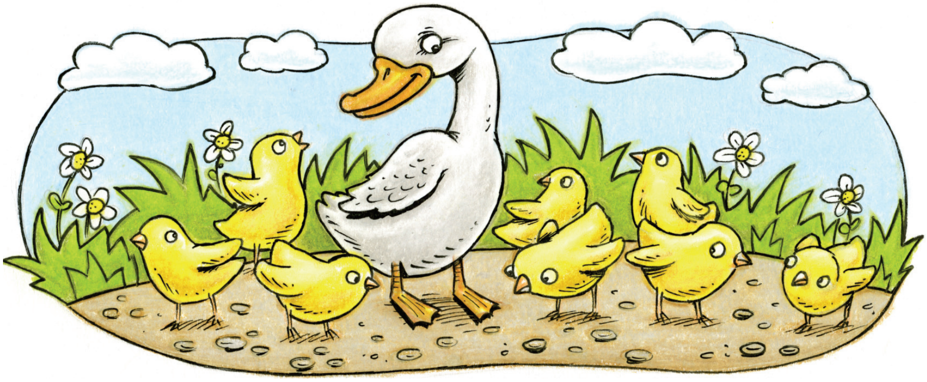
12.4

# The Effects of Outliers



## Focus on...

After this lesson, you will be able to...

- ❑ explain the effects of outliers on measures of central tendency
- ❑ justify whether outliers should be included when determining measures of central tendency

Can you spot the outlier in the cartoon shown? Suppose you are asked to determine the mean mass of these babies. Should this outlier be removed from the data set?

Some outliers are caused by mistakes in data collection. Other outliers are just as important as the other data values. When there are outliers in a data set, the mean, median, and range can be different from what they are when the outliers are removed. People who work with data need to decide when outliers should and should not be used when calculating measures of central tendency.

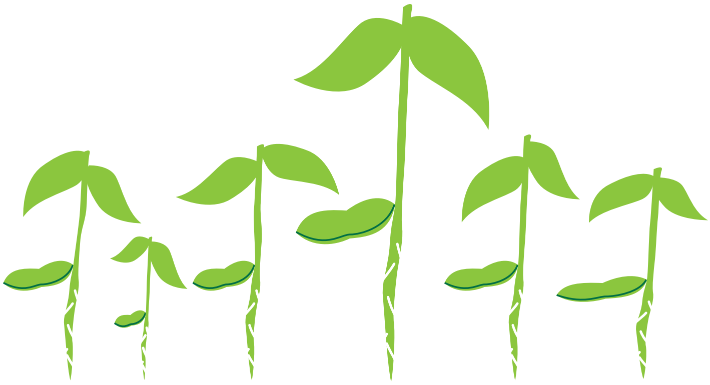
## Explore the Math

### How do outliers affect measures of central tendency?

**Materials**

- ruler

For a grade 7 science project, students were asked to plant and observe the growth of six bean seedlings. Students were given specific instructions to give their planted seeds 1 h of light and 30 mL of water per day. After two weeks the students brought their plants back to school.



**1.** With a ruler, measure the heights of the six bean seedlings shown. What are the heights?

**2.** Copy the following table into your notebook.

| Plant Height | Mean (cm) | Median (cm) | Largest Value | Smallest Value | Range |
|---|---|---|---|---|---|
| **With Outlier** | | | | | |
| **Without Outlier** | | | | | |

**3.** Complete the following calculations. Record your answers in the first row of your table.

**a)** Determine the mean and median heights for the plants. Round your answers to the nearest tenth of a centimetre.

**b)** What is the highest seedling height?

**c)** What is the lowest seedling height?

**d)** What is the range in heights?

**4.** Identify a possible outlier value.

**5.** Remove the outlier from your data. Repeat the calculations from #3. Record these answers in the second row of your table.

**6. a)** How has the median changed by removing the outlier?

**b)** How has the mean changed by removing the outlier?

**7.** What are some possible reasons why the one plant grew so much more than the other five? Compare your reasons with those of a classmate.

## Reflect on Your Findings

**8. a)** Which value is affected more by the presence of an outlier, the median or the mean? Explain.

**b)** Should the outlier value be included in the data for the science experiment? Explain why or why not.

## Example 1: Identify Outliers

Shannon practised shooting baskets every day last week to prepare for a basketball tournament. She recorded the number of baskets she made each day out of 25 shots.

| Day | Number of Baskets |
|---|---|
| Sunday | 14 |
| Monday | 17 |
| Tuesday | 16 |
| Wednesday | 20 |
| Thursday | 5 |
| Friday | 22 |
| Saturday | 18 |

**a)** What is the range of baskets scored?

**b)** What are the median and mean numbers of baskets scored?

**c)** Identify any possible outliers. Should the outlier(s) be removed from the data set? Explain why or why not.

### *Solution*

**a)** The highest and lowest values are 22 and 5.

$$\begin{aligned}\text{Range} &= 22 - 5\\ &= 17\end{aligned}$$

**b)** Arrange the data in order: ~~5~~, ~~14~~, ~~16~~, 17, ~~18~~, ~~20~~, ~~22~~

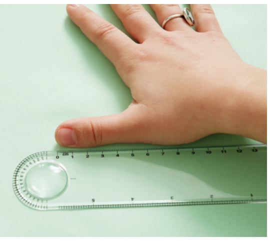
The median number of baskets scored is 17.

$$\begin{aligned}\text{Mean} &= \frac{5 + 14 + 16 + 17 + 18 + 20 + 22}{7}\\ &= \frac{112}{7}\\ &= 16\end{aligned}$$

The mean number of baskets scored is 16.

**c)** The value 5 could be considered an outlier. This value is significantly different from the other values. But this value should not be removed from the data set because it is just as important as the other data values. It is probably not an error in measurement. It may simply represent a poorer performance that day.

## Example 2: Identify Outliers and Determine Their Effects



In a science experiment, students were asked to measure the length of their right thumb from the first knuckle to the end of their thumb. The table shows the lengths that were measured by ten different students.

| Student | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Length (cm) | 3.2 | 2.7 | 3.1 | 2.8 | 2.8 | 2.7 | 31 | 3.3 | 2.6 | 3.0 |

**a)** What is the range?

**b)** What are the median and the mean?

**c)** Identify any possible outlier(s). Should the outlier(s) be removed from the data set? Explain why or why not.

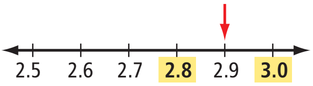
**d)** How would removing the outlier affect the median and the mean?

### Solution

**a)** The highest value is 31 cm.
The lowest value is 2.6 cm.
The range 28.4 cm.

$$\begin{aligned}\text{Range} &= 31 - 2.6\\ &= 28.4\end{aligned}$$

**b)** To find the median, arrange the numbers in order.
2.6, 2.7, 2.7, 2.8, 2.8, 3.0, 3.1, 3.2, 3.3, 31
With ten values, the median will be halfway between the fifth and sixth values.
The median is 2.9 cm.



$$\begin{aligned}\text{Mean} &= \frac{3.2 + 2.7 + 3.1 + 2.8 + 2.8 + 2.7 + 31 + 3.3 + 2.6 + 3.0}{10}\\ &= \frac{57.2}{10}\\ &= 5.72\end{aligned}$$

The mean is 5.72 cm.

**c)** The value of 31 is an outlier. It is about ten times greater than the other measurements and is most likely an error in either measurement or recording. The student may have measured in millimetres instead of centimetres. Or, it is possible that the student forgot to place the decimal point between the digits 3 and 1. In this case, the outlier should be removed.

**d)** Remove the outlier value of 31 from the ordered list of values from part b). Recalculate the median and the mean.
2.6, 2.7, 2.7, 2.8, 2.8, 3.0, 3.1, 3.2, 3.3
Since there are only nine values now, the median will be the fifth value. The median is 2.8 cm.

$$\begin{aligned}\text{Mean} &= \frac{2.6 + 2.7 + 2.7 + 2.8 + 2.8 + 3.0 + 3.1 + 3.2 + 3.3}{9}\\ &\approx 2.9\end{aligned}$$

The mean is approximately 2.9.

The median changes from 2.9 to 2.8. The mean changes from 5.7 to 2.9. The mean is affected more by removing the outlier.

### Show You Know

The following times were recorded, in seconds, for the runners in a race:
20.2, 16.5, 40.4, 18.5, 21.4, 20.5, 17.1, 24.5, 19.0

**a)** What is the range of times?

**b)** What are the median and mean times?

**c)** Identify any possible outliers. Should the outlier(s) be removed from the data set? Explain why or why not.

## Key Ideas

- Outliers can affect all measures of central tendency.
- When a small set of data has an outlier, the mean is usually affected more by the outlier than the median.
- Some outliers are just as important as the other data values, while others are better removed from the data set.

## Communicate the Ideas

**1.** Brian's bowling scores are 135, 132, 128, 316, 135, and 138. Identify a possible outlier in his scores. Should you remove it from the data set? Explain your reasoning.

**2. a)** Give an example of a situation where an outlier would be used when reporting on measures of central tendency.

**b)** Give an example of a situation where an outlier would not be used when reporting on measures of central tendency.

## Practise

*For help with #3 to #5, refer to Examples 1 and 2 on pages 442–443.*

**3.** The table shows the percent of students surveyed that had at least one song on their MP3 players by the musicians listed.

| Musician | Students With at Least One Song |
|---|---|
| Snoop Dogg | 42% |
| Shania Twain | 38% |
| Britney Spears | 6% |
| Kanye West | 40% |
| Led Zeppelin | 41% |
| Avril Lavigne | 38% |
| U2 | 88% |

**a)** What is the range?

**b)** What are the median and mean?

**c)** Identify any possible outliers. Should the outlier(s) be removed from the data set? Explain why or why not.

**4.** Two grade 7 students randomly ask the ages of the first eight people to pass them in the hallway. They record the following ages:

7, 11, 8, 8, 52, 9, 9, 10

**a)** What is the range?

**b)** What are the median and mean age?

**c)** Identify any possible outliers. Should the outlier(s) be removed from the data set? Explain why or why not.

**5.** Sharon recorded the following prices for five different brands of canned tomatoes on the grocery store shelf:

\$1.29, \$1.69, \$1.59, \$9.61, \$1.39

**a)** What is the range?

**b)** What are the median and the mean?

**c)** Identify any possible outlier(s). Should the outlier(s) be removed from the data set? Explain why or why not.

**d)** How would removing the outlier(s) affect the median and the mean?

## Apply

6. A medical study was conducted to learn the effect of caffeine on heart rate. Participants were asked to drink one 250-mL cup of coffee and record their number of heartbeats over a 15-second interval. The following data were collected.

   Heartbeats in 15 seconds:
   33, 35, 30, 70, 33, 31, 36, 40, 37, 29

   a) What is the range?

   b) What are the median and the mean?

   c) Identify any possible outlier(s). Should the outlier(s) be removed from the data set? Explain why or why not.

   d) How would removing the outlier(s) affect the median and the mean?

7. David had the following scores on his eight weekly spelling tests:

   70%, 80%, 80%, 70%, 100%, 80%, 20%, 90%

   a) What is the range?

   b) What are his median and mean scores?

   c) Identify any possible outlier(s). Should the outlier(s) be removed from the data set? Explain why or why not.

   d) How would removing the outlier(s) affect the median and the mean?

   e) Describe two different ways that you could determine David's overall mark to be between 75% and 85%.

## Extend

8. A set of nine numbers has two outliers. The mean and median both equal 50 if you include or exclude the outliers. What are the possible nine numbers?

## MATH LINK

In a gymnastics competition, each performance was judged by eight judges on a scale from 0.25 to 10.00. In order to calculate the gymnast's overall performance, the top score and bottom score were removed and the mean of the remaining scores was determined. This value is called the trimmed mean.

Jordan recorded the following scores for her friend's performance.

| Judge | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Score | 8.25 | 7.50 | 9.75 | 8.50 | 6.50 | 7.75 | 8.00 | 8.25 |

Round your answers to 2 decimal places.

a) Using all the scores, what is the median? mean? highest score? lowest score? range?

b) Remove the top and bottom scores. What is the new median? mean? range?

c) Which value in part b) has changed the most?

d) Would you consider the highest and lowest scores to be outliers in this example? Explain.