

Analyzing Data Sets for One Variable

8



People didn't fret over sugar amounts, carbohydrates, nor calories in the late 1800s. However, that didn't stop two famous brothers from offering a healthy alternative to the "most important meal of the day."



- 8.1 Start Your Day the Right Way**
Graphically Representing Data455
- 8.2 Which Measure Is Better?**
Determining the Best Measure of Center
for a Data Set469
- 8.3 You Are Too Far Away!**
Calculating IQR and Identifying Outliers.479
- 8.4 Whose Scores Are Better?**
Calculating and Interpreting Standard Deviation . . .489
- 8.5 Putting the Pieces Together**
Analyzing and Interpreting Data505

Start Your Day the Right Way

8.1

Graphically Representing Data

LEARNING GOALS

In this lesson, you will:

- Represent and interpret data displayed on dot plots.
- Represent and interpret data displayed on histograms.
- Represent and interpret data displayed on box-and-whisker plots.

KEY TERMS

- dot plot
- discrete data
- data distribution
- symmetric distribution
- skewed right distribution
- skewed left distribution
- box-and-whisker plot
- five number summary
- histogram
- bin
- frequency
- continuous data

Many nutritional experts call breakfast the most important meal of the day, and many people start their day with a bowl of cereal. However, cereal was not always an option. In the late 1800s, most people's diets consisted mainly of meat products, including breakfasts of pork and beef. However, John Harvey Kellogg and his brother William Keith Kellogg, both of whom worked at a health spa, began creating vegetarian-based breakfast options for their guests using grains. It was actually by mistake that they created some of the first flakes of wheat cereal. This mistake was an immediate success! Just a few years later, the Kellogg Company was selling more than one million cases of cereal a year.

Some of today's cereals still contain the healthy whole grains that the Kelloggs used in their original recipe. However, there are many other cereals that contain other ingredients that are not quite as healthy. What are some healthy cereal options in the stores today? What are some cereals that might not be considered as healthy? What is the difference between these two types of cereal?

PROBLEM 1 How Much Sugar Is Too Much?


Ms. Romano is a health coach and nutritionist. Recently, she encouraged Matthew to eat a healthier breakfast and recommended a cereal with less sugar. There are many different cereals and it seems like the amount of sugar in each type varies widely. Matthew took a trip to the grocery store and recorded the sugar amount that each cereal has in one serving.

Cereal Name	Sugar Amount in One Serving (grams)
Cocoa Rounds	13
Flakes of Corn	4
Frosty Flakes	11
Grape Nuggets	7
Golden Nuggets	10
Honey Nut Squares	10
Raisin Branola	7
Healthy Living Flakes	7
Wheatleys	8
Healthy Living Crunch	6
Multi-Grain Squares	7
All Branola	5
Munch Crunch	12
Branola Flakes	5
Complete Flakes	4
Corn Crisps	3
Rice Crisps	4
Shredded Wheatleys	1
Puffs	22
Fruit Circles	11

1. Analyze the data collected. What conclusions can you draw about the sugar amount in different types of cereal?

It may be difficult to properly analyze data in a table. One way to better organize the data is to create a graph. A **dot plot** is a graph that shows how *discrete data* are distributed using a number line. **Discrete data** are data that has only a finite number of values or data that can be “counted.” Dot plots are best used to organize and display the number of occurrences of a small number of data points.

Remember, if a value occurs more than once, place an “x” above the number each time the value occurs.



2. Construct a dot plot to represent the sugar amount in one serving of each breakfast cereal. Label the number line using intervals that will include all the data values. Place an “x” above the number that represents each data value. Make sure you name your dot plot.



3. Analyze the dot plot. What conclusions can you draw about the sugar amounts in one serving of breakfast cereal from the dot plot?

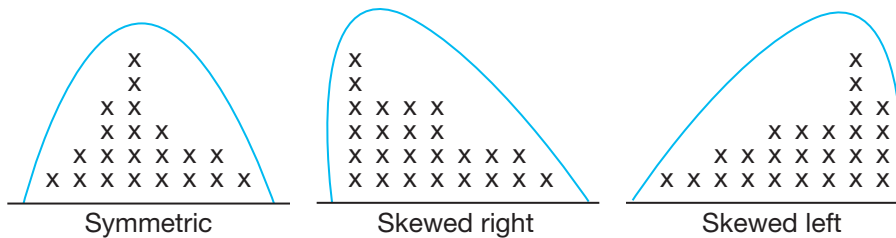
4. Jordan states that those numbers on the number line that do not contain any data values should be eliminated. Toni disagrees and says that all the numbers on the number line must be included even if there are no data values for that particular number. Who is correct? Explain your reasoning.



When you analyze a graphical display, you can look at several characteristics of the graph to draw conclusions. For example, you can ask yourself:

- What is the overall shape of the graph? Does it have any interesting patterns?
- Where is the approximate middle, or center, of the graph?
- How spread out are the data values on the graph?

The overall shape of a graph is called the *data distribution*. The **data distribution** is the way in which the data is spread out or clustered together. The shape of the distribution can reveal a lot of information about the data. There are many different distributions, but the most common are *symmetric*, *skewed right*, and *skewed left* as shown.



5. Describe the properties of a data distribution that is:
- symmetric.
 - skewed right.
 - skewed left.

In a **symmetric distribution** of data, the left and right halves of the graph are nearly mirror images of each other. There is often a “peak” in the middle of the graph.

In a **skewed right distribution** of data, the peak of the data is to the left side of the graph. There are only a few data points to the right side of the graph.

In a **skewed left distribution** of data, the peak of the data is to the right side of the graph. There are only a few data points to the left side of the graph.

6. Describe the distribution of the sugar amount in one serving of breakfast cereal. Explain what this means in terms of the problem situation.



7. Do you think the conclusion you came to in Question 6 is true of all breakfast cereals? Why or why not?

PROBLEM 2 Boxing It Up

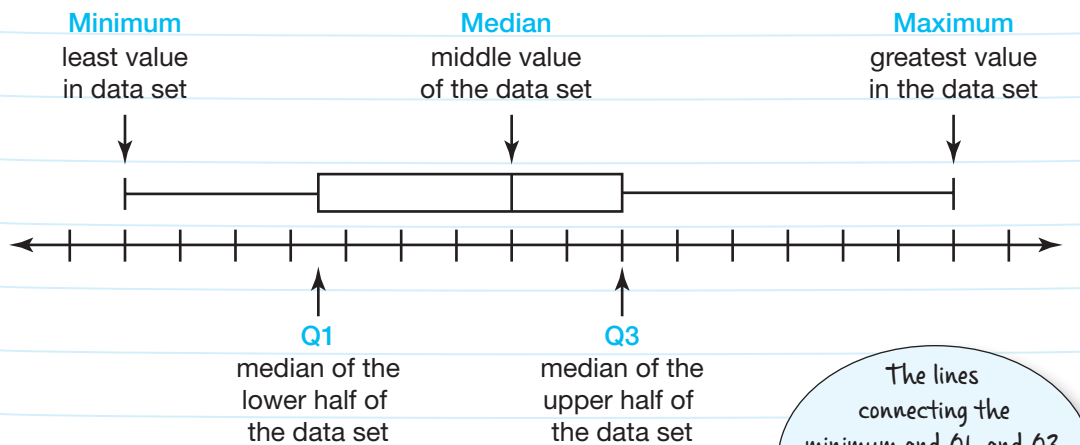


Another graphical representation that displays the distribution of quantitative data is a *box-and-whisker plot*. A **box-and-whisker plot** displays the data distribution based on a *five number summary*. The **five number summary** consists of the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value.

Quantitative data is just another term for numerical data.



The five number summary is used to create a box-and-whisker plot. Each vertical line of the box-and-whisker plot represents a value from the summary.



The lines connecting the minimum and Q1, and Q3 and the maximum are known as the whiskers.

There are four sections of the graphical display: minimum to Q1, Q1 to median, median to Q3, and Q3 to maximum. Each section of the box-and-whisker plot represents 25 percent of the data set.





1. Determine each percent of data values for the given sections of the box-and-whisker plot shown in the worked example. Explain your reasoning for each.

a. Less than Q1
Greater than Q1

b. Less than Q3
Greater than Q3

c. Less than the median
Greater than the median



d. Between Q1 and Q3



2. Construct a box-and-whisker plot of the sugar amount in one serving of each breakfast cereal from Problem 1, *How Much Sugar Is Too Much?*

Before you start constructing, list the data values in order.



3. Analyze the five number summary and box-and-whisker plot. What conclusions can you draw about the sugar amount in one serving of breakfast cereal from these representations?

Interpret the data in terms of percents.



4. Describe the data distribution shown in the box-and-whisker plot. Interpret the meaning of the distribution in terms of this problem situation.

5. Damon states that more breakfast cereals have over 10 grams of sugar per serving than have under 5 grams of sugar per serving because the whisker connecting the maximum and Q3 is longer than the whisker connecting the minimum and Q1. Is Damon correct? Explain why or why not.



PROBLEM 3 Weekend Gamers

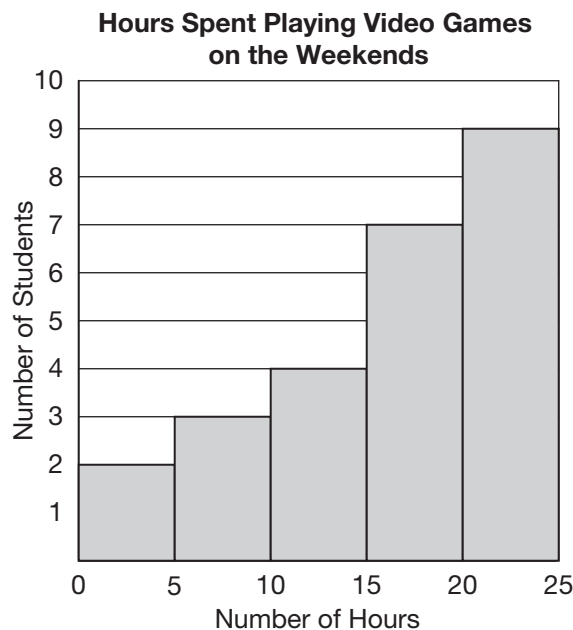


Another way to display quantitative data is to create a *histogram*. A **histogram** is a graphical way to display quantitative data using vertical bars. The width of a bar in a histogram represents an interval of data and is often referred to as a **bin**. A bin is represented by intervals of data instead of showing individual data values. The value shown on the left side of the bin is the least data value in the interval.

The height of each bar indicates the **frequency**, which is the number of data values included in any given bin.

Histograms are effective in displaying large amounts of *continuous data*. **Continuous data** is data which can take any numerical value within a range.

The histogram shown represents the data distribution for the number of hours students spend playing video games on the weekends. The data is gathered to the nearest half-hour.





1. What conclusions can you draw from the histogram about the number of hours students spend playing video games on weekends?
2. Jonae and Tyler must identify the greatest value represented in the bin beginning with 15. Their responses are shown.



Jonae

The bin that begins with the interval 15 includes all data values from 15 to 20.



Tyler

The bin that begins with the interval 15 includes all data values from 15 thru, but not including 20.

- a. Explain why Tyler's answer is correct and why Jonae's answer is incorrect.
 - b. Represent the bin that contains 15 as an inequality.
3. Analyze the histogram.
 - a. How many students play 5 to 9.5 hours of video games on weekends? Explain your reasoning.
 - b. How many total students are included in the data? Explain your reasoning.

- c.** Marcel states that between 0 and 5 students spend 2 hours playing video games on weekends. Is Marcel's statement correct? Explain why or why not.



- d.** How many students play 22 hours of video games on the weekends? Explain your reasoning.
- e.** What percent of the students play 10 or more hours of video games on the weekends? Explain your reasoning.



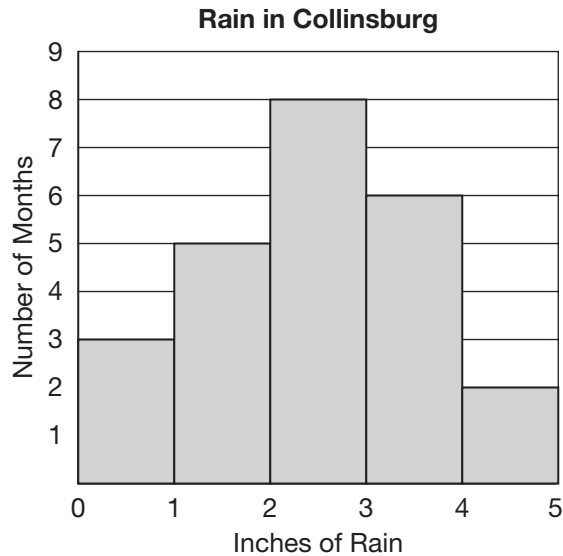
- 4.** Describe the data distribution displayed by the histogram. Interpret its meaning in terms of this problem situation.

Talk the Talk



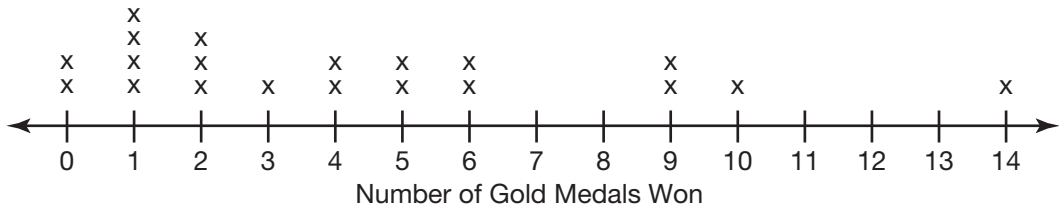
Analyze each data representation to answer the questions. Justify your reasoning using the characteristics of each representation.

1.



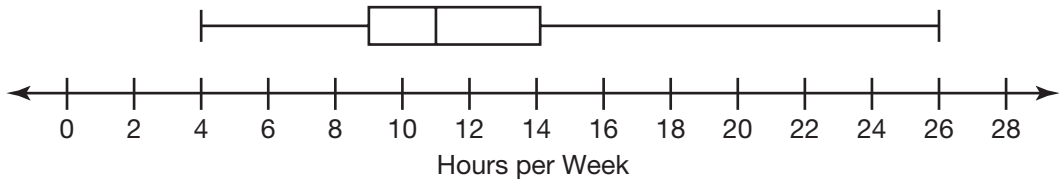
- Describe the information represented in the histogram.
- How many months are represented on the histogram? Describe how you determined your answer.
- Identify the intervals represented by each bin.
- How many months had 4 or more inches of rain?
- Describe the data distribution and interpret its meaning in terms of this problem situation.

2. Participants Who Won Gold Medals at the Special Olympics



- a. Describe the information represented in the dot plot.
- b. How many participants are represented in the dot plot?
- c. How many participants won 10 or more medals?
- d. Describe the data distribution and interpret its meaning in terms of this problem situation.

3. Volunteers Hours at the Local Animal Shelter



- a. Describe the information represented in the box-and-whisker plot.
- b. How many people are represented on the box-and-whisker plot?
- c. What percent of the people volunteered 14 or more hours?

- d. What percent of people volunteered less than 11 hours?
- e. How many hours did the middle 50 percent of the people volunteer?
- f. Describe the data distribution and interpret its meaning in terms of this problem situation.
4. Analyze each visual display shown. Describe what information each display provides. Be sure to include advantages and limitations and any specific characteristics for each visual display.
- table
 - dot plot
 - five number summary
 - box-and-whisker plot
 - histogram



Be prepared to share your solutions and methods.

Which Measure Is Better?

Determining the Best Measure of Center for a Data Set

LEARNING GOALS

In this lesson, you will:

- Calculate and interpret the mean of a data set.
- Calculate and interpret the median of a data set.
- Estimate the mean and median of a data set from its data distribution.
- Determine which measure of central tendency (mean or median) is best to use for a data set.

KEY TERMS

- statistic
- measure of central tendency

You have probably been able to recite your ABCs since you started school. Now you may even be learning a new language that might use new letters. Some languages have different alphabets, where each letter represents sounds that are unique to that language even if the letters are the same as English. There are also some alphabets, such as the Russian alphabet or the Chinese alphabet, which use different letter symbols altogether.

Today, you will get the opportunity to learn new letters from another alphabet. The letters of the Greek alphabet are often used in mathematics to represent different mathematical ideas. You should already know the letter pi (π), which represents the ratio of the circumference of a circle to its diameter. By the time you finish this chapter you will know at least two more Greek letters! Keep an eye out for them as you work through the lessons.

PROBLEM 1 How Sweet It Is



Previously you analyzed a data set by creating a graphical representation of the data. However, you can also analyze a data set by describing numerical characteristics, or **statistics**, of the data. A statistic that describes the “center” of a data set is called a *measure of central tendency*. A **measure of central tendency** is the numerical values used to describe the overall clustering of data in a set. Two measures of central tendency that are typically used to describe a set of data are the mean and the median.

The arithmetic mean, or mean, represents the sum of the data values divided by the number of values. A common notation for the mean is \bar{x} , which is read “x bar.”



The E-like symbol is actually the Greek letter sigma and in mathematical terms it means the “summation” or “sum of.”



The formula shown represents the mean of a data set.

$$\text{mean} \rightarrow \bar{x} = \frac{\sum x}{n}$$

the sum of the data values

the number of data values

The mean of the data set 5, 10, 9, 7, 5 can be written using this formula.

$$\bar{x} = \frac{5 + 10 + 9 + 7 + 5}{5}$$

$$\bar{x} = 7.2$$

The mean of this data set is 7.2.

Why don't I write the sigma when writing the data values in the formula?



Recall that Lesson 8.1 *How Much Sugar Is Too Much?*, Matthew collected data on the sugar amount in one serving of various breakfast cereals. The data collected is shown.

Cereal Name	Sugar Amount in One Serving (grams)	Cereal Name	Sugar Amount in One Serving (grams)
Cocoa Rounds	13	Multi-Grain Squares	7
Flakes of Corn	4	All Branola	5
Frosty Flakes	11	Munch Crunch	12
Grape Nuggets	7	Branola Flakes	5
Golden Nuggets	10	Complete Flakes	4
Honey Nut Squares	10	Corn Crisps	3
Raisin Branola	7	Rice Crisps	4
Healthy Living Flakes	7	Shredded Wheatleys	1
Wheatleys	8	Puffs	22
Healthy Living Crunch	6	Fruit Circles	11

1. Represent the sugar amount in different cereals using the formula for the mean. Then determine the mean of the data set.



You can use your graphing calculator to determine the mean of a data set.

Be sure to check that your lists are clear of old data. Delete any data that might be in your lists before entering new data.

Do the values need to be entered in order?

Step 1: Press **STAT** then press **ENTER** to select **1:Edit**.

Step 2: Enter the data values for the data set in List 1.

Step 3: Press **STAT** and scroll over to highlight **CALC**. Press **ENTER** to select **1:1-Var Stats**. Press **ENTER** again.

Step 4: The calculator should now show many values relating to the data set. You can scroll down for more values including the five number summary.



2. Enter the data set for the sugar amount in various breakfast cereals into a graphing calculator. Then for each given symbol, state what it represents and its calculated value.

a. \bar{x}

b. Σx

c. n



3. Compare your answers in Question 2 with the answers you wrote using the formula for determining the mean in Question 1. What do you notice?



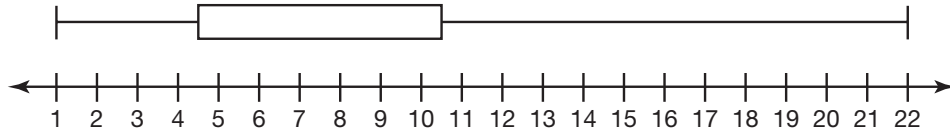
4. Determine the median sugar amount in grams in one serving of cereal. Interpret the meaning in terms of this problem situation.

Does the order of the data matter when determining the median?



5. The box-and-whisker plot you constructed in the previous lesson is shown. Locate and label the mean and median values on the dot plot.

Sugar in Breakfast Cereals




6. Compare the mean and median. Which measure best represents the data set?

Constructing a box-and-whisker plot can take some time when using paper and pencil. Technology can make constructing a box-and-whisker plot more efficient.



You can use a graphing calculator to construct a box-and-whisker plot.

- Step 1:** Press **STAT** and then press **ENTER** to select **1:Edit**.
- Step 2:** Enter the data values of the data set in List 1.
- Step 3:** Press **2nd** and **STAT PLOT**, which is above the **Y=** button.
- Step 4:** Select **1:** and press **ENTER**. Then highlight **PLOT 1** and press **ENTER** to turn Plot 1 on. Then scroll down to **Type:** and select the box-and-whisker icon.  Press **ENTER**.
- Step 5:** Make sure the **XList** is using the correct list. Then press **GRAPH**.

7. Let's consider the data set without the value of 22.
- Remove the value of 22 from the data set. Use your graphing calculator to create a box-and-whisker plot for the new data set.
 - Plot above the given box-and-whisker plot your new box-and-whisker plot on the same graph in Question 5.
 - How does the removal of the value 22 affect the distribution of the data set?



- d. Did the mean and median change with the removal of the value 22? Does your choice for the best measure of center from Question 6 still hold true?

PROBLEM 2 Does Height Really Matter?



The Mountain View High School basketball team has its first game of the season on their home court. Coach Maynard doesn't know much about the visiting team, but he does have a list of the heights of their top ten players. Coach Maynard wants to compare the heights of his top ten players to those on the visiting team.

Home Team Heights (inches)	Visiting Team Heights (inches)
69	68
70	68
67	68
68	69
66	69
65	67
70	72
70	71
71	66
71	67



1. Represent the data for each team on a dot plot.



2. Analyze each dot plot you created.
 - a. Describe the data distribution of each graph and explain what it means in terms of the players' heights on each team.

- b. Based on the dot plots, predict whether the mean or median will be greater for each data set. Explain your reasoning.

- c. Verify your prediction by calculating the mean and median heights for each team. Was your prediction correct?

- d. Which measure of central tendency best describes each data set? Explain your reasoning.



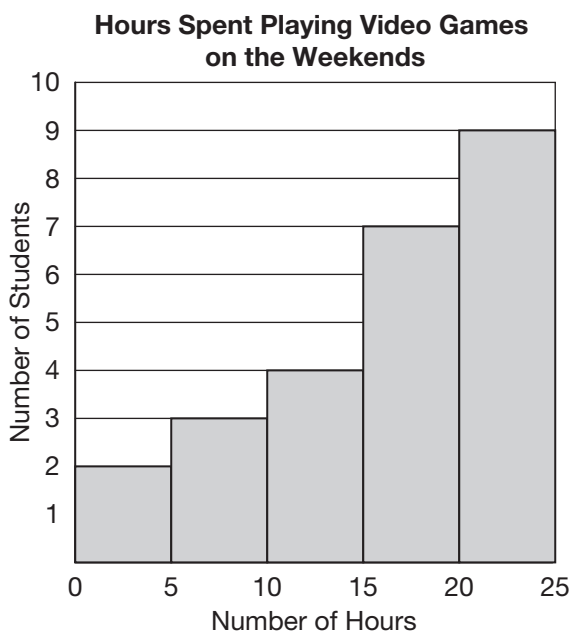
3. Describe the relationship that seems to exist between the data distribution and the values of the mean and median.



When the distribution of data is approximately symmetric, the mean is generally the more appropriate measure of center to use. When the distribution of data is skewed left or skewed right, the median is the more appropriate measure of center to use. The reason why the mean is more appropriate in a symmetric data distribution is due to the fact that most data points are close to the mean. There are not many if any data values that are much greater or lesser than the mean. In a skewed left or right distribution, most data values are closer to the median with few data points being much greater or lesser than the median. Therefore, the median is not affected by these values.



4. The histogram from Lesson 8.1 *Weekend Gamers* is shown.



- a. Predict whether the mean or median number of hours spent playing video games will be greater. Explain your reasoning.

Remember to use characteristics of the graph to explain your reasoning.



- b. Suppose the two measures of central tendency for the given histogram are 16.1 hours and 17.5 hours. Which value is the mean and which value is the median? Explain your reasoning.

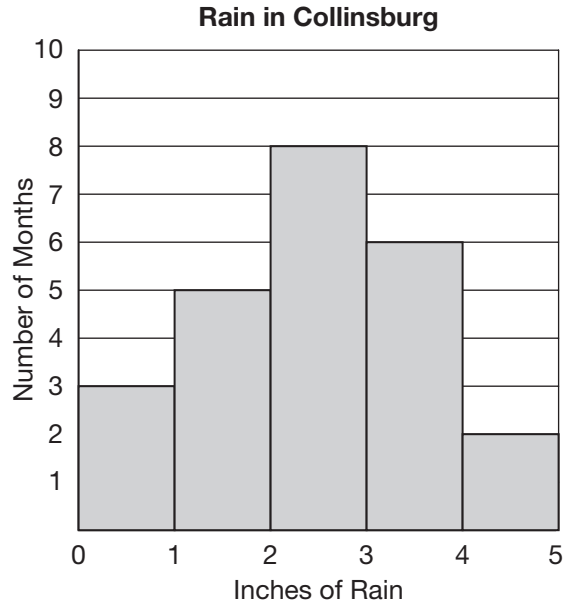


Talk the Talk



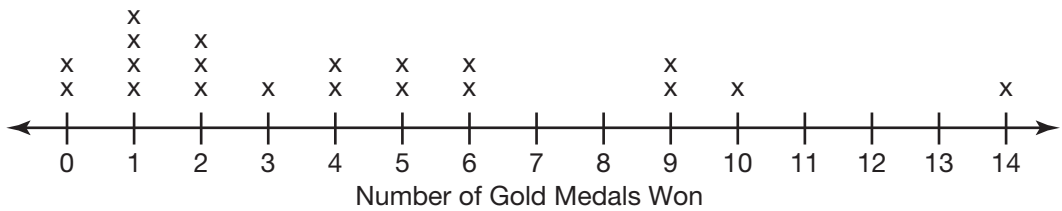
- Identify which measure of central tendency would be most appropriate to describe the for each given graph. Then determine the mean and median if possible. If it is not possible, explain why not.

a.



b.

Participants Who Won Gold Medals at the Special Olympics

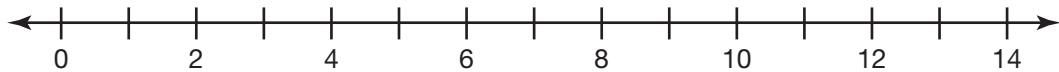


c. **Volunteer Hours at the Local Animal Shelter**



2. Consider the data set 0, 10, 10, 12, 14.

a. Construct and label a dot plot of the data.



b. Calculate the mean and median. Which measure do you think best represents the data set?

c. Remove the value of 0 from the data set. How does this affect the distribution of the data set?

d. Recalculate the mean and median without the value of 0. Does your choice in part (b) for the best measure of central tendency still hold true? Explain why or why not.

Why do you think the value 0 was selected to be removed from the data set?



Be prepared to share your solutions and methods.

You Are Too Far Away!

Calculating IQR and Identifying Outliers

LEARNING GOALS

In this lesson, you will:

- Calculate and interpret the interquartile range (IQR) of a data set.
- Determine if a data set contains outliers.

KEY TERMS

- interquartile range (IQR)
- outlier
- lower fence
- upper fence

Everywhere in our world there are boundaries that show where something begins and ends. The walls to your classroom are boundaries. The lanes on the road are boundaries. There are boundaries on sports fields and boundaries for each state and country. But what about the universe? Is there a boundary to show where the universe begins and ends?

That is a question that astronomers and physicists have debated for quite some time. For example, in the early 1900s, astronomer Harlow Shapely claimed that the entire universe was located within the Milky Way galaxy (the same galaxy where the Earth is located). He determined the galaxy was 300,000 light-years in diameter and in his opinion, could be thought of as the boundary of the universe. It was not until 1925, when Edwin Hubble showed that there are stars located much farther than 300,000 light years away. At this point, most scientists agreed that the universe must be larger than the Milky Way galaxy.

So we know the universe is larger than the Milky Way galaxy, but will we ever know just how large it is? Scientists have studied the idea that the universe is actually expanding for quite some time. What does this mean in terms of boundaries? Do you think we will ever know the size of the universe?

PROBLEM 1 Touchdown!


Coach Petersen's Middletown 9th grade football team is having a tough season. The team is struggling to win games. He is trying to determine why his team has only won a few times this year. The table shows the points scored in games in 2011 and 2012.

Points Scored (2011)	10	13	17	20	22	24	24	27	28	29	35
Points Scored (2012)	0	7	17	17	18	24	24	24	25	27	45

1. Analyze the data sets in the table.
 - a. In which year do you think the football team performed better? Explain your reasoning.



- b. Calculate the five number summary for each year.

- c. Construct box-and-whisker plots of each year's scores using the same number line for each.



- d. Evita states that because the medians are the same, both teams performed equally well. Is she correct? Explain your reasoning.



- e. What conclusions can you draw about the points scored each year?



Another measure of data distribution Coach Petersen can use to compare the teams is the *interquartile range* or *IQR*. The **interquartile range, IQR**, measures how far the data is spread out from the median. The IQR gives a realistic representation of the data without being affected by very high or very low data values. The IQR often helps show consistency within a data set. The IQR is the range of the middle 50 percent of the data. It is calculated by subtracting $Q3 - Q1$.

2. Calculate the IQR for the points scored each year. Then interpret the IQR for each year.

PROBLEM 2 Get Outta Here!



Another useful statistic when analyzing data is to determine if there are any *outliers*. An **outlier** is a data value that is significantly greater or lesser than other data values in a data set. It is important to identify outliers because outliers can often affect the other statistics of the data set such as the mean.

An outlier is typically calculated by multiplying the IQR by 1.5 and then determining if any data values are greater or lesser than that calculated distance away from Q1 or Q3. By calculating $Q1 - (IQR \cdot 1.5)$ and $Q3 + (IQR \cdot 1.5)$, you are determining a lower and upper limit for the data. Any value outside of these limits is an outlier. The value of $Q1 - (IQR \cdot 1.5)$ is known as the **lower fence** and the value of $Q3 + (IQR \cdot 1.5)$ is known as the **upper fence**.

Remember in the last lesson, How Sweet It Is, you were asked to remove the data value of 22 and then redraw the box-and-whisker. The value 22 was an outlier. Do you remember the affect?



Let's analyze the data set given to see how outliers can be represented on a box-and-whisker plot.

2, 5, 6, 6, 7, 9, 10, 11, 12, 12, 14, 28, 30

Minimum = 2, Q1 = 6, Median = 10, Q3 = 13, Maximum = 30

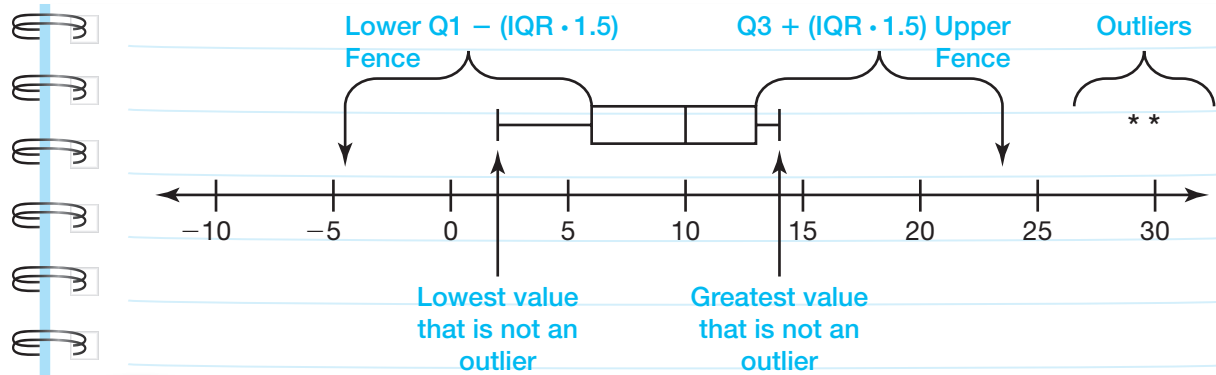
IQR = 7

	Lower Fence:	Upper Fence:
	$= Q1 - (IQR \cdot 1.5)$	$= Q3 + (IQR \cdot 1.5)$
	$= 6 - (7 \cdot 1.5)$	$= 13 + (7 \cdot 1.5)$
	$= -4.5$	$= 23.5$

Using the five number summary and IQR, calculate the upper and lower fence to determine if there are any outliers in the data set.

There are no values less than -4.5 .	Both 28 and 30 are greater than 23.5.
--	---------------------------------------

If there are outliers, the whisker will end at the lowest or highest value that is not an outlier.	Since 28 and 30 are both outliers, 14 is the greatest data value that is not an outlier.
--	--



There are no low outliers so the whisker still ends at the minimum value.



On a box-and-whisker plot, it is common to denote outliers with an asterisk.



Recall the data sets from Problem 1, *Touchdown!* The five number summary and IQR for each data set is shown.

2011:

Minimum = 10
 Q1 = 17
 Median = 24
 Q3 = 28
 Maximum = 45
 IQR = 11

2012:

Minimum = 0
 Q1 = 17
 Median = 24
 Q3 = 25
 Maximum = 45
 IQR = 8

1. Use the formulas to determine if there are any outliers in either data set.
 - a. Determine the upper and lower fence for each year's data set.
 - b. Identify any outliers in either set of data. Explain your reasoning.



2. Remove any outliers for the 2012 data set and, if necessary, reconstruct and label the box-and-whisker plot(s). Compare the IQR of the original data to your new calculations. What do you notice?



PROBLEM 3 Hurry Up!



Brenda needs to get the oil changed in her car, but she hates to wait! Quick Change and Speedy Oil are two garages near Brenda's house. She decides to check an online site that allows customers to comment on the service at different local businesses and record their wait times. Brenda chooses 12 customers at random for each garage. The wait times for each garage are shown.

Wait Times (minutes)							
Quick Change				Speedy Oil			
10	60	22	15	5	60	45	24
12	24	20	18	40	26	55	30
16	23	22	15	32	85	45	30

Don't forget to label each dot plot!



1. Create a box-and-whisker plot of each data set.



2. Calculate and interpret the IQR for each data set.
3. Describe each data distribution and explain its meaning in terms of this problem situation.
4. Identify the measure of central tendency that best represents each data set. Explain your reasoning.
5. Identify any outliers in the data sets.

6. Remove any outliers in each data set and, if necessary, reconstruct the box-and-whisker plot. Compare the IQR of the original data to your new calculations. What do you notice?

7. Does your choice for the best measure of center from Question 4 still hold true?



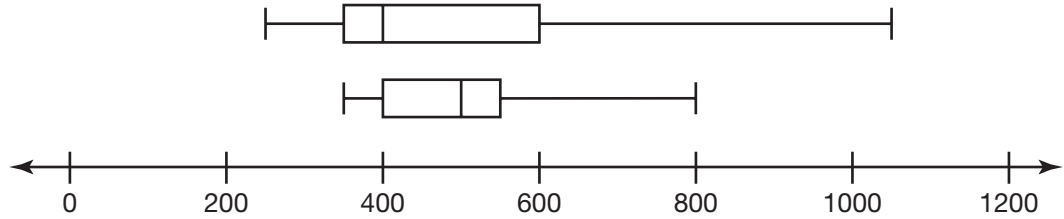
8. Based on the data gathered, which garage should Brenda choose if she is in a hurry?

Talk the Talk



1. Why is the IQR not affected by extremely high or low data values in a data set? Explain your reasoning.

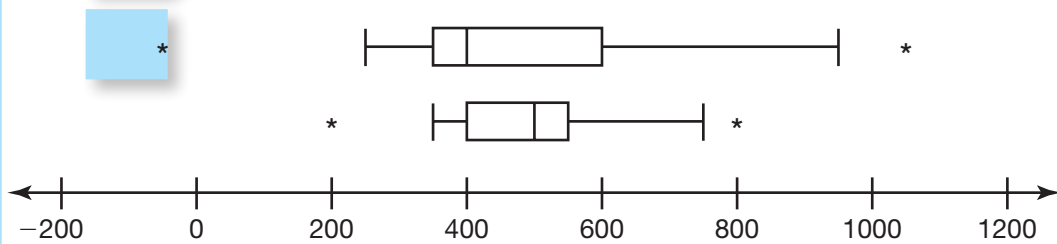
2. Use the two box-and-whisker plots shown to answer each question.



- a. Estimate the five number summary for each box plot to the nearest 50.
- b. Based on your estimates, calculate the IQR of both box-and-whisker plots.

- c. Determine if there are any outliers in either data set shown in the box-and-whisker plots.

- d. Lydia was told to assume that each data set has one outlier and that there are data values at the upper and lower fences. Lydia recreated the two box plots from Question 2 to represent the outliers. Her box plots are shown.



Are Lydia's box plots correct? Explain why or why not.



Be prepared to share your solutions and methods.

Whose Scores Are Better?

Calculating and Interpreting Standard Deviation

LEARNING GOALS

In this lesson, you will:

- Calculate and interpret the standard deviation of a data set.
- Compare the standard deviation of data sets.

KEY TERMS

- standard deviation
- normal distribution

How many times this year have you asked about your grade in a class? Most students who are serious about their learning and the future are interested in their progress in classes. Some students may even keep track of their own grades throughout the semester. But did you know that every country in the world has its own grading system?

Most likely your school uses letter grades from A to E or F which represent a percent of the points you earned in a class. However, if you went to school in Tunisia, your grades would range from 0 (worst) to 20 (best) and any score below a 10 is a fail. In Denmark, a 7-step-scale is used which ranges from 12 (excellent) to -3 (unacceptable). The grading in Denmark is also very strict with very few students receiving a 12 grade. In some schools in Italy, grades vary from 2 to 8 and each teacher can apply his or her own grading customs. The grades between 5 and 6 could range from $5+$, $5++$, $5\frac{1}{2}$, $5/6$, $6--$, $6-$. The symbols on these grades have no real mathematical meaning so calculating grades is somewhat arbitrary. Lately though there has been some push to try to get these schools to use a more uniform system like 1 through 10.

Are you familiar with any other grading scales or techniques teachers use in the classroom? Do you think some grading scales are easier or harder than others? Do you think anything else other than earned points can be used to determine a grade?

PROBLEM 1 Spelling S U C C E S S

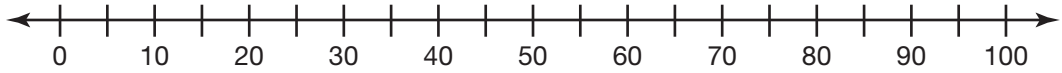
Ms. Webb is determining which student she should add to the spelling bee roster that will represent Tyler High School. The chart shows the 10 most recent scores for three students.

Jack	Aleah	Tymar
33	20	5
32	42	10
30	45	12
50	51	40
49	49	45
50	47	55
35	58	88
73	53	60
71	55	90
77	80	95

1. Determine the mean and median for each student's spelling bee scores.

2. What conclusions can you draw about the data from the mean and median scores?

3. Construct box-and-whisker plots of each student's spelling bee scores using the same number line.



4. Interpret the test scores of each student.

- Jack

- Aleah

- Tymer



5. Do you think these three students performed about the same on all the tests? Why or why not?

PROBLEM 2 So . . . Who Did Better?



You have learned about the spread of data values from the IQR, which is based on the median.

However, is there a way to measure the spread of data from another measure of central tendency?

Standard deviation is a measure of how spread out the data is from the mean. A formula can be used to determine the standard deviation of a data set. A lower standard deviation represents data that are more tightly clustered. A higher standard deviation represents data that are more spread out from the mean.

So, if the IQR is the measure of how spread out data is from the median, and standard deviation is the measure of how spread out data is from the mean, I wonder which will be affected by outliers?



The symbol to the left of the equals sign is a lower case sigma. This represents the standard deviation.

The formula to determine standard deviation of a population is represented as:

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

where σ is the standard deviation, x_i represents each individual data value, \bar{x} represents the mean of the data set, and n is the number of data points.



Let's look at each part of the standard deviation formula separately.



Follow the steps to determine the standard deviation. Let's use the data set 6, 4, 10, 8 where $\bar{x} = 7$.



First, think of each data value as its own term labeled as x_1, x_2 , and so on.

$$x_1 = 6$$

$$x_2 = 4$$

$$x_3 = 10$$

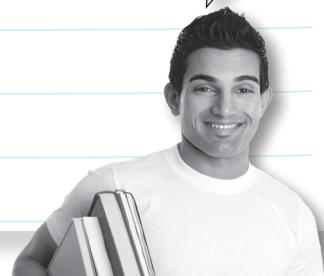
$$x_4 = 8$$



The first part of the formula identifies the terms to be added. Since n represents the total number of values and $i = 1$, add all the values that result from substituting in the first term to the fourth term.

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

This part of the formula just gives you information. You will not sum anything until after the next step.



Next, evaluate the expressions to be added. Take each term and subtract it from \bar{x} and then square each difference.

$$(x_i - \bar{x})^2$$

$$(6 - 7)^2 = 1$$

$$(4 - 7)^2 = 9$$

$$(10 - 7)^2 = 9$$

$$(8 - 7)^2 = 1$$

Now determine the sum of the squared values and divide the sum by the number of data values.

$$\frac{1 + 9 + 9 + 1}{4} = \frac{20}{4} = 5$$

Finally, calculate the square root of the quotient.

$$\sigma = \sqrt{5}$$

$$\sigma \approx 2.24$$

The standard deviation is approximately 2.24.

So the standard deviation for the given data set is approximately 2.24. It is important to note that if the data values have a unit of measure, the standard deviation of the data set also uses the same unit of measure.

1. Do you think the standard deviation for each student's spelling bee scores will be the same? If yes, explain your reasoning. If no, predict who will have a higher or lower standard deviation.



2. Now, let's use the standard deviation formula to determine the standard deviation of Jack's spelling bee scores.

a. Identify the data values you will use to determine the standard deviation. Explain your reasoning.

b. Determine the \bar{x} value.

c. Complete the table to represent each part of the formula. The data values have been put in ascending order.

x_i	$(x_i - \bar{x})^2$
30	$(30 - 50)^2 = 400$
32	
33	
35	
49	
50	
50	
71	
73	
77	
$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$	
$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$	



3. Determine the standard deviation for Jack's spelling bee scores and interpret the meaning.



You can use a graphing calculator to determine the standard deviation of a data set.

These steps are similar to the steps used to determine the mean on the graphing calculator.

Step 1: Press **STAT** and then **ENTER** to select **1:Edit**. Enter each data set into its own List.

Step 2: Press **STAT** then scroll to the right to highlight **CALC**. Press **ENTER** to select **1:Var-Stats**. Press **ENTER**.

Step 3: Your screen should display **1-Var Stats**. Press **2ND** then the list you want the calculator to use for these calculations.

Step 4: Your calculator should display the same data values as when you determined the mean. However, this time use the value for σ_x .

You can enter Jack's data in L_1 , Aleah's data in L_2 , and Tymar's data in L_3 .

So to determine the standard deviation of Jack's data you would enter **2nd L₁**.



4. Use a graphing calculator to determine the standard deviation of Aleah's and Tymar's spelling bee scores.

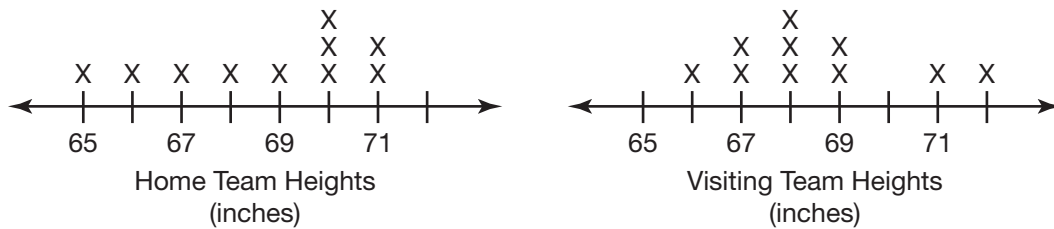
5. Was the prediction you made in Question 1 correct? What do the standard deviations tell you about each student's spelling bee scores?



6. Which student do you think Ms. Webb should add to the spelling bee roster? Use the standard deviation for the student you recommend to add to the roster to justify your answer.

PROBLEM 3 Working as a Team


Recall Lesson 8.2, *Does Height Really Matter?* The Mountain View High School basketball team has its first game of the season and Coach Maynard is comparing the heights of the home team's top ten players to the heights of the visiting team's top ten players. The dot plots of the data are given.



1. Predict which team has the greatest standard deviation in their heights. Explain how you determined your answer.



2. Determine the standard deviation of the heights of each team. Describe what this means in terms of this problem situation. How does this information help Coach Maynard?

PROBLEM 4 68–95–99: The Combination to Standard Deviation



So far, you have determined the standard deviation for different data sets. You have also interpreted the standard deviations to make decisions given a problem situation. Standard deviation can also be represented graphically by graphing a data set.

Recall that Ms. Webb is the spelling bee coach in Problem 1, *Spelling S U C C E S S*. Her class is preparing for their first spelling bee scrimmage. Ms. Webb needs to determine which student should be the spelling bee captain. Ms. Webb believes the captain should have the greatest mean score of the team. The two top spelling bee students' scores are shown.

Maria	Heidi
81	81
73	68
94	60
86	109
70	82
68	88
97	60
93	102
81	78
67	69
85	84
77	103
79	92
103	60
90	108



1. Analyze each student's spelling bee scores.
 - a. Determine the mean spelling bee score for each student.

Ms. Webb wants to also use the standard deviation to help her determine which student is a more consistent speller.

- b. Determine the standard deviation of Maria's scores. Then determine the value of the spelling bee scores that are 1 standard deviation from the mean. Explain how you determined her spelling bee point values.

Make sure to use the mean to the nearest hundredths place.



- c. Determine the standard deviation of Heidi's scores. Then determine the value of the spelling bee scores that are 1 standard deviation from the mean. Explain how you determine her spelling bee point values.

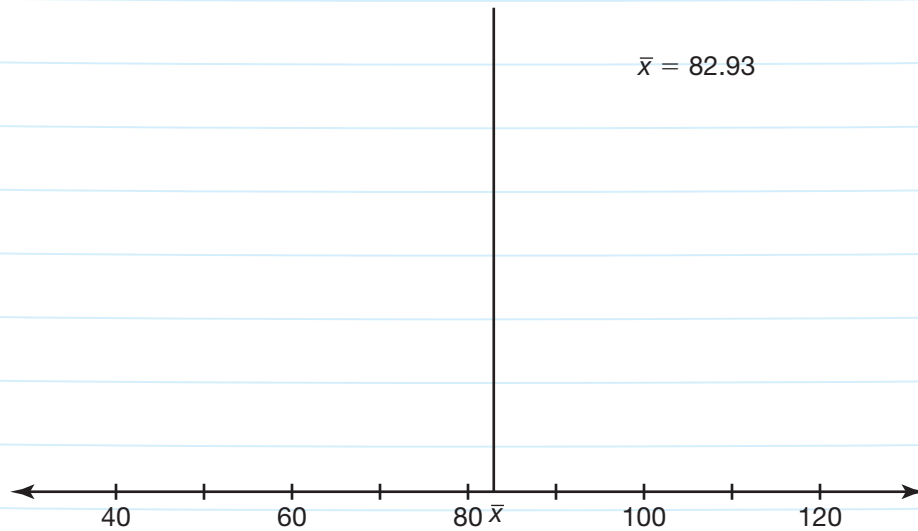


You have calculated 1 standard deviation for the data sets in previous problem situations. However, you can also determine different numbers of standard deviations. For example, 2 standard deviations or greater are calculated by multiplying the standard deviation by the number of standard deviations you are determining. Therefore, if a data set has a standard deviation of 15, then 2 standard deviations would be 30, and 3 standard deviations would be 45.

When you determine the standard deviation of a data set, you can represent it graphically. You can also determine the general percent of data values that are within 1 standard deviation, and the percent of data values that lie within 2 standard deviations in *normal distributions*. A **normal distribution** is a collection of many data points that form a bell-shaped curve.

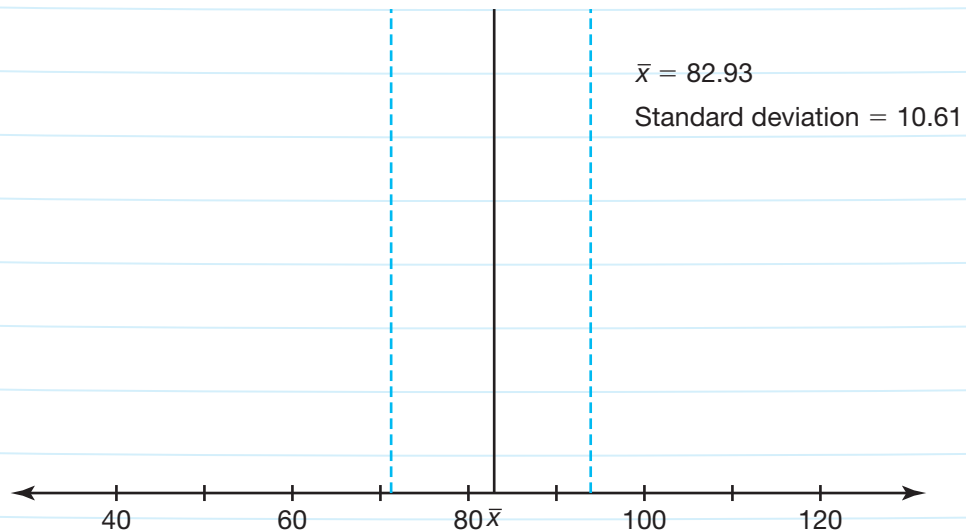
The mean of Maria's spelling bee scores is 82.93 points and 1 standard deviation is 10.61 points.

To graph the normal distribution of Maria's spelling bee scores, first graph the mean on a number line as: $x = 82.93$.



Next, graph 1 standard deviation from the mean. For Maria's spelling bee scores, the standard deviation is 10.61. Therefore, the values of the standard deviation from the mean are 72.32 and 93.54.

Use a dotted lines as $x = 72.32$ and $x = 93.54$.



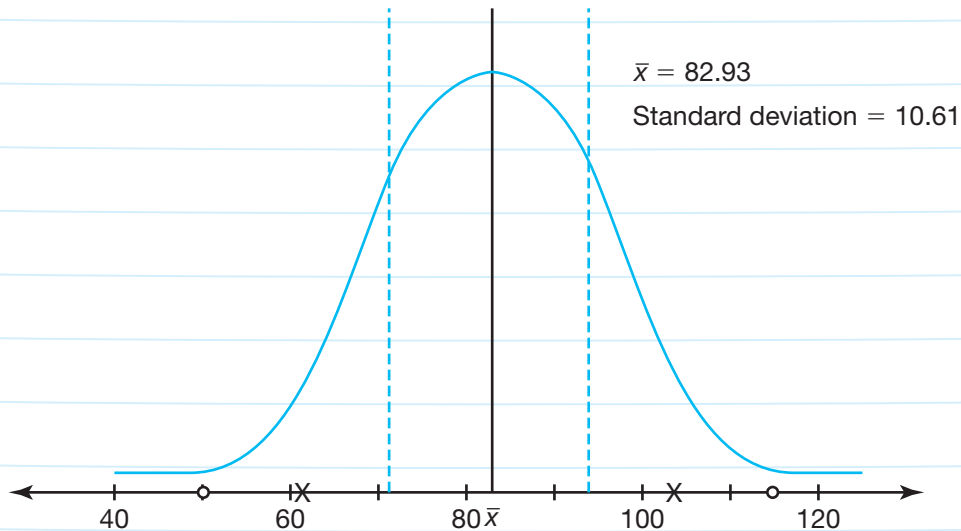
Then graph 2 standard deviations and 3 standard deviations from the mean. To determine 2 standard deviations, multiply the standard deviation by 2. To determine 3 standard deviations, multiply the standard deviation by 3.

For Maria's scores, 2 standard deviations would be 21.22 and 3 standard deviations would be 31.83.

Mark an "x" for the two points as 61.71 and 104.15 represent a standard deviation of 2.

Mark an "o" for the two points 114.76 and 51.10 for a standard deviation of 3.

Finally draw a smooth curve starting from the far left minimum value. The smooth curve should resemble a bell-shaped curve.



- Describe some observations you can make about the graph of Maria's spelling bee scores.

3. Plot each of Maria's scores on the graph of the worked example. Mark an "x" for the approximate location on the number line for each score.

To plot the scores in the graph, mark x's like you would with dot plots.

- a. Determine how many spelling bee scores are within 1 standard deviation of the mean for Maria's spelling bee scores.

- b. Determine how many spelling bee scores are within 2 standard deviations of the mean for Maria's spelling bee scores.



- c. Determine how many spelling bee scores are within 3 standard deviations of the mean for Maria's spelling bee scores.

Within the graph of a normal distribution, you can predict the percent of data points that are within one, two, or three standard deviations from the mean. Generally, 68% of the data points of a data set will fall within one standard deviation of the mean; while 95% of the data points of a data set will fall within two standard deviations of the mean; and 99% of the data points of a data set will fall within three standard deviations of the mean.



4. Analyze the number of data points you determined lie within 1, 2, or 3 standard deviations.
- a. What percent of data points from Maria's spelling bee scores fall within 1 standard deviation of the mean? Explain how you determined your answer.

- b. What percent of data points from Maria's spelling bee scores fall within 2 standard deviations from the mean? Explain how you determined your answer.

- c. What percent of data points from Maria's spelling bee scores fall within 3 standard deviations from the mean? Explain how you determined your answer.



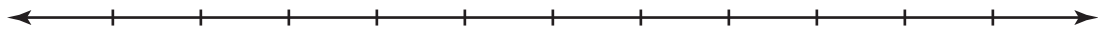
- d. Did the prediction about the percent of data points that fall within 1, 2, or 3 standard deviations match Maria's data set? Why do you think it did or did not?



It is important to note that the guideline regarding 68%, 95% and 99% is simply a guideline. In fact, there may be some data sets in which all of the data points lie within two standard deviations of the mean while other data sets may actually need four or greater standard deviations to encapsulate the entire data set. It is also important to know that because standard deviation is based on the mean of a data set, outliers may affect the standard deviation of the data set.



5. Graph 1, 2, and 3 standard deviations on the number line shown for Heidi's scores using a bell-shaped curve.



6. What similarities and differences do you notice between Maria's spelling score graph and Heidi's spelling score graph?



7. Advise Ms. Webb whom she should choose to captain the spelling bee team given the information about each student's standard deviation.

Talk the Talk



Mean and median are both measures of central tendency.

1. Identify which is more resistant to outliers, and which is more sensitive to outliers. Explain your reasoning.

The interquartile range and the standard deviation both measure the spread of data.

2. Identify which is more resistant to outliers, and which is more sensitive to outliers. Explain your reasoning.



Be prepared to share your solutions and methods.

Putting the Pieces Together

Analyzing and Interpreting Data

LEARNING GOALS

In this lesson, you will:

- Analyze and interpret data graphically and numerically.
- Determine which measure of central tendency and spread is most appropriate to describe a data set.

KEY TERMS

- stem-and-leaf plot
- side-by-side stem-and-leaf plot

Taking a trip on an airplane is always exciting. However, the process of flying can sometimes be frustrating. One of the most challenging tasks is boarding the plane before take-off. The most common method used to board passengers is boarding people by zone or row so that passengers in the back of the plane board first. This seems like it should be the most efficient way to board because people in the front won't be blocking the way. However, this is not necessarily the case. An astrophysicist used a computer simulation to try and determine the best method for loading passengers. After many simulations he found that passengers in even-numbered window seats near the back should board first, followed by even-numbered window seats in the middle, and even-numbered window seats in the front. This trend then continues through even-numbered middle seats, and even-numbered aisle seats. The whole process is then repeated with odd numbered seats.

So why does this work? It seems that allowing passengers a row between each other gives them more space to load their luggage and allows them to move if a passenger needs to get past them. This is not the only method that works, but it is the simplest for passengers to understand. Do you think airlines should try to change their methods for loading to this one? How much time do you really think it would save?

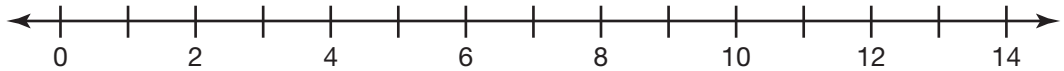
PROBLEM 1 Go For the Gold

When a participant takes part in the Special Olympics, each person receives a number. The chart shown represents the first twenty people labeled by their participation number and the number of gold medals each participant won.

Participation Number	Gold Medals Won
001	6
002	14
003	1
004	6
005	0
006	0
007	9
008	1
009	1
010	9
011	5
012	10
013	1
014	2
015	2
016	5
017	4
018	3
019	4
020	2



1. Analyze the data. Calculate the mean and standard deviation, and then interpret the meaning of each in terms of this problem situation.
2. Construct a box-and-whisker plot of the data and include any outliers.



3. Interpret the IQR.
4. Which measure of central tendency and spread should you use to describe this data? Explain your reasoning.
5. What conclusions can you draw about the number of gold medals participants won?

6. Shelly states the median and standard deviation should be used to describe the data because the standard deviation is less than the IQR. Is Shelly correct? Explain why or why not.



PROBLEM 2 Flying High



Data were collected from two rival airlines measuring the difference in the stated departure times, and the times the flights *actually* departed. The average departure time differences were recorded for each month for one year. The results are shown in the *side-by-side stem-and-leaf plot* given.

Difference in Departure Times (minutes)						
My Air Airlines				Fly High Airlines		
	5	0	0	7	8	
9	5	1	1	4	5	6
6	0	0	2	4	7	9
4	3	3	3	0	2	
		0	4	5	9	

$$2|4 = 24 \text{ minutes}$$

A **stem-and-leaf plot** is a graphical method used to represent ordered numerical data. Once the data is ordered, the stem and leaves are determined. Typically, the stem is all the digits in a number except the right most digit, which is the leaf. A **side-by-side stem-and-leaf plot** allows a comparison of two data sets. The two data sets share the same stem, but have leaves to the left and right of the stem.

Oh I remember stem-and-leaf plots! There should be a key somewhere which represents the value of each data point.



1. Describe the distribution of each data set.



2. Based on the shape of the data, calculate an appropriate measure of central tendency and spread for each data set.

3. What conclusions can you draw from the measure of central tendency and spread you calculated?



4. You are scheduling a flight for an important meeting and you must be there on time. Which airline would you schedule with? Explain your reasoning.

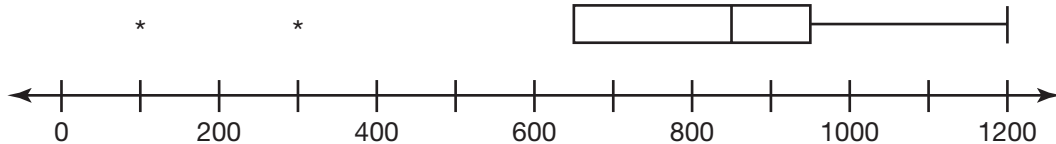
Talk the Talk

When analyzing data it is important to use both graphs and numbers to describe the data.

- The mean describes the average data point.
- The median describes the middle data point.
- Standard deviation describes the spread of the data from the mean.
- The interquartile range (IQR) describes the spread of the data from the median.
- For data that is symmetric, the mean is the most appropriate measure of central tendency and the standard deviation is the most appropriate measure of spread.
- For data that is skewed, the median is the most appropriate measure of central tendency and the IQR is the most appropriate measure of spread.



1. Analyze the box-and-whisker plot shown.



- a. Amina's teacher wants her students to create a list of data values that could result in the box plot shown. Amina states that she can just use the data values graphed as her list. She lists 100, 300, 700, 850, 950, and 1200 as her list. Is Amina's thinking correct? If yes, will this work for all box-and-whisker plots. If no, explain why not.

- b. Create a list of values that when graphed would result in the given box-and-whisker plot shown.

- c. Describe the data using an appropriate measure of central tendency and spread.

2. A data set ranges from 10 to 20. A value of 50 is added to the data set.
- Explain how the mean and median are affected by this new value.
 - Which measure of central tendency and spread would you use to describe the original data set before the new value is added? Explain your reasoning.
 - Which measure of central tendency and spread would you use to describe the data set after the new value is added? Explain your reasoning.



Be prepared to share your solutions and methods.

Chapter 8 Summary

KEY TERMS

- dot plot (8.1)
- discrete data (8.1)
- data distribution (8.1)
- symmetric distribution (8.1)
- skewed right distribution (8.1)
- skewed left distribution (8.1)
- box-and-whisker plot (8.1)
- five number summary (8.1)
- histogram (8.1)
- bin (8.1)
- frequency (8.1)
- continuous data (8.1)
- statistic (8.2)
- measure of central tendency (8.2)
- interquartile range (IQR) (8.3)
- outlier (8.3)
- lower fence (8.3)
- upper fence (8.3)
- standard deviation (8.4)
- normal distribution (8.4)
- stem-and-leaf plot (8.5)
- side-by-side stem-and-leaf plot (8.5)

8.1 Representing and Interpreting Data Displayed on Dot Plots

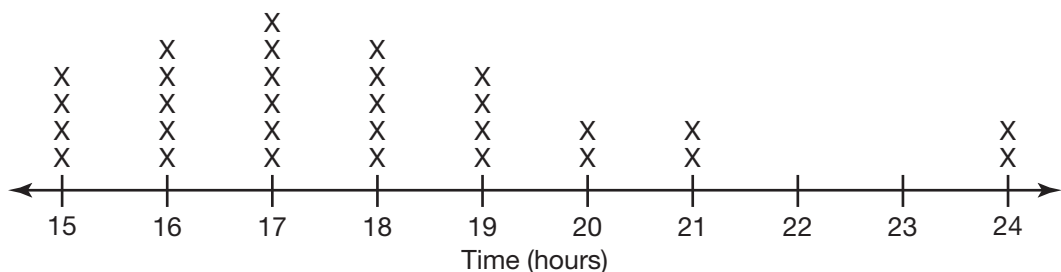
A dot plot is a graph that shows how discrete data are graphed using a number line. Discrete data are data that have only a finite number of values. Dot plots are best used to organize and display a small number of data points. The overall shape of the graph is called the distribution of the data, which is the way in which the data are spread out or clustered together. The most common distributions are symmetric, skewed right, and skewed left.

Example

A random sample of 30 college students was asked how much time he or she spent on homework during the previous week. The following times (in hours) were obtained:

16, 24, 18, 21, 18, 16, 18, 17, 15, 21, 19, 17, 17, 16, 19, 18, 15, 15, 20, 17, 15, 17, 24, 19, 16, 20, 16, 19, 18, 17

Time Spent on Homework in College



The data are skewed right.

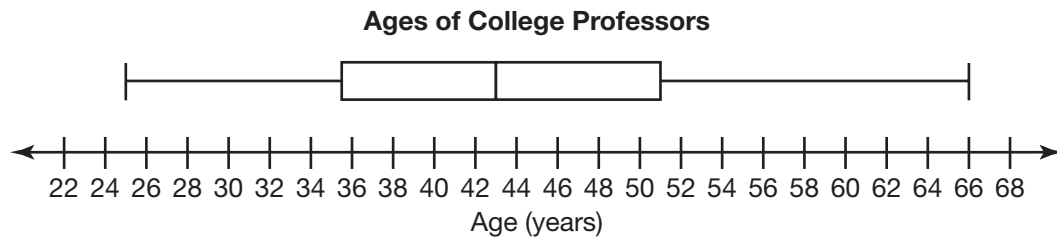
8.1 Representing and Interpreting Data Displayed on Box-and-Whisker Plots

A box-and-whisker plot displays the distribution of data based on a five number summary. The five number summary consists of the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value.

Example

The ages of 40 randomly selected college professors are given:

63, 48, 42, 42, 38, 59, 41, 44, 45, 28, 54, 62, 51, 44, 63, 66, 59, 46, 51, 28, 37, 66, 42, 40, 30, 31, 48, 32, 29, 42, 63, 37, 36, 47, 25, 34, 49, 30, 35, 50



Five-Number Summary:

- Lower bound: 25
- First quartile (Q1): 35.5
- Median: 43
- Third quartile (Q3): 51
- Upper bound: 66

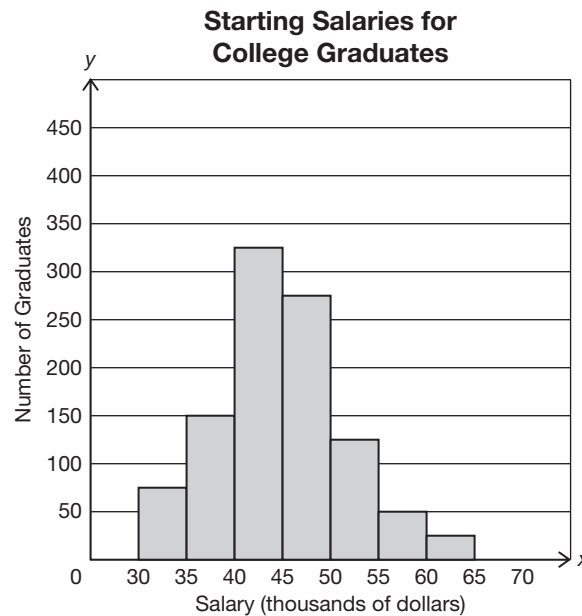
The following information can be determined from the box-and-whisker plot and five number summary:

- 50% of the professors are younger than 43 years old and 50% of the professors are older than 43 years old
- 25% of the professors are younger than 35.5 years old and 75% of the professors are older than 35.5 years old
- 75% of the professors are younger than 51 years old and 25% of the professors are older than 51 years old
- The middle 50% of the professors are between 35.5 years old and 51 years old.

A histogram is a graphical way to display quantitative data using vertical bars. The width of a bar represents an interval of data, and the height of the bar indicates the frequency. Histograms are effective in displaying large amounts of continuous data, which are data that can take any numerical value within a range.

Example

The histogram shows the starting salaries for college graduates based on a random sample of graduates.



The histogram shows that 325 graduates earned at least \$40,000 but less than \$45,000, 75 graduates earned at least \$30,000 but less than \$35,000, and only 25 graduates earned at least \$60,000 but less than \$65,000.

8.2 Calculating the Mean and Median of a Data Set

The measures of central tendency describe the “center” of the data set. Two measures of central tendency that are typically used to describe a set of data are the mean and the median. The arithmetic mean, or mean, represents the sum of the data values divided by the number of values. The median is the middle value of the data values.

Example

The number of home runs hit by each of the 12 batters for the York High School varsity baseball team is 0, 4, 8, 12, 14, 17, 19, 19, 23, 25, 28, and 48.

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ \bar{x} &= \frac{0 + 4 + 8 + 12 + 14 + 17 + 19 + 19 + 23 + 25 + 28 + 48}{12} \\ \bar{x} &= \frac{217}{12} \\ \bar{x} &\approx 18.08\bar{3}\end{aligned}$$

The mean of the data set is approximately 18 home runs.

0, 4, 8, 12, 14, 17, 19, 19, 23, 25, 28, 48

$$\begin{aligned}\text{median} &= \frac{17 + 19}{2} \\ &= \frac{36}{2} \\ &= 18\end{aligned}$$

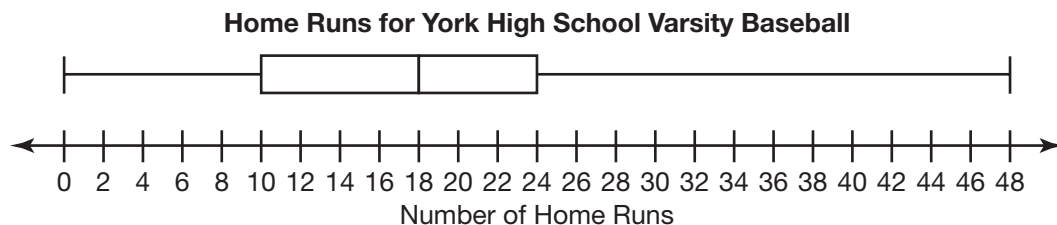
The median of the data set is 18 home runs.

8.2 Determining the Measure of Center which Best Represents a Data Set

The mean and median are two measures of central tendency which can be used to describe data. The distribution of the data set can be used to determine which measure is more appropriate. If the data is symmetric, the mean is more appropriate. If the data is skewed, the median is more appropriate because it is closer to most of the data points.

Example

The number of home runs hit by each of the 12 batters for the York High School varsity baseball team is represented on the box-and-whisker plot.



The data are skewed right, so the median would be the most appropriate measure of central tendency to describe the data.

Using the Interquartile Range to Determine if a Data Set Contains Outliers

The interquartile range, IQR, measures how far the data are spread out from the median. The IQR gives a realistic representation of the data without being affected by very high or very low data. The IQR is the range of the middle 50 percent of the data and is calculated by subtracting $Q3 - Q1$. An outlier is a data value that is significantly greater or lesser than the other data values. An outlier is typically calculated by multiplying the IQR by 1.5 and then determining if any data values are more than that distance away from $Q1$ or $Q3$.

Example

The data set represents the calorie count of 9 commercial breakfast sandwiches.

212, 361, 201, 203, 227, 224, 188, 192, 198

The five-number summary is:

- Minimum = 188
- First quartile = 195
- Median = 203
- Third quartile = 225.5
- Maximum = 361

$$\text{IQR} = 224 - 195$$

$$\text{IQR} = 29$$

The upper and lower fence are:

$$\begin{aligned} \text{Lower Fence} &= Q1 - (\text{IQR} \cdot 1.5) \\ &= 195 - (29 \cdot 1.5) \\ &= 151.5 \end{aligned}$$

$$\begin{aligned} \text{Upper Fence} &= Q3 + (\text{IQR} \cdot 1.5) \\ &= 225.5 + (29 \cdot 1.5) \\ &= 269 \end{aligned}$$

There are no values less than 151.5. One value, 361, is greater than 269. Therefore, 361 is an outlier.

Calculating and Interpreting the Standard Deviation of a Data Set

Standard deviation is a measure of how spread out the data are from the mean. A smaller standard deviation represents data that are more tightly clustered. A larger standard deviation represents data that are more spread out from the mean. The formula to determine standard deviation of a population is represented as:

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

A graphing calculator can also be used to determine the standard deviation.

Example

The data sets give the ages of 6 recent U.S. Presidents and the ages of the first 6 U.S. Presidents at their inauguration.

Recent Presidents	
President	Age
Carter	52
Reagan	69
G. H. W. Bush	64
Clinton	46
G. W. Bush	54
Obama	47

$$\sigma \approx 8.48$$

First Presidents	
President	Age
Washington	57
J. Adams	61
Jefferson	57
Madison	57
Monroe	58
J. Q. Adams	57

$$\sigma \approx 1.46$$

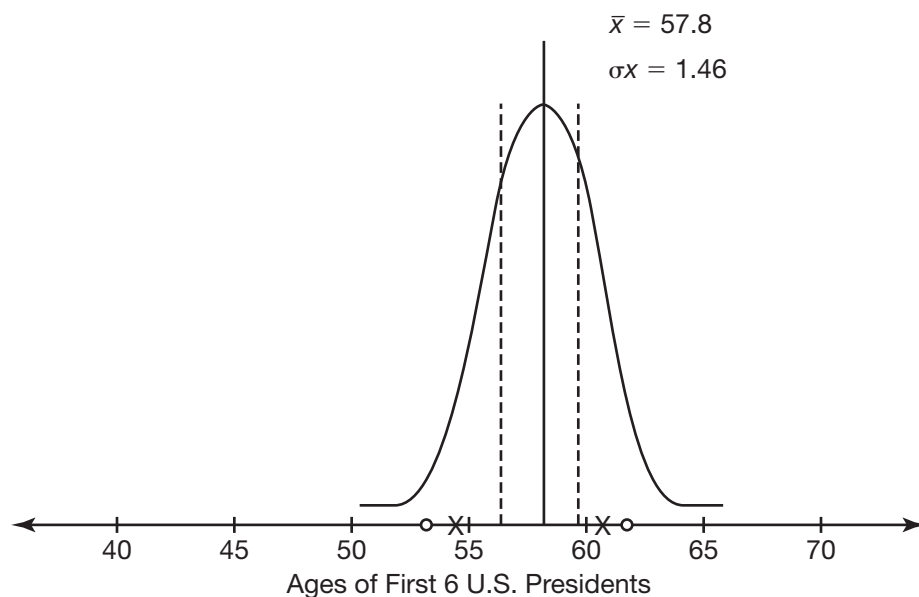
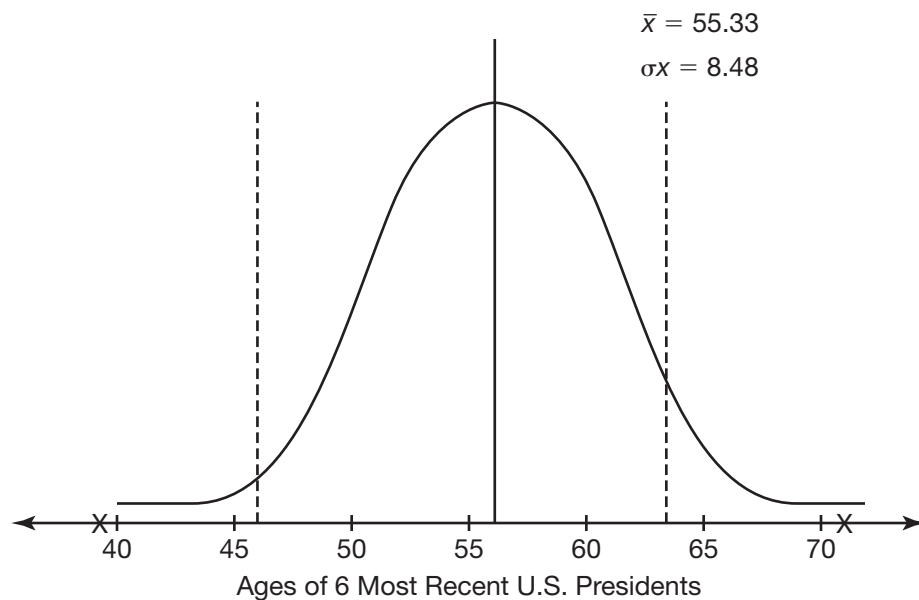
The ages of the 6 recent U.S. Presidents are more spread out than the ages of the first 6 Presidents because that data set has a higher standard deviation.

When you determine the standard deviation of a data set, you can represent it graphically in most normal distributions. A normal distribution is a function of a data distribution of many data points that form a bell-shaped curve.

By graphing the standard deviation, you can quickly determine which data set has a greater or lesser standard deviation. If a data set has a greater standard deviation, the data are more spread out from the mean in most normal distributions. If a data set has a lesser standard deviation, the data will be more clustered about the mean.

Example

The graph representing the ages of the 6 most recent U.S. Presidents is more spread out from the mean. The graph representing the ages of the first 6 U.S. presidents is clustered about the mean.



Determining which Measure of Center and Spread is Most Appropriate to Describe a Data Set

A stem-and-leaf plot is a graphical method used to represent ordered numerical data. A side-by-side stem-and-leaf plot allows a comparison of two data sets.

Example

The data sets give the ages of Oscar winners from 1999 through 2005 at the time of the award.

Year	Best Actor	Age
1999	Kevin Spacey	40
2000	Russell Crowe	36
2001	Denzel Washington	47
2002	Adrien Brody	29
2003	Sean Penn	43
2004	Jamie Foxx	37
2005	Philip Seymour Hoffman	38

Year	Best Actress	Age
1999	Hilary Swank	25
2000	Julia Roberts	33
2001	Halle Berry	35
2002	Nicole Kidman	35
2003	Charlize Theron	28
2004	Hilary Swank	30
2005	Reese Witherspoon	29

A side-by-side stem-and-leaf plot can be used to represent these data sets.

Ages of Oscar Winners From 1999 Through 2005 (years)							
Best Actors				Best Actresses			
		9	2	5	8	9	
8	7	6	3	0	3	5	5
7	3	0	4				

The data sets are relatively symmetric, so the mean and standard deviation are more appropriate to analyze the data.

For the Best Actors, the mean age is approximately 38.57 years. The standard deviation is approximately 5.26 years.

For the Best Actresses, the mean age is approximately 30.71 years. The standard deviation is approximately 3.49 years.

The average age of the Best Actors is about 8 years older than the Best Actresses. The ages for the Best Actors are more spread out than the ages for the Best Actresses.