Lesson 1: Posing Statistical Questions

Statistics is about using data to answer questions. In this module, the following four steps will summarize your work with data:

Step 1: Pose a question that can be answered by data.

Step 2: Determine a plan to collect the data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the data and summaries.

You will be guided through this process as you study these lessons. This first lesson is about the first step – what is a statistical question, and what does it mean that a question can be answered by data?

Classwork

Example 1: What is a Statistical Question?

Jerome, a 6th grader at Roosevelt Middle School, is a huge baseball fan. He loves to collect baseball cards. He has cards of current players and of players from past baseball seasons. With his teacher's permission, Jerome brought his baseball card collection to school. Each card has a picture of a current or past major league baseball player, along with information about the player. When he placed his cards out for the other students to see, they asked Jerome all sorts of questions about his cards. Some asked:

- How many cards does Jerome have altogether?
- What is the typical cost of a card in Jerome's collection?
- Where did Jerome get the cards?

Exercises 1–5

1. For each of the following, determine whether or not the question is a statistical question. Give a reason for your answer.
   a. Who is my favorite movie star?

   b. What are the favorite colors of 6th graders in my school?

c. How many years have students in my school's band or orchestra played an instrument?

d. What is the favorite subject of 6th graders at my school?

e. How many brothers and sisters does my best friend have?

2. Explain why each of the following questions is not a statistical question.
   a. How old am I?

   b. What's my favorite color?

   c. How old is the principal at our school?

3. Ronnie, a 6th grader, wanted to find out if he lived the farthest from school. Write a statistical question that would help Ronnie find the answer.

4. Write a statistical question that can be answered by collecting data from students in your class.

5. Change the following question to make it a statistical question: "How old is my math teacher?"

Example 2: Types of Data

We use two types of data to answer statistical questions: numerical data and categorical data. If we recorded the age of 25 baseball cards, we would have numerical data. Each value in a numerical data set is a number. If we recorded the team of the featured player for 25 baseball cards, you would have categorical data. Although you still have 25 data values, the data values are not numbers. They would be team names, which you can think of as categories.

Exercises 6–7

6. Identify each of the following data sets as categorical (C) or numerical (N).

   a. Heights of 20 6^{th} graders _____

   b. Favorite flavor of ice cream for each of 10 6^{th} graders _____

   c. Hours of sleep on a school night for 30 6^{th} graders _____

   d. Type of beverage drank at lunch for each of 15 6^{th} graders _____

   e. Eye color for each of 30 6^{th} graders _____

   f. Number of pencils in each desk of 15 6^{th} graders _____

7. For each of the following statistical questions, students asked Jerome to identify whether the data are numerical or categorical. Explain your answer, and list four possible data values.

   a. How old are the cards in the collection?

   b. How much did the cards in the collection cost?

   c. Where did you get the cards?

Lesson Summary

A **statistical question** is one that can be answered by collecting data that vary (i.e., not all of the data values are the same).

There are two types of data: numerical and categorical. In a **numerical data set**, every value in the set is a number. **Categorical data sets** can take on non-numerical values, such as names of colors, labels, etc. (e.g., "large," "medium," or "small").

Statistics is about using data to answer questions. In this module, the following 4 steps will summarize your work with data:

Step 1: Pose a question that can be answered by data.

Step 2: Determine a plan to collect the data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the data and the summaries.

Problem Set

1. For each of the following, determine whether the question is a statistical question. Give a reason for your answer.
   a. How many letters are in my last name?
   b. How many letters are in the last names of the students in my 6th grade class?
   c. What are the colors of the shoes worn by the students in my school?
   d. What is the maximum number of feet that roller coasters drop during a ride?
   e. What are the heart rates of the students in a 6th grade class?
   f. How many hours of sleep per night do 6th graders usually get when they have school the next day?
   g. How many miles per gallon do compact cars get?

2. Identify each of the following data sets as categorical (C) or numerical (N). Explain your answer.
   a. Arm spans of 12 6^{th} graders
   b. Number of languages spoken by each of 20 adults
   c. Favorite sport of each person in a group of 20 adults
   d. Number of pets for each of 40 3^{rd} graders
   e. Number of hours a week spent reading a book for a group of middle school students

3. Rewrite each of the following questions as a statistical question.
   a. How many pets does your teacher have?
   b. How many points did the high school soccer team score in its last game?
   c. How many pages are in our math book?
   d. Can I do a handstand?

4. Write a statistical question that would be answered by collecting data from the 6th graders in your classroom.

5. Are the data you would collect to answer that question categorical or numerical? Explain your answer.

Lesson 2: Displaying a Data Distribution

Classwork

Example 1: Heart Rate

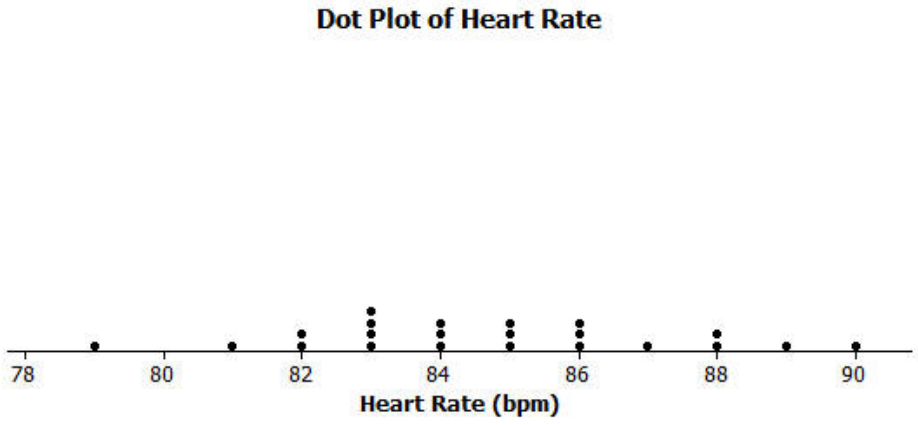
Mia, a 6th grader at Roosevelt Middle School, was thinking about joining the middle school track team. She read that Olympic athletes have lower resting heart rates than most people. She wondered about her own heart rate and how it would compare to other students. Mia was interested in investigating the statistical question: "What are the heart rates of the students in my 6th grade class?"

Heart rates are expressed as bpm (or beats per minute). Mia knew her resting heart rate was 80 beats per minute. She asked her teacher if she could collect the heart rates of the other students in her class. With the teacher's help, the other 6th graders in her class found their heart rates and reported them to Mia. Following are the heart rates (in beats per minute) for the 22 other students in Mia's class:

89 87 85 84 90 79 83 85 86 88 84 81 88 85 83 83 86 82 83 86 82 84

To learn about the heart rates, a good place to start is to make a graph of the data. There are several different graphs that could be used, including the three types of graphs that you will learn in this module: dot plots, histograms, and box plots. In this lesson, you will learn about dot plots.

Mia noticed that there were many different heart rates. She decided to make a *dot plot* to show the different heart rates. She drew a number line and started numbering from 78 to 92. She then placed a dot above the number on the number line for each heart rate. If there was already a dot above a number she added another dot above the one already there. She continued until she had added one dot for each heart rate.



Exercises 1–10

1. What was the heart rate for the student with the lowest heart rate?

2. What was the heart rate for the student with the highest heart rate?

3. How many students had a heart rate greater than 86?

4. What fraction of the students had a heart rate less than 82?

5. What is the most common heart rate?

6. What heart rate describes the center of the data?

7. What heart rates are the most unusual heart rates?

8. If Mia's teacher asked what the typical heart rate is for 6^{th} graders in the class, what would you tell Mia's teacher?

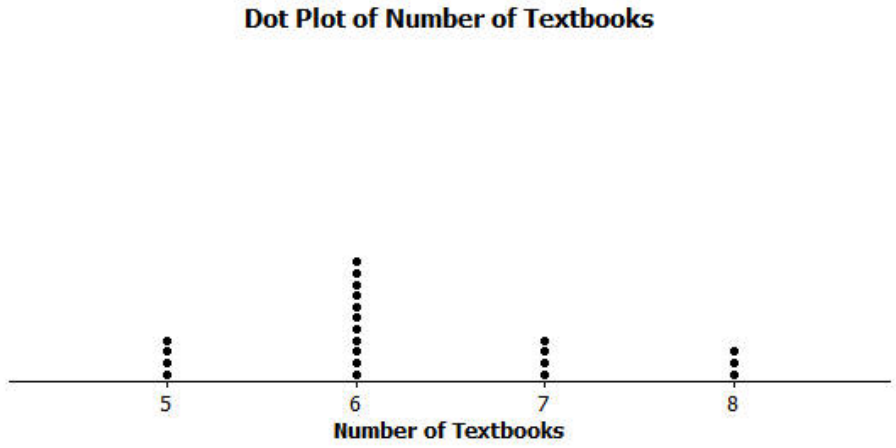
9. On the dot plot add a dot for Mia's heart rate.

10. How does Mia's heart rate compare with the heart rates of the other students in the class?

Example 2: Seeing the Spread in Dot Plots

Mia's class collected data to answer several other questions about her class. After they collected data, they drew dot plots of their findings.

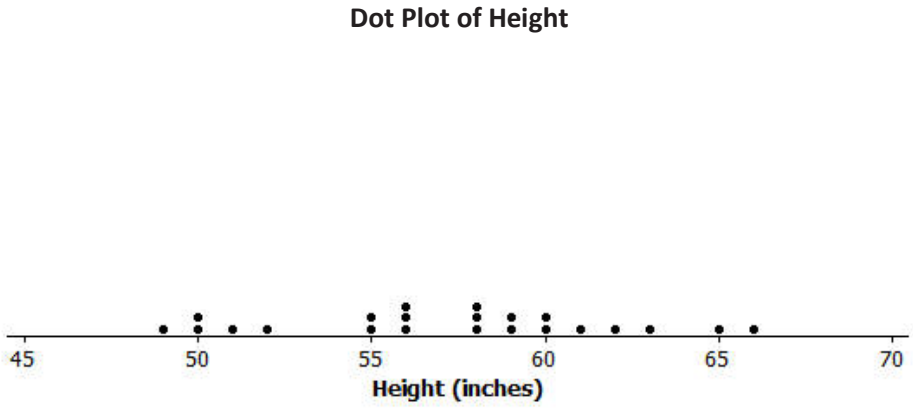
Here is a dot plot showing the data collected to answer the question: "How many textbooks are in the desks of 6th graders?"



When the students thought about this question, many said that they all had about the same number of books in their desk since they all take the same subjects in school.

The class noticed that the graph was not very spread out since there were only four different answers that students gave, with most of the students answering that they had 6 books in their desk.

Another student wanted to ask the question: "How tall are the 6th graders in our class?" When students thought about this question, they thought that the heights would be spread out since there were some shorter students and some very tall students in class. Here is a dot plot of the students' heights:



Exercises 11–14

Listed are four statistical questions and four different dot plots of data collected to answer these questions. Match each statistical question with the appropriate dot plot. Explain each of your choices.

Statistical Question:

11. What are the ages of 4th graders in our school?

12. What are the heights of the players on the 8th grade boys' basketball team?

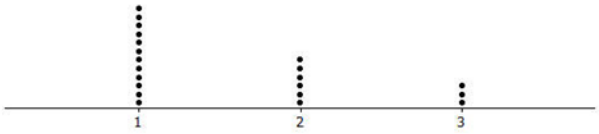
13. How many hours do 6th graders in our class watch TV on a school night?

14. How many different languages do students in our class speak?

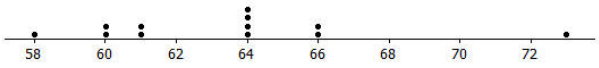
Dot plot A

Dot plot B

Dot plot C



Dot plot D



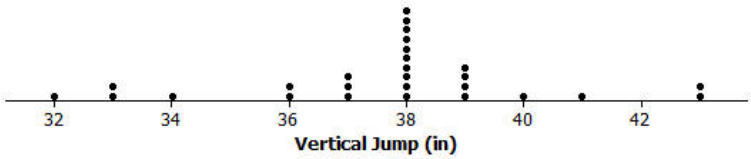
Lesson Summary

In this lesson, numerical data collected to answer a statistical question were shown in a *dot plot*. In a dot plot, a data value is represented by a dot over a number line. The number of dots over the number line at a particular value tells how many of the data points have that value. A dot plot can help you find the smallest and largest values, see how spread out the data are, and see where the center of the data is.

Problem Set

1. The dot plot below shows the vertical jump of some NBA players. A vertical jump is how high a player can jump from a standstill.

Dot Plot of Vertical Jump



a. What statistical question do you think could be answered using these data?

b. What was the highest vertical jump by a player?

c. What was the lowest vertical jump by a player?

d. What is the most common vertical jump?

e. How many players jumped that high?

f. How many players jumped higher than 40 inches?

g. Another NBA player jumped 33 inches. Add a dot for this player on the dot plot. How does this player compare with the other players?

2. Listed are two statistical questions and two different dot plots of data collected to answer these questions. Match each statistical question with its dot plot. Explain each of your choices.

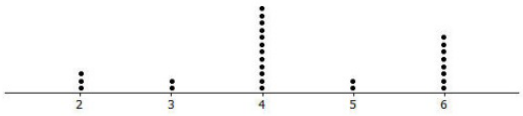
   Statistical questions:

   a. What is the number fish (if any) that students in class have in an aquarium at their home?

   b. How many pockets do the 6th graders have in the pants that they are wearing at school on a particular day?

Dot Plot A

Dot Plot B



3. Read each of the following statistical questions. Write a description of what the dot plot of the data collected to answer the question might look like. Your description should include a description of the spread of the data and the center of the data.

   a. What is the number of hours 6th grade students are in school during a typical school day?

   b. What is the number of video games owned by the 6th graders in our class?

Lesson 3: Creating a Dot Plot

Classwork

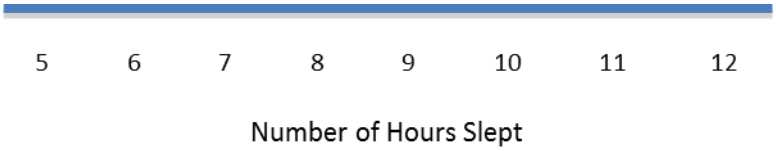
Example 1: Hours of Sleep

Robert, a 6th grader at Roosevelt Middle School, usually goes to bed around 10:00 p.m. and gets up around 6:00 a.m. to get ready for school. That means that he gets about 8 hours of sleep on a school night. He decided to investigate the statistical question: How many hours per night do 6th graders usually sleep when they have school the next day?

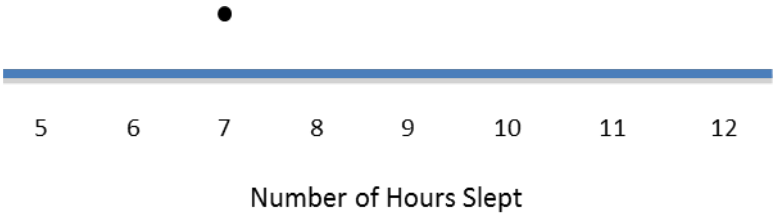
Robert took a survey of 29 6th graders and collected the following data to answer the question:

7 8 5 9 9 9 7 7 10 10 11 9 8 8 8 12 6 11 10 8 8 9 9 9 8 10 9 9 8

Robert decided to make a dot plot of the data to help him answer his statistical question. Robert first drew a number line and labeled it from 5 to 12 to match the lowest and highest number of hours slept.



He then placed a dot above 7 for the first piece of data he collected. He continued to place dots above the numbers until each number was represented by a dot.



Exercises 1–9

1. Complete Robert's dot plot by placing a dot above the number on the number line for each number of hours slept. If there is already a dot above a number, then add another dot above the dot already there.

2. What are the least and the most hours of sleep reported in the survey of 6th graders?

3. What is the most common number of hours slept?

4. How many hours of sleep describes the center of the data?

5. Think about how many hours of sleep you usually get on a school night. How does your number compare with the number of hours of sleep from the survey of 6th graders?

Here are the data for the number of hours 6th graders sleep when they don't have school the next day:

7 8 10 11 5 6 12 13 13 7 9 8 10 12 11 12 8 9 10 11 10 12 11 11 11 12 11 11 10

6. Make a dot plot of the number of hours slept when there is no school the next day.

7. How many hours of sleep with no school the next day describe the center of the data?

8. What are the least and most hours slept with no school the next day reported in the survey?

9. Do students sleep longer when they don't have school the next day than they do when they do have school the next day? Explain your answer using the data in both dot plots.

Example 2: Building and Interpreting a Frequency Table

A group of 6th graders investigated the statistical question: "How many hours per week do 6th graders spend playing a sport or outdoor game?"

Here are the data the students collected from a sample of 26 6th graders showing the number of hours per week spent playing a sport or a game outdoors:

3 2 0 6 3 3 3 1 1 2 2 8 12 4 4 4 3 3 1 1 0 0 6 2 3 2

To help organize the data, the students placed the number of hours into a frequency table. A frequency table lists items and how often each item occurs.

To build a frequency table, first draw three columns. Label one column "Number of Hours Playing a Sport/Game," label the second column "Tally," and the third column "Frequency." Since the least number of hours was 0, and the most was 12, list the numbers from 0 to 12 under the "Number of Hours" column.

Number of Hours Playing a Sport/Game	Tally	Frequency
0		
1		
2		
3	|	
4		
5		
6		
7		
8		
9		
10		
11		
12		

As you read each number of hours from the survey, place a tally mark opposite that number. The table shows a tally mark for the first number 3.

Exercises 10–15

10. Complete the tally mark column.

Number of hours	Tally	Frequency
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

11. For each number of hours, find the total number of tally marks and place this in the frequency column.

12. Make a dot plot of the number of hours playing a sport or playing outdoors.

13. What number of hours describes the center of the data?

14. How many 6th graders reported that they spend eight or more hours a week playing a sport or playing outdoors?

15. The 6th graders wanted to answer the question, "How many hours do 6th graders spend per week playing a sport or playing an outdoor game?" Using the frequency table and the dot plot, how would you answer the 6th graders question?

Lesson Summary

This lesson described how to make a *dot plot.* This plot starts with a number line labeled from the smallest to the largest value. Then, a dot is placed above the number on the number line for each value in your data.

This lesson also described how to make a *frequency table.* A frequency table consists of three columns. The first column contains all the values of the data listed in order from smallest to largest. The second column is the tally column, and the third column is the number of tallies for each data value.

Problem Set

1. The data below is the number of goals scored by a professional indoor soccer team over their last 23 games.

   8 16 10 9 11 11 10 15 16 11 15 13 8 9 11 9 8 11 16 15 10 9 12

   a. Make a dot plot of the number of goals scored.
   b. What number of goals describes the center of the data?
   c. What is the least and most number of goals scored by the team?
   d. Over the 23 games played, the team lost 10 games. Circle the dots on the plot that you think represent the games that the team lost. Explain your answer.

2. A 6th grader rolled two number cubes 21 times. The student found the sum of the two numbers that he rolled each time. The following are the sums of the 21 rolls of the two number cubes:

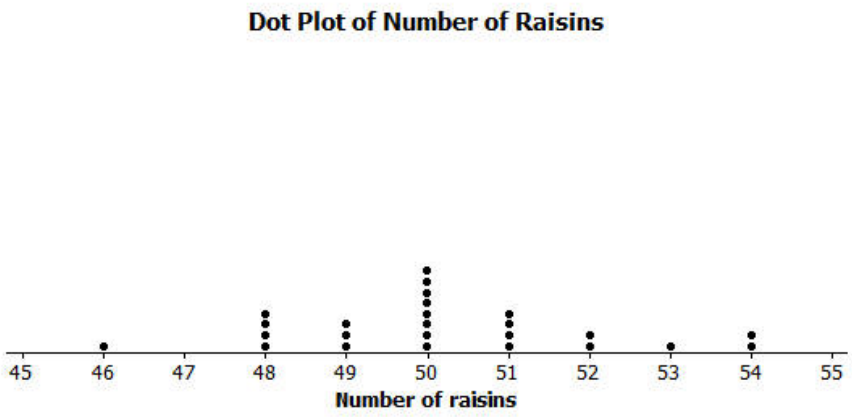
   9 2 4 6 5 7 8 11 9 4 6 5 7 7 8 8 7 5 7 6 6

   a. Complete the frequency table.

Sum rolled	Tally	Frequency
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

   b. What sum describes the center of the data?
   c. What was the most common sum of the number cubes?

3. The dot plot below shows the number of raisins in 25 selected small boxes of raisins.



a. Complete the frequency table.

Number of Raisins	Tally	Frequency
46		
47		
48		
49		
50		
51		
52		
53		
54		

b. Another student opened up a box of raisins and reported that it had 63 raisins. Did this student have the same size box of raisins? Why or why not?

Lesson 4: Creating a Histogram

Classwork

Example 1: Frequency Table with Intervals

The boys and girls basketball teams at Roosevelt Middle School wanted to raise money to help buy new uniforms. They decided to sell hats with the school logo on the front to family members and other interested fans. To obtain the correct hat size, the students had to measure the head circumference (distance around the head) of the adults who wanted to order a hat. The following data represents the head circumferences, in millimeters (mm), of the adults:

$513, 525, 531, 533, 535, 535, 542, 543, 546, 549, 551, 552, 552, 553, 554, 555, 560, 561, 563, 563, 563, 565,$

$565, 568, 568, 571, 571, 574, 577, 580, 583, 583, 584, 585, 591, 595, 598, 603, 612, 618$

The hats come in six sizes: XS, S, M, L, XL, and XXL. Each hat size covers a span of head circumferences. The hat manufacturer gave the students the table below that shows the interval of head circumferences for each hat size. The interval $510 - < 530$ represents head circumferences from 510 to 530, not including 530.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	$510 - < 530$		
S	$530 - < 550$		
M	$550 - < 570$		
L	$570 - < 590$		
XL	$590 - < 610$		
XXL	$610 - < 630$		

Exercises 1–4

1. If someone has a head circumference of 570, what size hat would they need?

2. Complete the tally and frequency columns in the table to determine the number of each size hat the students need to order for the adults who wanted to order a hat.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	$510 - < 530$		
S	$530 - < 550$		
M	$550 - < 570$		
L	$570 - < 590$		
XL	$590 - < 610$		
XXL	$610 - < 630$	||	2

3. What hat size does the data center around?

4. Describe any patterns that you observe in the frequency column?

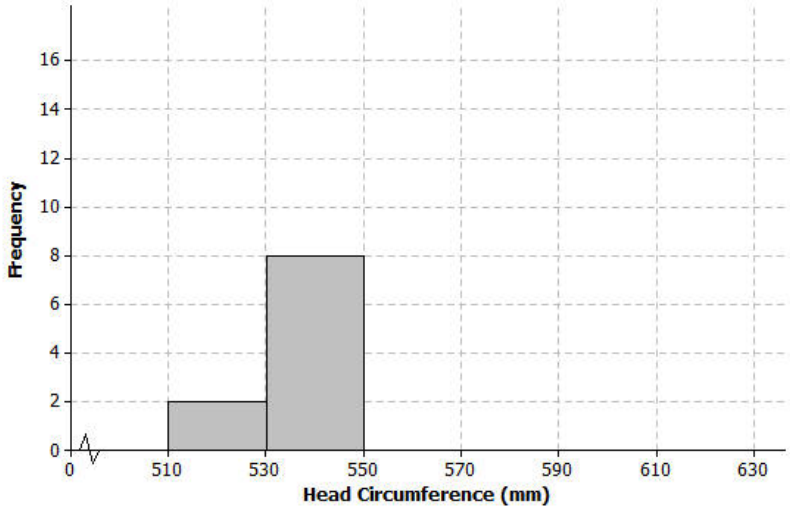
Example 2: Histogram

One student looked at the tally column and said that it looked somewhat like a bar graph turned on its side. A histogram is a graph that is like a bar graph, except that the horizontal axis is a number line that is marked off in equal intervals.

To make a histogram:

- Draw a horizontal line and mark the intervals.
- Draw a vertical line and label it "frequency."
- Mark the frequency axis with a scale that starts at 0 and goes up to something that is greater than the largest frequency in the frequency table.
- For each interval, draw a bar over that interval that has a height equal to the frequency for that interval.

The first two bars of the histogram have been drawn below.



Exercises 5–9

5. Complete the histogram by drawing bars whose heights are the frequencies for those intervals.

6. Based on the histogram, describe the center of the head circumferences.

7. How would the histogram change if you added head circumferences of 551 and 569?

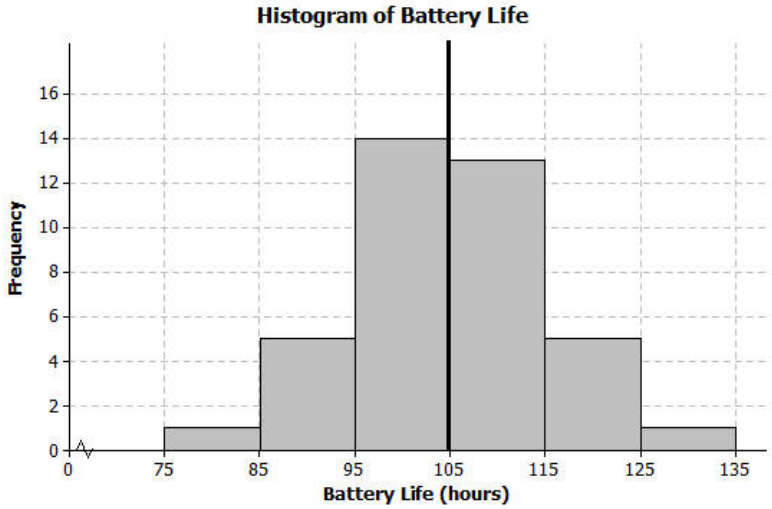
8. Because the 40 head circumference values were given, you could have constructed a dot plot to display the head circumference data. What information is lost when a histogram is used to represent a data distribution instead of a dot plot?

9. Suppose that there had been 200 head circumference measurements in the data set. Explain why you might prefer to summarize this data set using a histogram rather than a dot plot.

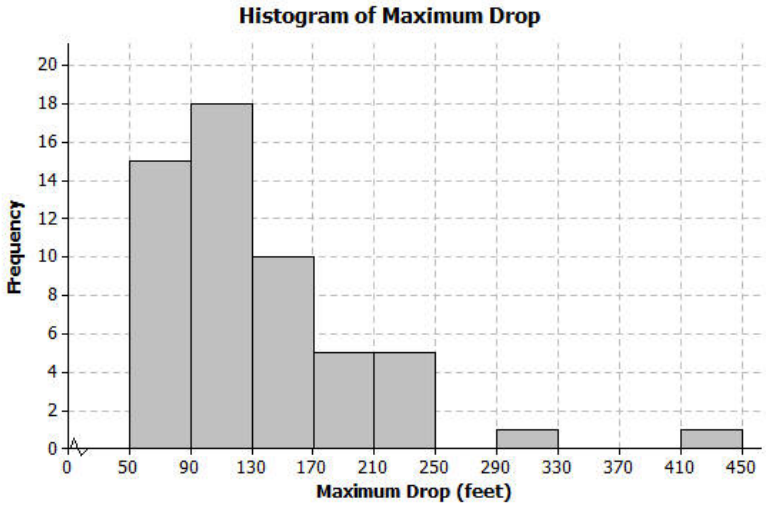
Example 3: Shape of the Histogram

A histogram is useful to describe the shape of the data distribution. It is important to think about the shape of a data distribution because depending on the shape, there are different ways to describe important features of the distribution, such as center and variability.

A group of students wanted to find out how long a certain brand of AA batteries lasted. The histogram below shows the data distribution for how long (in hours) that some AA batteries lasted. Looking at the shape of the histogram, notice how the data "mounds" up around a center of approximately 105. We would describe this shape as mound shaped or symmetric. If we were to draw a line down the center, notice how each side of the histogram is approximately the same or mirror images of each other. This means the graph is approximately symmetrical.

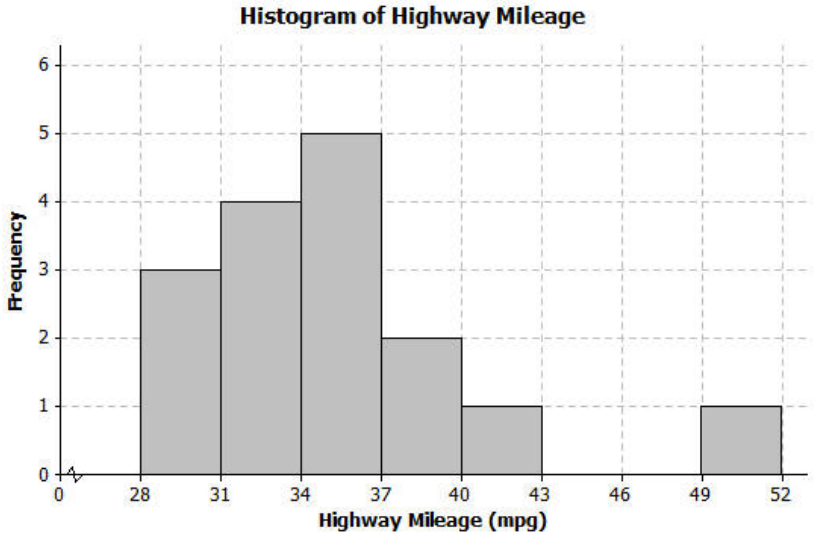


Another group of students wanted to investigate the maximum drop length for roller coasters. The histogram below shows the maximum drop (in feet) of a selected group of roller coasters. This histogram has a skewed shape. Most of the data are in the intervals from 50 to 170. But there are two values that are unusual (or not typical) when compared to the rest of the data. These values are much higher than most of the data.



Exercises 10–12

10. The histogram below shows the highway miles per gallon of different compact cars.



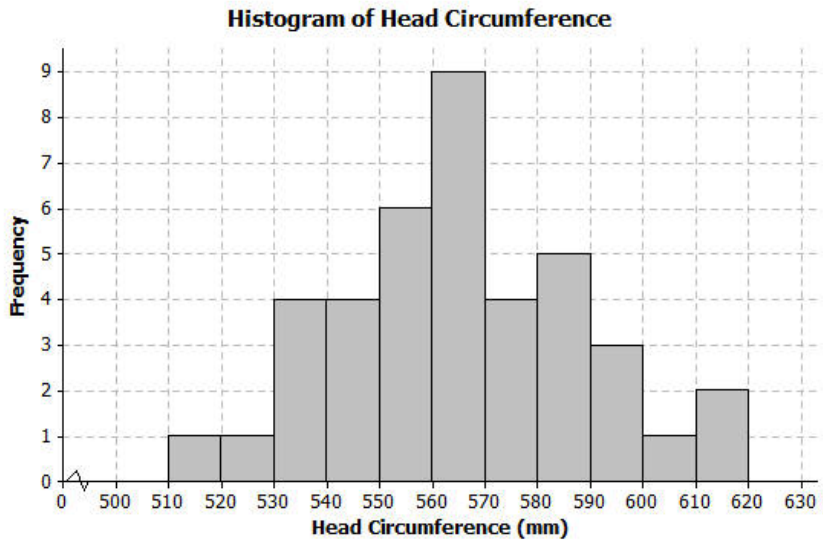
a. Describe the shape of the histogram as approximately symmetric, skewed left, or skewed right.

b. Draw a vertical line on the histogram to show where the "typical" number of miles per gallon for a compact car would be.

c. What does the shape of the histogram tell you about miles per gallon for compact cars?

11. Describe the shape of the head circumference histogram that you completed in Exercise 5 as approximately symmetric, skewed left, or skewed right.

12. Another student decided to organize the head circumference data by changing the width of each interval to be 10 instead of 20. Below is the histogram that the student made.



a. How does this histogram compare with the histogram of the head circumferences that you completed in Exercise 5?

b. Describe the shape of this new histogram as approximately symmetric, skewed left, or skewed right.

c. How many head circumferences are in the interval from 570 to 590?

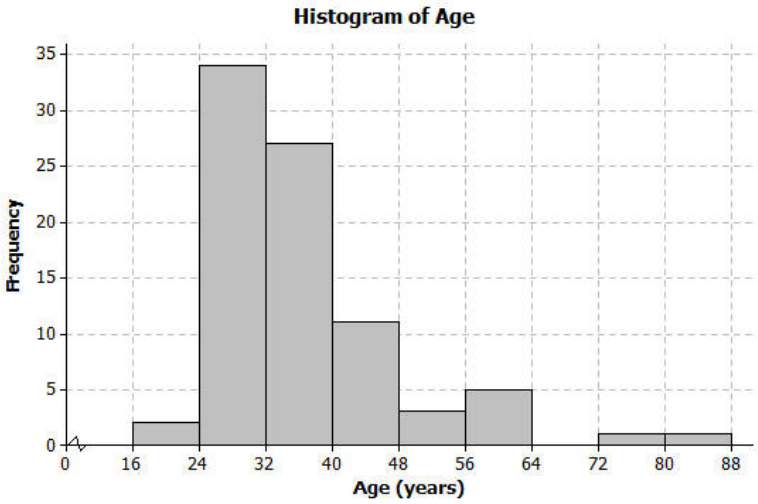
d. In what interval would a head circumference of 571 be included? In what interval would a head circumference of 610 be included?

Lesson Summary

A histogram is a graph that represents the number of data values falling in an interval with a bar. The horizontal axis shows the intervals and the vertical axis shows the frequencies (how many data values are in the interval). Each interval should be the same width, and the bars should touch each other.

Problem Set

1. The following histogram shows ages of the actresses whose performances have won in the Best Leading Actress category at the annual Academy Awards (Oscars).



   a. Which age interval contains the most actresses? How many actresses are represented in that interval?
   b. Describe the shape of the histogram.
   c. What does the shape tell you about the ages of actresses who win the Oscar for best actress award?
   d. Which interval describes the center of the ages of the actresses?
   e. An age of 72 would be included in which interval?

2. The frequency table below shows the seating capacity of arenas for NBA basketball teams

Number of Seats	Tally	Frequency
$17000 - < 17500$	||	2
$17500 - < 18000$	|	1
$18000 - < 18500$	$\cancel{\text{IIII}}$ |	6
$18500 - < 19000$	$\cancel{\text{IIII}}$	5
$19000 - < 19500$	$\cancel{\text{IIII}}$	5
$19500 - < 20000$	$\cancel{\text{IIII}}$	5
$20000 - < 20500$	||	2
$20500 - < 21000$	||	2
$21000 - < 21500$		0
$21500 - < 22000$		0
$22000 - < 22500$	|	1

a. Draw a histogram of the number of seats in NBA arenas. Use the histograms you have seen throughout this lesson to help you in the construction of your histogram.

b. What is the width of each interval? How do you know?

c. Describe the shape of the histogram.

d. Which interval describes the center of the number of seats?

3. Listed are the grams of carbohydrates in hamburgers at selected fast food restaurants.

33 40 66 45 28 30 52 40 26 42
42 44 33 44 45 32 45 45 52 24

a. Complete the frequency table with intervals of width 5.

Number of Carbohydrates (grams)	Tally	Frequency
$20 - < 25$		
$25 - < 30$		
$30 - < 35$		
$35 - < 40$		
$40 - < 45$		
$45 - < 50$		
$50 - < 55$		
$55 - < 60$		
$60 - < 65$		
$65 - < 70$		

b. Draw a histogram of the carbohydrate data.

c. Describe the center and shape of the histogram.

d. In the frequency table below, the intervals are changed. Using the carbohydrate data above, complete the frequency table with intervals of width 10.

Number of Carbohydrates (grams)	Tally	Frequency
$20 - < 30$		
$30 - < 40$		
$40 - < 50$		
$50 - < 60$		
$60 - < 70$		

e. Draw a histogram.

4. Use the histograms that you constructed in question 3 parts (b) and (e) to answer the following questions.
   a. Why are there fewer bars in the histogram in question 3 part (e) than the histogram in part (b)?
   b. Did the shape of the histogram in question 3 part (e) change from the shape of the histogram in part (b)?
   c. Did your estimate of the center change from the histogram in question 3 part (b) to the histogram in part (e)?

Lesson 5: Describing a Distribution Displayed in a Histogram

Classwork

Example 1: Relative Frequency Table

In Lesson 4, we investigated the head circumferences that the boys and girls basketball teams collected. Below is the frequency table of the head circumferences that they measured.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency
XS	$510 - < 530$	||	2
S	$530 - < 550$	𝍸 |||	8
M	$550 - < 570$	𝍸 𝍸 𝍸	15
L	$570 - < 590$	𝍸 ||||	9
XL	$590 - < 610$	||||	4
XXL	$610 - < 630$	||	2
		Total	40

Isabel, one of the basketball players, indicated that most of the hats were small, medium, or large. To decide if Isabel was correct, the players added a relative frequency column to the table. **Relative frequency** is the value of the frequency in an interval divided by the total number of data values.

Exercises 1–4

1. Complete the relative frequency column in the table below.

Hat Sizes	Interval of Head Circumferences (mm)	Tally	Frequency	Relative Frequency
XS	$510 - < 530$	||	2	$\frac{2}{40} = 0.05$
S	$530 - < 550$	𝍸 |||	8	$\frac{8}{40} = 0.20$
M	$550 - < 570$	𝍸 𝍸 𝍸	15	
L	$570 - < 590$	𝍸 ||||	9	
XL	$590 - < 610$	||||	4	
XXL	$610 - < 630$	||	2	
		Total	40	

2. What is the total of the relative frequency column?

3. Which interval has the greatest relative frequency? What is the value?

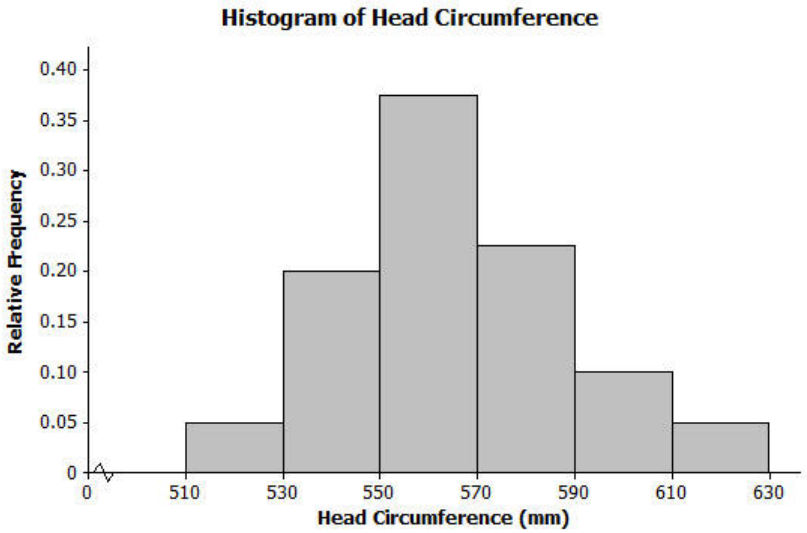
4. What percent of the head circumferences is between 530 and 589? Show how you determined the answer.

Example 2: Relative Frequency Histogram

The players decided to construct a histogram using the relative frequencies instead of the frequencies.

They noticed that the relative frequencies in the table ranged from close to 0 to about 0.40. They drew a number line and marked off the intervals on that line. Then, they drew the vertical line and labeled it relative frequency. They added a scale to this line by starting at 0 and counting by 0.05 until they reached 0.40.

They completed the histogram by drawing the bars so the height of each bar matched the relative frequency for that interval. Here is the completed relative frequency histogram:



Exercises 5–6

5. Answer the following questions.

   a. Describe the shape of the relative frequency histogram of head circumferences from Example 2.

   b. How does the shape of this histogram compare with the frequency histogram you drew in Exercise 5 of Lesson 4?

   c. Isabel said that most of the hats that needed to be ordered were small, medium, and large. Was she right? What percent of the hats to be ordered is small, medium, or large?

6. Here is the frequency table of the seating capacity of arenas for the NBA basketball teams.

Number of seats	Tally	Frequency	Relative Frequency
$17{,}000 - < 17{,}500$	||	2	
$17{,}500 - < 18{,}000$	|	1	
$18{,}000 - < 18{,}500$	$\cancel{\text{IIII}}$ |	6	
$18{,}500 - < 19{,}000$	$\cancel{\text{IIII}}$	5	
$19{,}000 - < 19{,}500$	$\cancel{\text{IIII}}$	5	
$19{,}500 - < 20{,}000$	$\cancel{\text{IIII}}$	5	
$20{,}000 - < 20{,}500$	||	2	
$20{,}500 - < 21{,}000$	||	2	
$21{,}000 - < 21{,}500$		0	
$21{,}500 - < 22{,}000$		0	
$22{,}000 - < 22{,}500$	|	1	

   a. What is the total number of NBA arenas?

   b. Complete the relative frequency column. Round to the nearest thousandth.

c. Construct a relative frequency histogram. Round to the nearest thousandth.

d. Describe the shape of the relative frequency histogram.

e. What percent of the arenas has a seating capacity between $18{,}500$ and $19{,}999$ seats?

f. How does this relative frequency histogram compare to the frequency histogram that you drew in problem 2 of the Problem Set in Lesson 4?

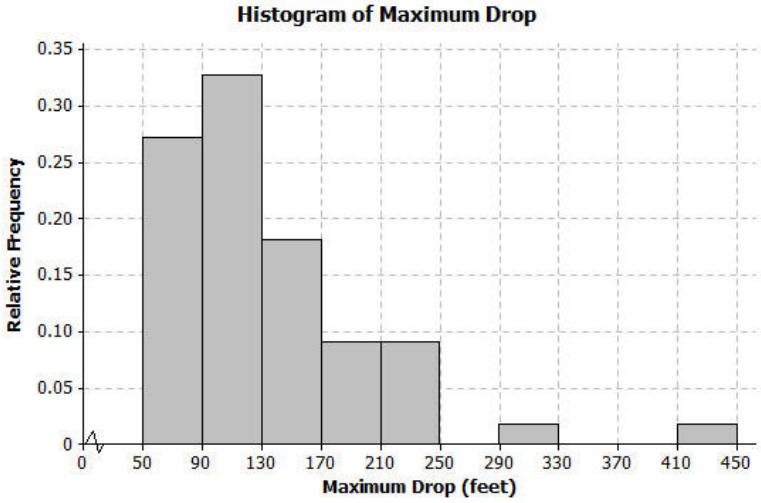
Lesson Summary

A **relative frequency histogram** uses the same data as a frequency histogram but compares the frequencies for each interval frequency to the total number of items. For example, if the first interval contains 8 out of the total of 32 items, the relative frequency of the first interval $\frac{8}{32}$ or $\frac{1}{4} = 0.25$.

The only difference between a frequency histogram and a relative frequency histogram is that the vertical axis uses relative frequency instead of frequency. The shapes of the histograms are the same as long as the intervals are the same.

Problem Set

1. Below is a relative frequency histogram of the maximum drop (in feet) of a selected group of roller coasters.



   a. Describe the shape of the relative frequency histogram.
   b. What does the shape tell you about the maximum drop (in feet) of roller coasters?
   c. Jerome said that more than half of the data is in the interval from $50 - 130$ feet. Do you agree with Jerome? Why or why not?

2. The frequency table below shows the length of selected movies shown in a local theater over the past 6 months.

Length of Movie (min)	Tally	Frequency	Relative Frequency
$80-<90$	|	1	0.036
$90-<100$	||||	4	0.143
$100-<110$	𝍸 ||	7	0.25
$110-<120$	𝍸	5	0.179
$120-<130$	𝍸 ||	7	0.25
$130-<140$	|||	3	0.107
$140-<150$	|	1	0.036

a. Complete the relative frequency column. Round to the nearest thousandth.

b. What percent of the movie lengths is greater than or equal to 130 minutes?

c. Draw a relative frequency histogram.

d. Describe the shape of the relative frequency histogram.

e. What does the shape tell you about the length of movie times?

3. The table below shows the highway mile per gallon of different compact cars.

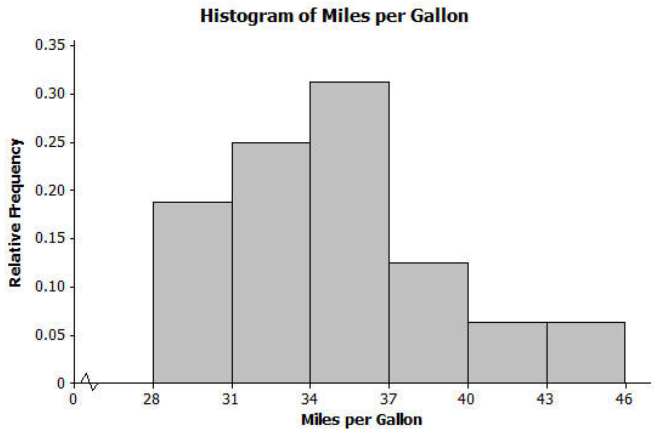
Mileage	Tally	Frequency	Relative Frequency
$128-<31$	|||	3	0.188
$31-<34$	||||	4	0.250
$34-<37$	𝍸	5	0.313
$37-<40$	||	2	0.125
$40-<43$	|	1	0.063
$43-<46$		0	0
$46-<49$		0	0
$49-<52$	|	1	0.063

a. What is the total number of compact cars?

b. Complete the relative frequency column. Round to the nearest thousandth.

c. What percent of the cars gets between 31 and up to but not including 37 miles per gallon on the highway?

d. Juan drew the relative frequency histogram of the miles per gallon of the compact cars, shown on the right. Do you agree with the way Juan drew the histogram? Explain your answer.



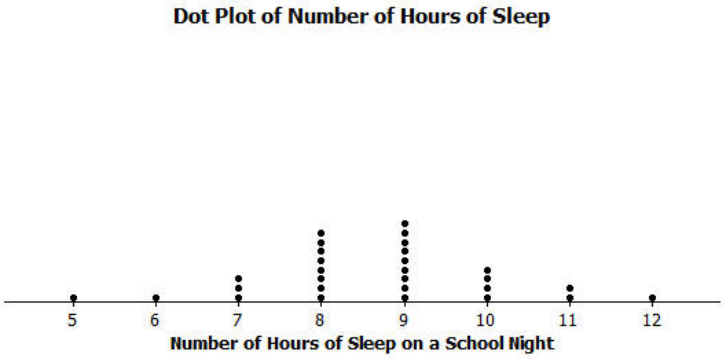
Lesson 6: Describing the Center of a Distribution Using the Mean

Classwork

Example 1

Recall that in Lesson 3, Robert, a 6th grader at Roosevelt Middle School, investigated the number of hours of sleep sixth grade students get on school nights. Today, he is to make a short report to the class on his investigation. Here is his report.

"I took a survey of 29 6th graders asking them 'How many hours of sleep per night do you usually get when you have school the next day?' The first thing I had to do was to organize the data. I did this by drawing a dot plot.



Part of our lessons last week was to identify what we thought was a centering point of the data, the spread of the data, and the shape of the data. So, for my data, looking at the dot plot, I would say that the typical number of hours sixth-grade students sleep get when they have school the next day is around 8 or 9 because that is what most students said and the values are kind of in the middle. I also noticed that the data were spread out from the center by about three or four hours in both directions. The shape of the distribution is kind of like a mound."

Michelle is Robert's classmate. She liked his report but has a really different thought about determining the center of the number of hours of sleep. Her idea is to even out the data in order to determine a typical or center value.

Exercises 1–6

Suppose that Michelle asks ten of her classmates for the number of hours they usually sleep when there is school the next day.

Suppose they responded (in hours): 8 10 8 8 11 11 9 8 10 7

1. How do you think Robert would organize his data? What do you think Robert would say is the center of these ten data points? Why?

2. Do you think his value is a good measure to use for the "center" of Michelle's data set? Why or why not?

Michelle's "center" is called the mean. She finds the total number of hours of sleep for each of the ten students. That is 90 hours. She has 90 Unifix cubes (Snap cubes). She gives each of the ten students the number of cubes that equals the number of hours of sleep each had reported. She then asks each of the ten students to connect their cubes in a stack and put their stacks on a table to compare them. She then has them share their cubes with each other until they all have the same number of cubes in their stacks when they are done sharing.

3. Work in a group. Each group of students gets 90 cubes. Make ten stacks of cubes representing the number of hours of sleep for each of the ten students. Using Michelle's Method, how many *cubes* are in each of the ten stacks when they are done sharing?

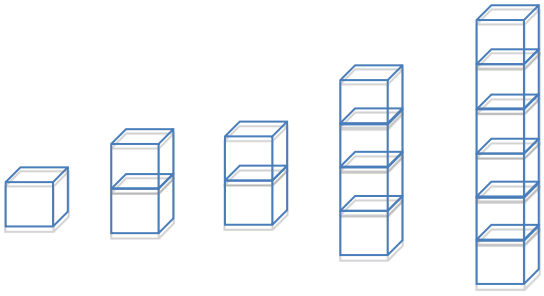
4. Noting that one cube represents one hour of sleep, interpret your answer to Exercise 3 in terms of "number of hours of sleep." What does this number of cubes in each stack represent? What is this value called?

5. Suppose that the student who told Michelle he slept 7 hours changes his data entry to 8 hours. You will need to get one more cube from your teacher. What does Michelle's procedure now produce for her center of the new set of data? What did you have to do with that extra cube to make Michelle's procedure work?

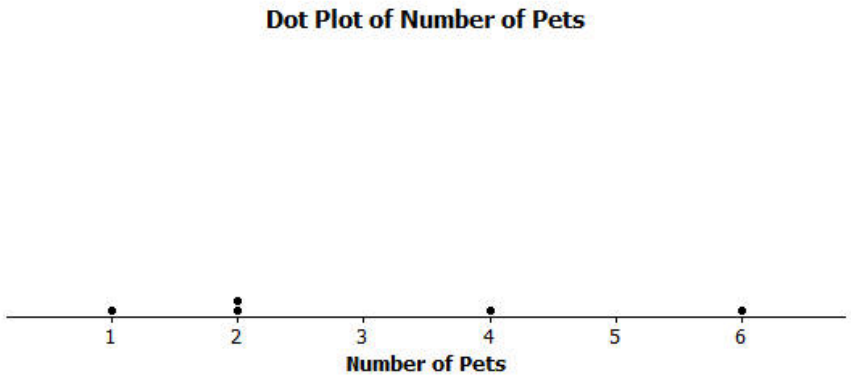
6. Interpret Michelle's "fair share" procedure by developing a mathematical formula that results in finding the fair share value without actually using cubes. Be sure that you can explain clearly how the fair share procedure and the mathematical formula relate to each other.

Example 2

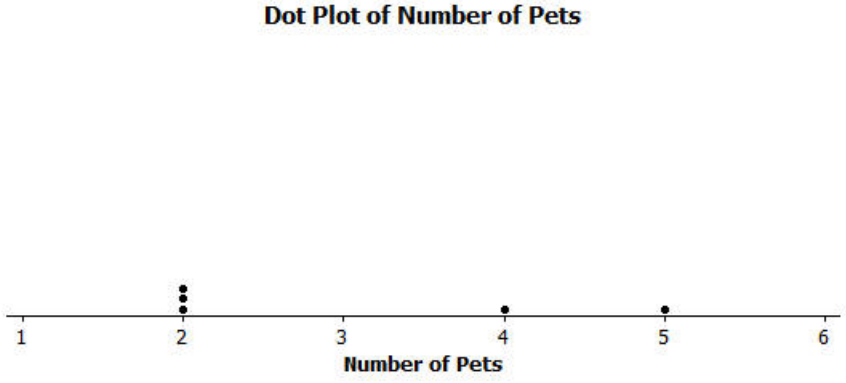
Suppose that Robert asked five sixth graders how many pets each had. Their responses were 2, 6, 2, 4, 1. Robert showed the data with cubes as follows:



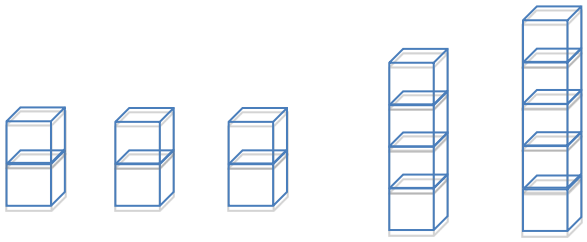
Note that one student has one pet, two students have two pets each, one student has four pets, and one student has six pets. Robert also represented the data set in the following dot plot.



Robert wants to illustrate Michelle's fair share method by using dot plots. He drew the following dot plot and said that it represents the result of the student with six pets sharing one of her pets with the student who has one pet.



Robert also represented the data with cubes as shown below.



Exercises 7–10

Now continue distributing the pets based on the following steps.

7. Robert does a fair share step by having the student with five pets share one of her pets with one of the students with two pets.
   a. Draw the cubes representation that shows Robert's fair share step.

   b. Draw the dot plot that shows Robert's fair share step.

8. Robert does another fair share step by having one of the students who has four pets share one pet with one of the students who has two pets.
   a. Draw the cubes representation that shows Robert's fair share step.

   b. Draw the dot plot that shows Robert's fair share step.

9. Robert does a final fair share step by having the student who has four pets share one pet with the student who has two pets.
   a. Draw the cubes representation that shows Robert's final fair share step.

   b. Draw the dot plot representation that shows Robert's final fair share step.

10. Explain in your own words why the final representations using cubes and a dot plot show that the mean number of pets owned by the five students is 3 pets.

Lesson Summary

In this lesson, you developed a method to define the center of a data distribution. The method was called the "fair share" method and the center of a data distribution that it produced is called the mean of the data set. The reason it is called the fair share value is that if all the subjects were to have the same data value, it would be the mean value.

Mathematically the "fair share" term comes from finding the total of all of the data values and dividing the total by the number of data points. The arithmetic operation of division divides a total into equal parts.

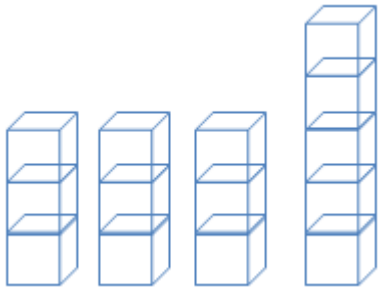
Problem Set

1. A game was played where ten tennis balls are tossed into a basket from a certain distance. The number of successful tosses for six students were: $4, 1, 3, 2, 1, 7$.
   a. Draw a representation of the data using cubes where one cube represents one successful toss of a tennis ball into the basket.
   b. Draw the original data set using a dot plot.

2. Find the mean number of successful tosses for this data set by Michelle's fair share method. For each step, show the cubes representation and the corresponding dot plot. Explain each step in words in the context of the problem. You may move more than one successful toss in a step, but be sure that your explanation is clear. You must show two or more steps.

Step described in words	"Fair Share" cube representation	Dot plot

3. The number of pockets in the clothes worn by four students to school today is $4, 1, 3, 6$. Paige produces the following cube representation as she does the fair share process. Help her decide how to finish the process of 3, 3, 3, and 5 cubes.



4. Suppose that the mean number of chocolate chips in 30 cookies is 14 chocolate chips.
   a. Interpret the mean number of chocolate chips in terms of fair share.
   b. Describe the dot plot representation of the fair share mean of 14 chocolate chips in 30 cookies.

5. Suppose that the following are lengths (in millimeters) of radish seedlings grown in identical conditions for three days: 12 11 12 14 13 9 13 11 13 10 10 14 16 13 11.
   a. Find the mean length for these 15 radish seedlings.
   b. Interpret the value from part (a) in terms of the "fair share" center length.

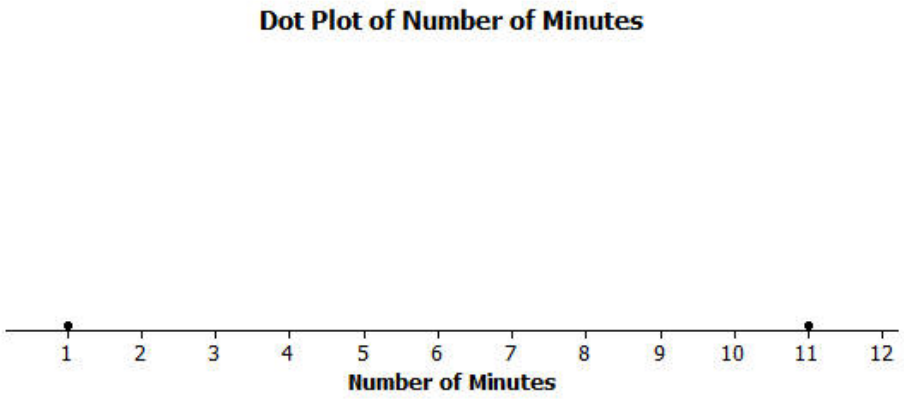
Lesson 7: The Mean as a Balance Point

In Lesson 3, Robert gave us an informal interpretation of the center of a data set. In Lesson 6, Michelle developed a more formal interpretation of the center as a "fair share" mean, a value that every person in the data set would have if they all had the same value. In this lesson, Sabina will show us how to interpret the mean as a "balance point."

Classwork

Example 1: The Mean as a Balance Point

Sabina wants to know how long it takes students to get to school. She asks two students how long it takes them to get to school. It takes one student 1 minute and the other student 11 minutes. Sabina represents these data on a ruler putting a penny at 1 and another at 11 and shows that the ruler balances on the eraser end of a pencil at 6. Note that the mean of 1 and 11 is also 6. Sabina thinks that there might be a connection between the mean of two data points and where they balance on a ruler. She thinks the mean may be the balancing point. What do you think? Will Sabina's ruler balance at 6? Is the mean of 1 and 11 equal to 6? Sabina shows the result on a dot plot.



Sabina decides to move the penny at 1 to 4 and the other penny from 11 to 8 on the ruler, noting that the movement for the two pennies is the same distance but in opposite directions. She notices that the ruler still balances at 6. Sabina decides that if data points move the same distance but in opposite directions, the balancing point on the ruler does not change. Does this make sense? Notice that this implies that the mean of the time to get to school for two students who take 4 minutes and 8 minutes to get to school is also 6 minutes.

Sabina continues by moving the penny at 4 to 6. To keep the ruler balanced at 6, how far should Sabina move the penny at 8 and in what direction? Since the penny at 4 moved two to the right, to maintain the balance the penny at 8 needs to move two to the left. Both pennies are now at 6, and the ruler clearly balances there. Note that the mean of these two values (6 minutes and 6 minutes) is still 6 minutes.

Exercises 1–2

Now it is your turn to try balancing two pennies on a ruler.

1. Tape one penny at 2.5 on your ruler.
   a. Where should a second penny be taped so that the ruler will balance at 6?

   b. How far is the penny at 2.5 from 6? How far is the other penny from 6?

   c. Is the mean of the two locations of the pennies equal to 6?

2. Move the penny that is at 2.5 two inches to the right.
   a. Where will the point be placed?

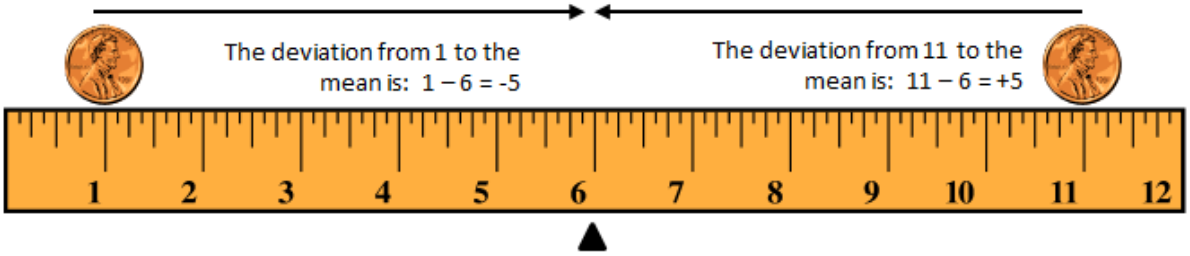
   b. What do you have to do with the other data point to keep the balance point at 6?

   c. What is the mean of the two new data points? Is it the same value as the balancing point of the ruler?

Example 2: Understanding Deviations

In the above example using two pennies, it appears that the balance point of the ruler occurs at the mean location of the two pennies. We computed the distance from the balance point to each penny location and treated the distances as positive numbers. In statistics, we calculate a difference by subtracting the mean from the data point and call it the deviation of a data point from the mean. So, points to the left of the mean are less than the mean and have a negative deviation. Points to the right of the mean are greater than the mean and have a positive deviation.

Let's look at Sabina's initial placement of pennies at 1 and 11 with a mean at 6 on the graph below. Notice that the deviations are $+5$ and -5. What is the sum of the deviations?



Similarly, when Sabina moved the pennies to 4 and 8, the deviation of 4 from 6 is $4 - 6 = -2$, and the deviation of 8 from 6 is $8 - 6 = +2$. Here again, the sum of the two deviations is 0, since $-2 + 2 = 0$. It appears that for two data points the mean is the point when the sum of its deviations is equal to 0.

Exercises 3–4

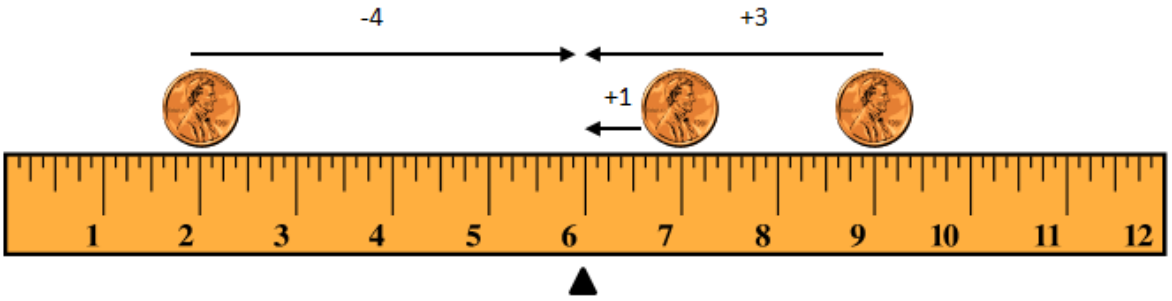
Refer back to Exercise 2, where one penny was located at 2.5 and the mean was at 6.

3. Where was the second penny located?

4. Calculate the deviations of the two pennies and show that the sum of the deviations is 0.

Example 3: Balancing More than Two Points

Sabina wants to know what happens if there are more than two data points. Suppose there are three students. One student lives 2 minutes from school, and another student lives 9 minutes from school. If the mean time for all three students is 6 minutes, she wonders how long it takes the third student to get to school. She tapes pennies at 2 and 9 and by experimenting finds the ruler balances with a third penny placed at 7. To check what she found, she calculates deviations.



The data point at 2 has a deviation of -4 from the mean. The data point at 7 has a deviation of $+1$ from the mean. The data point at 9 has a deviation of $+3$ from the mean. The sum of the three deviations is 0, since $-4 + 1 + 3 = 0$. So, the mean is indeed 6 minutes.

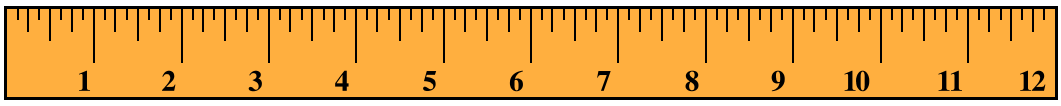
Robert says that he found out that the third penny needs to be at 7 without using his ruler. He put 2 and 9 on a dot plot. He says that the sum of the two deviations for the points at 2 and 9 is -1, since $-4 + 3 = -1$. So, he claims that the third data point would require a deviation of $+1$ to make the sum of all three deviations equal to 0. That makes the third data point 1 minute above the mean of 6 minutes, which is 7 minutes.

Exercises 5–7

Imagine you are balancing pennies on a ruler.

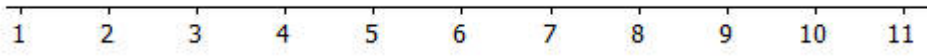
5. Suppose you place one penny each at 3, 7, and 8 on your ruler.

   a. Sketch a picture of the ruler. At what value do you think the ruler will balance? Mark the balancing point with the symbol Δ.



   b. What is the mean of 3, 7, and 8? Does your ruler balance at the mean?

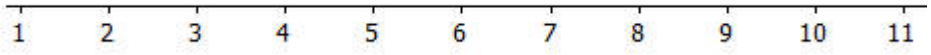
c. Show part (a) on a dot plot. Mark the balancing point with the symbol Δ.



d. What are the deviations from each of the data points to the balancing point? What is the sum of the deviations? What is the value of the mean?

6. Now suppose you place a penny each at 7 and 9 on your ruler.

a. Draw a dot plot representing these two pennies.

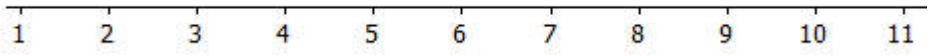


b. Estimate where to place a third penny on your ruler so that the ruler balances at 6 and mark the point on the dot plot above. Mark the balancing point with the symbol Δ.

c. Explain why your answer in part (b) is true by calculating the deviations of the points from 6. Is the sum of the deviations equal 0?

7. Suppose you place two pennies at 7 and one penny at 9 on your ruler.

   a. Draw a dot plot representing these three pennies.

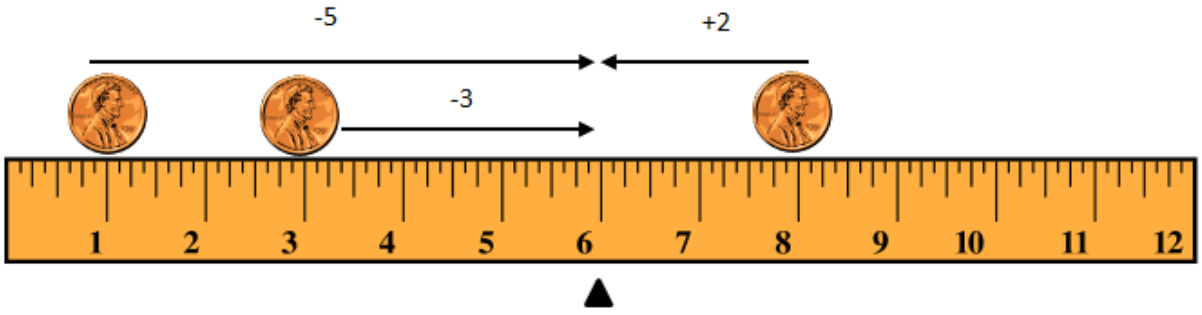


   b. Estimate where to place a fourth penny on your ruler so that the ruler balances at 6 and mark the point on the dot plot above. Mark the balancing point with the symbol Δ.

   c. Explain why your answer in part (b) is true by calculating the deviations of the points from 6. Does the sum of the deviations equal 0?

Example 4: Finding the Mean

Not all data distributions on a ruler are going to have a "fair share" mean, or "balance point" of 6. What if the data were 1, 3, and 8? Will your ruler balance at 6? Why not?



Notice that the deviation of 1 from 6 is -5. The deviation of 3 from 6 is -3. The deviation of 8 from 6 is $+2$. The sum of the deviations is -6, since $-5 + (-3) + 2 = -6$. The sum should be 0. Therefore, the mean is not at 6. Is the mean greater than 6 or less than 6? The sum of the deviations is negative. To decrease the negative deviations and increase the positive deviations, the balance point would have to be less than 6.

Let's see if the balance point is at 5. The deviation of 1 from 5 is -4. The deviation of 3 from 5 is -2. The deviation of 8 from 5 is $+3$. The sum of the three deviations is -3, since $-4 + (-2) + 3 = -3$. That's closer to 0 than before.

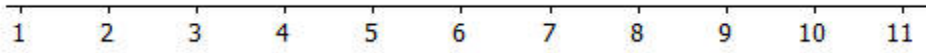
Let's keep going and try 4 as the balance point. The deviation of 1 from 4 is -3. The deviation of 3 from 4 is -1. The deviation of 8 from 4 is $+4$. The sum of the deviations is 0, since $-3 + (-1) + 4 = 0$. The balancing point of the data distribution of 1, 3, and 8 shown on your ruler or on a dot plot is at 4. The mean of 1, 3, and 8 is 4.

Exercise 8

Use what you have learned about the mean to answer the following questions.

8. Recall in Lesson 6 that Michelle asked ten of her classmates for the number of hours they usually sleep when there is school the next day. Their responses (in hours) were $8, 10, 8, 8, 11, 11, 9, 8, 10, 7$.

   a. It's hard to balance ten pennies. Instead of actually using pennies and a ruler, draw a dot plot that represents the data set.



   b. Use your dot plot to find the balance point. What is the sum of the deviations of the data points from the fair share mean of 9 hours?

Lesson Summary

In this lesson, the "balance" process was developed to provide another way in which the mean characterizes the "center" of a distribution.

- The mean is the balance point of the data set when the data are shown as dots on a dot plot (or pennies on a ruler).
- The difference formed by subtracting the mean from a data point is called its deviation.
- The mean can be defined as the value that makes the sum of all deviations in a distribution equal to zero.
- The mean is the point that balances the sum of the positive deviations with the sum of the negative deviations.

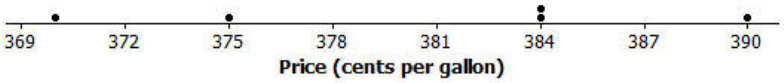
Problem Set

1. The number of pockets in the clothes worn by four students to school today is $4, 1, 3, 4$.
   a. Perform the "fair share" process to find the mean number of pockets for these four students. Sketch the cube representations for each step of the process.
   b. Find the sum of the deviations to prove the mean found in part (a) is correct.

2. The times (rounded to the nearest minute) it took each of six classmates to run a mile are $7, 9, 10, 11, 11$, and 12 minutes.
   a. Draw a dot plot representation for the times. Suppose that Sabina thinks the mean is 11 minutes. Use the sum of the deviations to show Sabina that the balance point of 11 is too high.
   b. Sabina now thinks the mean is 9 minutes. Use the sum of the deviations to verify that 9 is too small to be the mean number of pockets.
   c. Sabina asks you to find the mean by using the balancing process. Demonstrate that the mean is 10 minutes.

3. The prices per gallon of gasoline (in cents) at five stations across town on one day are shown in the following dot plot. The price for a sixth station is missing, but the mean price for all six stations was reported to be 380 cents per gallon. Use the "balancing" process to determine the price of a gallon of gasoline at the sixth station?

Dot Plot of Price (cents per gallon)



4. The number of phones (landline and cell) owned by the members of each of nine families is $3, 5, 5, 5, 6, 6, 6, 6, 8$.
   a. Use the mathematical formula for the mean (sum the data points and divide by the number of data points) to find the mean number of phones owned for these nine families.
   b. Draw a dot plot of the data and verify your answer in part (a) by using the "balancing" process and finding the sum of the deviations.

Lesson 8: Variability in a Data Distribution

Classwork

Example 1: Comparing Two Distributions

Robert's family is planning to move to either New York City or San Francisco. Robert has a cousin in San Francisco and asked her how she likes living in a climate as warm as San Francisco. She replied that it doesn't get very warm in San Francisco. He was surprised, and since temperature was one of the criteria he was going to use to form his opinion about where to move, he decided to investigate the temperature distributions for New York City and San Francisco. The table below gives average temperatures (in degrees Fahrenheit) for each month for the two cities.

City	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
New York City	39	42	50	61	71	81	85	84	76	65	55	47
San Francisco	57	60	62	63	64	67	67	68	70	69	63	58

Exercises 1–2

Use the table above to answer the following:

1. Calculate the annual mean monthly temperature for each city.

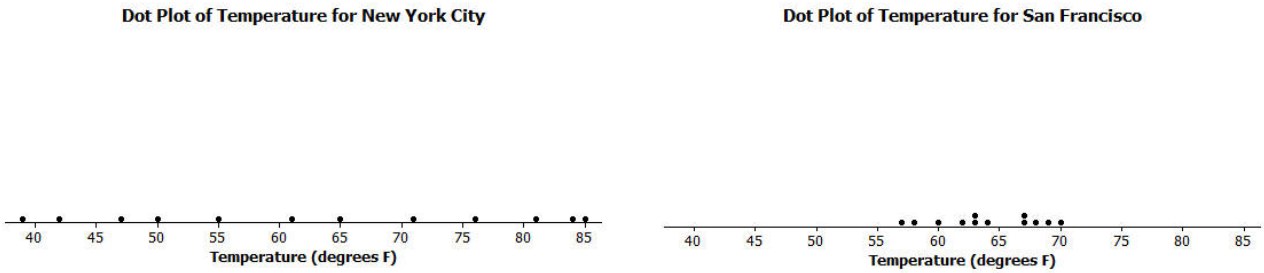
2. Recall that Robert is trying to decide to which city he wants to move. What is your advice to him based on comparing the overall annual mean monthly temperatures of the two cities?

Example 2: Understanding Variability

In Exercise 2, you found the overall mean monthly temperatures in both the New York City distribution and the San Francisco distribution to be about the same. That didn't help Robert very much in making a decision between the two cities. Since the mean monthly temperatures are about the same, should Robert just toss a coin to make his decision? Is there anything else Robert could look at in comparing the two distributions?

Variability was introduced in an earlier lesson. Variability is used in statistics to describe how spread out the data in a distribution are from some focal point in the distribution (such as the mean). Maybe Robert should look at how spread

out the New York City monthly temperature data are from its mean and how spread out the San Francisco monthly temperature data are from its mean. To compare the variability of monthly temperatures between the two cities, it may be helpful to look at dot plots. The dot plots for the monthly temperature distributions for New York City and San Francisco follow.



Exercises 3–7

Use the dot plots above to answer the following:

3. Mark the location of the mean on each distribution with the balancing Δ symbol. How do the two distributions compare based on their means?

4. Describe the variability of the New York City monthly temperatures from the mean of the New York City temperatures.

5. Describe the variability of the San Francisco monthly temperatures from the mean of the San Francisco monthly temperatures.

6. Compare the amount of variability in the two distributions. Is the variability about the same, or is it different? If different, which monthly temperature distribution has more variability? Explain.

7. If Robert prefers to choose the city where the temperatures vary the least from month to month, which city should he choose? Explain.

Example 3: Using Mean and Variability in a Data Distribution

The mean is used to describe the "typical" value for the entire distribution. Sabina asks Robert which city he thinks has the better climate? He responds that they both have about the same mean, but that the mean is a better measure or a more precise measure of a typical monthly temperature for San Francisco than it is for New York City. She's confused and asks him to explain what he means by this statement.

Robert says that the mean of 63 degrees in New York City (64 in San Francisco) can be interpreted as the typical temperature for any month in the distributions. So, 63 or 64 degrees should represent all of the months' temperatures fairly closely. However, the temperatures in New York City in the winter months are in the 40s and in the summer months are in the 80s. The mean of 63 isn't too close to those temperatures. Therefore, the mean is not a good indicator of typical monthly temperature. The mean is a much better indicator of the typical monthly temperature in San Francisco because the variability of the temperatures there is much smaller.

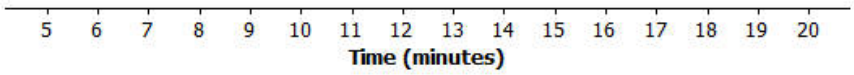
Exercises 8–14

Consider the following two distributions of times it takes six students to get to school in the morning, and to go home from school in the afternoon.

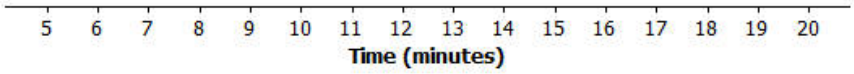
	Time (minutes)					
Morning	11	12	14	14	16	17
Afternoon	6	10	13	18	18	19

8. To visualize the means and variability, draw dot plots for each of the two distributions.

Morning



Afternoon



9. What is the mean time to get from home to school in the morning for these six students?

10. What is the mean time to get from school to home in the afternoon for these six students?

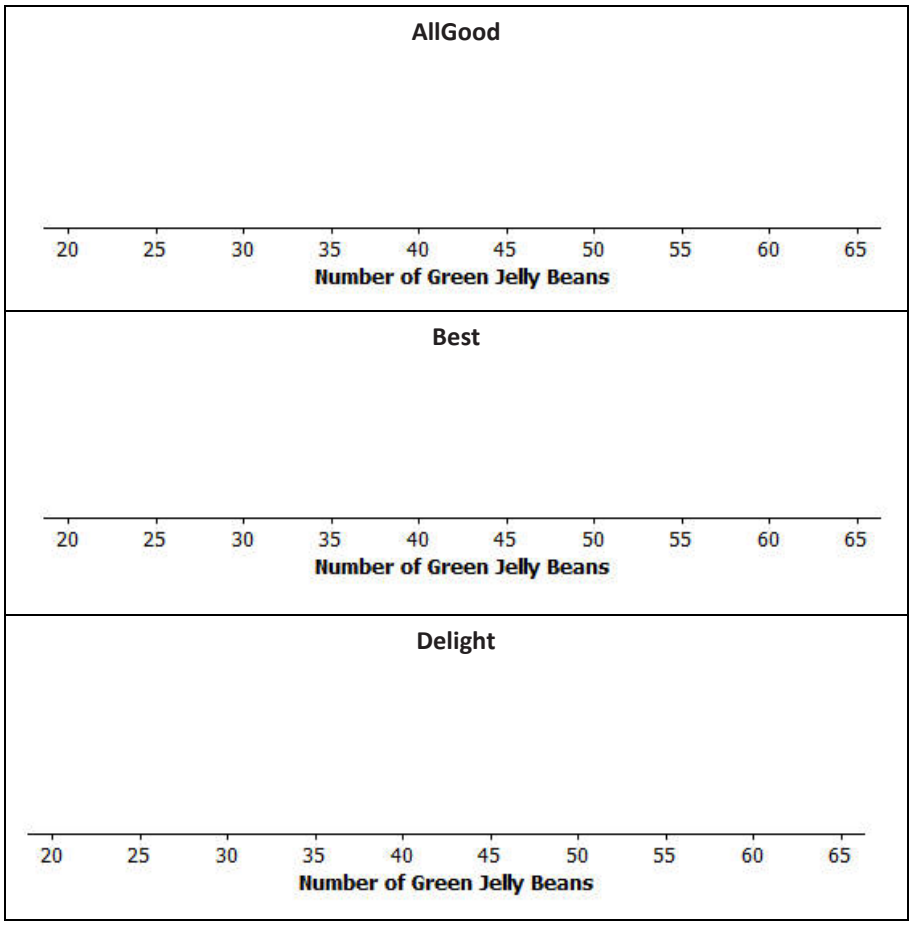
11. For which distribution does the mean give a more precise indicator of a typical value? Explain your answer.

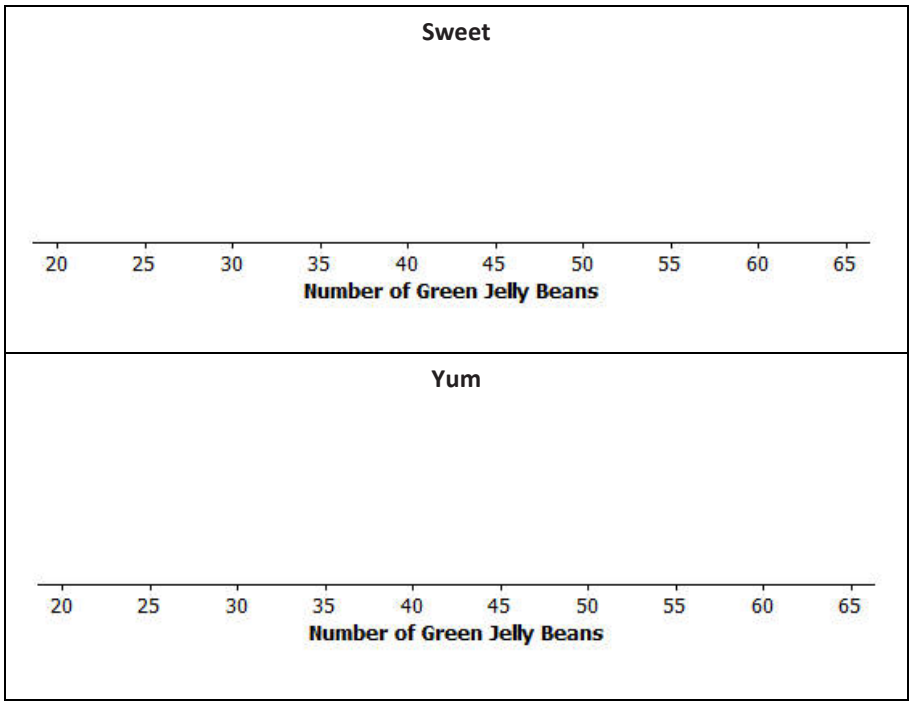
Distributions can be ordered according to how much the data values vary around their means.

Consider the following data on the number of green jellybeans in seven bags of jellybeans from each of five different candy manufacturers (AllGood, Best, Delight, Sweet, Yum). The mean in each distribution is 42 green jellybeans.

	1	2	3	4	5	6	7
AllGood	40	40	41	42	42	43	46
Best	22	31	36	42	48	53	62
Delight	26	36	40	43	47	50	52
Sweet	36	39	42	42	42	44	49
Yum	33	36	42	42	45	48	48

12. Draw a dot plot of the distribution of number of green jellybeans for each of the five candy makers. Mark the location of the mean on each distribution with the balancing Δ symbol.





13. Order the candy manufacturers from the one you think has least variability to the one with most variability. Explain your reasoning for choosing the order.

14. For which company would the mean be considered a better indicator of a typical value (based on least variability)?

Lesson Summary

We can compare distributions based on their means, but variability must also be considered. The mean of a distribution with small variability (not a lot of spread) is considered to be a better indication of a typical value than the mean of a distribution with greater variability (wide spread).

Problem Set

1. The number of pockets in the clothes worn by seven students to school yesterday were $4, 1, 3, 4, 2, 2, 5$. Today those seven students each had three pockets in their clothes.
   a. Draw one dot plot for what the students wore yesterday, and another dot plot for what the students wore today. Be sure to use the same scales. Show the means by using the balancing Δ symbol.
   b. For each distribution, find the mean number of pockets worn by the seven students.
   c. For which distribution is the mean number of pockets a better indicator of what is "typical?" Explain.

2. The number of minutes (rounded) it took to run a certain short cross-country route was recorded for each of five students. The resulting data were $9, 10, 11, 14$, and 16 minutes. The number of minutes (rounded to the nearest minute) it took the five students to run a different cross-country route was also recorded, resulting in the following data: $6, 8, 12, 15$, and 19 minutes.
   a. Draw dot plots for the two distributions of the time it takes to run a cross-country route. Be sure to use the same scale on both dot plots.
   b. Do the distributions have the same mean?
   c. In which distribution is the mean a better indicator of the typical amount of time taken to run its cross-country route? Explain.

3. The following table shows the prices per gallon of gasoline (in cents) at five stations across town as recorded on Monday, Wednesday, and Friday of a certain week.

Day	R&C	Al's	PB	Sam's	Ann's
Monday	359	358	362	359	362
Wednesday	357	365	364	354	360
Friday	350	350	360	370	370

   a. The mean price per day over the five stations is the same for the three days. Without doing any calculation and simply looking at Friday's prices, what must the mean price be?
   b. In which daily distribution is its mean a better indicator of the typical price per gallon for the five stations? Explain.

Lesson 9: The Mean Absolute Deviation (MAD)

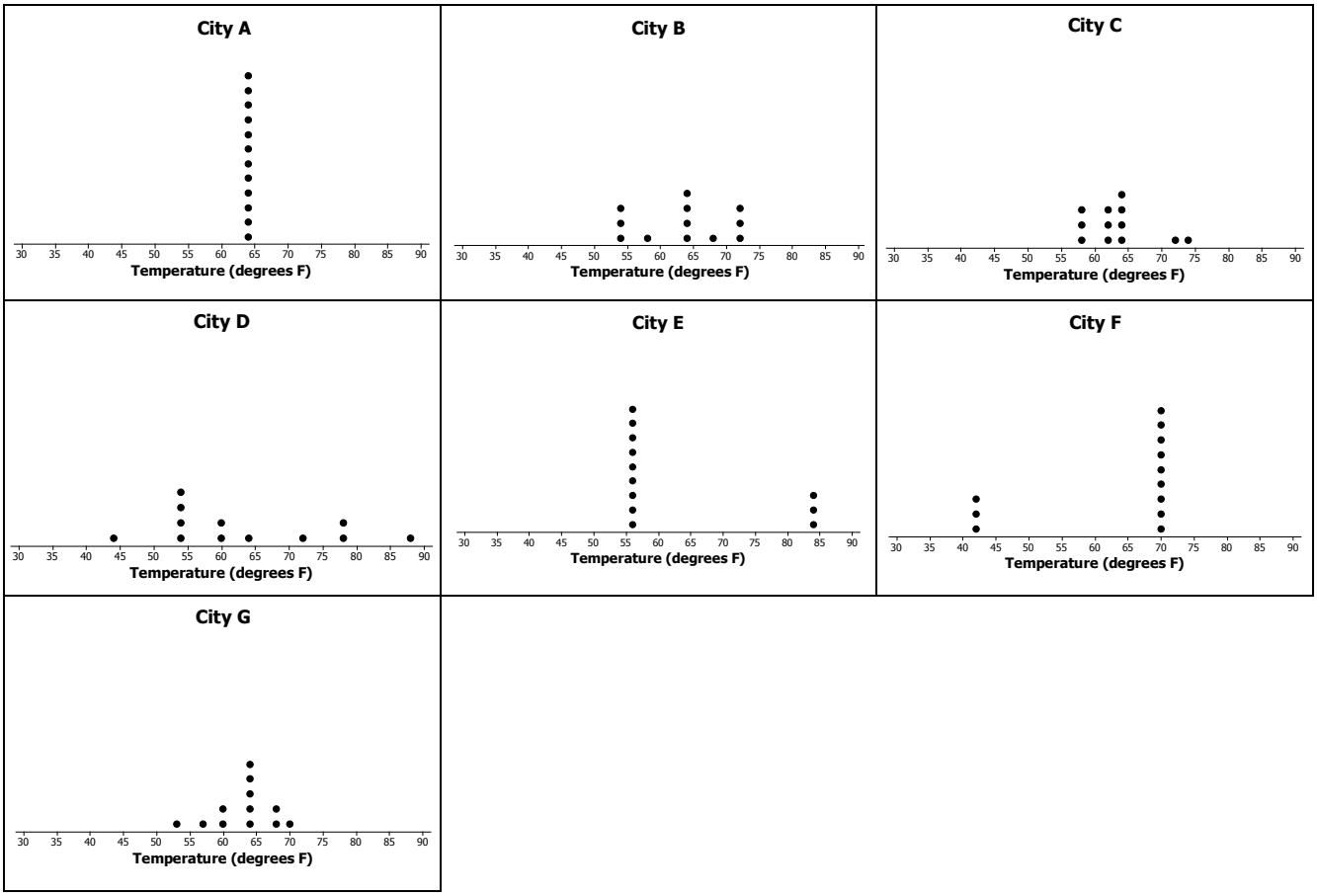
Classwork

Example 1: Variability

In Lesson 8, Robert tried to decide to which of two cities he would rather move, based on comparing their mean annual temperatures. Since the mean yearly temperature for New York City and San Francisco turned out to be about the same, he decided instead to compare the cities based on the variability in their monthly temperatures from the overall mean. He looked at the two distributions and decided that the New York City temperatures were more spread out from their mean than were the San Francisco temperatures from their mean.

Exercises 1–3

The following temperature distributions for seven other cities all have a mean temperature of approximately 63 degrees. They do not have the same variability. Consider the following dot plots of the mean yearly temperatures of the seven cities in degrees Fahrenheit.



1. Which distribution has the smallest variability of the temperatures from its mean of $\mathbf{63}$ degrees? Explain your answer.

2. Which distribution(s) seems to have the most variability of the temperatures from the mean of $\mathbf{63}$ degrees? Explain your answer.

3. Order the seven distributions from least variability to most variability. Explain why you listed the distributions in the order that you chose.

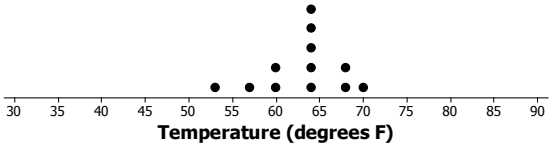
Example 2: Measuring Variability

Based on just looking at the distributions, there are different orderings of variability that seem to make some sense. Sabina is interested in developing a formula that will give a number that measures the variability in a data distribution. She would then use the formula for each data set and order the distributions from lowest to highest. She remembers from a previous lesson that a deviation is found by subtracting the mean from a data point. The formula was summarized as: deviation = data point – mean. Using deviations to develop a formula measuring variability is a good idea to consider.

Exercises 4–6

The dot plot for the temperatures in City G is shown below. Use the dot plot and the mean temperature of 63 degrees to answer the following questions.

City G

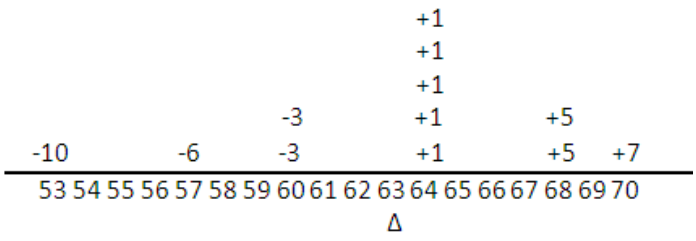


4. Fill in the following table for City G temperature deviations.

Temp	Deviation	Result
53	$53-63$	-10
57	$57-63$	-6
60	$60-63$	-3
60	$60-63$	-3
64	$64-63$	$+1$
64	$64-63$	$+1$
64		
64		
64		
68		
68		
70		
Sum		

5. Why should the sum of your deviations column be equal to zero? (Hint: Recall the balance interpretation of the mean of a data set.)

6. Another way to graph the deviations is to write them on a number line as follows. What is the sum of the positive deviations (the deviations to the right of the mean)? What is the sum of the negative deviations (the deviations to the left of the mean)? What is the total sum of the deviations?



Example 3: Finding the Mean Absolute Deviation (MAD)

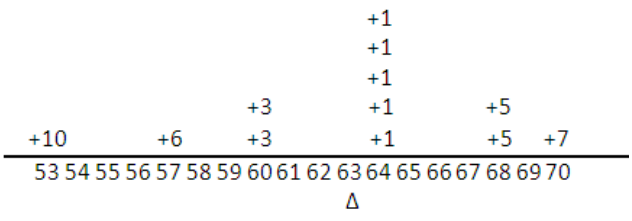
By the balance interpretation of the mean, the sum of the deviations for any data set will always be zero. Sabina is disappointed that her idea of developing a measure of variability using deviations isn't working. She still likes the concept of using deviations to measure variability, but the problem is that the sum of the positive deviations is cancelling out the sum of the negative deviations. What would you suggest she do to keep the deviations as the basis for a formula but to avoid the deviations cancelling out each other?

Exercises 7–8

7. One suggestion to possibly help Sabina is to take the absolute value of the deviations.
   a. Fill in the following table.

Temp	Deviation	Result	Abs
53	$53-63$	-10	$+10$
57	$57-63$	-6	$+6$
60	$60-63$	-3	$+3$
60	$60-63$	-3	$+3$
64	$64-63$	$+1$	$+1$
64	$64-63$	$+1$	$+1$
64			
64			
64			
68			
68			
70			

   b. From the following graph, what is the sum of the absolute deviations?



c. Sabina suggests that the mean of the absolute deviations could be a measure of the variability in a data set. Its value is the average distance that all the data values are from the mean temperature. It is called the Mean Absolute Deviation and is denoted by the letters, MAD. Find the MAD for this data set of City G temperatures. Round to the nearest tenth.

d. Find the MAD for each of the temperature distributions in all seven cities, and use the values to order the distributions from least variability to most variability. Recall that the mean for each data set is 63 degrees. Does the list that you made in Exercise 2 by just looking at the distributions match this list made by ordering MAD values?

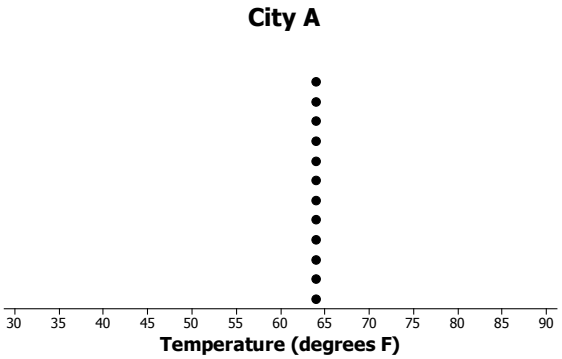
e. Which of the following is a correct interpretation of the MAD?

i. The monthly temperatures in City G are spread 3.7 degrees from the approximate mean of 63 degrees.

ii. The monthly temperatures in City G are, on average, 3.7 degrees from the approximate mean temperature of 63 degrees.

iii. The monthly temperatures in City G differ from the approximate mean temperature of 63 degrees by 3.7 degrees.

8. The dot plot for City A temperatures follows.



a. How much variability is there in City A's temperatures? Why?

b. Does the MAD agree with your answer in part (a)?

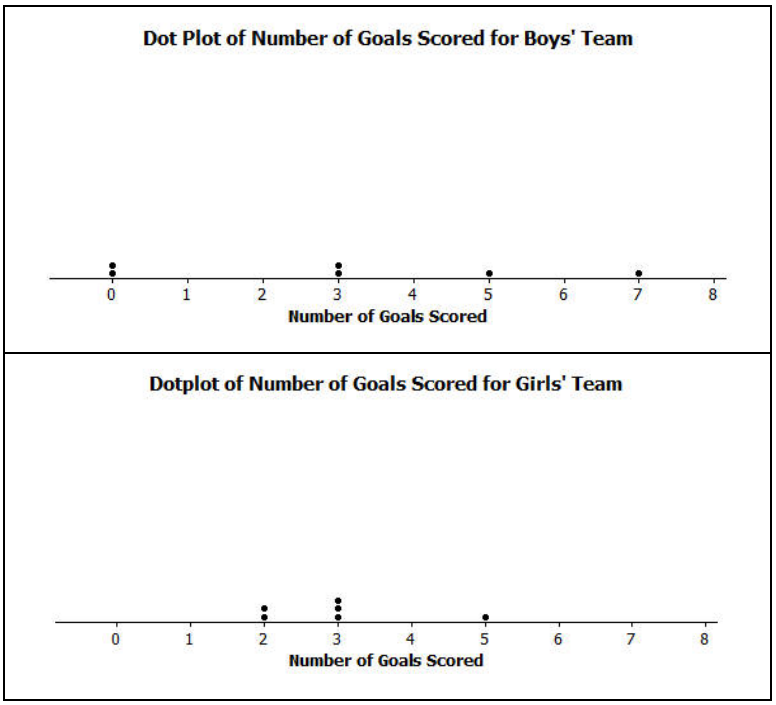
Lesson Summary

In this lesson, a formula was developed that measures the amount of variability in a data distribution.

- The absolute deviation of a data point is how far away that data point is from the mean.
- The Mean Absolute Deviation (MAD) is computed by finding the mean of the absolute deviations in the distribution.
- The value of MAD is the average distance that all the data values are from the mean.
- A small MAD indicates that the distribution has very little variability.
- A large MAD indicates that the data points are spread far away from the mean.

Problem Set

1. Suppose the dot plot on the left shows the number of goals a boys' soccer team has scored in six games so far this season, and the dot plot on the right shows the number of goals a girls' soccer team has scored in six games so far this season. The mean for both of these teams is 3.



   a. Before doing any calculations, which dot plot has the larger MAD? Explain how you know.

b. Use the following tables to find the MAD number of goals for each distribution. Round your calculations to the nearest hundredth.

Boys' Team		
#Goals	Deviations	Absolute Deviations
0	-3	
0	-3	
3	$3-3=0$	
3		
5		
7		
Sum		

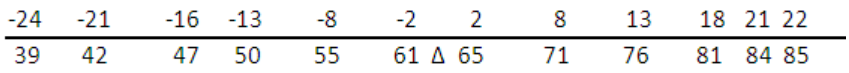
Girls' Team		
#Goals	Deviations	Absolute Deviations
2		
2		
3		
3		
3		
5		
Sum		

c. Based on the computed MAD values, for which distribution is the mean a better indication of a typical value? Explain your answer.

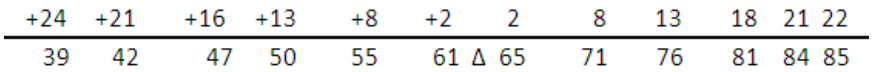
2. Recall Robert's problem of deciding whether to move to New York City or to San Francisco. The table of temperatures (in degrees Fahrenheit) and deviations for the New York City distribution is as follows:

NYC	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temp	39	42	50	61	71	81	85	84	76	65	55	47
Deviation	-24	-21	-13	-2	8	18	22	21	13	2	-8	-16

a. The dot plot below is written with the deviations above each of the monthly temperatures. What is the sum of all of the deviations? Are you surprised? Explain.



b. The absolute deviations for the monthly temperatures are shown below. Use this information to calculate the MAD. Explain the MAD in words for this problem.



c. Complete the following table and then use the values to calculate the MAD for the San Francisco data distribution.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Temp	57	60	62	63	64	67	67	68	70	69	63	58
Deviations				-1	0	$+3$						
Absolute Deviations												

d. Comparing the MAD values for New York City and San Francisco, which city would Robert choose to move to if he is interested in having a lot of variability in monthly temperatures? Explain using the MAD.

3. Consider the following data of the number of green jellybeans in seven bags sampled from five different candy manufacturers (Awesome, Delight, Finest, Sweeties, YumYum). Note that the mean in each distribution is 42 green jellybeans.

	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7
Awesome	40	40	41	42	42	43	46
Delight	22	31	36	42	48	53	62
Finest	26	36	40	43	47	50	52
Sweeties	36	39	42	42	42	44	49
YumYum	33	36	42	42	45	48	48

a. Complete the following table of the deviations of the number of green jellybeans from the mean number of green jellybeans in the seven bags.

	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7
Awesome	-2	-2	-1	0	0	$+1$	$+4$
Delight	-20	-11	-6				
Finest	-16						
Sweeties							
YumYum							

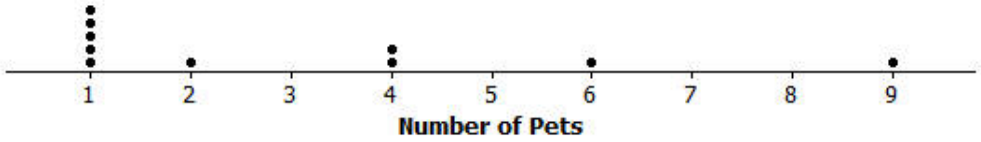
b. Based on what you learned about MAD, which manufacturer do you think will have the lowest MAD? Calculate the MAD for the manufacturer you selected.

Lesson 10: Describing Distributions Using the Mean and MAD

Classwork

Example 1: Describing Distributions

In Lesson 9, Sabina developed the mean absolute deviation (MAD) as a number that measures variability in a data distribution. Using the mean and MAD with a dot plot allows you to describe the center, spread, and shape of a data distribution. For example, suppose that data on the number of pets for ten students is shown in the dot plot below.



There are several ways to describe the data distribution. The mean number of pets each student has is three, which is a measure of center. There is variability in the number of pets the students have, which is an average of 2.2 pets from the mean (the MAD). The shape of the distribution is heavy on the left and it thins out to the right.

Exercises 1–4

1. Suppose that the weights of seven middle-school students' backpacks are given below.
   a. Fill in the following table.

Student	Alan	Beth	Char	Damon	Elisha	Fred	Georgia
Weight (lbs.)	18	18	18	18	18	18	18
Deviations							
Absolute Deviations							

   b. Draw a dot plot for these data and calculate the mean and MAD.

c. Describe this distribution of weights of backpacks by discussing the center, spread, and shape.

2. Suppose that the weight of Elisha's backpack is 17 pounds, rather than 18.
   a. Draw a dot plot for the new distribution.

   b. Without doing any calculation, how is the mean affected by the lighter weight? Would the new mean be the same, smaller, or larger?

   c. Without doing any calculation, how is the MAD affected by the lighter weight? Would the new MAD be the same, smaller, or larger?

3. Suppose that in addition to Elisha's backpack weight having changed from 18 to 17 lb., Fred's backpack weight is changed from 18 to 19 lb.
   a. Draw a dot plot for the new distribution.

b. Without doing any calculation, what would be the value of the new mean compared to the original mean?

c. Without doing any calculation, would the MAD for the new distribution be the same, smaller, or larger than the original MAD?

d. Without doing any calculation, how would the MAD for the new distribution compare to the one in Exercise 2?

4. Suppose that seven second-graders' backpack weights were:

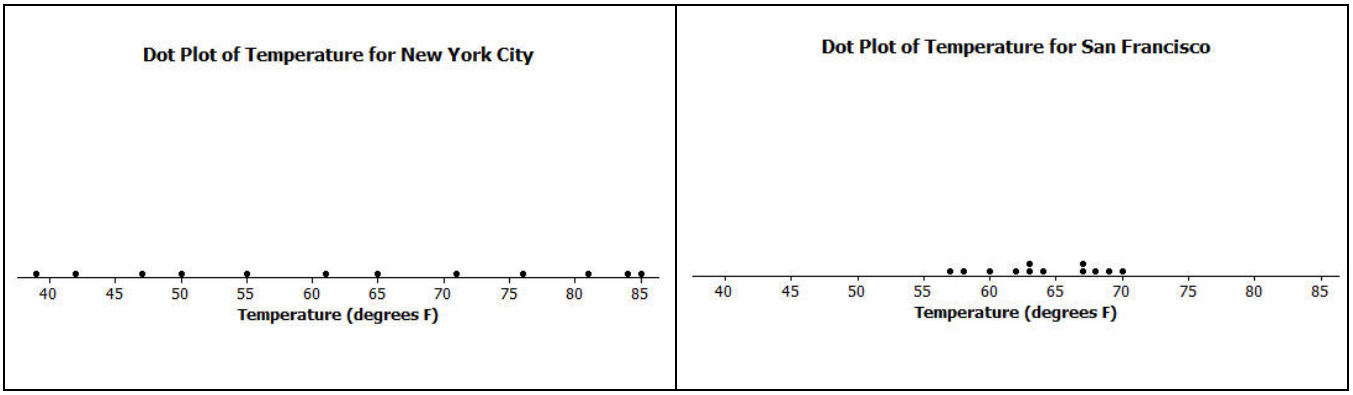
Student	Alice	Bob	Carol	Damon	Ed	Felipe	Gale
Weight (lbs.)	5	5	5	5	5	5	5

a. How is the distribution of backpack weights for the second-graders similar to the original distribution for sixth-graders given in Exercise 1?

b. How are the distributions different?

Example 2: Using the Mean Versus the MAD

Decision-making by comparing distributions is an important function of statistics. Recall that Robert is trying to decide whether to move to New York City or to San Francisco based on temperature. Comparing the center, spread, and shape for the two temperature distributions could help him decide.



From the dot plots, Robert saw that monthly temperatures in New York City were spread fairly evenly from around 40 degrees to the 80s, but in San Francisco the monthly temperatures did not vary as much. He was surprised that the mean temperature was about the same for both cities. The MAD of 14 degrees for New York City told him that, on average, a month's temperature was 14 degrees above or below 63 degrees. That is a lot of variability, which was consistent with the dot plot. On the other hand, the MAD for San Francisco told him that San Francisco's monthly temperatures differ, on average, only 3.5 degrees from the mean of 64 degrees. So, the mean doesn't help Robert very much in making a decision, but the MAD and dot plot are helpful.

Which city should he choose if he loves hot weather and really dislikes cold weather?

Exercises 5–7

5. Robert wants to compare temperatures for Cities B and C.

	J	F	M	A	M	J	J	A	S	O	N	D
City B	54	54	58	63	63	68	72	72	72	63	63	54
City C	54	44	54	61	63	72	78	85	78	59	54	54

a. Draw a dot plot of the monthly temperatures for each of the cities.

b. Verify that the mean monthly temperature for each distribution is 63 degrees.

c. Find the MAD for each of the cities. Interpret the two MADs in words and compare their values.

6. How would you describe the differences in the shapes of the monthly temperature distributions of the two cities?

7. Suppose that Robert had to decide between Cities D, E, and F.

	J	F	M	A	M	J	J	A	S	O	N	D	Mean	MAD
City D	54	44	54	59	63	72	78	87	78	59	54	54	63	10.5
City E	56	56	56	56	56	84	84	84	56	56	56	56	63	10.5
City F	42	42	70	70	70	70	70	70	70	70	70	42	63	10.5

a. Draw dot plots for each distribution.

b. Interpret the MAD for the distributions. What does this mean about variability?

c. How will Robert decide to which city he should move? List possible reasons Robert might have for choosing each city.

Lesson Summary

A data distribution can be described in terms of its center, spread, and shape.

- The center can be measured by the mean.
- The spread can be measured by the mean absolute deviation (MAD).
- A dot plot shows the shape of the distribution.

Problem Set

1. Draw a dot plot of the times that five students studied for a test if the mean time they studied was two hours and the MAD was zero hours.

2. Suppose the times that five students studied for a test is as follows:

Student	Aria	Ben	Chloe	Dellan	Emma
Time (hrs.)	1.5	2	2	2.5	2

   Michelle said that the MAD for this data set is 0 because the dot plot is balanced around 2. Without doing any calculation, do you agree with Michelle? Why or why not?

3. Suppose that the number of text messages eight students receive on a typical day is as follows:

Student	1	2	3	4	5	6	7	8
Number	42	56	35	70	56	50	65	50

   a. Draw a dot plot for the number of text messages received on a typical day by these eight students.

   b. Find the mean number of text messages these eight students receive on a typical day.

   c. Find the MAD number of text messages and explain its meaning using the words of this problem.

   d. Describe the shape of this data distribution.

   e. Suppose that in the original data set, Student 3 receives an additional five more text messages per day, and Student 4 receives five fewer messages per day.

      i. Without doing any calculation, does the mean for the new data set stay the same, increase, or decrease as compared to the original mean? Explain your reasoning.

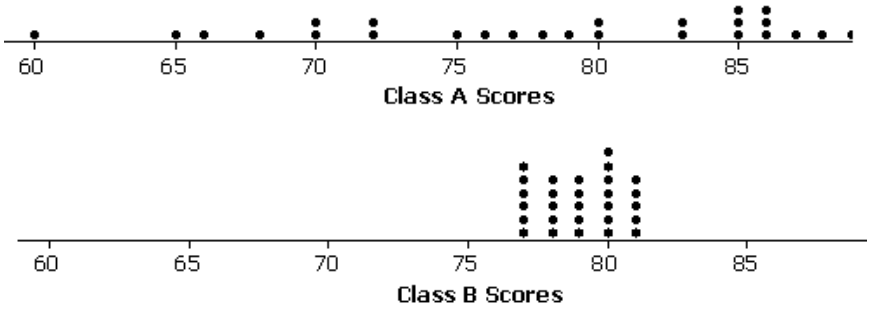
      ii. Without doing any calculation, does the MAD for the new data set stay the same, increase, or decrease as compared to the original MAD? Explain your reasoning.

Lesson 11: Describing Distributions Using the Mean and MAD

Classwork

Example 1: Comparing Distributions with the Same Mean

In Lesson 10, a data distribution was characterized mainly by its center (mean) and variability (MAD). How these measures help us make a decision often depends on the context of the situation. For example, suppose that two classes of students took the same test and their grades (based on 100 points) are shown in the following dot plots. The mean score for each distribution is 79 points. Would you rather be in Class A or Class B if you had a score of 79?



Exercises 1–6

1. Looking at the dot plots, which class has the greater MAD? Explain without actually calculating the MAD.

2. If Liz had one of the highest scores in her class, in which class would she rather be? Explain your reasoning.

3. If Logan scored below average, in which class would he rather be? Explain your reasoning.

Exercises 4–6

Your little brother asks you to replace the battery in his favorite remote control car. The car is constructed so that it is difficult to replace its battery. Your research of the lifetimes (in hours) of two different battery brands (A and B) shows the following data for 20 batteries from each brand:

A	12	14	14	15	16	17	17	18	19	20	21	21	23	23	24	24	24	25	26	27
B	18	18	19	19	19	19	19	19	20	20	20	20	20	21	21	21	21	22	22	22

4. To help you decide which battery to purchase, start by drawing a dot plot for each brand.

5. Find the mean battery life for each brand and compare them.

6. Looking at the variability of each data set shown in its dot plot, give one reason you would choose Brand A. What is one reason you would choose Brand B? Explain your reasoning.

Example 2: Comparing Distributions with Different Means

You have been comparing distributions that have the same mean, but different variability. As you have seen, deciding whether large variability or small variability is best depends on the context and on what is being asked. For example, in Exercise 2, Liz preferred to be in the distribution with more variability because she had one of the highest scores in the class. Thus, her score would have been higher had she been in Class A than had she been in Class B. Logan, on the other hand, preferred the class with lesser variability (i.e., Class B), since his score was below average.

If two data distributions have different means, how does a measure of variability play a part in making decisions?

Exercises 7–9

Suppose that you wanted to answer the following question: Are field crickets better predictors of atmospheric temperature than katydids are? Both species of insect make chirping sounds by rubbing their front wings together.

The following data are the number of chirps (per minute) for 10 insects each. All the data were taken on the same evening at the same time.

Insect	1	2	3	4	5	6	7	8	9	10
Crickets	35	32	35	37	34	34	38	35	36	34
Katydids	66	62	61	64	63	62	68	64	66	64

7. Draw dot plots for these two data distributions using the same scale, going from 30 to 70. Visually, what conclusions can you draw from the dot plots?

8. Calculate the mean and MAD for each distribution.

9. The outside temperature T can be predicted by counting the number of chirps made by these insects.

   a. For crickets, T is found by adding 40 to its mean number of chirps per minute. What value of T is being predicted by the crickets?

   b. For katydids, T is found by adding 161 to its mean number of chirps per minute and then dividing the sum by 3. What value of T is being predicted by the katydids?

   c. The temperature was 75 degrees when these data were recorded, so using the mean from each data set gave an accurate prediction of temperature. If you were going to use the number of chirps from a single cricket or a single katydid to predict the temperature, would you use a cricket or a katydid? Explain how variability in the distributions of number of chirps played a role in your decision.

Lesson Summary

This lesson focused on comparing two data distributions based on center and variability. It is important to consider the context when comparing distributions. In decision-making, drawing dot plots and calculating means and MADs can help you make informed decisions.

Problem Set

1. Two classes took the same mathematics test. Summary measures for the two classes are as follows:

	Mean	MAD
Class A	78	2
Class B	78	10

   a. Suppose that you received the highest score in your class. Would your score have been higher if you were in Class A or Class B? Explain your reasoning.

   b. Suppose that your score was below the mean score. In which class would you prefer to have been? Explain your reasoning.

2. Eight tomato plants each of two varieties, LoveEm and Wonderful, are grown under the same conditions. The numbers of tomatoes produced from each plant of each variety are shown:

Plant	1	2	3	4	5	6	7	8
LoveEm	27	29	27	28	31	27	28	27
Wonderful	31	20	25	50	32	25	22	51

   a. Draw dot plots to help you decide which variety is more productive.

   b. Calculate the mean number of tomatoes produced for each variety. Which one produces more tomatoes on average?

   c. If you want to be able to accurately predict the number of tomatoes a plant is going to produce, which variety should you choose – the one with the smaller MAD, or the one with the larger MAD? Explain your reasoning.

   d. Calculate the MAD of each plant variety.

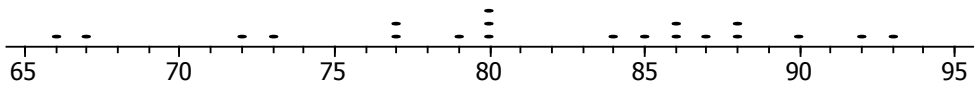
Lesson 12: Describing the Center of a Distribution Using the Median

How do we summarize a data distribution? What provides us with a good description of the data? The following exercises help us to understand how a numerical summary answers these questions.

Classwork

Example 1: The Median—A Typical Number

Suppose a chain restaurant (Restaurant A) advertises that a typical number of french fries in a large bag is 82. The graph shows the number of french fries in selected samples of large bags from Restaurant A.



Number of French Fries in a Large Bag (Restaurant A)

Exercises 1–3

1. You just bought a large bag of fries from the restaurant. Do you think you have 82 french fries? Why or why not?

2. How many bags were in the sample?

3. Which of the following statements would seem to be true given the data? Explain your reasoning.
   a. Half of the bags had more than 82 fries in them.
   b. Half of the bags had fewer than 82 fries in them.
   c. More than half of the bags had more than 82 fries in them.
   d. More than half of the bags had fewer than 82 fries in them.
   e. If you got a random bag of fries, you could get as many as 93 fries.

Example 2: The Median

Sometimes it is useful to know what point separates a data distribution into two equal parts, where one part represents the larger "half" of the data values and the other part represents the smaller "half" of the data values. This point is called the **median**. When the data are arranged in order from smallest to largest, the same number of values will be above the median as are below the median.

Exercises 4–7

4. Suppose you were trying to convince your family that you needed a new pair of tennis shoes. After checking with your friends, you argued that half of them had more than four pairs of tennis shoes, and you only had two pairs. Give another example of when you might want to know that a data value is a half-way point? Explain your thinking.

5. Use the information from the dot plot in Example 1. The median number of fries was 82.
   a. What percent of the bags have more fries than the median? Less than the median?

   b. Suppose the bag with 93 fries was miscounted and there were only 85 fries. Would the median change? Why or why not?

   c. Suppose the bag with 93 fries really only had 80 fries. Would the median change? Why or why not?

6. The owner of the chain decided to check the number of french fries at another restaurant in the chain. Here is the data for Restaurant B: 82, 83, 83, 79, 85, 82, 78, 76, 76, 75, 78, 74, 70, 60, 82, 82, 83, 83, 83.
   a. How many bags of fries were counted?

   b. Sallee claims the median is 75 as she sees that 75 is the middle number in the data set listed above. She thinks half of the bags had fewer than 75 fries. Do you think she would change her mind if the data were plotted in a dot plot? Why or why not?

   c. Jake said the median was 83. What would you say to Jake?

d. Betse argued that the median was halfway between 60 and 85 or 72.5. Do you think she is right? Why or why not?

e. Chris thought the median was 82. Do you agree? Why or why not?

7. Calculate the mean and compare it to the median. What do you observe about the two values? If the mean and median are both measures of center, why do you think one of them is lower than the other?

Exercises 8–10: Finding Medians from Frequency Tables

8. A third restaurant (Restaurant C) tallied a sample of bags of french fries and found the results below.

Number of fries	Frequency
75	||
76	|
77	||
78	|||
79	|||||
80	||||
81	|
82	|
83	
84	|||
85	|||
86	|

a. How many bags of fries did they count?

b. What is the median number of fries for the sample of bags from this restaurant? Describe how you found your answer.

9. Robere decided to divide the data into four parts. He found the median of the whole set.
   a. List the 13 values of the bottom half. Find the median of these 13 values.

   b. List the 13 values of the top half. Find the median of these 13 values.

10. Which of the three restaurants seems most likely to really have 82 fries in a typical bag? Explain your thinking.

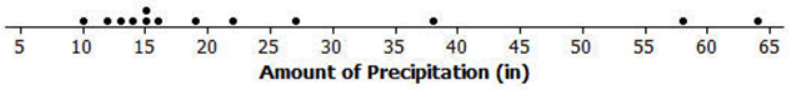
Lesson Summary

In this lesson, you learned about a summary measure for a set of data called the median. To find a median you first have to order the data. The **median** is the midpoint of a set of ordered data; it separates the data into two parts with the same number of values below as above that point. For an even number of data values, you find the average of the two middle numbers; for an odd number of data values, you use the middle value. It is important to note that the median might not be a data value and that the median has nothing to do with a measure of distance. Medians are sometimes called a measure of the center of a frequency distribution but do not have to be the middle of the spread or range (maximum-minimum) of the data.

Problem Set

1. The amount of precipitation in the western states in the U.S. is given in the table as well as the graph.

State	Amount of Precipitation (in.)
WA	38.4
OR	27.4
CA	22.2
MT	15.3
ID	18.9
WY	12.9
NV	9.5
UT	12.2
CO	15.9
AZ	13.6
NM	14.6
AK	58.3
HI	63.7



Data Source: http://www.currentresults.com/Weather/US/average-annual-state-precipitation.php

a. How do the amounts vary across the states?

b. Find the median. What does the median tell you about the amount of precipitation?

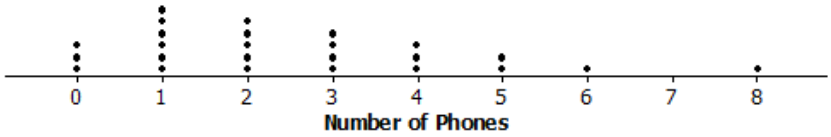
c. Use the median and the range to describe the average monthly precipitation in western states in the U.S.

d. Do you think the mean or median would be a better description of the typical amount of precipitation? Explain your thinking.

2. Identify the following as true or false. If a statement is false, give an example showing why.
   a. The median is always equal to one of the values in the data set.
   b. The median is the midpoint between the smallest and largest values in the data set.
   c. At most, half of the values in a data set have values less than the median.
   d. In a data set with 25 different values, if you change the two smallest values of a data set to smaller values, the median will not be changed.
   e. If you add 10 to every element of a data set, the median will not change.

3. Make up a data set such that the following is true:
   a. The set has 11 different values and the median is 5.
   b. The set has 10 values and the median is 25.
   c. The set has 7 values and the median is the same as the smallest value.

4. The dot plot shows the number of landline phones that a sample of people have in their homes.



   a. How many people were in the sample?
   b. Why do you think three people have no landline phones in their homes?
   c. Find the median number of phones for the people in the sample.
   d. Use the median and the range (maximum-minimum) to describe the distribution of the number of phones.

5. The salaries of the Los Angeles Lakers for the 2012–2013 basketball season are given below.

Player	Salary ($)
Kobe Bryant	$27,849,149
Dwight Howard	$19,536,360
Pau Gasol	$19,000,000
Steve Nash	$8,700,000
Metta World Peace	$7,258,960
Steve Blake	$4,000,000
Jordan Hill	$3,563,600
Chris Duhon	$3,500,000
Jodie Meeks	$1,500,000
Earl Clark	$1,240,000
Devin Ebanks	$1,054,389
Darius Morris	$962,195
Antawn Jamison	$854,389
Robert Sacre	$473,604
Darius Johnson-Odom	$203,371

Data Source: www.basketball-reference.com/contracts/LAL.html

a. Just looking at the data, what do you notice about the salaries?

b. Find the median salary, and explain what it tells you about the salaries.

c. Find the median of the lower half of the salaries and the median of the upper half of the salaries.

d. Find the width of each of the following intervals. What do you notice about the size of the interval widths, and what does that tell you about the salaries?

   i. minimum salary to median of the lower half:

   ii. median of the lower half to the median of the whole set:

   iii. median of the whole set to the median of the upper half:

   iv. median of the upper half to the highest salary:

6. Use the salary table from above to answer the following.

a. If you were to find the mean salary, how do you think it would compare to the median? Explain your reasoning.

b. Which measure do you think would give a better picture of a typical salary for the Lakers, the mean or the median? Explain your thinking.

Lesson 13: Describing Variability using the Interquartile Range (IQR)

The median was used to describe the typical value of our data in Lesson 12. Clearly, not all of the data is described by the value. How do we find a description of how the data vary? What is a good way to indicate how the data vary when we use a median as our typical value? These questions are developed in the following exercises.

Classwork

Exercises 1–4

1. In Lesson 12, you thought about the claim made by a chain restaurant that the typical number of French fries in a large bag was 82. Then, you looked at data on the number of fries in a bag from three of the restaurants.
   a. How do you think the data was collected and what problems might have come up in collecting the data?

   b. What scenario(s) would give counts that might not be representative of typical bags?

2. In Exercise 7 of Lesson 12, you found the median of the top half and the median of the bottom half of the counts for each of the three restaurants. These were the numbers you found: Restaurant A – 87.5 and 77; Restaurant B – 82 and 79; Restaurant C – 84 and 78. The difference between the medians of the two halves is called the interquartile range or IQR.
   a. What is the IQR for each of the three restaurants?

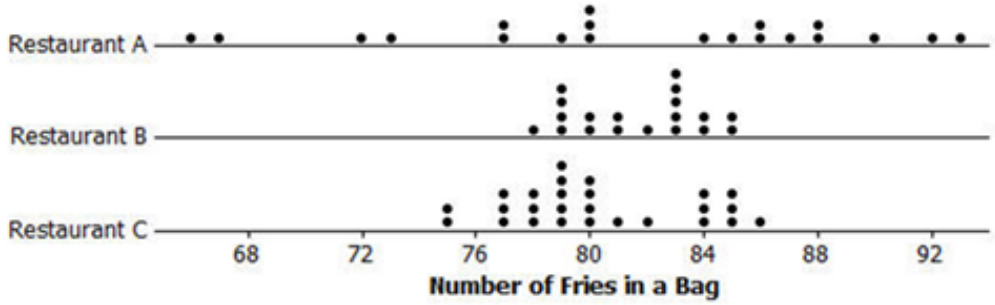
b. Which of the restaurants had the smallest IQR, and what does that tell you?

c. About what fraction of the counts would be between the quartiles? Explain your thinking.

3. The medians of the lower and upper half of a data set are called quartiles. The median of the top half of the data is called the upper quartile; the median of the bottom half of the data is called the lower quartile. Do these names make sense? Why or why not?

4.

a. Mark the quartiles for each restaurant on the graphs below.



b. Does the IQR help you decide which of the three restaurants seems most likely to really have 82 fries in a typical bag? Explain your thinking.

Example 1: Finding the IQR

Read through the following steps. If something does not make sense to you, make a note and raise it during class discussion. Consider the data: $1, 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, 17$

Creating an IQR:

I. Order the data: The data is already ordered.

$$1, 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, 17$$

II. Find the minimum and maximum: The minimum data point is 1, and the maximum is 17.

$$(1), 1, 3, 4, 6, 6, 7, 8, 10, 11, 11, 12, 15, 15, 17, 17, (17)$$

III. Find the median: There are 17 data points so the 9^{th} one from the smallest or from the largest will be the median.

$$1, 1, 3, 4, 6, 6, 7, 8, (10), 11, 11, 12, 15, 15, 17, 17, 17$$

median

IV. Find the lower quartile and upper quartile: The lower quartile (Q1) will be half way between (the mean) of the 4^{th} and 5^{th} data points (4 and 6), or 5 and the upper quartile (Q3) will be half way between the 13^{th} and the 14^{th} data points (15 and 15), or 15.

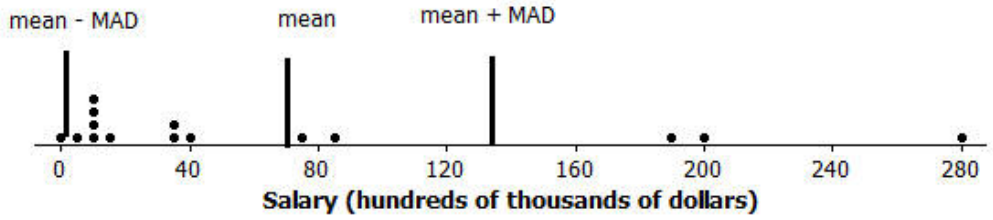
$$1, 1, 3, (4, 6), 6, 7, 8, 10, 11, (11, 12, 15, 15), 17, 17, 17$$

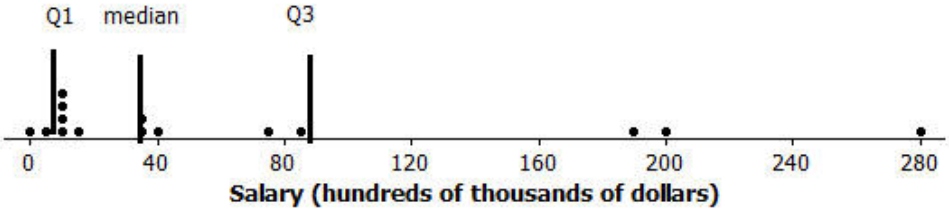
Q1 is 5 Q3 is 15

V. Find the difference between Q3 and Q1: The $IQR = 15 - 5 = 10$.

Exercises 5–6

5. When should you use the IQR? The data for the 2012 salaries for the Lakers basketball team are given in the two plots below (see problem 5 in the Problem Set from Lesson 12).





a. The data are given in hundreds of thousands of dollars. What would a salary of 40 hundred thousand dollars be?

b. The vertical lines on the top plot show the mean and the mean ± the MAD. The bottom plot shows the median and the IQR. Which interval is a better picture of the typical salaries? Explain your thinking.

6. Create three different contexts for which a set of data collected related to those contexts could have an IQR of 20. Define a median for each context. Be specific about how the data might have been collected and the units involved. Be ready to describe what the median and IQR mean in each case.

a.

b.

c.

Lesson Summary

One of our goals in statistics is to summarize a whole set of data in a short concise way. We do this by thinking about some measure of what is typical and how the data are spread relative to what is typical.

In earlier lessons, you learned about the MAD as a way to measure the spread of data about the mean. In this lesson, you learned about the IQR as a way to measure the spread of data around the median.

To find the IQR, you order the data, find the median of the data, and then find the median of the lower half of the data (the lower quartile) and the median of the upper half of the data (the upper quartile). The IQR is the difference between the upper quartile and the lower quartile, which is the length of the interval that includes the middle half of the data, because the median and the two quartiles divide the data into four sections, with about $\frac{1}{4}$ of the data in each section. Two of the sections are between the quartiles, so the interval between the quartiles would contain about 50% of the data.

Small IQRs indicate that the middle half of the data are close to the median; a larger IQR would indicate that the middle half of the data is spread over a wider interval relative to the median.

Problem Set

1. The average monthly high temperatures (in °F) for St. Louis and San Francisco are given in the table below.

	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec
St. Louis	40	45	55	67	77	85	89	88	81	69	56	43
San Francisco	57	60	62	63	64	67	67	68	70	69	63	57

Data Source: www.weather.com/weather/wxclimatology/monthly/graph/USCA0987

www.weather.com/weather/wxclimatology/monthly/graph/USMO0787

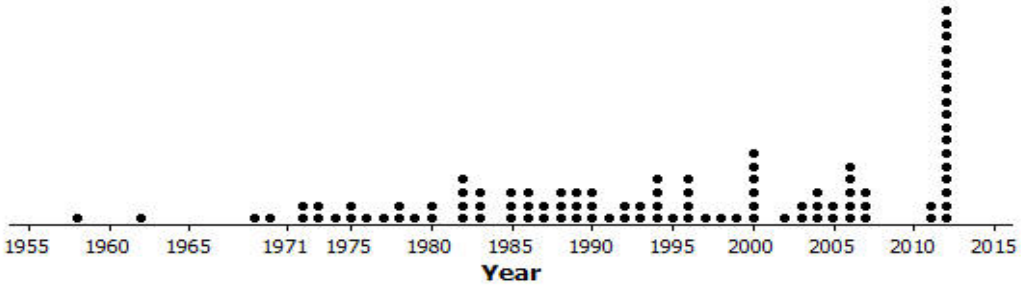
a. How do you think the data might have been collected?

b. Do you think it would be possible for $\frac{1}{4}$ of the temperatures in the month of July for St. Louis to be $95°$ or above? Why or why not?

c. Make a prediction about how the sizes of the IQR for the temperatures for each city compare. Explain your thinking.

d. Find the IQR for the average monthly high temperature for each city. How do the results compare to your conjecture?

2. The plot below shows the years in which each of 100 pennies were made.



   a. What does the stack of 17 dots at 2012 representing 17 pennies tell you about the "age" of the pennies in 2014?

   b. Here is some information about the sample of pennies. The mean year they were made is 1994; the first year any of the pennies were made was 1958; the newest pennies were made in 2012; Q1 is 1984, the median is 1994, and Q3 is 2006; the MAD is 11.5 years. Use the information to indicate the years in which the middle half of the pennies was made.

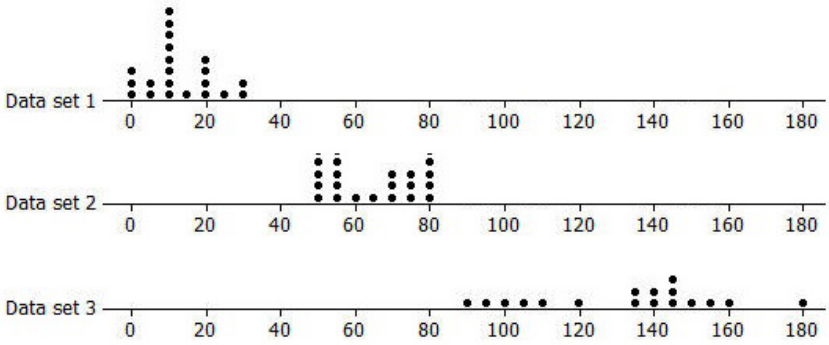
3. Create a data set with at least 6 elements such that it has the following:

   a. A small IQR and a big range (maximum-minimum).

   b. An IQR equal to the range.

   c. The lower quartile is the same as the median.

4. Rank the following three data sets by the value of the IQR.



5. Here are the counts of the fries in each of the bags from Restaurant A:
80, 72, 77, 80, 90, 85, 93, 79, 84, 73, 87, 67, 80, 86, 92, 88, 86, 88, 66, and 77.

   a. Suppose one bag of fries had been overlooked in the sample and that bag had only 50 fries. Would the IQR change? Explain your reasoning.

   b. Will adding another data value always change the IQR? Give an example to support your answer.

Lesson 14: Summarizing a Distribution Using a Box Plot

A box plot is a graph that is used to summarize a data distribution. What does the box plot tell us about the data distribution? How does the box plot indicate the variability of the data distribution?

Classwork

Example 1: Time to Get to School

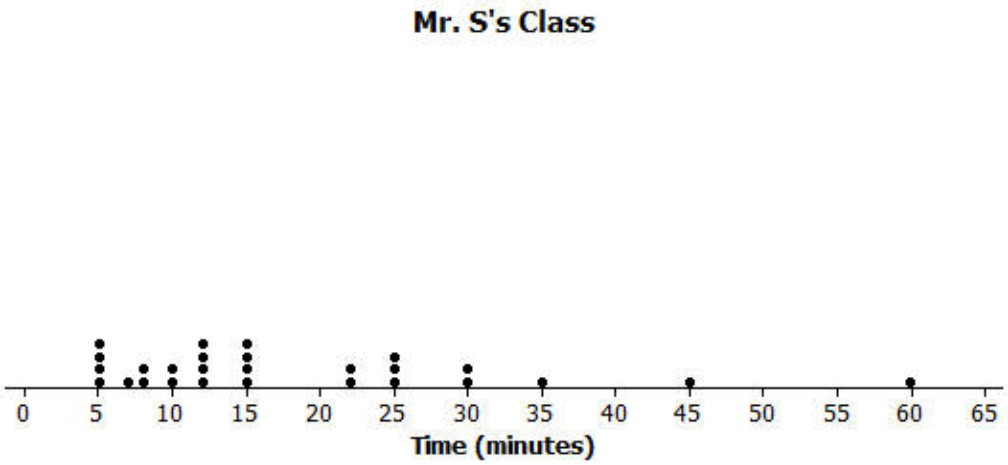
What is the typical amount of time it takes for a person in your class to get to school? The amount of time it takes to get to school in the morning varies for each person in your class. Take a minute to answer the following questions. Your class will use this information to create a dot plot.

Write your name and an estimate of the number of minutes it took you to get to school today on a post-it note.

What were some of the things you had to think about when you made your estimate?

Exercises 1–4

Here is a dot plot of the estimates of the times it took students in Mr. S's class to get to school one morning.



1. Put a line in the dot plot that seems to separate the shortest times and the longest times.

2. Put another line in the plot that separates those who seem to live really close to school and one that marks off those who took a long time to get to school.

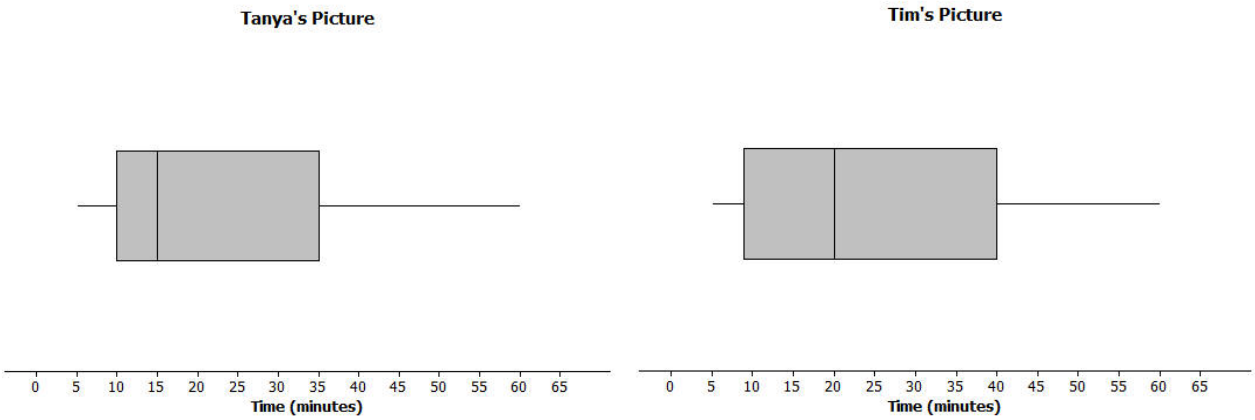
3. Your plot should be divided into four sections. Record the number of times in each of the four sections.

4. Share your marked up dot plot with some of your classmates. Compare how each of you divided the plot into four sections.

Exercises 5–7: Time to Get to School

The teacher asked the class to make a representation that would summarize the times it took students in Mr. S's class to get to school and how they are spread out. Tim decided to get rid of the dots and just use a picture of the divisions he made of the shortest times and the longest times. He put a box around the two middle sections.

Tanya thought that was a good idea and made a picture of the way she had divided the times. Here are their pictures.



5. What do the pictures tell you about the length of time it takes the students to get to school?

6. What don't the pictures tell you about the length of time it takes the students to get to school?

7. How do the two pictures compare?

Example 2: Making a Box Plot

Mr. S suggested that to be sure everyone had the same picture, statisticians developed a standard procedure for making the cut marks for the sections.

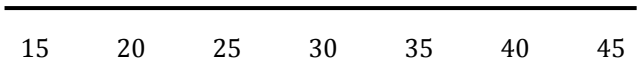
Mr. S. wrote the following on the board:

To make a box plot

- Find the median of all of the data
- Find Q1, the median of the bottom half of the data, and Q3, the median of the top half of the data.
- Draw a box that goes from Q1 to Q3, the two middle sections.
- Draw a line segment connecting the minimum value to the box and one that connects the maximum value to the box.

Now use the given number line to make a box plot of the data below.

20, 21, 25, 31, 35, 38, 40, 42, 44



The 5-number summary is as follows:

Min = 20
Q1 = 23
Median = 35
Q3 = 41
Max = 44

Exercises 8–11: A Human Box Plot

Consider again your post-it note that you used to write down the number of minutes it takes you to get to school. If possible, you and your classmates will form a human box plot of the number of minutes it takes your class to get to school.

8. Find the median of the group. Does someone represent the median? If not, who is the closest to the median?

9. Find the maximum and minimum of the group. Who are they?

10. Find Q1 and Q3 of the group. Does anyone represent Q1 or Q3? If not, who is the closest to Q1? Who is closest to Q3?

11. Sketch the box plot for this data set below.

Lesson Summary

The focus of this lesson is moving from a plot that shows all of the data values (dot plot) to one that summarizes the data with five points (box plot).

You learned how to make a box plot by doing the following:

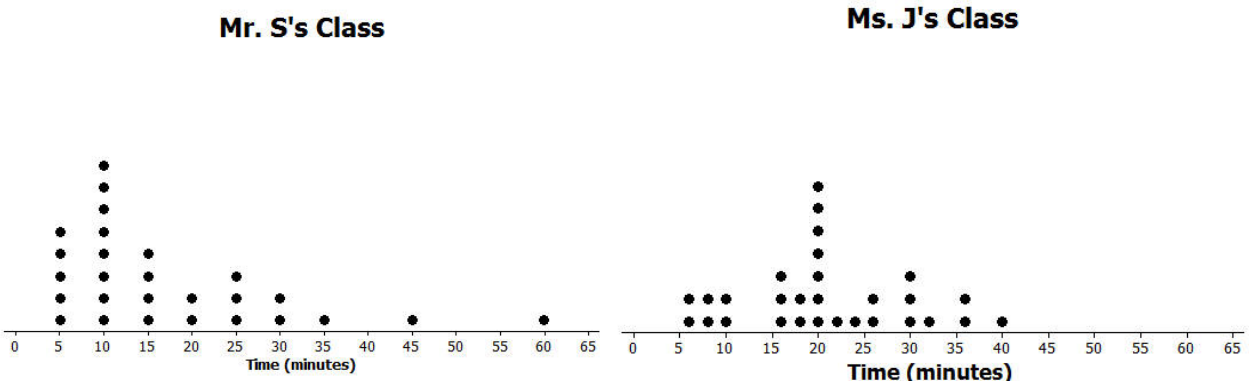
- Finding the median of all of the data
- Finding Q1, the median of the bottom half of the data, and Q3, the median of the top half of the data.
- Drawing a box that goes from Q1 to Q3, the two middle sections.
- Drawing a line segment connecting the minimum value to the box and one that connects the maximum value to the box.

You also learned important characteristics of box plots:

- $\frac{1}{4}$ of the data are in each of the sections of the plot.
- The length of the interval for a section does not indicate either how the data are grouped in that interval or how many values are in the interval.

Problem Set

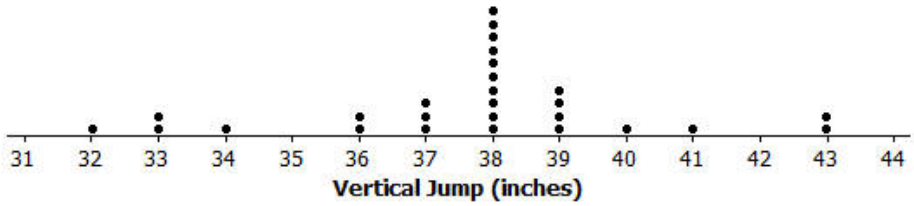
1. Dot plots for the amount of time it took students in Mr. S's and Ms. J's classes to get to school are below



   a. Make a box plot of the times for each class.

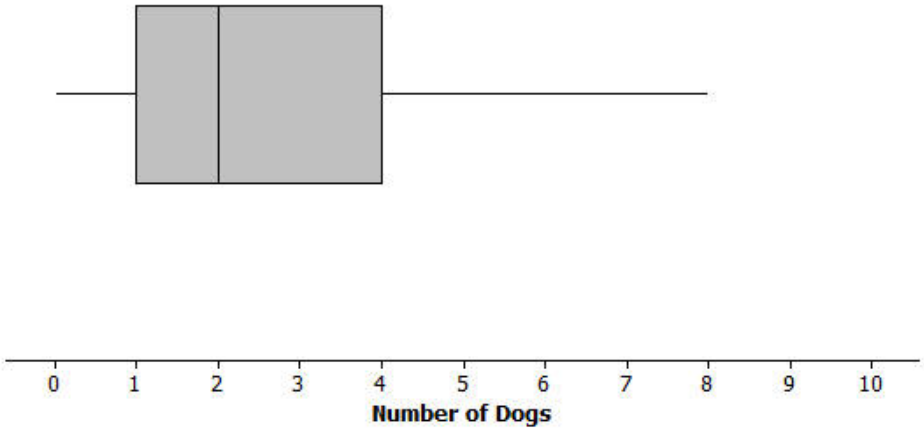
   b. What is one thing you can see in the dot plot that you cannot see in the box plot? What is something that is easier to see in the box plot than in the dot plot?

2. The dot plot below shows the vertical jump of some NBA players. A vertical jump is how high a player can jump from a standstill. Draw a box plot of the heights for the vertical jumps of the NBA players above the dot plot.



3. The mean daily temperatures in °F for the month of February for a certain city are as follows:
4, 11, 14, 15, 17, 20, 30, 23, 20, 35, 35, 31, 34, 23, 15, 19, 39, 22, 15, 15, 19, 39, 22, 23, 29, 26, 29, 29
   a. Make a box plot of the temperatures.
   b. Make a prediction about the part of the United States you think the city might be located. Explain your reasoning.
   c. Describe the data distribution of temperature. Include a description of the center and spread.

4. The plot below shows the results of a survey of households about the number of dogs they have. Identify the following statements as true or false. Explain your reasoning in each case.



   a. The maximum number of dogs per house is 8.
   b. At least $\frac{1}{2}$ of the houses have 2 or more dogs.
   c. All of the houses have dogs.
   d. Half of the houses surveyed have between 2 and 4 dogs.
   e. Most of the houses surveyed have no dogs.

Lesson 15: More Practice with Box Plots

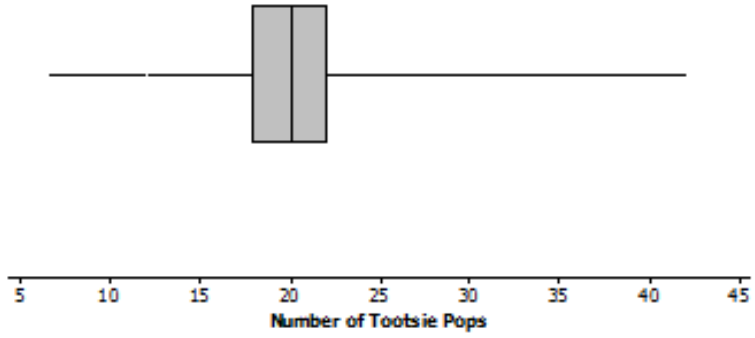
You reach into a jar of Tootsie Pops. How many Tootsie Pops do you think you could hold in one hand? Do you think the number you could hold is greater than or less than what other students can hold? Is the number you could hold a typical number of Tootsie Pops? This lesson examines these questions.

Classwork

Example 1: Tootsie Pops

As you learned earlier, the five numbers that you need to make a box plot are the minimum, the lower quartile, the median, the upper quartile, and the maximum. These numbers are called the 5-number summary of the data.

Ninety-four people were asked to grab as many Tootsie Pops as they could hold. Here is a box plot for these data. Are you surprised?



Exercises 1–5

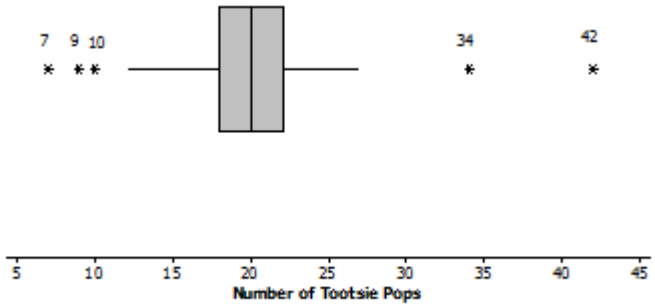
1. What might explain the variability in how many Tootsie Pops those 94 people were able to hold?

2. Estimate the values in the 5-number summary from the box plot.

3. Describe how the box plot can help you understand the difference in the number of Tootsie Pops people could hold.

4. Here is Jayne's description of what she sees in the plot. Do you agree or disagree with her description? Explain your reasoning.
"One person could hold as many as 42 Tootsie Pops. The number of Tootsie Pops people could hold was really different and spread about equally from 7 to 42. About one half of the people could hold more than 20 Tootsie Pops."

5. Here is a different plot of the same data on the number of Tootsie Pops 94 people could hold.



a. Why do you suppose the five values are separate points and are labeled?

b. Does knowing these data values change anything about your responses to Exercises 1 to 4 above?

Exercises 6–10: Maximum Speeds

The maximum speeds of selected birds and land animals are given in the tables below.

Bird	Speed (mph)
Peregrine falcon	242
Swift bird	120
Spine-tailed swift	106
White-throated needletail	105
Eurasian hobby	100
Pigeon	100
Frigate bird	95
Spur-winged goose	88
Red-breasted merganser	80
Canvasback duck	72
Anna's Hummingbird	61.06
Ostrich	60

Land animal	Speed (mph)
Cheetah	75
Free-tailed bat (in flight)	60
Pronghorn antelope	55
Lion	50
Wildebeest	50
Jackrabbit	44
African wild dog	44
Kangaroo	45
Horse	43.97
Thomson's gazelle	43
Greyhound	43
Coyote	40
Mule deer	35
Grizzly bear	30
Cat	30
Elephant	25
Pig	9

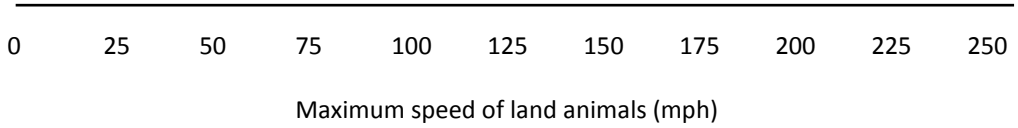
Data Source: Natural History Magazine, March 1974, copyright 1974; The American Museum of Natural History; and James G. Doherty, general curator, The Wildlife Conservation Society; *http://www.thetravelalmanac.com/lists/animals-speed.htm; http://en.wikipedia.org/wiki/Fastest_animals*

6. As you look at the speeds, what strikes you as interesting?

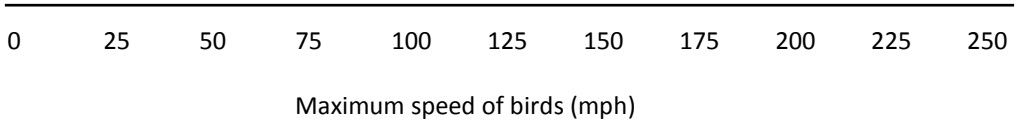
7. Do birds or land animals seem to have the greatest variability in speeds? Explain your reasoning.

8. Find the 5-number summary for the speeds in each data set. What do the 5-number summaries tell you about the distribution of speeds for each data set?

9. Use the 5-number summaries to make a box plot for each of the two data sets.



Maximum speed of land animals (mph)

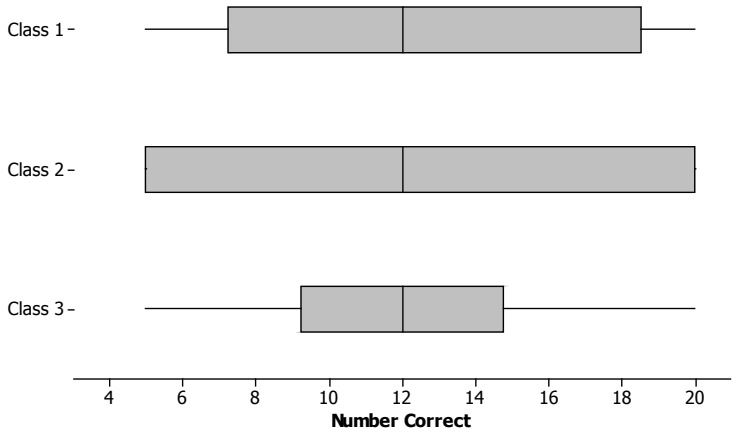


Maximum speed of birds (mph)

10. Write several sentences to tell someone about the speeds of birds and land animals.

Exercises 11–15: What is the Same and What is Different?

Consider the following box plots, which show the number of questions different students in three different classes got correct on a 20-question quiz.



11. Describe the variability in the scores of the three classes.

12. a. Estimate the interquartile range for each of the three sets of scores.

b. What fraction of students does the interquartile range represent?

c. What does the value of the IQR tell you about how the scores are distributed?

13. The teacher asked students to draw a box plot with a minimum value at 34 and a maximum value at 64 that had an interquartile range of 10. Jeremy said he could not draw just one because he did not know where to put the box on the number line. Do you agree with Jeremy? Why or why not?

14. Which class do you believe performed the best? Be sure to use the data from the box plots to back up your answer.

15. a. Find the IQR for the three data sets in the first two examples: maximum speed of birds, maximum speed of land animals, and number of Tootsie Pops.

    b. Which data set had the highest percentage of data values between the lower quartile and the upper quartile? Explain your thinking.

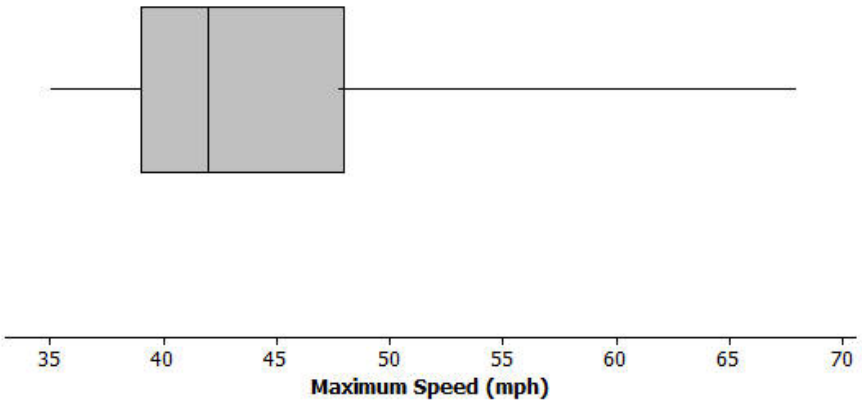
Lesson Summary

In this lesson, you learned about the 5-number summary for a set of data: minimum, lower quartile, median, upper quartile, and maximum. You made box plots after finding the 5-number summary for two sets of data (speeds of birds and land animals), and you estimated the 5-number summary from box plots (number of Tootsie Pops people can hold, class scores). You also found the interquartile range (IQR), which is the difference between the upper quartile and lower quartile. The IQR, the length of the box in the box plot, indicates how closely the middle half of the data is bunched around the median. (Note that because sometimes data values repeat and the same numerical value may fall in two sections of the plot, it is not always exactly half. This happened with the two speeds of 50 mph – one went into the top quarter of the data and the other into the third quarter – the upper quartile was 50.)

You also practiced describing a set of data using the 5-number summary, making sure to be as precise as possible- avoiding words like "a lot" and "most" and instead saying about one half or three fourths.

Problem Set

1. The box plot below summarizes the maximum speeds of certain kinds of fish.



   a. Estimate the 5-number summary from the box plot.
   b. The fastest fish is the sailfish at 68 mph followed by the marlin at 50 mph. What does this tell you about the spread of the fish speeds in the top quarter of the plot?
   c. Use the 5-number summary and the IQR to describe the speeds of the fish.

2. Suppose you knew that the interquartile range for the number of hours students spent playing video games during the school week was 10. What do you think about each of the following statements? Explain your reasoning.
   a. About half of the students played video games for 10 hours during a school week.
   b. All of the students played at least 10 hours of video games during the school week.
   c. About half of the class could have played video games from 10 to 20 hours a week or from 15 to 25 hours.

3. Suppose you know the following for a data set: minimum value is 130, the lower quartile is 142, the IQR is 30, half of the data are less than 168, and the maximum value is 195.
   a. Think of a context for which these numbers might make sense.
   b. Sketch a box plot.
   c. Are there more data values above or below the median? Explain your reasoning.

4. The speeds for the fastest dogs are given in the table below.

Breed	Speed (mph)
Greyhound	45
African Wild Dog	44
Saluki	43
Whippet	36
Basanji	35
German Shepherd	32
Vizsla	32
Doberman Pinscher	30

Breed	Speed (mph)
Irish Wolfhound	30
Dalmatian	30
Border Collie	30
Alaskan Husky	28
Giant Schnauzer	28
Jack Russell Terrier	25
Australian Cattle Dog	20

Data Source: http://www.vetstreet.com/our-pet-experts/meet-eight-of-the-fastest-dogs-on-the-planet; http://canidaepetfood.blogspot.com/2012/08/which-dog-breeds-are-fastest.html

   a. Find the 5-number summary for this data set and use it to create a box plot of the speeds.
   b. Why is the median not in the center of the box?
   c. Write a few sentences telling your brother or sister about the speed of the fastest dogs.

Lesson 16: Understanding Box Plots

Classwork

Exercise 1: Supreme Court Chief Justices

The Supreme Court is the highest court of law in the United States, and it makes decisions that affect the whole country. The Chief Justice is appointed to the Court and will be a justice the rest of his or her life unless he or she resigns or becomes ill. Some people think that this gives the Chief Justice a very long time to be on the Supreme Court. The first Chief Justice was appointed in 1789.

The table shows the years in office for each of the Chief Justices of the Supreme Court as of 2013:

Name	Years	Appointed in
John Jay	6	1789
John Rutledge	1	1795
Oliver Ellsworth	4	1796
John Marshall	34	1801
Roger Brooke Taney	28	1836
Salmon P. Chase	9	1864
Morrison R. Waite	14	1874
Melville W. Fuller	22	1888
Edward D. White	11	1910
William Howard Taft	9	1921
Charles Evens Hughes	11	1930
Harlan Fiske Stone	5	1941
Fred M. Vinson	7	1946
Earl Warren	16	1953
Warren E. Burger	17	1969
William H. Rehnquist	19	1986
John G. Roberts	8	2005

Data Source: http://en.wikipedia.org/wiki/List_of_Justices_of_the_Supreme_Court_of_the_United_States

1. Use the table to answer the following:
   a. Which Chief Justice served the longest term and which served the shortest term? How many years did each of these Chief Justices serve?

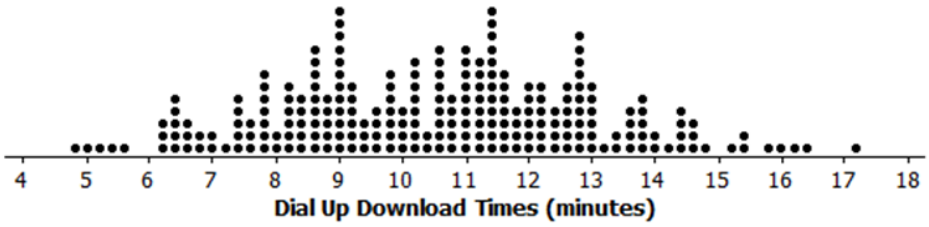
b. What is the median number of years these Chief Justices have served on the Supreme Court? Explain how you found the median and what it means in terms of the data.

c. Make a box plot of the years the justices served. Describe the shape of the distribution and how the median and IQR relate to the box plot.

d. Is the median half way between the least and the most number of years served? Why or why not?

Exercises 2–3: Downloading Songs

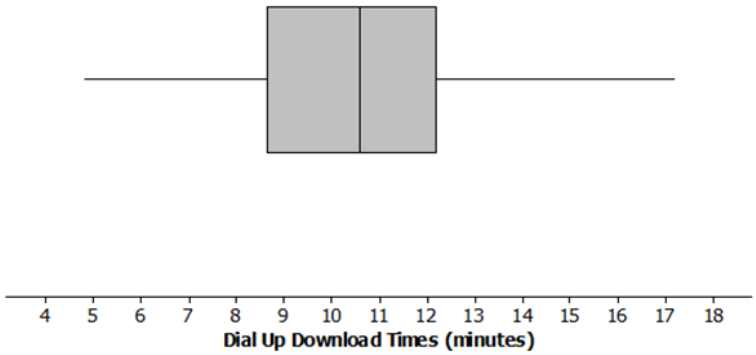
2. A broadband company timed how long it took to download 232 four-minute songs on a dial up connection. The dot plot below shows their results.



a. What can you observe about the download times from the dot plot?

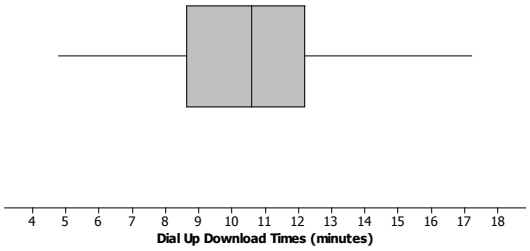
b. Is it easy to tell whether or not 12.5 minutes is in the top quarter of the download times?

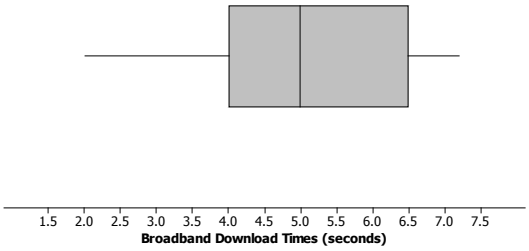
c. The box plot of the data is shown below. Now answer parts (a) and (b) above using the box plot.



d. What are the advantages of using a box plot to display a large set of data? What are the disadvantages?

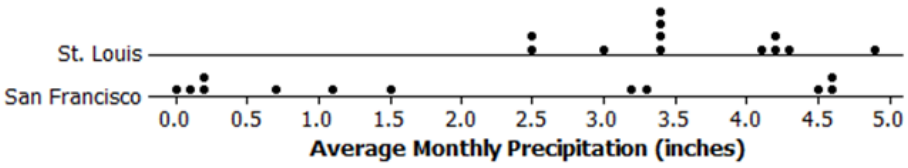
3. Molly presented the plots below to argue that using a dial up connection would be better than using a broadband connection. She argued that the dial up connection seems to have less variability around the median even though the overall range seems to be about the same for the download times using broadband. What would you say?





Exercises 4–5: Rainfall

4. Data on average rainfall for each of the twelve months of the year were used to construct the two dot plots below.



a. How many data points are in each dot plot? What does each data point represent?

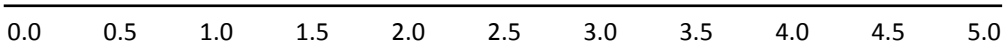
b. Make a conjecture about which city has the most variability in the average monthly amount of precipitation and how this would be reflected in the IQRs for the data from both cities.

c. Based on the dot plots, what are the approximate values of the interquartile ranges (IQR) of the amount of average monthly precipitation in inches for each city? Use each IQR to compare the cities.

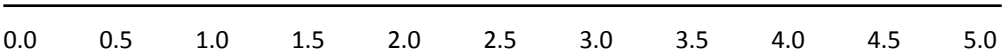
d. In an earlier lesson, the average monthly temperatures were rounded to the nearest degree Fahrenheit. Would it make sense to round the amount of precipitation to the nearest inch? Why or why not?

5. Use the data from Exercise 4 to answer the following.

   a. Make a box plot of the amount of precipitation for each city.



Average Monthly Precipitation in St. Louis (inches)



Average Monthly Precipitation in San Francisco (inches)

   b. Compare the percent of months that have above 2 inches of precipitation for the two cities. Explain your thinking.

   c. How do the top fourths of the average monthly precipitation in the two cities compare?

   d. Describe the intervals that contain the smallest 25% of the average monthly precipitation amounts for each city.

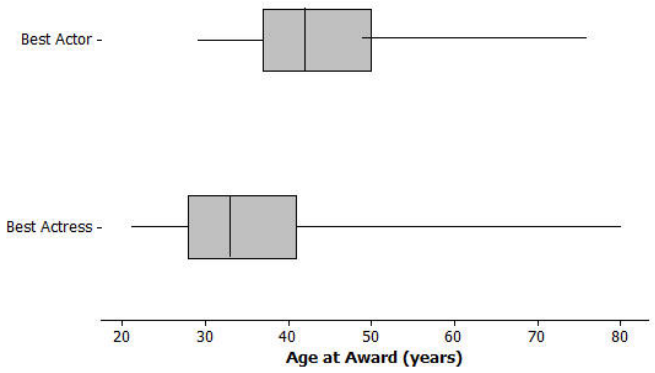
   e. Think about the dot plots and the box plots. Which representation do you think helps you the most in understanding how the data vary?

Lesson Summary

In this lesson, you reviewed what you know about box plots, the 5-number summary of the data used to construct a box plot, and the IQR. Box plots are very useful for comparing data sets and for working with large amounts of data. When you compare two or more data sets using box plots; however, you have to be sure that the scales and units are the same.

Problem Set

1. The box plots below summarize the ages at the time of the award for leading actress and leading actor Academy Award winners.



Data Source: http://en.wikipedia.org/wiki/List_of_Best_Actor_winners_by_age_at_win
http://en.wikipedia.org/wiki/List_of_Best_Actress_winners_by_age_at_win

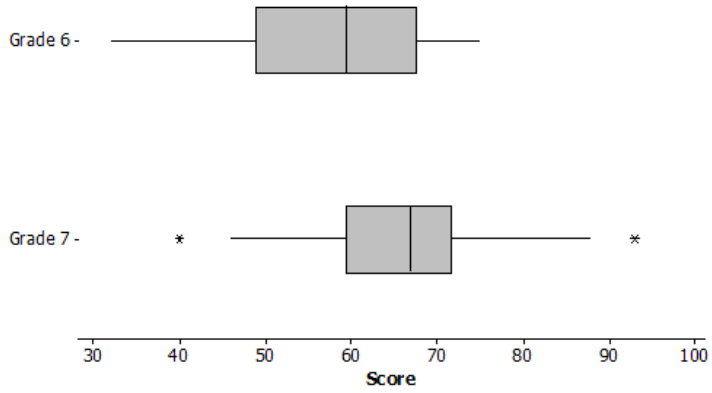
a. Do you think it is harder for an older woman to win an academy award for best actress than it is for an older man to win a best actor award? Why or why not?

b. The oldest female to win an academy award was Jessica Tandy in 1990 for *Driving Miss Daisy*. The oldest actor was Henry Fonda for *On Golden Pond* in 1982. How old were they when they won the award? How can you tell? Were they a lot older than most of the other winners?

c. The 2013 winning actor was Daniel Day-Lewis for *Lincoln*. He was 55 years old at that time. What can you say about the percent of male award winners who were older than Daniel Day-Lewis when they won their Oscar?

d. Use the information you can see in the box plots to write a paragraph supporting or refuting the claim that fewer older actresses than actors win academy awards.

2. The scores of sixth and seventh graders on a test about polygons and their characteristics are summarized in the box plots below.



   a. In which grade did the students do the best? Explain how you can tell.

   b. Why do you think two of the data values in grade seven are not part of the line segments?

   c. How do the median scores for the two grades compare? Is this surprising? Why or why not?

   d. How do the IQRs compare for the two grades?

3. A formula for IQR could be written as $Q3 - Q1 = IQR$. Suppose you knew the IQR and the Q1. How could you find the Q3?

4. Consider the statement, "Historically, the average length of service as Chief Justice on the Supreme Court has been less than 15 years; however, since 1970 the average length of service has increased." Use the data given in Exercise 1 to answer the following questions.

   a. Do you agree or disagree with the statement? Explain your thinking.

   b. Would your answer change if you used the median number of years rather than the mean?

Lesson 17: Developing a Statistical Project

Classwork

Statistical questions you investigated in this module included the following:

- How many hours of sleep do 6th graders typically get on a night when there is school the next day?
- What is the typical number of books read over the course of 6 months by a 6th grader?
- What is the typical heart rate of a student in a 6th grade class?
- How many hours does a 6th grader typically spend playing a sport or a game outdoors?
- What is the head circumference of adults interested in buying baseball hats?
- How long is the battery life of a certain brand of batteries?
- How many pets do students have?
- How long does it take a student to get to school?
- What is a typical daily temperature of New York City?
- What is the typical weight of a backpack for students at a certain school?
- What is the typical number of french fries in a large order from a fast food restaurant?
- What is the typical number of minutes a student spends on homework each day?
- What is the typical height of a vertical jump for a player in the NBA?

What do these questions have in common?

Why do several of these questions include the word "typical"?

A Review of a Statistical Study

Recall from the very first lesson in this module that a statistical question is a question answered by data that you anticipate will vary.

Let's review the steps of a statistical investigation.

Step 1: Pose a question that can be answered by data.

Step 2: Collect appropriate data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the numerical summaries and graphs.

The first step is to pose a statistical question. Select one of the above questions and write it in the following Statistical Study Review Template.

The second step is to collect the data. In all of these investigations, you were given data. How do you think the data for the question you selected in Step 1 was collected? Write your answer in the summary below for Step 2.

The third step involves the various ways you summarize the data. List the various ways you summarized data for Step 3.

Step 1: Statistical question.
Step 2: Collect data.
Step 3: Summarize the data.

Finally, the fourth step is to answer the statistical question. The answer to the statistical question was the focus of the investigation in each of the lessons. Describing a data distribution in terms of shape, center, and spread or, depending on the shape of the data distribution, calculating the mean or the median of the data often answer statistical questions.

Project (Exploratory Challenge)

Now it is your turn to answer a statistical question based on data you collect. Before you collect data, explore possible statistical questions. For each question, indicate data that you would collect and summarize to answer the question. Also indicate how you plan to collect the data.

Think of questions that could be answered by data collected from members of your class or school or data that could be collected from recognized websites (e.g., The American Statistical Association and the project Census At Schools). Check with your teacher if you are planning to work with data from an outside source such as one of the above websites. Your teacher will need to approve both your question and your plan to collect data before data are collected.

As a class, explore possibilities of a statistical investigation. Record some of the ideas discussed by your class using the following table.

Possible statistical questions	What data would be collected and how would the data be collected?

After discussing several of the above possibilities of a statistical project, prepare a statistical question and a plan to collect data to present to your teacher. After your teacher approves your question and data collection plan, begin collecting the data. Carefully organize your data as you begin developing the summaries to answer your statistical question. In future lessons, you will be directed to begin creating a poster or an outline of a presentation that will be shared with your teacher and other members of your class.

For this lesson, complete the following to present to your teacher:

1. The statistical question for my investigation is as follows:

2. Here is the plan I propose to collect my data. (Include the exact questions you may ask an individual or a clear description of what you plan to measure or count.)

Lesson Summary

A statistical study involves a four-step investigative process:

- Pose questions that can be answered by data.
- Design a plan for collecting appropriate data and then use the plan to collect data.
- Analyze the data.
- Interpret results and draw valid conclusions from the data to the question posed.

Problem Set

Your teacher will outline steps you are expected to complete in the next several days to develop this project. Keep in mind that the first step in developing your project is a statistical question. With one of the statistical questions posed in this lesson or with a new one developed in this lesson, organize your question and plan to collect and summarize data. Complete the process as outlined by your teacher.

Lesson 18: Connecting Graphical Representations and Numerical Summaries

It can be difficult to understand a data set by just looking at raw data. Imagine that Joaquin's project is on finding the typical weight of bears. He found an article about bears that provided an unsorted list of the weights of 250 bears with no numerical summaries or graphs of these data. It would be very difficult to quickly draw some conclusions. Joaquin decided to design his project using this data.

Now consider the case where the article provides you with a statement, "the median weight for the bears studied was 305 pounds." This is useful, but sometimes even numerical summaries alone cannot completely convey interesting aspects of a data distribution. Often, readers want to have a concise and useful summary of the information that is both numerical and visual.

In the next couple of lessons, you will begin to take the graphical representations and numerical summaries you learned and apply them to different situations. While working through these lessons, keep in mind your own statistical question. Think about which graphs will best showcase your data and which numerical summaries will represent the data you are collecting.

Classwork

Example 1: Summary Information from Graphs

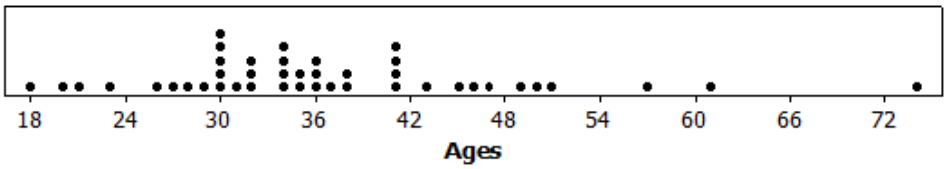
Recall that a *dot plot* includes a dot on a scale or number line for each observation in a data set. Dots are stacked on top of one another when there are multiple occurrences of a data value. Recall also that a *histogram* similarly uses a scale or number line to present the frequency or relative frequency of groups of data based on intervals of equal width. For each interval, the height of the bar is proportional to the number of observations in the interval; the taller the bar, the greater the number of observations in that interval. This means that when both graphs are generated for a given data set, the two graphs will display some similarities.

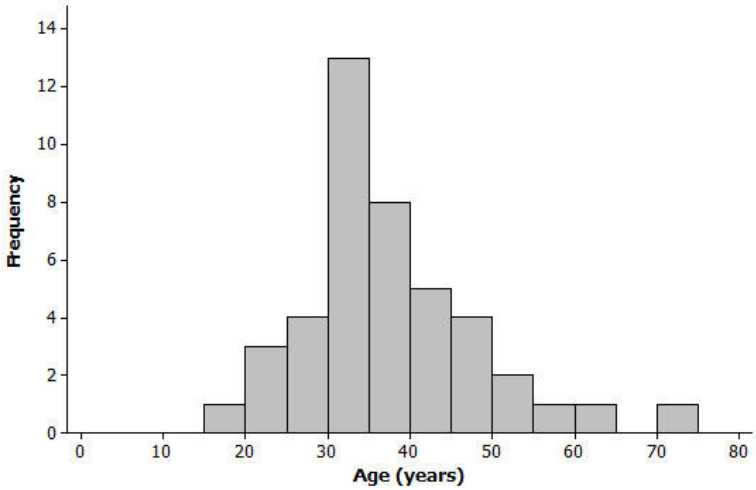
Here is a data set of the ages (in years) of 43 participants in a recent local 5-kilometer race.

20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here are some summary statistics, a histogram, and a dot plot for the data:

Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74; Mean = 36.81, MAD = 8.1





Exercises 1–7

1. Based on the histogram, would you describe the shape of the data distribution as approximately symmetric or as skewed? Would you have reached this same conclusion looking at the dot plot?

2. Is it easier to see the shape of the data distribution from the histogram or the dot plot?

3. What is something you can see in the dot plot that is not as easy to see in the histogram?

4. Do the dot plot and the histogram seem to be centered in about the same place?

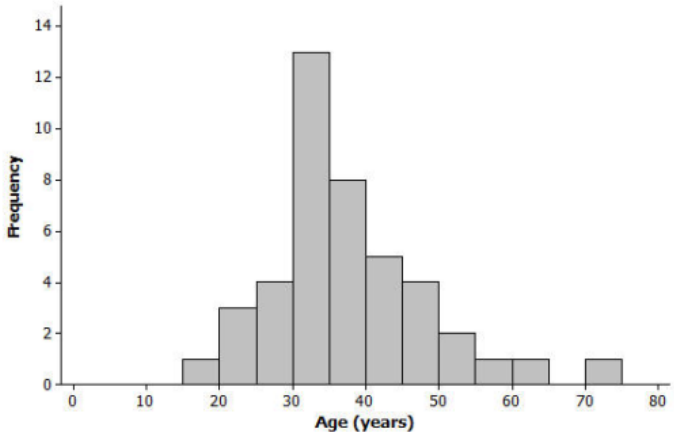
5. Do both the dot plot and the histogram convey information about the variability in the age distribution?

6. If you did not have the original data set and only had the dot plot and the histogram, would you be able to find the value of the median age from the dot plot?

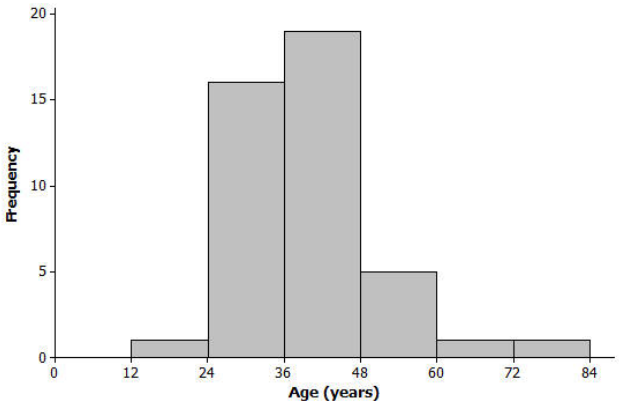
7. Explain why you would only be able to estimate the value of the median if you only had a histogram of the data.

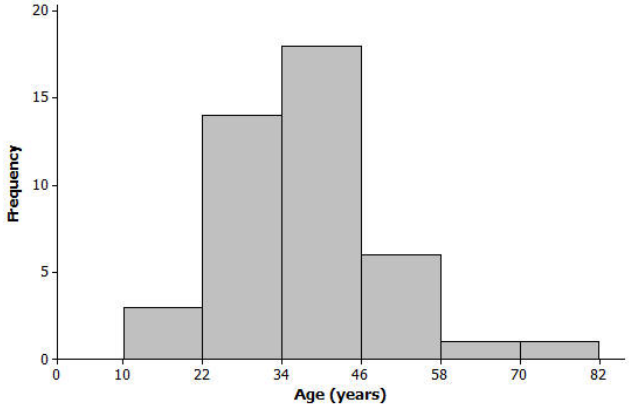
Exercises 8–13: Graphs and Numerical Summaries

8. Suppose that a newspaper article was written about the race and the histogram of the ages from Example 1 was shown in the article. The writer stated, "The race attracted many older runners this year; the median age was 45." Explain how we would know that this is an incorrect statement based on just the histogram.

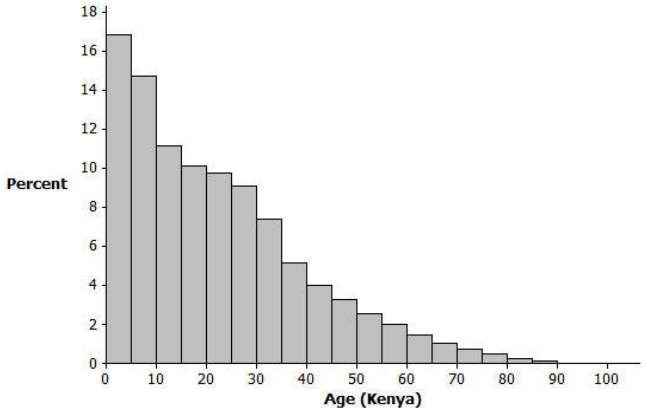


9. One of the histograms below is another valid histogram for the runners' ages. Select the correct histogram, and explain how you determined which graph is valid (and which one is incorrect) based on the summary measures and dot plot.





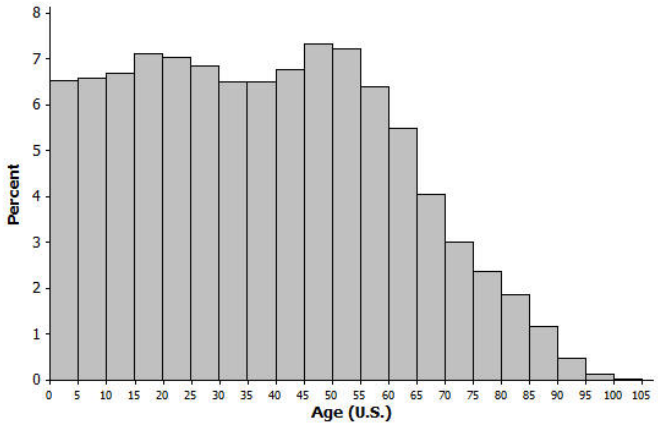
10. The histogram below represents the age distribution of the population of Kenya in 2010.



a. How do we know from the graph above that the first quartile (Q1) of this age distribution is between 5 and 9 years of age?

b. Someone believes that the median age is near 30. Explain how the graph supports this belief, OR explain why the graph does not support this belief.

11. The histogram below represents the age distribution of the population of the United States in 2010. Based on the histogram, which of the following ranges do you think includes the median age for the United States: 20–29, 30–39, or 40–49? Why?

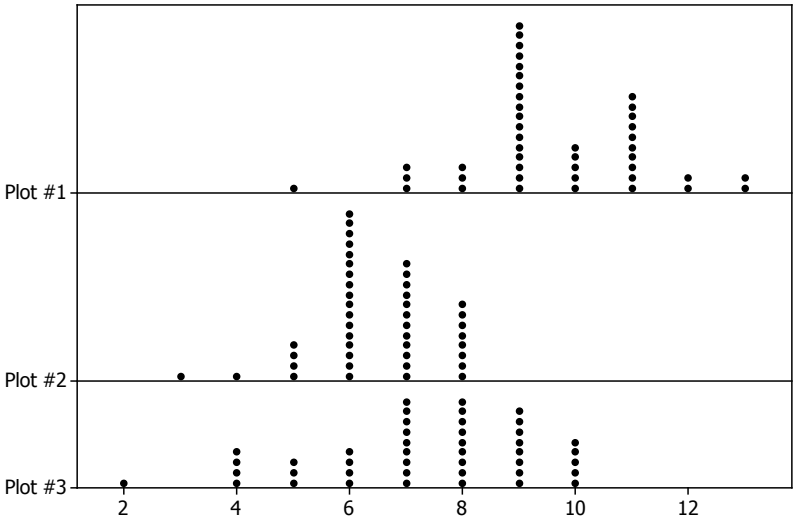


12. Use the histograms from Exercises 10 and 11 to answer the following:

    a. Which country's age distribution (Kenya or United States) has a third quartile in the 50s? How did you decide?

    b. If someone believed that the average age of a person living in the United States was greater than the average age of a person living in Kenya, how could you support that claim by comparing the histograms?

13. Match the following sets of summary measures with the corresponding dot plot. Only ONE dot plot matches each group of summary measures. Explain why you selected the dot plot or why the other dot plots would not represent the summary measures. Note: the same scale is used in each dot plot.



    a. Median = 8 and IQR = 3 Plot # _____

    b. Mean = 9.6 and MAD = 1.28 Plot # _____

    c. Median = 6 and Range = 5 Plot # _____

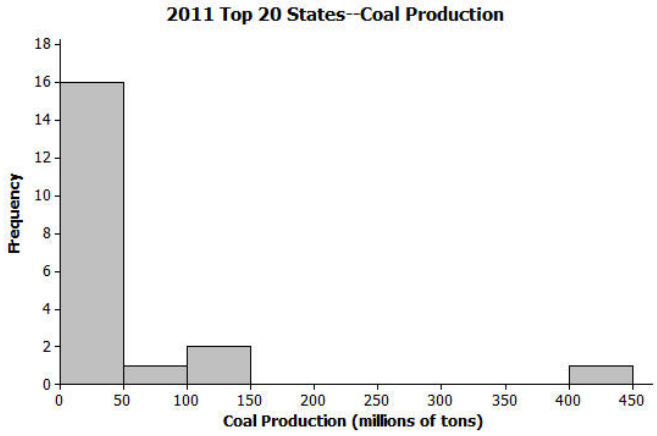
Lesson Summary

Generally, we can compute or approximate many values in a numerical summary for a data set by looking at a histogram or a dot plot for the data set. Thus, we can generally match a histogram or a dot plot to summary measures provided.

When making a histogram and a dot plot for the same data set, the two graphs will have similarities. However, some information may be more easily communicated by one graph as opposed to the other.

Problem Set

1. The following histogram shows the amount of coal produced (by state) for the 20 largest coal producing states in 2011. Many of these states produced less than 50 million tons of coal, but one state produced over 400 million tons (Wyoming). For the histogram, which ONE of the three sets of summary measures could match the graph? For each choice that you eliminate, list at least one reason for eliminating the choice.



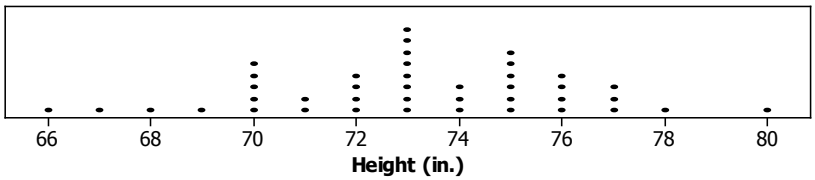
(U.S. Coal Production by State data as reported by the National Mining Association from http://www.nma.org/pdf/c_production_state_rank.pdf accessed May 5, 2013)

a. Minimum = 1, Q1 = 12, Median = 36, Q3 = 57, Maximum = 410; Mean = 33, MAD = 2.76

b. Minimum = 2, Q1 = 13.5, Median = 27.5, Q3 = 44, Maximum = 439; Mean = 54.6, MAD = 52.36

c. Minimum = 10, Q1 = 37.5, Median = 62, Q3 = 105, Maximum = 439; Mean = 54.6, MAD = 52.36

2. The heights (rounded to the nearest inch) of the 41 members of the 2012–2013 University of Texas Men's Swimming and Diving Team are shown in the dot plot below.

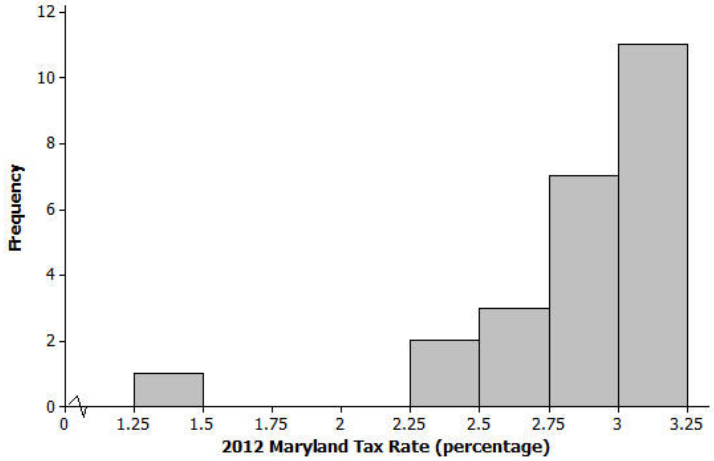


Data Source: http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html accessed April 30, 2013

a. Use the dot plot to determine the 5-number summary (minimum, lower quartile, median, upper quartile, and maximum) for the data set.

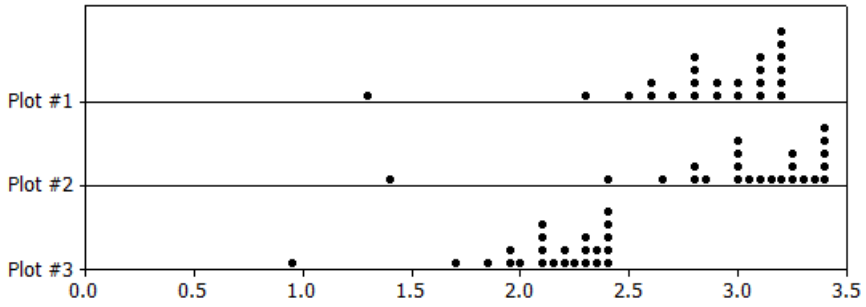
b. Based on this dot plot, make a histogram of the heights using the following classes: $66 - < 68$ inches, $68 - < 70$ inches, and so on.

3. According to the website of the Comptroller of Maryland, "Maryland's 23 counties and Baltimore City levy a local income tax … Local officials set the rates, which range between 1.25% and 3.20% for the current tax year (2012)." A histogram of the 24 tax rates (in percentages) appears below.



Data Source: http://taxes.marylandtaxes.com accessed May 5, 2013

Which ONE of the three dot plots below matches the "2012 Maryland Tax Rates" histogram above? Explain how you determined the correct dot plot.



4. For each of the following five sets of summary measures, indicate if the set of summary measures could match the "2012 Maryland Tax Rates" histogram above. For each set of summary measures that you eliminate, explain why you eliminated that choice.
   a. Mean = 1.01, MAD = 5.4
   b. Median = 2.93, IQR = 0.45
   c. Mean = 3.5, MAD = 1.1
   d. Median = 3.10, IQR = 2.15
   e. Minimum = 1.25, Maximum = 3.20

Lesson 19: Comparing Data Distributions

As you have seen in previous lessons, it can be difficult to understand a data set just by looking at raw data. Often, readers want to have a concise and useful summary.

This becomes extremely important when data distributions are compared to one another. While a reader may be interested in knowing if a typical adult male polar bear in Alaska is larger than a typical adult male grizzly bear in British Columbia, it would also be useful to be able to compare the variability and shape of the distributions of these two groups of bears as well. With summary graphs of the two distributions placed side-by-side, you can more easily assess and compare the characteristics of one distribution to the other distribution.

By this point, you should have completed the collection of data for your statistical question. This lesson will provide graphical representations of data distributions that are part of the summaries expected in your project.

Classwork

Example 1: Comparing Groups Using Box Plots

Recall that a *box plot* is a visual representation of a 5-number summary. It is drawn with careful reference to a number line, so the difference between any two values in the 5-number summary is represented visually as a distance. For example, the box of a box plot is drawn so that width of the box represents the IQR. The whiskers (the lines that extend from the box) are drawn such that the distance from the far end of one whisker to the far end of the other whisker represents the range. If two box plots (each representing a different distribution) were drawn side-by-side using the same scale, one could quickly compare the IQRs and ranges of the two distributions while also gaining a sense of the 5-number summary values for each distribution.

Here is a data set of the ages of 43 participants in a local 5-kilometer race (shown in a previous lesson).

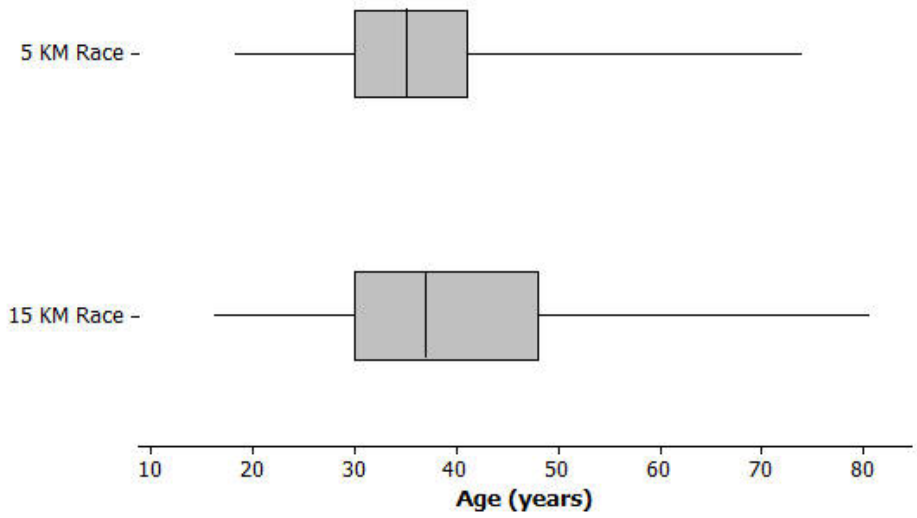
20	30	30	35	36	34	38	46
45	18	43	23	47	27	21	30
32	32	31	32	36	74	41	41
51	61	50	34	34	34	35	28
57	26	29	49	41	36	37	41
38	30	30					

Here is the 5-number summary for the data: Minimum = 18, Q1 = 30, Median = 35, Q3 = 41, Maximum = 74.

Later that year, the same town also held a 15-kilometer race. The ages of the 55 participants in that race appear below.

47	19	30	30	36	37	35	39
19	49	47	16	45	22	50	27
19	20	30	32	32	31	32	37
22	81	43	43	54	66	53	35
22	35	35	36	28	61	26	29
38	52	43	37	38	43	39	30
58	30	48	49	54	56	58	

Does the longer race appear to attract different runners in terms of age? Here are side-by-side box plots that may help answer that question. Side-by-side box plots are two or more box plots drawn



Exercises 1–6

1. Based on the side-by-side box plots, estimate the 5-number summary for the 15-kilometer race data set.

2. Do the two data sets have the same median? If not, which race had the higher median age?

3. Do the two data sets have the same IQR? If not, which distribution has the greater spread in the middle 50% of its distribution?

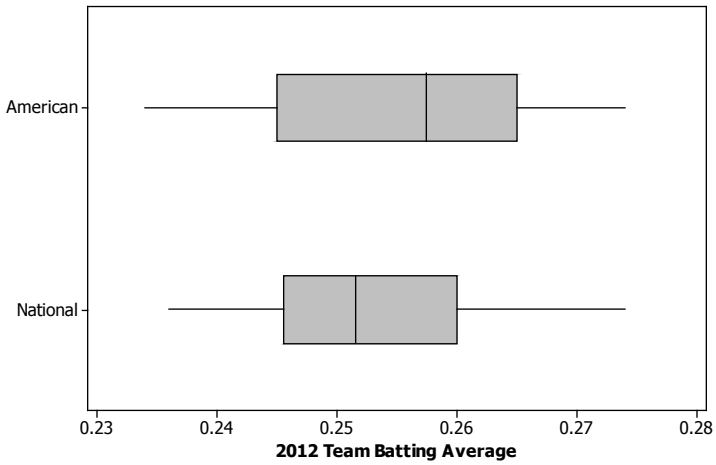
4. Which race had the smaller overall range of ages? What do you think the range of ages is for the 15-kilometer race?

5. Which race had the oldest participant? About how old was this participant?

6. Now consider just the youngest 25% of participants in the 15-kilometer race. How old was the youngest runner in this group? How old was the oldest runner in this group? How does that compare with the 5-kilometer race?

Exercises 7–12: Comparing Box Plots

In 2012, Major League Baseball was comprised of two leagues: an American League of 14 teams and a National League of 16 teams. It is believed that the American League teams would generally have higher values of certain offensive statistics such as batting average and on-base percentage. (Teams want to have high values of these statistics.) Use the following side-by-side box plots to investigate these claims. (Source: http://mlb.mlb.com/stats/sortable.jsp accessed May 13, 2013)

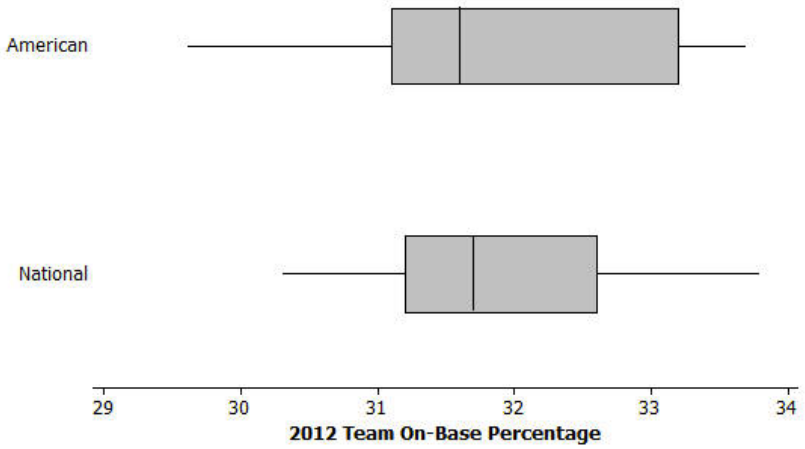


7. Was the highest American League team batting average very different from the highest National League team batting average? If so, approximately how large was the difference and which league had the higher maximum value?

8. Was the range of American League team batting averages very different or only slightly different from the range of National League team batting averages?

9. Which league had the higher median team batting average? Given the scale of the graph and the range of the data sets, does the difference between the median values for the two leagues seem to be small or large? Explain why you think it is small or large.

10. Based on the box plots below for on-base percentage, which 3 summary values (from the 5-number summary) appear to be the same or virtually the same for both leagues?



11. Which league's data set appears to have less variability? Explain.

12. Respond to the original statement: "It is believed that the American League teams would generally have higher values of … on-base percentage." Do you agree or disagree based on the graphs above? Explain.

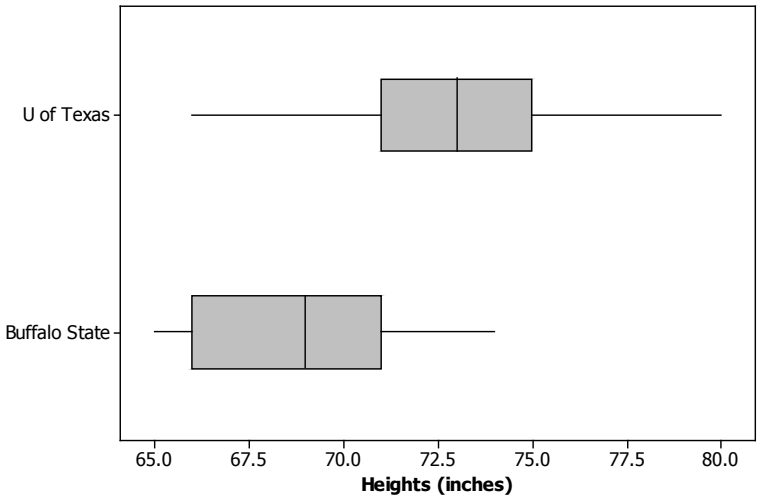
Lesson Summary

When comparing the distribution of a quantitative variable for two or more distinct groups, it is useful to display graphs of the groups' distributions side-by-side using the same scale. Generally, you can more easily notice, quantify, and describe the similarities and differences in the distributions of the groups.

Problem Set

1. College athletic programs are separated into divisions based on school size, available athletic scholarships, and other factors. A researcher is curious to know if members of swimming and diving programs in Division I (schools that offer athletic scholarships and tend to have large enrollment) are generally taller than the swimmers and divers in Division III programs (schools that do not offer athletic scholarships and tend to have smaller enrollment). To begin the investigation, the researcher creates side-by-side box plots for the heights (in inches) of members of the 2012–2013 University of Texas Men's Swimming and Diving Team (a Division I program) and the heights (in inches) of members of the 2012–2013 Buffalo State College Men's Swimming and Diving Team (a Division III program).

   (From http://www.texassports.com/sports/m-swim/mtt/tex-m-swim-mtt.html accessed April 30, 2013, all 41 member heights listed, and http://www.buffalostateathletics.com/roster.aspx?path=mswim& accessed May 15, 2013, 11 members on roster; only 10 heights were listed)



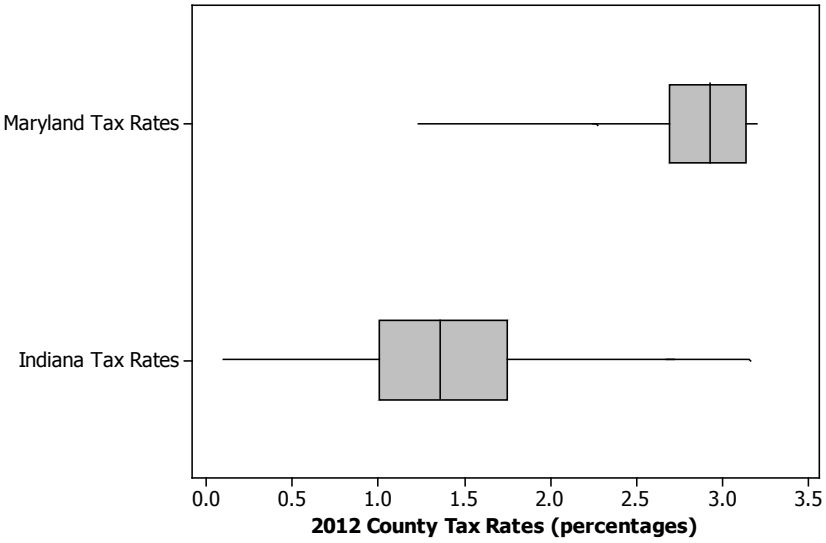
   a. Which data set has the smaller range?

   b. True or False: A team member of median height on the University of Texas team would be taller than a team member of median height on the Buffalo State College team.

   c. To be thorough, the researcher will examine many other college's sports programs to further investigate her claim that members of swimming and diving programs in Division I are generally taller than the swimmers and divers in Division III. But given the graph above, in this initial stage of her research, do you think that her claim might be valid? Carefully support your answer using comparative summary measures or graphical attributes.

2. Different states use different methods for determining a person's income tax. However, Maryland and Indiana both have systems where a person pays a different income tax rate based on the county in which he/she lives. Box plots summarizing the 24 different county tax rates for Maryland's 23 counties and Baltimore City (taxed like a county in this case) and the resident tax rates for 91 counties in Indiana in 2012 are shown below.

   (From http://taxes.marylandtaxes.com/Individual_Taxes/Individual_Tax_Types/Income_Tax/Tax_Information/Tax_Rates/Local_and_County_Tax_Rates.shtml accessed May 5, 2013 and www.in.gov/dor/files/12-county-rates.pdf accessed May 16, 2013)



   a. True or False: At least one Indiana county income tax rate is higher than the median county income tax rate in Maryland. Explain how you know.

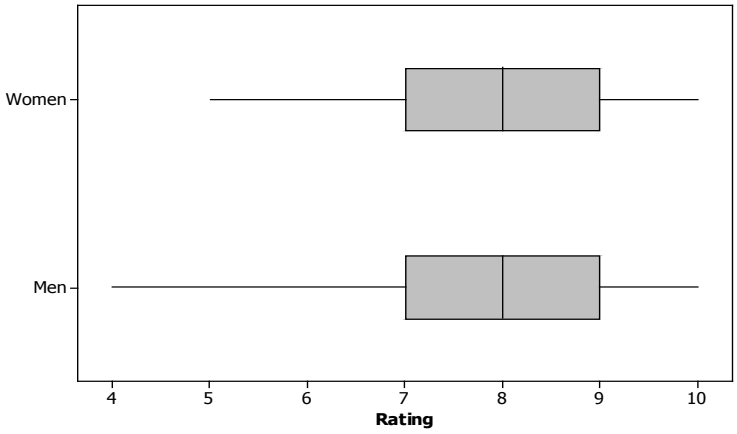
   b. True or False: The 24 Maryland county income tax rates have less variability than the 91 Indiana county income tax rates. Explain how you know.

   c. Which state appears to have typically lower county income tax rates? Explain.

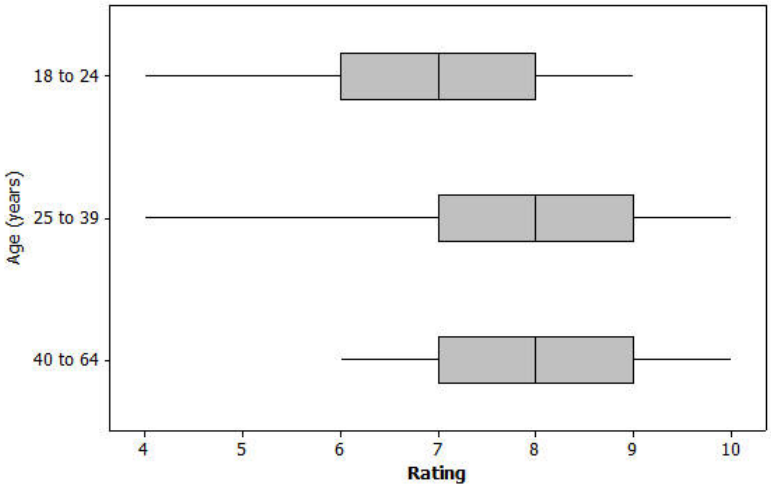
3. Many movie studios rely heavily on customer data in test markets to determine how a film will be marketed and distributed. Recently, previews of a soon to be released film were shown to 300 people. Each person was asked to rate the movie on a scale of 0 to 10, with 10 representing "best movie I've ever seen" and 0 representing "worst movie I've ever seen."

   Below are some side-by-side box plots that summarize the ratings based on certain demographic characteristics.

   For 150 women and 150 men:



   For 3 distinct age groups:



   a. Generally, does it appear that the men and women rated the film in a similar manner or in a very different manner? Write a few sentences explaining your answer using comparative information about center and spread from the graph.

   b. Generally, it appears that the film typically received better ratings from the older members of the group. Write a few sentences using comparative measures of center and spread or graphical attributes to justify this claim.

Lesson 20: Describing Center, Variability, and Shape of a Data Distribution from a Graphic Representation

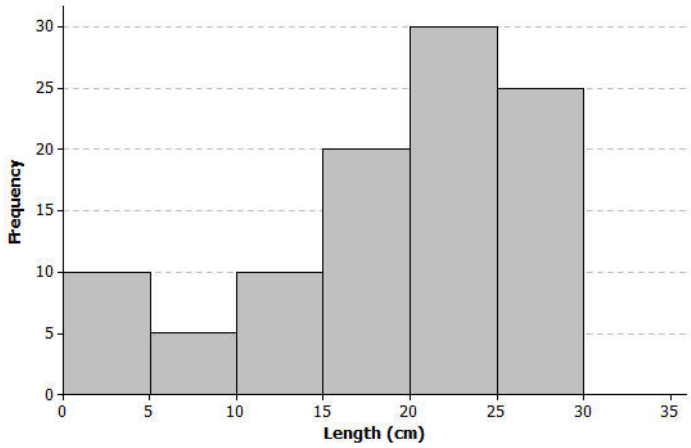
Great Lakes Yellow Perch are fish that live in each of the five Great Lakes and many other lakes in the eastern and upper Great Lakes regions of the United States and Canada. Both countries are actively involved in efforts to maintain a healthy population of perch in these lakes.

Classwork

Example 1: The Great Lakes Yellow Perch

Scientists collected data from many samples of yellow perch because they were concerned about the survival of the yellow perch. What data do you think researchers might want to collect about the perch?

Scientists captured yellow perch from a lake in this region. They recorded data on each fish, and then returned each fish to the lake. Consider the following histogram of data on the length (in centimeters) for a sample of yellow perch.



Exercises 1–11

Scientists were concerned about the survival of the yellow perch as they studied the histogram.

1. What statistical question could be answered based on this data distribution? How do you think the scientists collected these data?

2. Use the histogram to complete the following table:

Length of fish in centimeters (cm)	Number of fish
$0 - < 5$ cm	
$5 - < 10$ cm	
$10 - < 15$ cm	
$15 - < 20$ cm	
$20 - < 25$ cm	
$25 - < 30$ cm	

3. The length of each fish was measured and recorded before the fish was released back into the lake. How many yellow perch were measured in this sample?

4. Would you describe the distribution of the lengths of the fish in the sample as a skewed data distribution or as a symmetrical data distribution? Explain your answer.

5. What percentage of fish in the sample were less than 10 centimeters in length?

6. If the smallest fish in this sample were 2 centimeters in length, what is your estimate of an interval of lengths that would contain the lengths of the shortest 25% of the fish? Explain how you determined your answer.

7. If the length of the largest yellow perch was 29 centimeters, what is your estimate of an interval of lengths that would contain the lengths of the longest 25% of the fish?

8. Estimate the median length of the yellow perch in the sample. Explain how you determined your estimate.

9. Based on the shape of this data distribution, do you think the mean length of a yellow perch would be greater than, less than, or the same as your estimate of the median? Explain your answer.

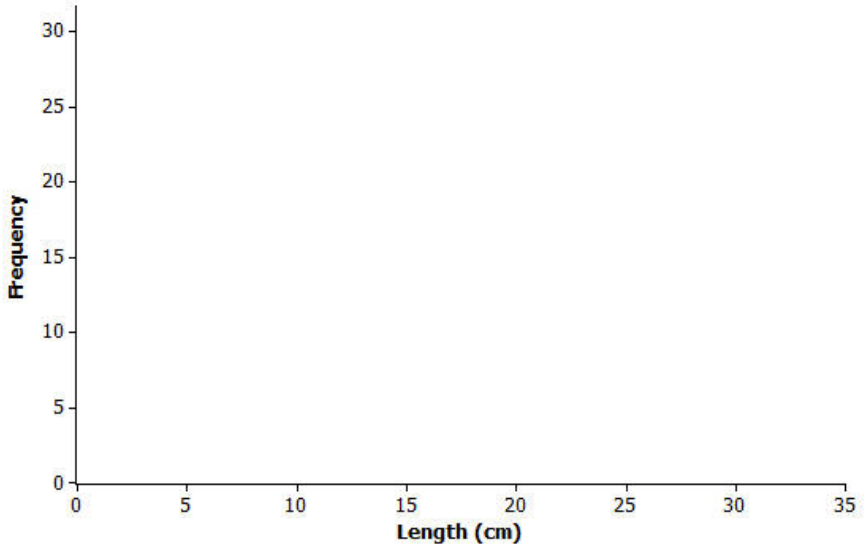
10. Recall that the mean length is the balance point of the distribution of lengths. Estimate the mean length for this sample of yellow perch.

11. The length of a yellow perch is used to estimate the age of the fish. Yellow perch typically grow throughout their lives. Adult yellow perch have lengths between 10 and 30 centimeters. How many of the yellow perch in this sample would be considered adult yellow perch? What percentage of the fish in the sample are adult fish?

Example 2: What Would a Better Distribution Look Like?

Yellow perch are part of the food supply of larger fish and other wild life in the Great Lakes region. Why do you think that the scientists worried when they saw the histogram of fish lengths given above?

Sketch a histogram representing a sample of 100 yellow perch lengths that you think would indicate the perch are not in danger of dying out?



Exercises 12–17: Estimating the Variability in Yellow Perch Lengths

You estimated the median length of yellow perch from the first sample in Exercise 8. It is also useful to describe variability in the length of yellow perch. Why might this be important? Consider the following questions:

12. In several previous lessons, you described a data distribution using the 5-number summary. Use the histogram and your answers to the questions in Exercise 2 to provide estimates of the values for the 5-number summary for this sample:

    Min or minimum value =

    Q1 value =

    Median =

    Q3 value =

    Max or maximum value =

13. Based on the 5-number summary, what is an estimate of the value of the interquartile range (IQR) for this data distribution?

14. Sketch a box plot representing the lengths of the yellow perch in this sample.

15. Which measure of center, the median or the mean, is closer to where the lengths of yellow perch tend to cluster?

16. What value would you report as a typical length for the yellow perch in this sample?

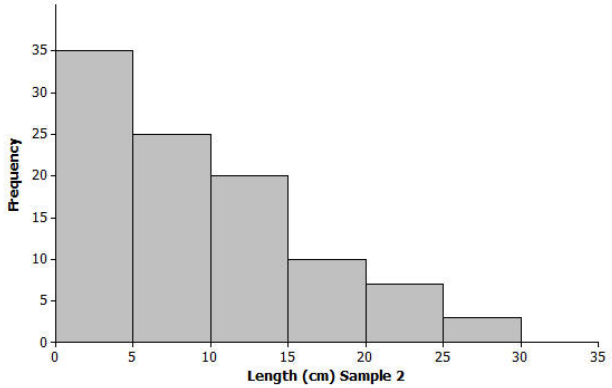
17. The mean absolute deviation (or MAD) or the interquartile range (IQR) are used to describe the variability of a data distribution. Which measure of variability would you use for this sample of perch? Explain your answer.

Lesson Summary

Data distributions are usually described in terms of shape, center, and spread. Graphical displays, such as histograms, dot plots, and box plots, are used to assess the shape. Depending on the shape of a data distribution, different measures of center and variability are used to describe the distribution. For a distribution that is skewed, the median is used to describe a typical value, whereas the mean is used for distributions that are approximately symmetric. The IQR is used to describe variability for a skewed data distribution, while the MAD is used to describe variability for distributions that are approximately symmetric.

Problem Set

Another sample of Great Lake yellow perch from a different lake was collected. A histogram of the lengths for the fish in this sample is shown below:



1. If the length of a yellow perch is an indicator of its age, how does this second sample differ from the sample you investigated in the exercises? Explain your answer.

2. Does this histogram represent a data distribution that is skewed or that is nearly symmetrical?

3. What measure of center would you use to describe a typical length of a yellow perch in this second sample? Explain your answer.

4. Assume the smallest perch caught was 2 centimeters in length, and the largest perch caught was 29 centimeters in length. Estimate the values in the 5-number summary for this sample:

   Min or minimum value =

   Q1 value =

   Median =

   Q3 value =

   Max or maximum value =

5. Based on the shape of this data distribution, do you think the mean length of a yellow perch from this second sample would be greater than, less than, or the same as your estimate of the median? Explain your answer.

6. Estimate the mean value of this data distribution.

7. What is your estimate of a typical length of a yellow perch in this sample? Did you use the mean length from problem 5 for this estimate? Explain why or why not.

8. Would you use the MAD or the IQR to describe variability in the length of Great Lakes yellow perch in this sample? Estimate the value of the measure of variability that you selected.

Lesson 21: Summarizing a Data Distribution by Describing Center, Variability, and Shape

Each of the lessons in this module is about data. What are data? What questions can be answered by data? How do you represent the data distribution so that you can understand and describe its shape? What does the shape tell us about how to summarize the data? What is a typical value of the data set? These questions, and many others, were part of your work in the exercises and investigations. There is still a lot to learn about what data tell us. You will continue to work with statistics and probability in grades seven and eight and throughout high school. You have already, however, started to learn how to uncover the stories behind data.

When you started this module, the four steps used to carry out a statistical study were introduced:

Step 1: Pose a question that can be answered by data.

Step 2: Collect appropriate data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the numerical summaries and graphs.

In this lesson, you will carry out these steps using a given data set.

Classwork

Exploratory Challenge: Annual Rainfall in the State of New York

The National Climate Data Center collects data throughout the United States that can be used to summarize the climate of a region. You can obtain climate data for a state, a city, a county, or a region. If you were interested in researching the climate in your area, what data would you collect? Explain why you think this data would be important as a statistical study of the climate in your area.

For this lesson, you will use yearly rainfall data for the state of New York that were compiled by the National Climate Data Center. The following data are the number of inches of rain (averaged over various locations in the state) for the years from 1983 to 2012 (30 years).

45	42	39	44	39	35	42	49	37	42	41	42	37	50	39
41	38	46	34	44	48	50	47	49	44	49	43	44	54	40

Use the four steps to carry out a statistical study using this data.

Step 1: Pose a question that can be answered by data.

What is a statistical question that you think can be answered with these data? Write your question in the template provided for this lesson.

Step 2: Collect appropriate data.

The data have already been collected for this lesson. How do you think these data were collected? Recall that the data are the number of inches of rain (averaged over various locations in the state) for the years from 1983 to 2012 (30 years). Write a summary of how you think the data were collected in the template for this lesson.

Step 3: Summarize the data with graphs and numerical summaries.

A good first step might be to summarize the data with a dot plot. What other graph might you construct? Construct a dot plot or another appropriate graph in the template for this lesson.

What numerical summaries will you calculate? What measure of center will you use to describe a typical value for these data? What measure of variability will you calculate and use to summarize the spread of the data? Calculate the numerical summaries and write them in the template for this lesson.

Step 4: Answer your statistical question using the numerical summaries and graphs.

Write a summary that answers the question you posed in the template for this lesson

Template for Lesson 21

Step 1: What is your statistical question?

Step 2: How do you think the data were collected?

Step 3: Construct graphs and calculate numerical summaries of the data.

Construct at least one graph of the data distribution. Calculate appropriate numerical summaries of the data. Also indicate why you selected these summaries.

Step 4: Answer your statistical question using your graphs and numerical summaries.

Lesson Summary

Statistics is about using data to answer questions. The four steps used to carry out a statistical study include posing a question that can be answered by data, collecting appropriate data, summarizing the data with graphs and numerical summaries, and using the data, graphs, and summaries to answer the statistical question.

Problem Set

In Lesson 17, you posed a statistical question and a plan to collect data to answer your question. You also constructed graphs and calculated numerical summaries of your data. Review the data collected and your summaries.

Based on directions from your teacher, create a poster or an outline for a presentation using your own data. On your poster, indicate your statistical question. Also, indicate a brief summary of how you collected your data based on the plan you proposed in Lesson 17. Include a graph that shows the shape of your data distribution, along with summary measures of center and variability. Finally, answer your statistical question based on the graphs and the numerical summaries.

Share the poster you will present in Lesson 22 with your teacher. If you are instructed to prepare an outline of the presentation, share your outline with your teacher.

Lesson 22: Presenting a Summary of a Statistical Project

A statistical study involves the following four-step investigative process:

Step 1: Pose a question that can be answered by data.

Step 2: Collect appropriate data.

Step 3: Summarize the data with graphs and numerical summaries.

Step 4: Answer the question posed in Step 1 using the numerical summaries and graphs.

Now it is your turn to be a researcher and to present your own statistical study. In Lesson 17, you posed a statistical question, proposed a plan to collect data to answer the question, and collected the data. In Lesson 21, you created a poster or an outline of a presentation that included the following: the statistical question, the plan you used to collect the data, graphs and numerical summaries of the data, and an answer to the statistical question based on your data. Use the following table to organize your presentation.

Points to consider:	Notes to include in your presentation:
(1) Describe your statistical question.	
(2) Explain to your audience why you were interested in this question.	
(3) Explain the plan you used to collect the data.	
(4) Explain how you organized the data you collected.	

(5)	Explain the graphs you prepared for your presentation and why you made this graph.	
(6)	Explain what measure of center and what measure of variability you selected to summarize your study. Explain what you selected these values.	
(7)	Describe what you learned from the data. (Be sure to include an answer to the question from step (1) above.)	

Closing Exercise

After you have presented your study, consider what your next steps are by answering the following questions:

1. What questions still remain after you concluded your statistical study?

2. What statistical question would you like to answer next as a follow-up to this study?

3. How would you collect the data to answer the new question you posed in (2)?

Lesson Summary

Statistics is about using data to answer questions. The four steps used to carry out a statistical study include posing a question that can be answered by data, collecting appropriate data, summarizing the data with graphs and numerical summaries, and using the data, graphs, and summaries to answer the statistical question.