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 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?

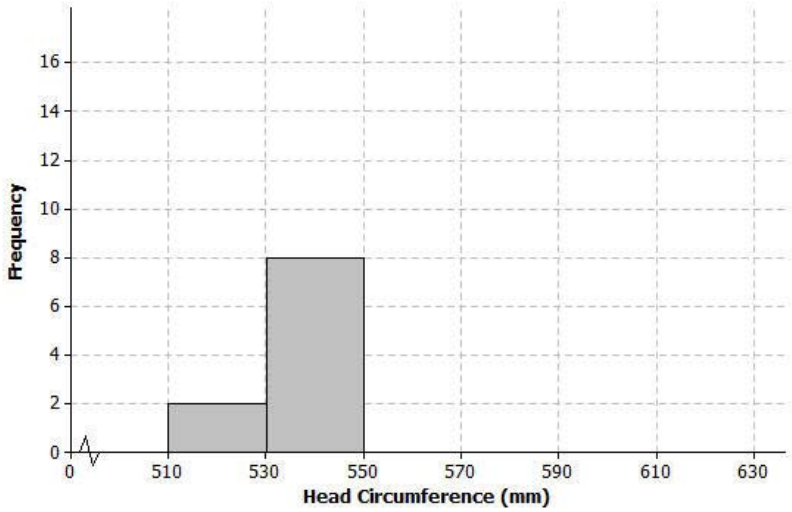
Example 2: Histogram

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?

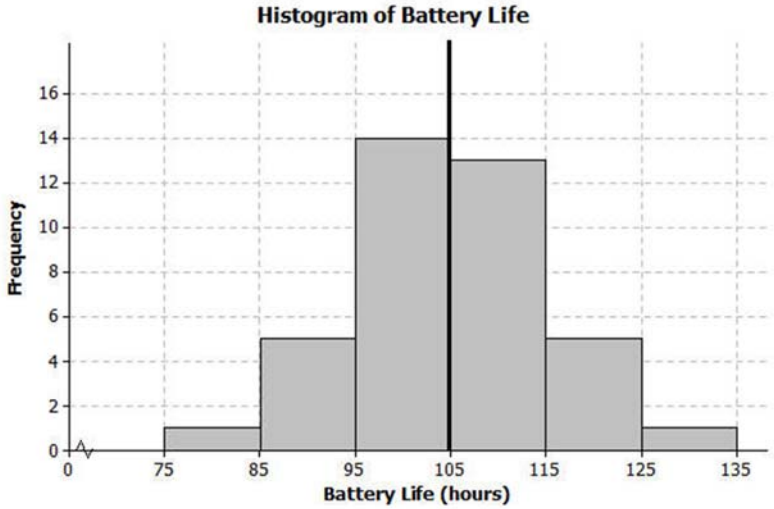
8. Because the 40 head circumference values were given, you could have constructed a dot plot to display the head circumference data. What information is lost when a histogram is used to represent a data distribution instead of a dot plot?

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.

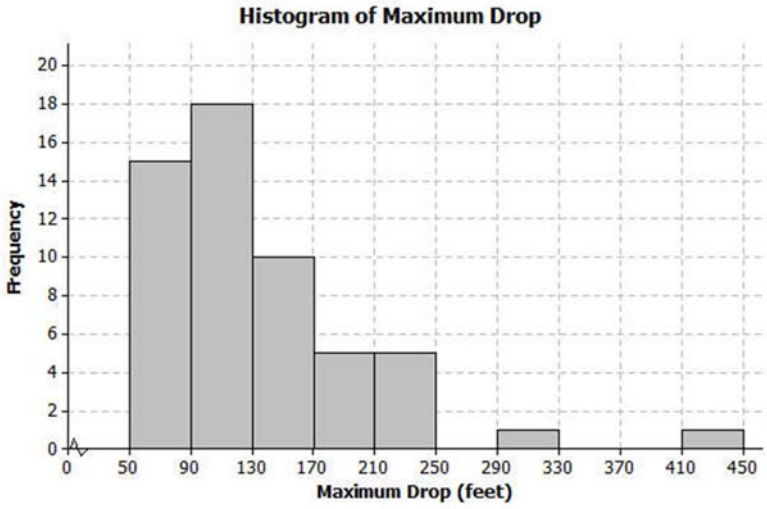
Example 3: Shape of the Histogram

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.

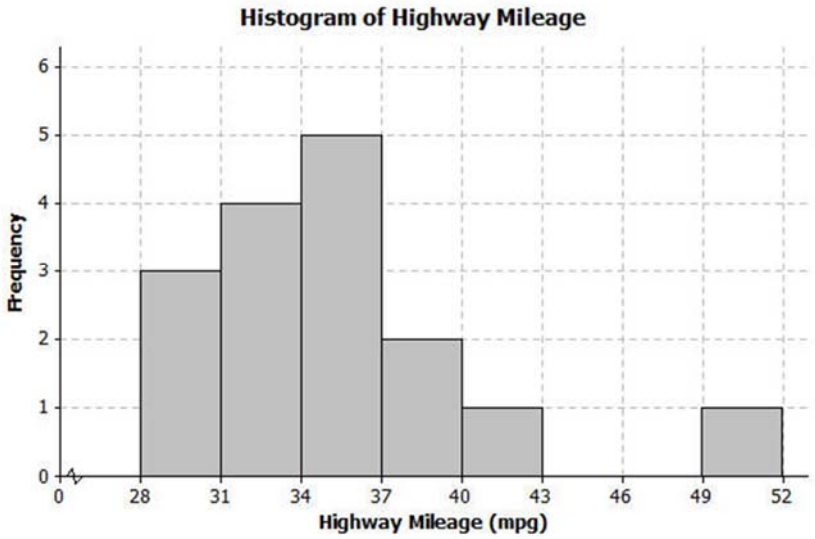


Another group of students wanted to investigate the maximum drop length for roller coasters. The histogram below shows the maximum drop (in feet) of a selected group of roller coasters. This histogram has a skewed shape. Most of the data are in the intervals from 50 to 170. But there are two values that 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.



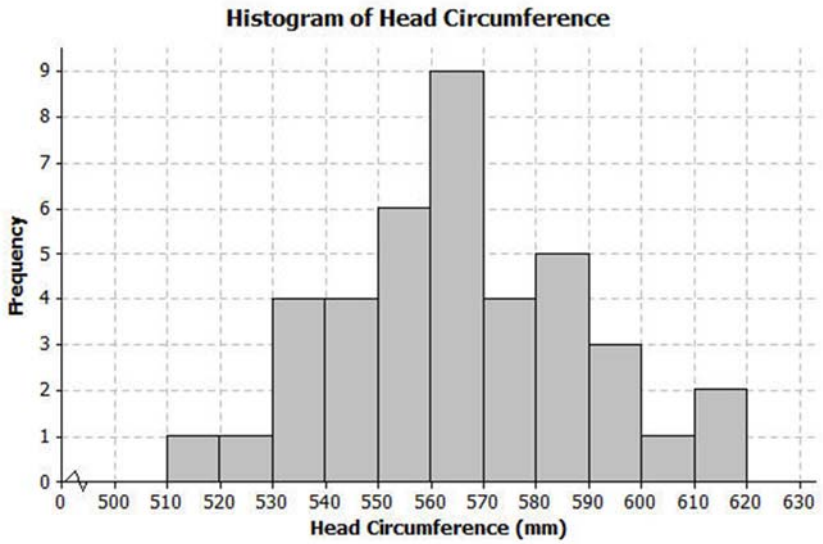
a. Describe the shape of the histogram as approximately symmetric, skewed left, or skewed right.

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?

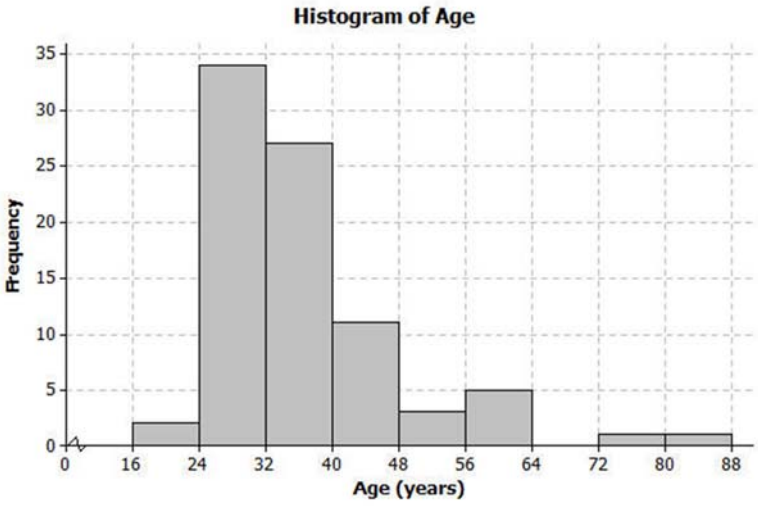
d. In what interval would a head circumference of 571 be included? In what interval would a head circumference of 610 be included?

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.

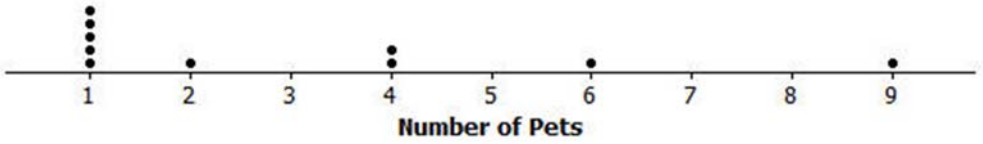
4. Use the histograms that you constructed in question 3 parts (b) and (e) to answer the following questions.
   a. Why are there fewer bars in the histogram in question 3 part (e) than the histogram in part (b)?
   b. Did the shape of the histogram in question 3 part (e) change from the shape of the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

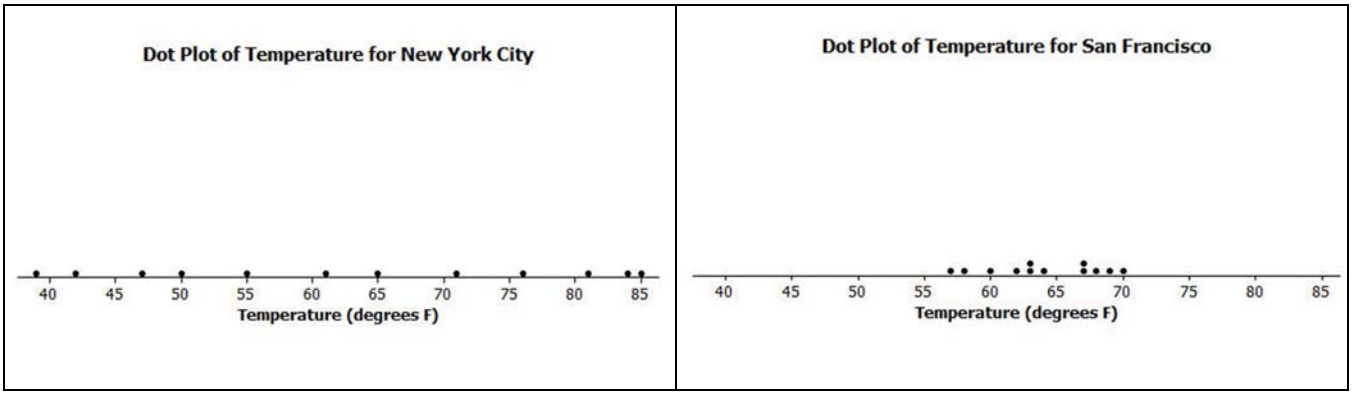
Student	Alice	Bob	Carol	Damon	Ed	Felipe	Gale
Weight (lbs.)	5	5	5	5	5	5	5

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 15: More Practice with Box Plots

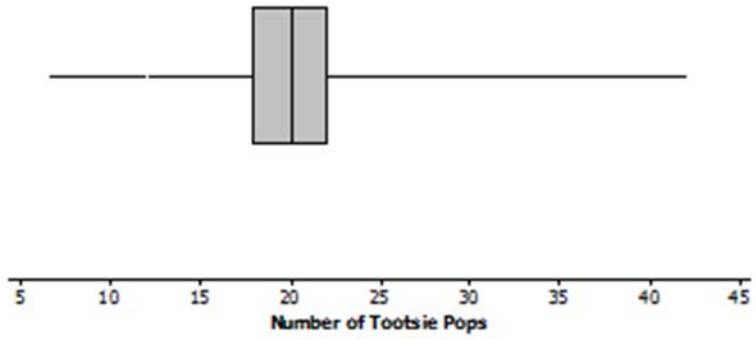
You reach into a jar of Tootsie Pops. How many Tootsie Pops do you think you could hold in one hand? Do you think the number you could hold is greater than or less than what other students can hold? Is the number you could hold a typical number of Tootsie Pops? This lesson examines these questions.

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

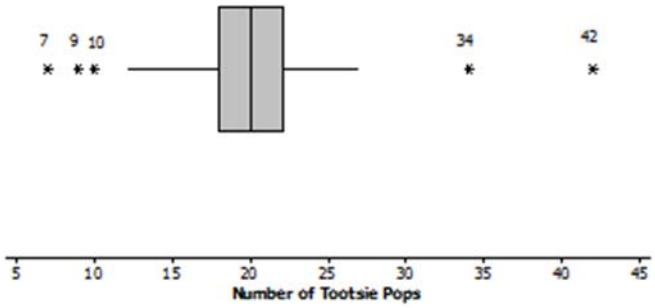
1. What might explain the variability in how many Tootsie Pops those 94 people were able to hold?

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

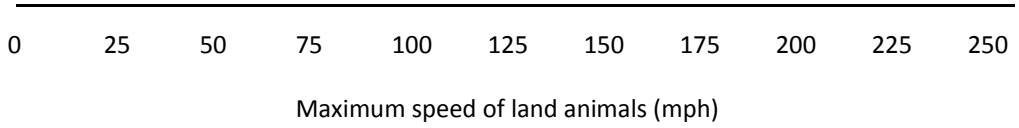
Data Source: Natural History Magazine, March 1974, copyright 1974; The American Museum of Natural History; and James G. Doherty, general curator, The Wildlife Conservation Society; *http://www.thetravelalmanac.com/lists/animals-speed.htm; http://en.wikipedia.org/wiki/Fastest_animals*

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

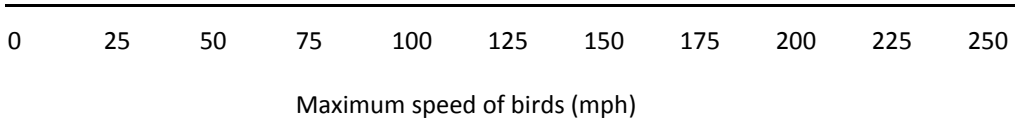
7. Do birds or land animals seem to have the greatest variability in speeds? Explain your reasoning.

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)

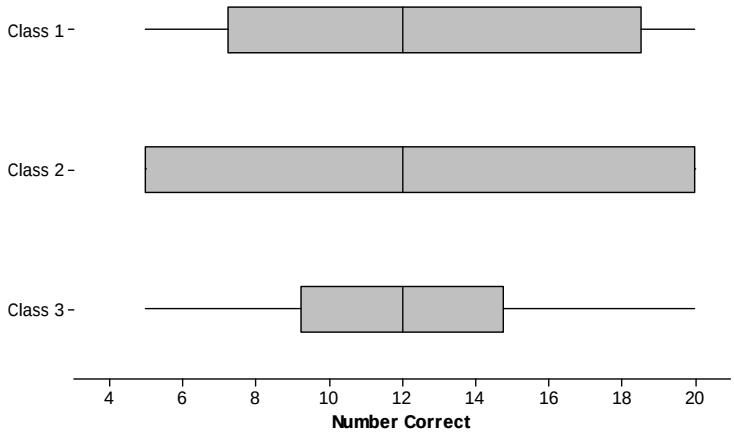


Maximum speed of birds (mph)

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.

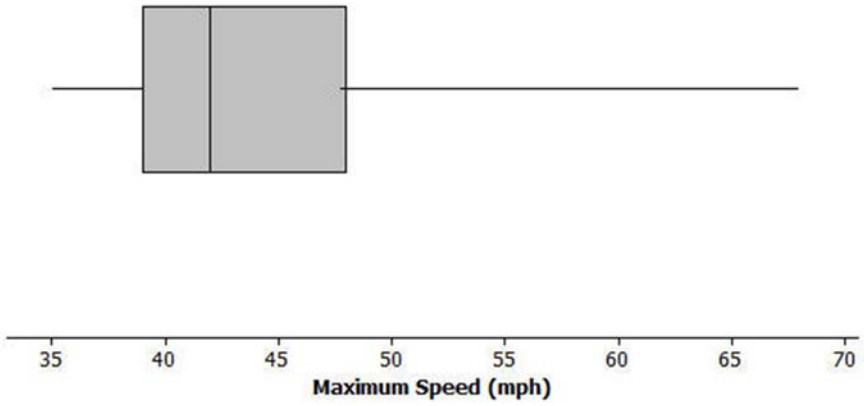
Lesson Summary

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

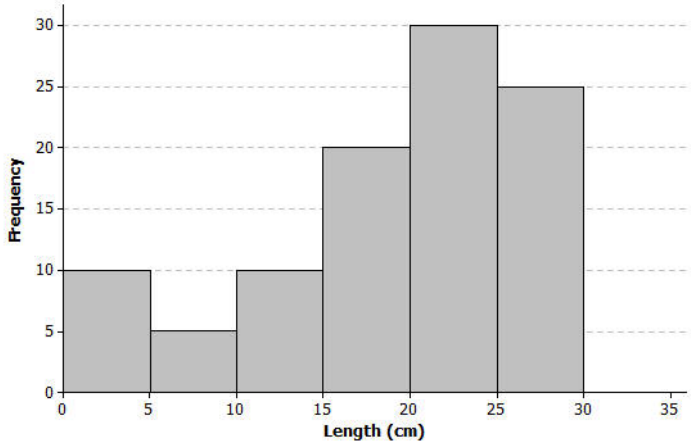
Great Lakes Yellow Perch are fish that live in each of the five Great Lakes and many other lakes in the eastern and upper Great Lakes regions of the United States and Canada. Both countries are actively involved in efforts to maintain a healthy population of perch in these lakes.

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.

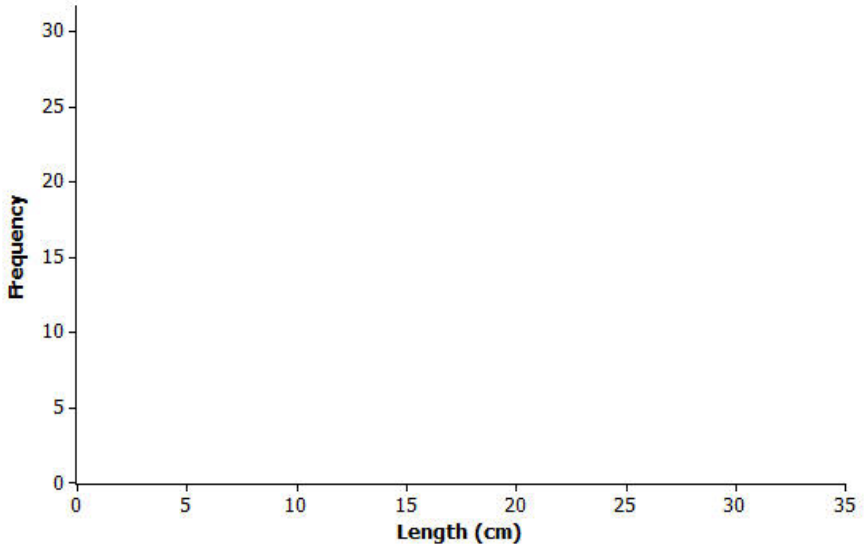
10. Recall that the mean length is the balance point of the distribution of lengths. Estimate the mean length for this sample of yellow perch.

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 =

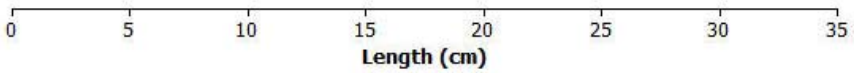
    Median =

    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?

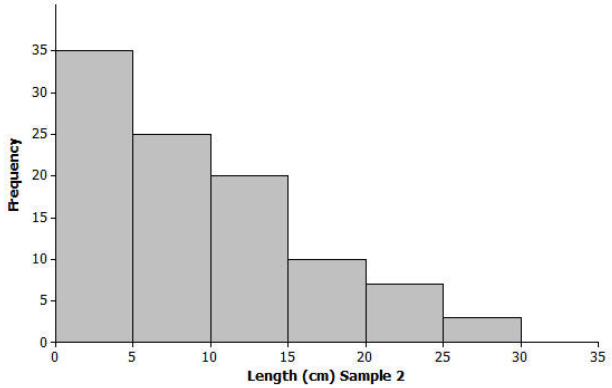
17. The mean absolute deviation (or MAD) or the interquartile range (IQR) are used to describe the variability of a data distribution. Which measure of variability would you use for this sample of perch? Explain your answer.

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.