2

# A Skewed Reality

## Determining the Better Measure of Center and Spread for a Data Set

## Warm Up

Calculate the mean of each data set.

1. 4, 4, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 12, 12

2. 0, 2, 10, 10, 11, 11, 11, 12, 12, 12, 13, 13

3. 40, 60, 60, 70, 70, 70, 80, 80, 100

4. 20, 20, 22, 23, 23, 24, 24, 24, 42, 50

## Learning Goals

- Calculate and interpret the mean and median of a data set.
- Determine which measure of central tendency is best to use for a data set.
- Calculate and interpret the interquartile range (IQR) of a data set.
- Determine whether a data set contains outliers.
- Calculate and interpret the standard deviation of a data set.
- Determine which measure of spread is best to use for a data set.

## Key Terms

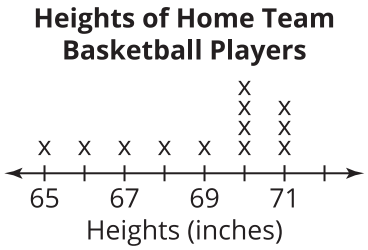
- statistics
- measure of central tendency
- interquartile range (IQR)
- data distribution
- outlier
- lower fence
- upper fence
- standard deviation

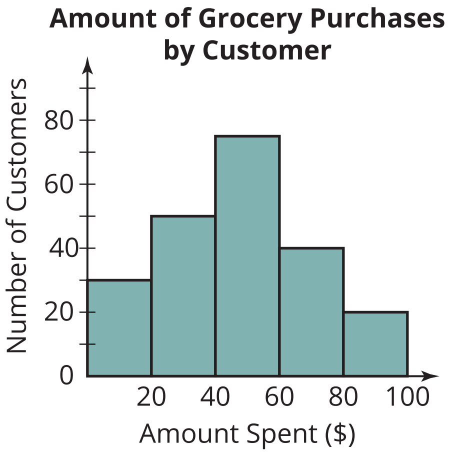
You have displayed and interpreted data sets using the statistical process. How can you further describe a data set using center, shape, and spread?

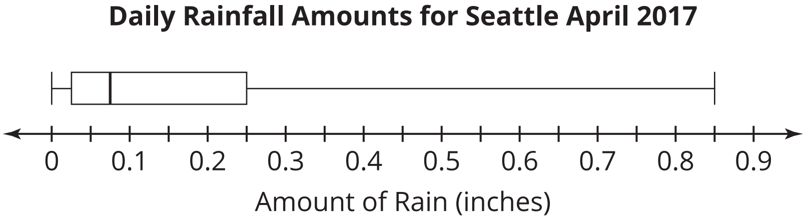
GETTING STARTED

## Make Your Mark

Consider each data display.







**1. Without doing any calculations, predict whether the mean or median will be greater for the data set represented by each display. Indicate your predictions by marking and labeling each measure of center on the number lines of the dot plot and the box-and-whisker plot, and within one or more bins of the histogram. Explain your reasoning.**

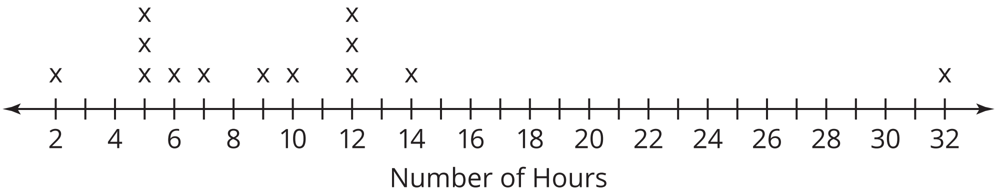
ACTIVITY

# 2.1 Median and IQR

You can 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 value 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.

A gym surveys its members about the average number of hours they spend at the gym each week. The data are recorded in the dot plot shown.

**Average Number of Hours Spent in Gym Each Week**



To describe the mean of a data set you need to calculate $\bar{x}$, which is read as "*x* bar."

Think about:

Just as you analyze data presented in a scatter plot to determine which type of regression equation best fits the data, you can analyze data in a display to determine which measure of center best fits the data.

## Worked Example

The formula shown represents the mean of a data set.

the sum of the data values

mean $\longrightarrow$ $\bar{x} = \frac{\Sigma x}{n}$

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.

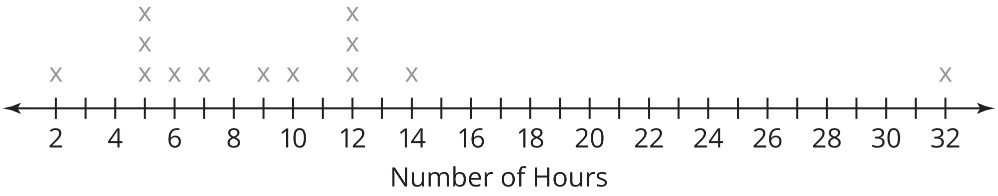
The E-like symbol is actually the Greek letter sigma and in mathematical terms it means the "summation" or "sum of."

The median of the data set from the worked example is 7, because the data in order from least to greatest are 5, 5, 7, 9, 10.

To describe the median of a data set, determine the middle number in a data set when the values are placed in order from least to greatest or greatest to least.

**1. Analyze the data collected from the gym members in the dot plot.**

**Average Number of Hours Spent in Gym Each Week**

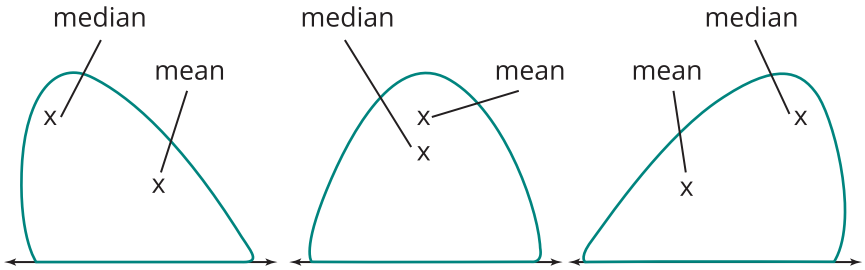


**a. Calculate the five-number summary for the data. Construct a box-and-whisker plot that displays the same data on top of the dot plot.**

**b. Calculate the mean of the data. Mark $\bar{x}$ above the point on the number line.**

**c. What do you notice about how the data are clustered?**

The overall shape of a graph is called the **data distribution**. Remember, there are three common distributions of data: skewed left, skewed right, and symmetric. The distribution of data can help you determine whether the mean or median is a better measure of center. Examine the diagrams shown.



**skewed right**

The mean of a data set is greater than the median when the data are skewed to the right.

The median is the best measure of center because the median is not affected by very large data values.

**symmetric**

The mean and median are equal when the data are symmetric.

**skewed left**

The mean of a data set is less than the median when the data are skewed to the left.

The median is the best measure of center because the median is not affected by very small data values.

**2. Which measure of central tendency would you choose to represent the data set? Explain your reasoning.**

The IQR is the range of the middle 50 percent of the data.

Another characteristic to consider when analyzing a graphical display is the spread, or variability, of the data. One common measure of spread is the *interquartile range* or *IQR*. The **interquartile range, IQR**, measures how far the data are spread out from the median. It is calculated by subtracting Q3 − Q1 in the five-number summary.

If the median is the better measure of center to use to describe a data set, then the IQR should be used to describe the spread. A box-and-whisker plot provides both of these pieces of information.

**3. Calculate the IQR of the data displayed in Question 1.**

Remember:

An **outlier** is a data value that is significantly greater or lesser than other data values in a data set.

Another useful statistic when analyzing data is to determine if there are any *outliers*. 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. The value of Q1 − (IQR · 1.5) is known as the **lower fence** and the value of Q3 + (IQR · 1.5) is known as the **upper fence**. Any value outside these limits is an outlier.

Let's analyze the data set from Question 1 to see how outliers can be represented on a box-and-whisker plot.

## Worked Example

2, 5, 5, 5, 6, 7, 9, 10, 12, 12, 12, 14, 32

Given this data set, the five-number summary is:

Minimum = 2, Q1 = 5, Median = 9, Q3 = 12, Maximum = 32

IQR = 7

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

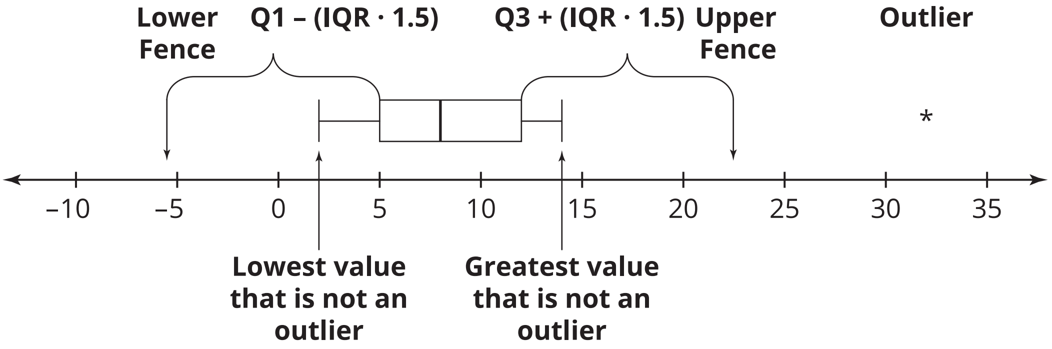
Lower Fence:
$= Q1 - (IQR \cdot 1.5)$
$= 5 - (7 \cdot 1.5)$
$= -5.5$
There are no values less than −5.5.

Upper Fence:
$= Q3 + (IQR \cdot 1.5)$
$= 12 + (7 \cdot 1.5)$
$= 22.5$
The value 32 is greater than 22.5.

If there are outliers, the whisker will end at the lowest or highest value that is not an outlier. Since 32 is an outlier, 14 is the greatest data value that is not an outlier.

Once the outlier is removed, the five-number summary is:

Minimum = 2, Q1 = 5, Median = 8, Q3 = 12, Maximum = 14



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

4. **Recalculate the IQR of the data from Question 1 with the outlier removed.**

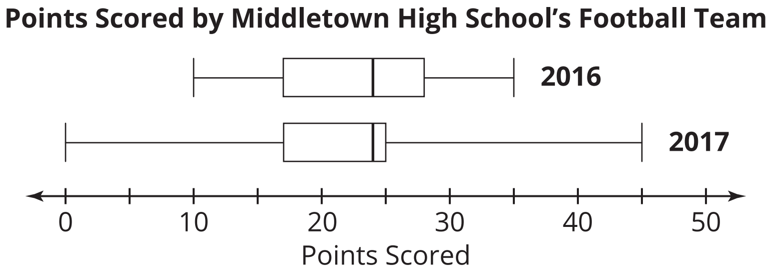
5. **Was the IQR affected by the outlier? Do you think this is true in all cases?**

ACTIVITY 2.2

# Using IQR to Identify Outliers

Coach Petersen's Middletown High School football team is struggling to win games this season. He is trying to determine why his team has won only a few times this year. The table shows the points scored in games in 2016 and 2017. The box-and-whisker plots represent and compare the data in the table.

| Points Scored (2016) | Points Scored (2017) |
|---|---|
| 10 | 0 |
| 13 | 7 |
| 17 | 17 |
| 20 | 17 |
| 22 | 18 |
| 24 | 24 |
| 24 | 24 |
| 27 | 24 |
| 28 | 25 |
| 29 | 27 |
| 35 | 45 |



1. **Which year do you think was better in terms of points scored?**

When comparing two data sets, if one data set appears symmetric and the other appears skewed, the median and IQR should be used to compare both data sets.

2. **Calculate and interpret the IQR for the points scored each year. What does the IQR tell you about which year was better?**

3. **Remove any outliers for the data sets 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?**

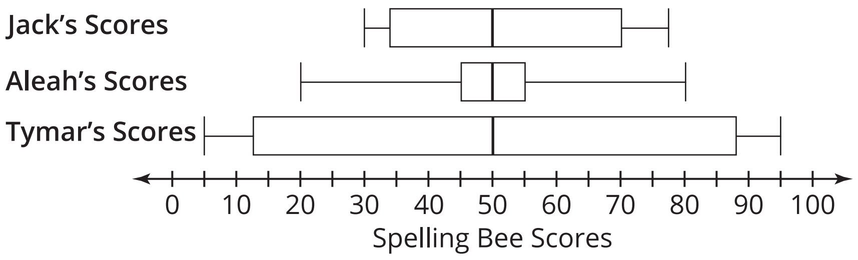
4. **Analyze the box-and-whisker plots with the outliers removed and compare the number of points scored each year.**

# ACTIVITY 2.3 Mean and Standard Deviation

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 |

The box-and-whisker plots display each of the student's spelling bee scores.



**1. Describe the shape of each student's data set.**

The reason why $n - 1$ is used in the formula is that statisticians have determined that it calculates a statistic that more closely represents the population.

You have learned about the spread of data values from the median, or the IQR. If you know the mean of a data set, you can calculate the spread using *standard deviation*. **Standard deviation** is a measure of how spread out the data are from the mean.

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

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

where $s$ is the standard deviation, $x_i$ represents each individual data value, $\overline{x}$ 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.

## Worked Example

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

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.

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

Next, evaluate the expressions to be added. Subtract $\overline{x}$ from each term and then square each difference.

$$(x_1 - \overline{x})^2$$
$$(6 - 7)^2 = 1$$
$$(4 - 7)^2 = 9$$
$$(10 - 7)^2 = 9$$
$$(8 - 7)^2 = 1$$

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

$$\frac{1 + 9 + 9 + 1}{4 - 1} = \frac{20}{3} \approx 6.7$$

Finally, calculate the square root of the quotient.

$$s = \sqrt{6.7}$$
$$s \approx 2.6$$

So the standard deviation for the given data set is approximately 2.6. It is important to note that if the data values have a unit of measure, the standard deviation of the data set also used the same unit of measure.

**2. 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.**

**Think about:**

The whole purpose of statistics is to make sense of a population using a sample. What is the sample in Ms. Webb's data? What is the population?

Each student's data set shows a symmetric distribution, so the mean is the better measure of center. Therefore, the standard deviation is the better measure to use to describe the spread.

**3. Use the standard deviation formula to determine the standard deviation of Jack's spelling bee scores.**

**a. Determine the $\bar{x}$ value.**

b. **Complete the table. The data values have been put in ascending order.**

| $x_i$ | $x_i - \bar{x}$ | $(x_1 - \bar{x})^2$ |
|---|---|---|
| 30 | | |
| 32 | | |
| 33 | | |
| 35 | | |
| 49 | | |
| 50 | | |
| 50 | | |
| 71 | | |
| 73 | | |
| 77 | | |
| **Sum** | | |

**Think about:**

The mean represents the balance point of the data values in the set. What sum should you expect to get when you add all the values for $x_i - \bar{x}$?

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

4. **Complete each table for Aleah's and Tymar's spelling bee scores.**

a. **Aleah**

| $x_i$ | $x_i - \bar{x}$ | $(x_1 - \bar{x})^2$ |
|---|---|---|
| 20 | | |
| 42 | | |
| 45 | | |
| 47 | | |
| 49 | | |
| 51 | | |
| 53 | | |
| 55 | | |
| 58 | | |
| 80 | | |
| **Sum** | | |

**b. Tymar**

| $x_i$ | $x_i - \overline{x}$ | $(x_1 - \overline{x})^2$ |
|---|---|---|
| 5 | | |
| 10 | | |
| 12 | | |
| 40 | | |
| 45 | | |
| 55 | | |
| 60 | | |
| 88 | | |
| 90 | | |
| 95 | | |
| **Sum** | | |

5. **Determine the standard deviation of Aleah's and Tymar's spelling bee scores.**

6. **Was the prediction you made in Question 2 correct? What do the standard deviations tell you about each student's spelling bee scores?**

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

To calculate the standard deviation:

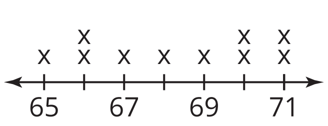
- Calculate the mean of the data set.
- Calculate the deviations from the mean.
- Add up the squared deviations.
- Divide by *n* − 1.
- Take the square root.

ACTIVITY 2.4

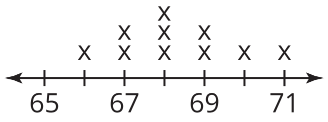
# Interpreting Standard Deviation

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.

**Home Team Heights (inches)**



**Visiting Team Heights (inches)**



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

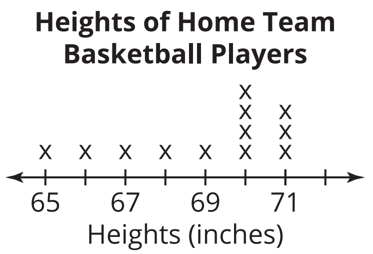
You can use technology to calculate the standard deviation of the data values for the sample of the population.

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?**

## TALK the TALK

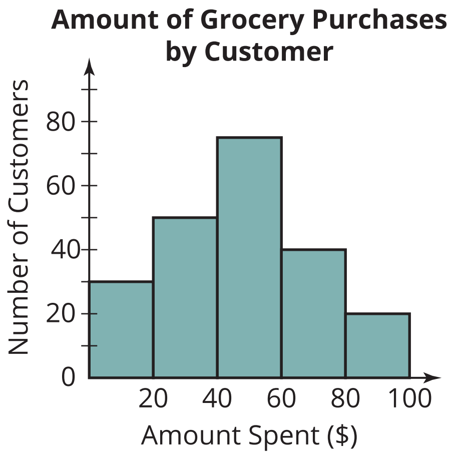
### Data on Display

Consider the dot plot from the Getting Started.



1. **Calculate the mean and median heights for the basketball players on the home team. Was your prediction correct?**

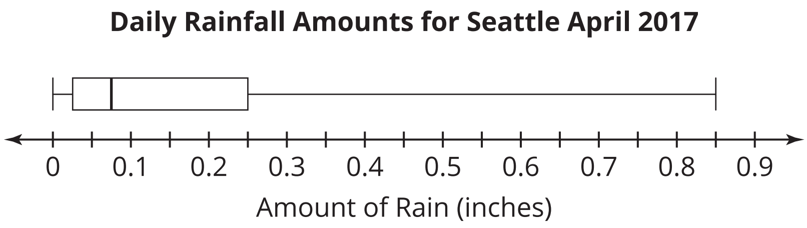
Consider the histogram from the Getting Started.



2. **The mean of the data set is 51 and the median of the data set is 50. How do these values compare to your prediction?**

NOTES

Consider the box-and-whisker plot from the Getting Started.



3. **The median of the data set is 0.07 and the mean of the data set is 0.16. How do these values compare to your prediction?**

4. **Determine which measure of center and which measure of spread would be most appropriate to use to describe each data set. Explain your reasoning.**

5. **How do you know which measure of center and measure of spread is most appropriate for a given data set?**

# Assignment

## Write

Match each definition to its corresponding term.

1. interquartile range (IQR)
2. standard deviation
3. lower fence
4. upper fence
5. statistic
6. measure of central tendency
7. outlier

a. a value calculated using the formula $Q1 - (IQR \cdot 1.5)$

b. numeric characteristics of a data set

c. a value that is much greater or lesser than other values in a data set

d. a value calculated using the formula $Q3 + (IQR \cdot 1.5)$

e. a measure of spread from the mean

f. a value used to describe the overall clustering of data in a set

g. a measure of spread from the median

## Remember

The median is the better measure of central tendency and the IQR is the better measure of spread to use to describe a data set that is skewed. The mean is the better measure of central tendency and the standard deviation is the better measure of spread to use to describe a data set that is symmetric. Outliers in a data set are calculated using the formula $Q1 - (IQR \cdot 1.5)$ to determine a lower fence and $Q3 + (IQR \cdot 1.5)$ to determine an upper fence. Any value outside these limits is an outlier.

## Practice

1. Consider each data set. Calculate the median, mean, IQR, and standard deviation of each set. Then, determine which measure of central tendency and which measure of spread is the most appropriate to use to describe the data set. Explain your reasoning.

   a. 1, 2, 2, 4, 8, 8, 8, 9, 9, 9, 10, 10, 10

   b. 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9

   c. 0, 1, 2, 10, 12, 12, 16, 16, 16, 16, 18, 18, 20

   d. 2, 2, 2, 3, 3, 4, 4, 8, 9, 9, 10, 10, 10

2. The five number summaries for the average monthly precipitation in millimeters during the summer for the Western and Midwestern states are provided.

| **West** | **Midwest** |
|---|---|
| Min = 7 | Min = 68 |
| Q1 = 22 | Q1 = 81.5 |
| Med = 33 | Med = 99.5 |
| Q3 = 49 | Q3 = 102.5 |
| Max = 107 | Max = 111 |

   a. Construct box-and-whisker plots of each area's monthly precipitation using the same number line for each.

   b. Describe the distribution of both box-and-whisker plots and explain what they mean in terms of the problem situation.

   c. Determine if there are outliers in either data set. Show your work and explain how you determined your answer.

   d. Chen is considering a long camping trip this summer and hopes to avoid the rain. Would you recommend that he camp in the West or the Midwest? Explain your reasoning.

## Stretch

Create a data set of 15 numbers where the mean and median are both 59 and the standard deviation is between 10 and 11. Then, add an outlier to your data set. How are the mean and standard deviation affected?

## Review

1. Alejandra has $900 to open a bank account. She wants to put her money in the bank where she will earn the most money over time.  Alejandra has a choice between the Platinum Bank that offers an account with 3% compound interest and the Diamond Bank that offers an account with 4% simple interest.
   a. What is the function used to calculate the balance in each account based on the year, *t*? Describe each function.
   b. In which bank should Alejandra deposit her money? Explain your reasoning.
2. The following is a list of seconds it takes swimmers to swim 50 yards freestyle.
   29, 27, 28, 24, 32, 30, 28, 29, 32, 26, 34, 30, 25, 27, 30, 29, 25, 28, 29, 32
   a. Construct a box-and-whisker plot based on the list of swimmers' times.
   b. What does the distribution of the box-and-whisker plot mean in terms of the swimmers' times?
3. Solve for *x* in each equation.
   a. $6^{5x-4} = 6^{4x}$  b. $9^{x} = 3^{3x+2}$