at lunchtime in 2010 and 2011.

   Results: Among adults, adolescents, and school age children, the mean actual calorie content of meals was 836 calories (SD 465), 756 calories (SD 455), and 733 calories (SD 359), respectively. Compared with the actual figures, participants underestimated calorie content by means of 175 calories, 259 calories, and 175 calories, respectively.

   Source: http://www.bmj.com/content/346/bmj.f2907

   a. Calculate the margin of error associated with the estimate of the mean number of actual calories in the meals eaten by each of the groups: adults, adolescents, and school age children.
   b. Write a sentence interpreting the margin of error for the adult group.
   c. Explain why the margin of error for the estimate of the mean number of actual calories in meals eaten by adults is smaller than the margin of error of the mean number of actual calories in meals eaten by school age children.
   d. Write a conclusion that the researchers could draw from this study.

2. The Gallup organization published the following results from a poll that it conducted.

   "By their own admission, many young Americans, aged 18 to 29, say they spend too much time using the Internet (59%), their cell phones or smartphones (58%), and social media sites such as Facebook (48%). Americans' perceptions that they spend 'too much' time using each of these technologies decline with age. Conversely, older Americans are most likely to say they spend too much time watching television, and among all Americans, television is the most overused technology tested.

   Results are based on telephone interviews conducted as part of Gallup Daily tracking April 9-10, 2012, with a random sample of $1,051$ adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia.

For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is ± 4 percentage points."

Source: http://www.gallup.com/poll/153863/Young-Adults-Admit-Time-Cell-Phones-Web.aspx

a. Write a newspaper headline that would capture the main idea from the poll.

b. Use the phrase from the article "their cell phones or smartphones (58%), " to calculate the margin of error. Show your work.

c. How do your results compare with the margin of error stated in the article?

d. Interpret the statement "the margin of sampling error is ± 3 percentage points."

e. What would happen to the margin of error if Gallup had surveyed 100 people instead of the $1,051$? The

3. Holiday Inn Resort Brand conducted the *Kid Classified* survey. $1,500$ parents and children nationwide were interviewed via an online survey.

The results of the survey state:

> "While many parents surveyed say they have some financial savings set aside specifically for vacation travel, **more than half of parents in the survey (52%) noted that saving enough money was the biggest challenge to planning a family vacation**, more so than coordination of family schedules (19%) or taking time off of work (12%)."

Source: http://www.lodgingmagazine.com/holiday-inn-resorts-catering-to-kids/

a. Calculate the margin of error associated with the estimate of the proportion of all parents who would say that saving enough money is the biggest challenge to planning a family vacation.

b. Write a sentence interpreting the margin of error.

c. Comment on how the survey was conducted.

Lesson 23: Experiments and the Role of Random Assignment

Classwork

Exercises 1–4: Experiments

Two studies are described below. One is an observational study, while the other is an experiment.

Study A:

A new dog food, specially designed for older dogs, has been developed. A veterinarian wants to test this new food against another dog food currently on the market to see if it improves dogs' health. Thirty older dogs were randomly assigned to either the "new" food group or the "current" food group. After they were fed either the "new" or "current" food for six months, their improvement in health was rated.

Study B:

The administration at a large school wanted to determine if there was a difference in the mean number of text messages sent by 9^{th} grade students and by 11^{th} grade students during a day. Each person in a random sample of thirty 9^{th} grade students was asked how many text messages he or she sent per day. Each person in another random sample of thirty 11^{th} grade students was asked how many text messages he or she sent per day. The difference in the mean number of texts per day was determined.

1. Which study is the experiment? Explain. Discuss the answer with your partner.

2. In your own words, describe what a subject is in an experiment.

3. In your own words, describe what a response variable is in an experiment.

4. In your own words, describe what a treatment is in an experiment.

Exercises 5–9: Random Selection and Random Assignment

Take another look at the two studies described above. Study A (the dog food study) is an experiment, while study B (text messages) is an observational study. The term *random sample* implies that a sample was randomly selected from a population. The terms *random selection* and *random assignment* have very different meanings.

Random selection refers to randomly selecting a sample from a population. Random selection allows generalization to a population and is used in well-designed observational studies. Sometimes, but not always, the subjects in an experiment are randomly selected.

Random assignment refers to randomly assigning the subjects in an experiment to treatments. Random assignment allows for cause and effect conclusions and is used in well-designed experiments.

In study B, the data were collected from two random samples of students.

5. Can the results of the survey be generalized to all 9$^{\text{th}}$ grade and all 11$^{\text{th}}$ grade students at the school? Why or why not? Discuss the answer with your partner.

6. Suppose there really is a difference in the mean number of texts sent by 9$^{\text{th}}$ grade students and by 11$^{\text{th}}$ grade students. Can we say that the grade level of the students is the cause of the difference in the mean number of texts sent? Why or why not? Discuss the answer with your partner.

In study A, the dogs were randomly assigned to one of the two types of food.

7. Suppose the dogs that were fed the new food showed improved health. Can we say that the new food is the cause of the improvement in the dogs' health? Why or why not? Discuss the answer with your partner.

8. Can the results of the dog food study be generalized to all dogs? To all older dogs? Why or why not? Discuss the answer with your partner.

The table below summarizes the differences between the terms random selection and random assignment.

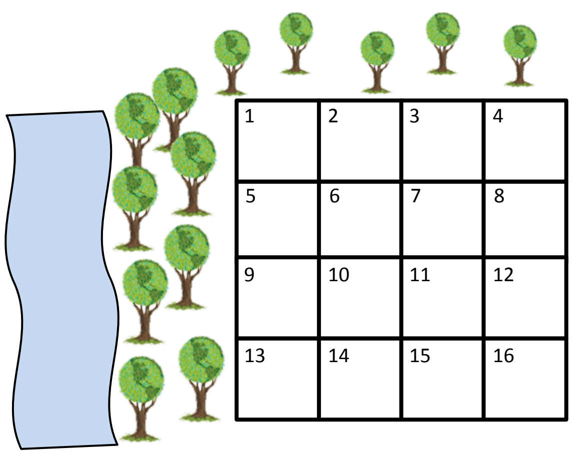
9. For each statement, put a check mark in the appropriate column(s) and explain your choices.

	Random Selection	Random Assignment
Used in experiments		
Used in observational studies		
Allows generalization to the population		
Allows a cause and effect conclusion		

Exercises 10–17

What is the purpose of random assignment in experiments? To answer this, consider the following investigation.

A researcher wants to determine if the yield of corn is different when the soil is treated with one of two different types of fertilizers, fertilizer A and fertilizer B. The researcher has 16 acres of land located beside a river that has several trees along its bank. There are also a few trees to the north of the 16 acres. The land has been divided into sixteen 1-acre plots. (See the diagram below.) These 16 plots are to be planted with the same type of corn but can be fertilized differently. At the end of the growing season, the corn yield will be measured for each plot, and the mean yields for the plots assigned to each fertilizer will be compared.



10. For the experiment, identify the following and explain each answer:

a. Subjects (Hint: Not always people or animals)

b. Treatments

c. Response variable

Next, you need to assign the plots to one of the two treatments. To do this, follow the instructions given by your teacher.

11. Write A (for fertilizer A) or B (for fertilizer B) in each of the 16 squares in the diagram so that it corresponds to your random assignment of fertilizer to plots.

Let's investigate the results of the random assignment of the fertilizer types to the plots.

12. On the diagram above, draw a vertical line down the center of the 16 plots of land.

13. Count the number of plots on the left side of the vertical line that will receive fertilizer A. Count the number of plots on the right side of the vertical line that will receive fertilizer A.

Left _________ Right _________

14. On the diagram above, draw a horizontal line through the center of the 16 plots of land.

15. Count the number of plots above the horizontal line that will receive fertilizer A. Count the number of plots below the horizontal line that will receive fertilizer A.

Above _________ Below _________

In experiments, random assignment is used as a way of ensuring that the groups that receive each treatment are as much alike as possible with respect to other factors that might affect the response.

16. Explain what this means in the context of this experiment.

17. Suppose that at the end of the experiment the mean yield for one of the fertilizers is quite a bit higher than the mean yield for the other fertilizer. Explain why it would be reasonable to say that the type of fertilizer is the cause of the difference in yield and not the proximity to the river or to the northern trees.

Lesson Summary

- An experiment is an investigation designed to compare the effect of two or more treatments on a response variable.
- A subject is a participant in the experiment.
- The response variable is a variable that is not controlled by the experimenter and that is measured as part of the experiment.
- The treatments are the conditions to which subjects are randomly assigned by the experimenter.
- Random selection refers to randomly selecting a sample from a population.
  - Random selection allows for generalization to a population.
- Random assignment refers to randomly assigning subjects to treatment groups.
  - Random assignment allows for cause and effect conclusions.
  - The purpose of random assignment in an experiment is to create similar groups of subjects for each of the treatments in the experiment.

Problem Set

For Problems 1 through 5, identify (i) the subjects, (ii) the treatments, and (iii) the response variable for each experiment.

1. A botanist was interested in determining the effects of watering (three days a week or daily) on the heat rating of jalapeño peppers. The botanist wanted to know which watering schedule would produce the highest heat rating in the peppers. He conducted an experiment, randomly assigning each watering schedule to half of twelve plots that had similar soil and full sun. The average final heat rating for the peppers grown in each plot was recorded at the end of the growing season.

2. A manufacturer advertises that its new plastic cake pan bakes cakes more evenly. A consumer group wants to carry out an experiment to see if the plastic cake pans do bake more evenly than standard metal cake pans. Twenty cake mixes (same brand and type) are randomly assigned to either the plastic pan or the metal pan. All of the cakes are baked in the same oven. The rating scale was then used to rate the evenness of each cake.

3. The city council of a large city is considering a new law that prohibits talking on a cell phone while driving. A consumer rights organization wants to know if talking on a cell phone while driving distracts a person's attention, causing them to make errors while driving. An experiment is designed that uses a driving simulator to compare the two treatments, driving while talking on a cell phone and driving while not talking on a cell phone. The number of errors made while driving on an obstacle course will be recorded for each driver. Each person in a random sample of 200 licensed drivers in the city was asked to participate in the experiment. All of the drivers agreed to participate in the experiment. Half of the drivers were randomly assigned to drive an obstacle course while talking on a phone. The remaining half was assigned to drive the obstacle course while not talking on a phone.

4. Researchers studied 208 infants whose brains were temporarily deprived of oxygen as a result of complications at birth (*The New England Journal of Medicine*, October 13, 2005). An experiment was performed to determine if reducing body temperature for three days after birth improved their chances of surviving without brain damage. Infants were randomly assigned to usual care or whole-body cooling. The amount of brain damage was measured for each infant.

5. The head of the quality control department at a printing company would like to carry out an experiment to determine which of three different glues results in the greatest binding strength. Copies of a book were randomly assigned to one of the three different glues.

6. In Problem 3, suppose that drivers who talked on the phone while driving committed more errors on the obstacle course than drivers who did not talk on the phone while driving. Can we say that talking on the cell phone while driving is the cause of the increased errors on the obstacle course? Why or why not?

7. Can the results of the experiment in Problem 3 be generalized to all licensed drivers in the city? Why or why not?

8. In Problem 4, one of the treatment groups was to use usual care for the infants. Why was this treatment group included in the experiment?

9. In Problem 5, why were copies of only one book used in the experiment?

Lesson 24: Differences Due to Random Assignment Alone

Classwork

Exercises 1–17

Twenty adult drivers were asked the following question:

"What speed is the fastest that you have driven?"

The table below summarizes the fastest speeds driven in miles per hour (mph).

70	60	70	95	50	60	80	75	55	90
110	65	65	65	55	70	75	70	65	40

1. What is the mean fastest speed driven?

2. What is the range of fastest speed driven?

3. Imagine that the fastest speeds were randomly divided into two groups. How would the means and ranges compare to one another? To the means and ranges of the whole group? Explain your thinking.

Let's investigate what happens when the fastest speeds driven are randomly divided into two equal-size groups.

4. Following the instructions from your teacher, randomly divide the 20 values in the table above into two groups of 10 values each.

											Mean
Group 1											
Group 2											

5. Do you expect the means of these two groups to be equal? Why or why not?

6. Compute the means of these two groups. Write the means in the chart above.

7. How do these two means compare to each other?

8. How do these two means compare to the mean fastest speed driven for the entire group (Exercise 1)?

9. Use the instructions provided for Exercise 4 to repeat the random division process two more times. Compute the mean of each group for each of the random divisions into two groups. Record your results in the tables below.

											Mean
Group 3											
Group 4											
Group 5											
Group 6											

10. Plot the means of all six groups on a class dot plot.

11. Based on the class dot plot, what can you say about the possible values of the group means?

12. What is the smallest possible value for a group mean? Largest possible value?

13. What is the largest possible range for the distribution of group means?

14. How does the largest possible range in the group means compare to the range of the original data set (Exercise 2)? Why is this so?

15. What is the shape of the distribution of group means?

16. Will your answer to the above question always be true? Explain.

17. When a single set of values is randomly divided into two equal groups, explain how the means of these two groups may be very different from each other and may be very different from the mean of the single set of values.

Lesson Summary

When a single set of values is randomly divided into two groups,

- The two group means will tend to differ just by chance.
- The distribution of random groups' means will be centered at the single set's mean.
- The range of the distribution of the random groups' means will be smaller than the range of the data set.
- The shape of the distribution of the random groups' means will be symmetrical.

Problem Set

In one high school, there are eight math classes during 2nd period. The number of students in each 2nd period math class is recorded below.

$$32 \qquad 27 \qquad 26 \qquad 23 \qquad 25 \qquad 22 \qquad 30 \qquad 19$$

This data set is randomly divided into two equal size groups, and the group means are computed.

1. Will the two groups means be the same? Why or why not?

The random division into two groups process is repeated many times to create a distribution of group mean class size.

2. What is the center of the distribution of group mean class size?

3. What is the largest possible range of the distribution of group mean class size?

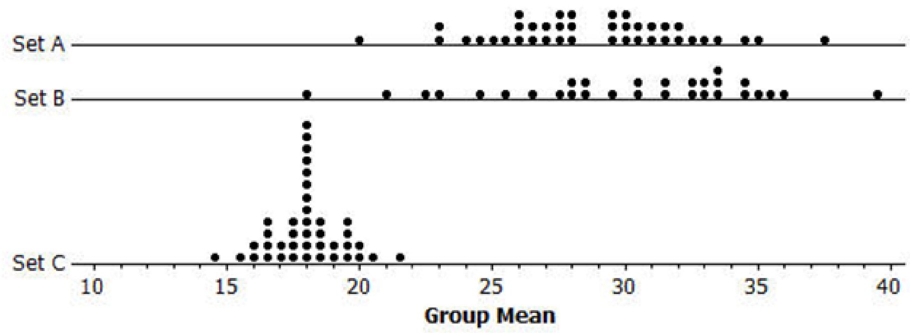
4. What possible values for the mean class size are more likely to happen than others? Explain why you chose these values.

There are 3 different sets of numbers: Set A, Set B, and Set C. Each set contains 10 numbers. In two of the sets, the 10 numbers were randomly divided into two groups of 5 numbers each, and the mean for each group was calculated. These two means are plotted on a dot plot. This procedure was repeated many times, and the dot plots of the group means are shown below.

The third set did not use the above procedure to compute the means.

For each set, the smallest possible group mean and the largest possible group mean were calculated, and these two means are shown in the dot plots below.

Use the dot plots below to answer Problems 5–8.



5. Which set is NOT one of the two sets that were randomly divided into two groups of 5 numbers? Explain.

6. Estimate the mean of the original values in Set A. Show your work.

7. Estimate the range of the group means shown in the dot plot for Set C. Show your work.

8. Is the range of the original values in Set C smaller or larger than your answer in Problem 7? Explain.

Lesson 25: Ruling Out Chance

Classwork

Opening Exercise

a. Explain in writing what you learned about randomly divided groups from the last lesson. Share your thoughts with a neighbor.

b. How could simulation be used to understand the "typical" differences between randomized groups?

Exercises 1–3: Random Assignments and Computing the Difference of the Group Means

Imagine that 10 tomatoes of varying shapes and sizes have been placed in front of you. These 10 tomatoes (all of the same variety) have been part of a nutrient experiment where the application of the nutrient is expected to yield larger tomatoes that weigh more. All 10 tomatoes have been grown under similar conditions regarding soil, water, and sunlight, but 5 of the tomatoes received the additional nutrient supplement. Using the weight data of these 10 tomatoes, you wish to examine the claim that the nutrient yields larger tomatoes on average.

1. Why would it be important in this experiment for the 10 tomatoes to all be of the same variety and grown under the same conditions (except for the treatment applied to 5 of the tomatoes)?

Here are the 10 tomatoes with their weights shown. They have been ordered from largest to smallest based on weight.

9.1	8.4	8.0	7.7	7.3	6.4	5.9	5.2	4.4	3.8
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10

For now, do not be concerned about which tomatoes received the additional nutrients. The object here is to randomly assign the tomatoes to two groups.

Imagine that someone assisting you uses a random-number generator or some other impartial selection device and randomly selects Tomatoes #1, 4, 5, 7, and 10 to be in Group A. By default, Tomatoes #2, 3, 6, 8, and 9 will be in Group B. The result is illustrated below.

Group A	Group B
9.1	8.4
7.7	8.0
7.3	6.4
5.9	5.2
3.8	4.4

2. Confirm that the mean for Group A is 6.76 ounces, and calculate the mean for Group B.

3. Calculate the difference between the mean of Group A and the mean of Group B (that is, calculate $\bar{x}_A - \bar{x}_B$).

Exercises 4–6: Interpreting the Value of a Difference

The statistic of interest that you care about is the difference between the mean of the 5 tomatoes in Group A and the mean of the 5 tomatoes in Group B. For now, call that difference "Diff." "Diff" = $\bar{x}_A - \bar{x}_B$.

4. Explain what a "Diff" value of "1.64 ounces" would mean in terms of which group has the larger mean weight and the number of ounces by which that group's mean weight exceeds the other group's mean weight.

5. Explain what a "Diff" value of "-0.4 ounces" would mean in terms of which group has the larger mean weight and the number of ounces by which that group's mean weight exceeds the other group's mean weight.

6. Explain what a "Diff" value of "0 ounces" would mean regarding the difference between the mean weight of the 5 tomatoes in Group A and the mean weight of the 5 tomatoes in Group B.

Exercises 7–8: Additional Random Assignments

7. Below is a second random assignment of the 10 tomatoes to two groups. Calculate the mean of each group, and then calculate the value of "Diff" for this second case. Also, interpret the "Diff" value in context using your responses to the previous questions as a guide.

Group A	Group B
9.1	7.7
8.4	5.9
8.0	5.2
7.3	4.4
6.4	3.8

8. Here is a third random assignment of the 10 tomatoes. Calculate the mean of each group, and then calculate the value of "Diff" for this case. Interpret the "Diff" value in context using your responses to the previous questions as a guide.

Group A	Group B
9.1	8.4
7.7	8.0
7.3	6.4
5.2	5.9
3.8	4.4

Lesson Summary

In this lesson, when the single group of observations was randomly divided into two groups, the means of these two groups differed by chance. These differences have a context based on the purpose of the experiment and the units of the original observations.

The differences varied. In some cases, the difference in the means of these two groups was very small (or "0"), but in other cases, this difference was larger. However, in order to determine which differences were typical and ordinary vs. unusual and rare, a sense of the center, spread, and shape of the distribution of possible differences is needed. In the following lessons, you will develop this distribution by executing repeated random assignments similar to the ones you saw in this lesson.

Problem Set

Six ping-pong balls are labeled as follows: $0, 3, 6, 9, 12, 18$. Three ping-pong balls will be randomly assigned to Group A; the rest will be assigned to Group B. “Diff” = $\bar{x}_A - \bar{x}_B$.

In the Exit Ticket problem, 4 of the 20 possible randomizations have been addressed.

1. Develop the remaining 16 possible random assignments to two groups, and calculate the "Diff" value for each.

   (Note: Avoid redundant cases; selecting "0, 3, and 6" for Group A is NOT a distinct random assignment from selecting "6, 0, and 3" – so do not record both!)

Group A Selection			"A" Mean	"B" Mean	"Diff"	
0	3	6	3	13	-10	Question #4
0	3	9	4	12	-8	
0	3	12	5	11	-6	
0	3	18	7	9	-2	
0	6	9	5	11	-6	
0	6	12	6	10	-4	
0	6	18	8	8	0	
0	9	12	7	9	-2	
0	9	18	9	7	2	
0	12	18	10	6	4	
3	6	9	6	10	-4	
3	6	12	7	9	-2	Question #1
3	6	18	9	7	2	
3	9	12	8	8	0	
3	9	18	10	6	4	
3	12	18	11	5	6	Question #2
6	9	12	9	7	2	
6	9	18	11	5	6	
6	12	18	12	4	8	
9	12	18	13	3	10	Question #3

2. Create a dot plot that shows the 20 "Diff" values obtained from the 20 possible randomizations. By visual inspection, what is the mean and median value of the distribution?

3. Based on your dot plot, what is the probability of obtaining a "Diff" value of "8 or higher"?

4. Would a “Diff” value of “8 or higher” be considered a difference that is likely to happen or one that is unlikely to happen? Explain.

5. Based on your dot plot, what is the probability of obtaining a “Diff” value of “-2 or smaller”?

6. Would a “Diff” value of “-2 or smaller” be considered a difference that is likely to happen or one that is unlikely to happen? Explain.

Lesson 26: Ruling Out Chance

Classwork

Opening Exercise

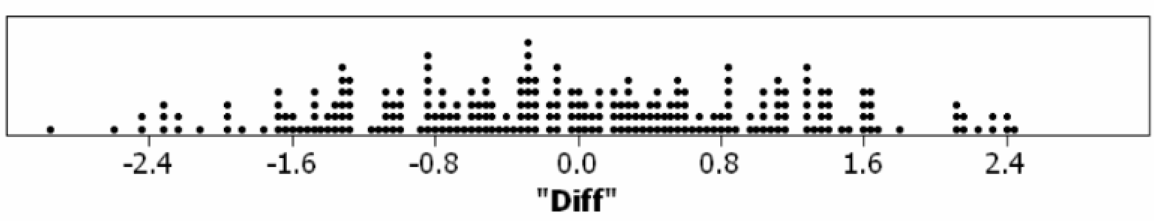
Previously, you considered the random assignment of 10 tomatoes into two distinct groups of 5 tomatoes each called Group A and Group B. With each random assignment, you calculated "Diff" = $\bar{x}_A - \bar{x}_B$, the difference between the mean weight of the 5 tomatoes in Group A and the mean weight of the 5 tomatoes in Group B.

a. Summarize in writing what you learned in the last lesson. Share your thoughts with a neighbor.

b. Recall that 5 of these 10 tomatoes are from plants that received a nutrient treatment in the hope of growing bigger tomatoes. But what if the treatment was ***not*** effective? What difference would you expect to find between the group means?

Exercises 1–2: The Distribution of "Diff" and Why "0" is Important

In the previous lesson, 3 instances of the tomato randomization were considered. Imagine that the random assignment was conducted an additional 247 times, and 250 "Diff" values were computed from these 250 random assignments. The results are shown graphically below in a dot plot where each dot represents the "Diff" value that results from a random assignment:



This dot plot will serve as your *randomization distribution* for the "Diff" statistic in this tomato randomization example. The dots are placed at increments of 0.04 ounces.

1. Given the distribution picture above, what is the *approximate* value of the median and mean of the distribution? Specifically, do you think this distribution is centered near a value that implies "No Difference" between Group A and Group B?

2. Given the distribution pictured above and based on the simulation results, determine the approximate probability of obtaining a "Diff" value in the cases described in (a), (b), and (c).

   a. of 1.64 ounces or more

   b. of -0.80 ounces or less

   c. within 0.80 ounces of 0 ounces

   d. How do you think these probabilities could be useful to people that are designing experiments?

Exercises 3-5: Statistically Significant "Diff" Values

In the context of a randomization distribution that is based upon the assumption that there is no real difference between the groups, consider a "Diff" value of X to be "statistically significant" if there is a low probability of obtaining a result that is as extreme as or more extreme than X.

3. Using that definition and your work above, would you consider any of the "Diff" values below to be statistically significant? Explain.
   a. 1.64 ounces

   b. -0.80 ounces

   c. Values within 0.80 ounces of 0 ounces

4. In the previous lessons, you obtained "Diff" values of 0.28 ounces, 2.44 ounces, and 0 ounces for 3 different tomato randomizations. Would you consider any of those values to be "statistically significant" for this distribution? Explain.

5. Recalling that "Diff" is the mean weight of the 5 Group A tomatoes minus the mean weight of the 5 Group B tomatoes, how would you explain the meaning of a "Diff" value of 1.64 ounces in this case?

Exercises 6-8: The Implication of Statistically Significant "Diff" Values

Keep in mind that for reasons mentioned earlier, the randomization distribution above is demonstrating what is likely to happen *by chance alone* if the treatment was not effective. As stated in the previous lesson, you can use this randomization distribution to assess whether or not the *actual* difference in means *obtained from your experiment* (the difference between the mean weight of the 5 actual control group tomatoes and the mean weight of the 5 actual treatment group tomatoes) is consistent with usual chance behavior. The logic is as follows:

- If the observed difference is “extreme” and not typical of chance behavior, it may be considered “statistically significant” and possibly not the result of chance behavior.
- If the difference is not the result of chance behavior, then maybe the difference did not just happen by chance alone.
- If the difference did not just happen by chance alone, maybe the difference you observed is caused by the treatment in question, which, in this case, is the nutrient. In the context of our example, a statistically significant "Diff" value provides evidence that the nutrient treatment did in fact yield heavier tomatoes on average.

6. For reasons that will be explained in the next lesson, for your tomato example, "Diff" values that are *positive* and statistically significant will be considered as good evidence that your nutrient treatment did in fact yield heavier tomatoes on average. Again, using the randomization distribution shown earlier in the lesson, which (if any) of the following "Diff" values would you consider to be statistically significant and lead you to think that the nutrient treatment did, in fact, yield heavier tomatoes on average? Explain for each case.

$$\text{Diff} = 0.4,\ \text{Diff} = 0.8,\ \text{Diff} = 1.2,\ \text{Diff} = 1.6,\ \text{Diff} = 2.0,\ \text{Diff} = 2.4$$

7. In the first random assignment in the previous lesson, you obtained a "Diff" value of 0.28 ounces. Earlier in this lesson, you were asked to consider if this might be a "statistically significant" value. Given the distribution shown in this lesson, if you had obtained a "Diff" value of 0.28 ounces ***in your experiment*** and the 5 Group A tomatoes had been the "treatment" tomatoes that received the nutrient, would you say that the "Diff" value was extreme enough to support a conclusion that the nutrient treatment yielded heavier tomatoes on average? Or, do you think such a "Diff" value may just occur by chance when the treatment is ineffective? Explain.

8. In the second random assignment in the previous lesson, you obtained a "Diff" value of 2.44 ounces. Earlier in this lesson, you were asked to consider if this might be a "statistically significant" value. Given the distribution shown in this lesson, if you had obtained a "Diff" value of 2.44 ounces ***in your experiment*** and the 5 Group A tomatoes had been the "treatment" tomatoes that received the nutrient, would you say that the "Diff" value was extreme enough to support a conclusion that the nutrient treatment yielded heavier tomatoes on average? Or do you think such a "Diff" value may just occur by chance when the treatment is ineffective? Explain.

Lesson Summary

In the previous lesson, the concept of randomly separating 10 tomatoes into 2 groups and comparing the means of each group was introduced. The randomization distribution of the difference in means that is created from multiple occurrences of these random assignments demonstrates what is likely to happen *by chance alone* if the nutrient treatment is *not* effective. When the results of your tomato growth experiment are compared to that distribution, you can then determine if the tomato growth experiment's results were typical of chance behavior.

If the results appear typical of chance behavior and near the center of the distribution (that is, not relatively very far from a "Diff" of 0), then there is little evidence that the treatment was effective. However, if it appears that the experiment's results are not typical of chance behavior, then, maybe, the difference you are observing didn't just happen by chance alone. It may indicate a statistically significant difference between the treatment group and the control group, and the source of that difference might be (in this case) the nutrient treatment.

Problem Set

In each of the 3 cases below, calculate the "Diff" value as directed, and write a sentence explaining what the "Diff" value means in context. Write the sentence for a general audience.

1. Group A: 8 dieters lost an average of 8 pounds.

   Group B: 8 non-dieters lost an average of 2 pounds over the same time period.

   Calculate and interpret "Diff" = the mean of Group A minus the mean of Group B.

2. Group A: 11 students were on average 0.4 seconds faster in their 100 meter run times after following a new training regimen.

   Group B: 11 students were on average 0.2 seconds slower in their 100 meter run times after not following any new training regimens.

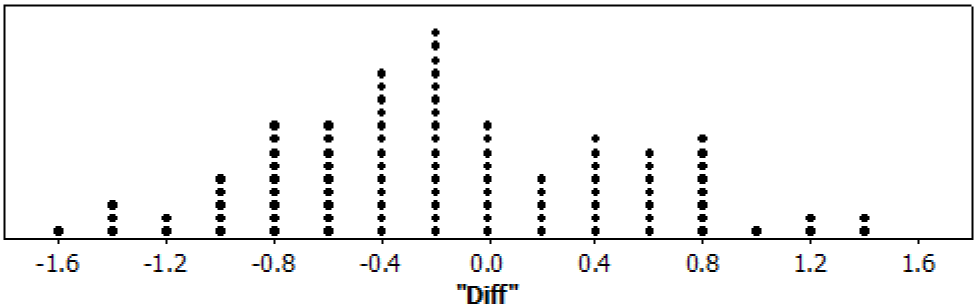
   Calculate and interpret "Diff" = the mean of Group A minus the mean of Group B.

3. Group A: 20 squash that have been grown in an irrigated field have an average weight of 1.3 pounds.

   Group B: 20 squash that have been grown in a non-irrigated field have an average weight of 1.2 pounds.

   Calculate and interpret "Diff" = the mean of Group A minus the mean of Group B.

4. Using the randomization distribution shown in the Exit Ticket, what is the probability of obtaining a "Diff" value of -0.6 or less?



5. Would a "Diff" value of -0.6 or less be considered a “statistically significant difference”? Why or why not?

6. Using the randomization distribution shown in the Exit Ticket, what is the probability of obtaining a "Diff" value of -1.2 or less?

7. Would a "Diff" value of -1.2 or less be considered a “statistically significant difference”? Why or why not?

Lesson 27: Ruling Out Chance

Classwork

Exercises 1–4: Carrying Out a Randomization Test

The following are the general steps for carrying out a randomization test to analyze the results of an experiment. The steps are also presented in the context of the "tomato" example of the previous lessons.

***Step 1*—Develop competing claims: No Difference vs. Difference**

One claim corresponds to no difference between the two groups in the experiment. This claim is called the *null hypothesis*.

- For the tomato example, the null hypothesis is that the nutrient treatment is not effective in increasing tomato weight. This is equivalent to saying that the average weight of treated tomatoes may be the same as the average weight of non-treated (control) tomatoes.

The competing claim corresponds to a difference between the two groups. This claim could take the form of a "different from," "greater than," or "less than" statement. This claim is called the *alternative hypothesis*.

- For the tomato example, the alternative hypothesis is that the nutrient treatment is effective in increasing tomato weight. This is equivalent to saying that the average weight of treated tomatoes *is greater than* the average weight of non-treated (control) tomatoes.

1. Previously, the statistic of interest that you used was the difference between the mean weight of the 5 tomatoes in Group A and the mean weight of the 5 tomatoes in Group B. That difference was called "Diff." "Diff" $=\bar{x}_A - \bar{x}_B$. If the treatment tomatoes are represented by Group A and the control tomatoes are represented by Group B, what type of statistically significant values of "Diff" would support the claim that the average weight of treated tomatoes *is greater than* the average weight of non-treated (control) tomatoes: negative values of "Diff," positive values of "Diff," or both? Explain.

***Step 2*—Take measurements from each group, and calculate the value of the "Diff" statistic from the experiment**.

For the tomato example, first measure the weights of the 5 tomatoes from the treatment group (Group A); next, measure the weights of the 5 tomatoes from the control group (Group B); finally, compute "Diff" $=\bar{x}_A - \bar{x}_B$, which will serve as the result from your experiment.

2. Assume that the following represents the two groups of tomatoes from the *actual* experiment. Calculate the value of "Diff" $=\boldsymbol{\bar{x}_A - \bar{x}_B}$. This will serve as the result from your experiment.

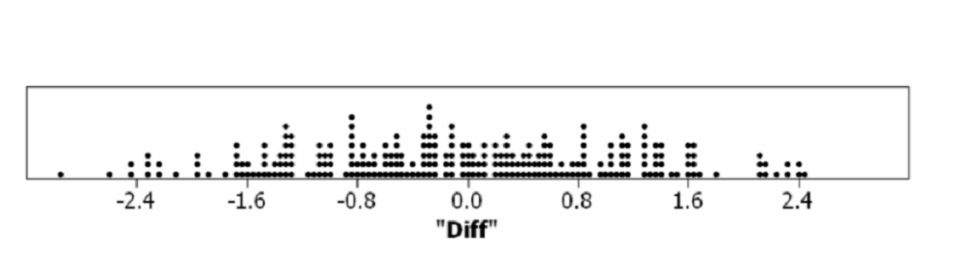
   These are the same 10 tomatoes used in previous lessons; the identification of which tomatoes are "treatment" vs. "control" is now revealed.

TREATMENT Group A	**CONTROL** Group B
9.1	7.7
8.4	6.4
8.0	5.2
7.3	4.4
5.9	3.8

Again, these tomatoes represent the **actual result** from your experiment. You will now create the randomization distribution by making repeated random assignments of these 10 tomatoes into 2 groups and recording the observed difference in means for each random assignment. This develops a randomization distribution of the many possible difference values that could occur under the assumption that there is no difference between the mean weights of tomatoes that receive the treatment and tomatoes that don't receive the treatment.

***Step 3*–Randomly assign the observations to two groups, and calculate the difference between the group means. Repeat this several times, recording each difference. This will create the randomization distribution for the "Diff" statistic.**

Examples of this technique were presented in a previous lesson. For the tomato example, the randomization distribution has already been presented in a previous lesson and is shown again here. The dots are placed at increments of 0.04 ounces.



***Step 4*—With reference to the randomization distribution (Step 3) and the inequality in your alternative hypothesis (Step 1), compute the probability of getting a "Diff" value as extreme as or more extreme than the "Diff" value you obtained in your experiment (Step 2).**

For the tomato example, since the treatment group is Group A, the "Diff" value of $\bar{x}_A - \bar{x}_B$ is $\bar{x}_{Treatment} - \bar{x}_{Control}$. Since the alternative claim is supported by $\bar{x}_{Treatment} > \bar{x}_{Control}$, you are seeking statistically significant "Diff" values that are positive since if $\bar{x}_{Treatment} > \bar{x}_{Control}$, then $\bar{x}_{Treatment} - \bar{x}_{Control} > 0$.

Statistically significant values of "Diff" that are negative, in this case, would imply that the treatment made the tomatoes smaller on average.

3. Using your calculation from Exercise 2, determine the probability of getting a "Diff" value as extreme as or more extreme than the "Diff" value you obtained for this experiment (in Step 2).

***Step 5*—Make a conclusion in context based on the probability calculation (Step 4).**

If there is a **small probability** of obtaining a "Diff" value as extreme as or more extreme than the "Diff" value you obtained in your experiment, then **the "Diff" value from the experiment is unusual** and not typical of chance behavior. Your experiment's results probably did not happen by chance, and the results probably occurred because of a statistically significant difference in the two groups.

- In the tomato experiment, if you think there is a statistically significant difference in the two groups, you have evidence that the treatment may in fact be yielding heavier tomatoes on average.

If there is **not a small probability** of obtaining a "Diff" value as extreme as or more extreme than the "Diff" value you obtained in your experiment, then **your "Diff" value from the experiment is NOT considered unusual** and could be typical of chance behavior. The experiment's results may have just happened by chance and not because of a statistically significant difference in the two groups.

- In the tomato experiment, if you don't think that there is a statistically significant difference in the two groups, then you do not have evidence that the treatment results in larger tomatoes on average.

In some cases, a specific cut-off value called a *significance level* might be employed to assist in determining how small this probability must be in order to consider results statistically significant.

4. Based on your probability calculation in Exercise 3, do the data from the tomato experiment support the claim that the treatment yields heavier tomatoes on average? Explain.

Exercises 5–10: Developing the Randomization Distribution

Although you are familiar with how a randomization distribution is created in the tomato example, the randomization distribution was provided for you. In this exercise, you will develop two randomization distributions based on the same group of 10 tomatoes. One distribution will be developed by hand and will contain the results of at least 250 random assignments. The second distribution will be developed using technology and will contain the results of at least 250 random assignments. Once the two distributions have been developed, you will be asked to compare the distributions

Manually Generated

Your instructor will provide you with specific guidance regarding how many random assignments you need to carry out. Ultimately, your class should generate at least 250 random assignments, compute the "Diff" value for each, and record these 250 or more "Diff" values on a class or individual dot plot.

5. To begin, write the 10 tomato weights on 10 equally sized slips of paper – one weight on each slip. Place the slips in a container and shake the container well. Remove 5 slips and assign those 5 tomatoes as Group A. The remaining tomatoes will serve as Group B.

6. Calculate the mean weight for Group A and the mean weight for Group B. Then, calculate "Diff" = $\bar{x}_A - \bar{x}_B$ for this random assignment.

7. Record your "Diff" value and add this value to the dot plot. Repeat as needed per your instructor's request until a manually generated randomization distribution of at least 250 differences has been achieved.

   (Note: This distribution will most likely be slightly different from the "tomato" randomization distribution given earlier in this lesson.)

Computer Generated

At this stage, you will be encouraged to use a web-based randomization testing applet/calculator to perform the steps above. The applet is located at http://www.rossmanchance.com/applets/AnovaShuffle.htm. To supplement the instructions below, a screenshot of the applet appears as the final page of this lesson.

Upon reaching the applet, do the following:

- Press the "Clear" button to clear the data under "Sample Data."
- Enter the tomato data exactly as shown below. When finished, press the "Use Data" button.

Group	Ounces
Treatment	9.1
Treatment	8.4
Treatment	8
Control	7.7
Treatment	7.3
Control	6.4
Treatment	5.9
Control	5.2
Control	4.4
Control	3.8

Once the data are entered, notice that dot plots of the two groups appear. Also, the "statistic" window below the data now says "difference in means," and an "Observed Diff" value of 2.24 is computed for the experiment's data (just as you computed in Exercise 2).

By design, the applet will determine the difference of means based on the first group name it encounters in the data set – specifically, it will use the first group name it encounters as the first value in the "difference of means" calculation. In other words, to compute the difference in means as $\bar{x}_{Treatment} - \bar{x}_{Control}$, a "Treatment" observation needs to appear prior to any "Control" observations in the data set as entered.

- Select the check box next to "Show Shuffle Options" and a dot plot template will appear.
- Enter "250" in the box next to "Number of Shuffles," and press the "Shuffle Responses" button. A randomization distribution based on 250 randomizations (in the form of a histogram) is created.

This distribution will most likely be *slightly* different from both the "tomato" randomization distribution that appeared earlier in this lesson and the randomization distribution that was manually generated in Exercise 7.

8. Write a few comments comparing the manually generated distribution and the computer generated distribution. Specifically, did they appear to have roughly the same shape, center, and spread?

The applet also allows you to compute probabilities. For this case:

- Under "Count Samples," select "Greater Than." Then, in the box next to "Greater Than," enter "2.2399."

  Since the applet computes the count value as *strictly* "greater than" and not "greater than or equal to," in order to obtain the probability of obtaining a value as extreme as or more extreme than the "Observed Diff" value of 2.24, you will need to enter a value just slightly below 2.24 to ensure that "Diff" observations of 2.24 are included in the count.

- Select the "Count" button. The probability of obtaining a "Diff" value of 2.24 or more in this distribution will be computed for you.

  The applet displays the randomization distribution in the form of a histogram, and it shades in red *all* histogram classes that contain *any* difference values that meet your "Count Samples" criteria. Due to the grouping and binning of the classes, some of the red shaded classes (bars) may also contain difference values that do not fit your "Count Samples" criteria. Just keep in mind that the "Count" value stated in red below the histogram will be exact; the red shading in the histogram may be approximate.

9. How did the probability of obtaining a "Diff" value of 2.24 or more using your computer generated distribution compare with the probability of obtaining a "Diff" value of 2.24 or more using your manually generated distribution?

10. Would you come to the same conclusion regarding the experiment using either the computer generated or manually generated distribution? Explain. Is this the same conclusion you came to using the distribution shown earlier in this lesson back in "Step 3"?

Lesson Summary

The following are the general steps for carrying out a randomization test to analyze the results of an experiment.

<u>Step 1</u>—Develop competing claims: No Difference vs. Difference

Develop the null hypothesis: This claim is that there is no difference between the two groups in the experiment.

Develop the alternative hypothesis: The competing claim is that there <u>**is**</u> a difference between the two groups. This difference could take the form of a "different from," "greater than," or "less than" statement depending on the purpose of the experiment and the claim being assessed.

<u>Step 2</u>—Take measurements from each group, and calculate the value of the "diff" statistic from the experiment.

This is the observed "Diff" value from the experiment.

<u>Step 3</u>—Randomly assign the observations to two groups, and calculate the difference between the group means. Repeat this several times, recording each difference.

This will create the *randomization distribution* for the "Diff" statistic under the assumption that there is no statistically significant difference between the two groups.

<u>Step 4</u>—With reference to the randomization distribution (from Step 3) and the inequality in your alternative hypothesis (from Step 1), compute the probability of getting a "Diff" value as extreme as or more extreme than the "Diff" value you obtained in your experiment (from Step 2).

<u>Step 5</u>—Make a conclusion in context based on the probability calculation (from Step 4).

<u>**Small probability:**</u> If the "Diff" value from the experiment is unusual and not typical of chance behavior, your experiment's results probably did not happen by chance. The results probably occurred because of a statistically significant difference in the two groups.

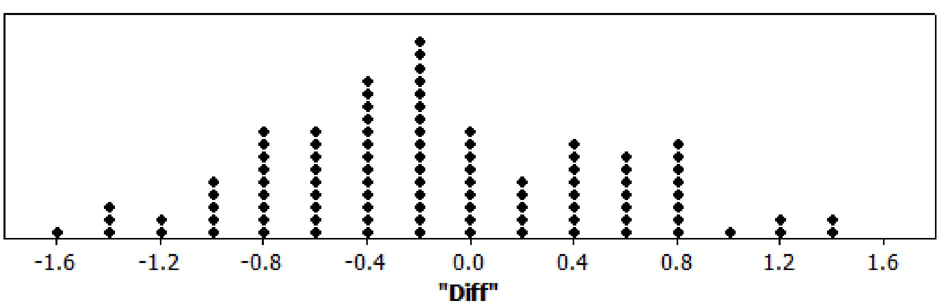
<u>**Not a small probability:**</u> If the "Diff" value from the experiment is NOT considered unusual and could be typical of chance behavior, your experiment's results may have just happened by chance and NOT because of a statistically significant difference in the two groups.

Note: The use of technology is strongly encouraged to assist in Steps 2 – 4.

Problem Set

1. Using the 20 observations that appear in "Exit Ticket" Question 2 for the "changes in pain" scores of 20 individuals, use the "Anova Shuffle" applet to develop a randomization distribution of the value "Diff" ($\overline{x}_A - \overline{x}_B$) based on 100 random assignments of these 20 observations into two groups of 10. Enter the data exactly as shown in Question 2. Describe similarities and differences between this new randomization distribution and the distribution shown in "Exit Ticket" Question 3.

Group	ChangeinScore
A	0
A	0
A	-1
A	-1
A	-2
A	-2
A	-3
A	-3
A	-3
A	-4
B	0
B	0
B	0
B	0
B	0
B	0
B	-1
B	-1
B	-1
B	-2



2. In a previous lesson, the burn times of 6 candles were presented. It is believed that candles from Group A will burn longer on average than candles from Group B. The data from the experiment (now shown with group identifiers) are provided below.

Group	Burntime
A	18
A	12
B	9
A	6
B	3
B	0

Perform a randomization test of this claim. Carry out all 5 steps, and use the "Anova Shuffle" applet to perform Steps 2 to 4. Enter the data exactly as presented above, and in Step 3, develop the randomization distribution based on 200 random assignments.

Lesson 28: Drawing a Conclusion from an Experiment

Classwork

In this lesson, you will be conducting all phases of an experiment: collecting data, creating a randomization distribution based on these data, and determining if there is a significant difference in treatment effects. In the next lesson, you will develop a report of your findings.

The following experiments are in homage to George E. P. Box, a famous statistician who worked extensively in the areas of quality control, design of experiments, and other topics. He earned the honor of Fellow of the Royal Society during his career and is a former president of the American Statistical Association. Several resources are available regarding his work and life including the book *Statistics for Experimenters: Design, Innovation, and Discovery* by Box, Hunter, and Hunter.

The experiments will investigate whether modifications in certain dimensions of a paper helicopter will affect its flight time.

Exercise 1: Build the Helicopters

In preparation for your data collection, you will need to construct 20 paper helicopters following the blueprint given at the end of this lesson. For consistency, use the same type of paper for each helicopter. For greater stability, you may want to use a piece of tape to secure the two folded body panels to the body of the helicopter. By design, there will be some overlap from this folding in some helicopters.

You will carry out an experiment to investigate the effect of wing length on flight time.

a. Construct 20 helicopters with wing length $=$ 4 inches, body length $=$ 3 inches. Label 10 each of these helicopters with the word "long."

b. Take the other 10 helicopters and cut one inch off each of the wings so that you have 10 helicopters with 3 inch wings. Label each of these helicopters with the word “short.”

c. How do you think wing length will affect flight time? Explain your answer.

Exploratory Challenge 1: Data Collection

Once you have built the 20 helicopters, each of them will be "flown" by dropping the helicopter from a fixed distance above the ground (preferably 12 feet or higher—record this height for use when presenting your findings later). For consistency, drop all helicopters from the same height each time, and try to perform this exercise in a space where possible confounding factors such as wind gusts and drafts from heating and air conditioning are eliminated.

a. Place the 20 helicopters in a bag, shake the bag, and randomly pull out one helicopter. Drop the helicopter from the starting height, and, using a stopwatch, record the amount of time it takes until the helicopter reaches the ground. Write down this "flight time" in the appropriate column in the table below. Repeat for the remaining 19 helicopters.

   Some helicopters might fly more smoothly than others; you may want to record relevant comments in your report.

Flight Time (seconds)

Long Wings	**Short Wings**

b. Why might it be important to randomize (impartially select) the order in which the helicopters were dropped?

   (This is *different from* the randomization you perform later when you are allocating observations to groups to develop the randomization distribution.)

Exploratory Challenge 2: Developing Claims and Using Technology

With the data in hand, you will now perform your analysis regarding the effect of “wing length.”

Experiment: Wing Length

In this experiment, you'll examine whether wing length makes a difference in flight time. You will compare the helicopters with long wings (wing length $= 4$ inches) to the helicopters with short wings (wing length $= 3$ inches). Since you are dropping the helicopters from the same height in the same location, using the same type of paper, the only difference in the two groups will be the different wing lengths.

Questions: Does a 1-inch addition in wing length appear to result in a change in average flight time? If so, do helicopters with longer wing length or shorter wing length tend to have longer flight times on average?

Carry out a complete randomization test to answer these questions. Show all 5 steps and use the “Anova Shuffle” Applet described in the previous lessons to assist both in creating the distribution and with your computations. Be sure to write a final conclusion that clearly answers the questions in context.

Lesson Summary

In previous lessons, you learned how to carry out a randomization test to decide if there was a statistically significant difference between two groups in an experiment. Throughout these previous lessons, certain aspects of proper experimental design were discussed. In this lesson, you were able to carry out a complete experiment and collect your own data. When an experiment is developed, you must be careful to minimize confounding effects that may compromise or invalidate findings. When possible, the treatment groups should be created so that the only distinction between the groups in the experiment is the treatment imposed.

Problem Set

One other variable that can be adjusted in the paper helicopters is "body width." See the blueprints for details.

1. Construct 10 helicopters using the blueprint from the lesson. Label each helicopter with the word "narrow."

2. Develop a blueprint for a helicopter that is identical to the helicopter of the blueprint used in class except for the fact that the body width will now be 1.75 inches.

3. Use the blueprint to construct 10 of these new helicopters, and label each of these helicopters with the word "wide."

4. Place the 20 helicopters in a bag, shake the bag, and randomly pull out one helicopter. Drop the helicopter from the starting height, and, using a stopwatch, record the amount of time it takes until the helicopter reaches the ground. Write down this "flight time" in the appropriate column in the table below. Repeat for the remaining 19 helicopters.

Flight Time (seconds)

Narrow Body	Wide Body

5. Question: Does a 0.5 inch addition in body width appear to result in a change in average flight time? If so, do helicopters with wider body width or narrower body width tend to have longer flight times on average? Carry out a complete randomization test to answer these questions. Show all 5 steps and use the "Anova Shuffle" Applet described in previous lessons to assist both in creating the distribution and with your computations. Be sure to write a final conclusion that clearly answers the questions in context.

Lesson 29: Drawing a Conclusion from an Experiment

Classwork

In this lesson, you will develop a comprehensive poster summarizing your experiments.

Characteristics of a Good Poster

Your instructor will provide you with specific instructions and a rubric for assessing your poster (taken from "Poster Judging Rubric" at the "Poster Competition and Project Competition" page of the American Statistical Association, www.amstat.org/education/posterprojects/pdfs/PosterJudgingRubric.pdf).

Generally speaking, the presentation of a statistical analysis and/or experiment should clearly state the question or purpose. The presentation should lead to the conclusion on a path that is easy to follow. The results of the study should be immediately obvious to the viewer. Any graphs included should be relevant to the question of interest and appropriate for the type of data collected.

Exploratory Challenge: Explaining the Experiment and Results

Your classwork will involve developing your poster. Your instructor will provide guidance as to groups, amount of time to spend, the rubric to be used for evaluation, etc. Your poster should address the results of both Experiments 1 and 2 regarding the effects of both body width and wing length.

In addition to the general concerns of colors, fonts to use, etc., in preparation for creating your poster, consider (and answer) these classwork questions:

- What was the objective of the experiment?
- How did you collect your data?
- What summary values and graphs should you present?
- How will you develop and present a summary of the experiment in a way that it is easy to follow and effortlessly leads the viewer to the conclusion?
- How will you explain "statistical significance"?

Lesson 30: Evaluating Reports Based on Data from an Experiment

Classwork

Exercises 1-7

Pericarditis is an inflammation (irritation and swelling) of the pericardium, the thin sac that surrounds the heart. When extra fluid builds up between the two layers of the pericardium, the heart's actions are restricted. An experiment reported in the article "A Randomized Trial of Colchicine for Acute Pericarditis" in *The New England Journal of Medicine* (October 2013) tested the effects of the drug colchicine on acute pericarditis.

Read the summary of the article, and answer the following questions.

Website: www.nejm.org/doi/full/10.1056/NEJMoa1208536

1. How many treatment groups are there?

2. What treatments are being compared?

3. Is there a placebo group? Explain

4. How many subjects are in each treatment group?

5. Do you think that the number of subjects in each treatment is enough? Explain.

6. What method was used to assign the subjects to the treatment groups? Explain why this is important.

Suppose newspaper reporters brainstormed some headlines for an article on this experiment. These are their suggested headlines:

A. "New Treatment Helps Pericarditis Patients"

B. "Colchicine Tends to Improve Treatment for Pericarditis"

C. "Pericarditis Patients May Get Help"

7. Which of the headlines above would be best to use for the article? Explain why.

Exercises 8-10

What you should look for when evaluating an experiment . . .

- Were the subjects randomly assigned to treatment groups?
- Was there a control group or a comparison group?
- Were the sample sizes reasonable large?
- Do the results show a cause and effect relationship?

Read the summary of the two articles below. Write a few sentences evaluating these articles using the guidelines above.

8. The study "Semantic Memory Functional MRI and Cognitive Function after Exercise Intervention in Mild Cognitive Impairment" (*Journal of Alzheimer's Disease*, November 2013) was performed to see if exercise would increase memory retrieval in older adults with mild cognitive impairment (associated with early memory loss).

   Website: http://iospress.metapress.com/content/xm8t241628h37h7t/

9. The article "Effects of Bracing in Adolescents with Idiopathic Scoliosis" (*New England Journal of Medicine*, October 2013) reports on the role of bracing patients with adolescent idiopathic scoliosis (curvature of the spine) for prevention of back surgery.

   Website: www.nejm.org/doi/full/10.1056/NEJMoa1307337

10. View the report by Tom Bemis (Market Watch, *Wall Street Journal*, August 13, 2013) about the type of car driven by a person and the person's driving behavior.

    Website: http://live.wsj.com/video/bmw-drivers-really-are-jerks-studies-find/29285015-BB1A-4E41-B0C0-0A41CB990F60.html#!29285015-BB1A-4E41-B0C0-0A41CB990F60

    Is the title "BMW Drivers Really Are Jerks" an accurate title for these reported studies? Why or why not? If not, suggest a better title.

Lesson Summary

- A cause and effect relationship can only be shown by a well-designed experiment.
- Randomly assigning the subjects to treatment groups evens out the effects of extraneous variables to create comparable treatment groups.
- A control group (which may be a placebo group) or a comparison group (a standard treatment) is sometimes included in an experiment so that you can evaluate the effect of the treatment.
- The number of subjects in each treatment (sample size) should be large enough for the random assignment to experimental groups to create groups with comparable variability between the subjects.

Problem Set

Read the following articles and summaries. Write a few sentences evaluating each one using the guidelines given in the lesson.

1. The article "Emerging Technology" (*Discover Magazine*, November 2005) reports a study on the effect of "infomania" on IQ scores.

   Website: discovermagazine.com/2005/nov/emerging-technology

2. In *The New England Journal of Medicine*, October 2013, the article "Increased Survival in Pancreatic Cancer With nab-Paclitaxel Plus Gemcitabine" reports on an experiment to test which treatment, nab-paclitaxel plus gemcitabine or gemcitabine alone, is the most effective in treating advanced pancreatic cancer.

   Website: www.nejm.org/doi/full/10.1056/NEJMoa1304369

3. Doctors conducted a randomized trial of hypothermia in infants with a gestational age of at least 36 weeks who were admitted to the hospital at or before six hours of age with either severe acidosis or perinatal complications and resuscitation at birth and who had moderate or severe encephalopathy. The trial, "Whole-Body Hypothermia for Neonates with Hypoxic–Ischemic Encephalopathy," tested two treatments, standard care or whole-body cooling for 72 hours.

   Website: www.nejm.org/doi/full/10.1056/NEJMcps050929