7.3 The Effects of Outliers on Average

Explore

Students in a Grade 7 class measured their pulse rates.
Here are their results in beats per minute:
97, 69, 83, 66, 78, 8, 55, 82, 47, 52, 67, 76, 84,
64, 72, 80, 72, 70, 69, 80, 66, 60, 72, 88, 88

➤ Calculate the mean, median, and mode for these data.
➤ Are there any numbers that are significantly different from the rest?
   If so, remove them.
   Calculate the mean, median, and mode again.
   Explain how the three averages are affected.

Reflect & Share

Compare your results with those of another pair of classmates.
How did you decide which numbers were significantly different?
Why do you think they are so different?

Connect

A number in a set of data that is significantly different from
the other numbers is called an **outlier**.

An outlier is much greater than or much less than most of the
numbers in the data set.

Outliers sometimes occur as a result of error in measurement or recording.
In these cases, outliers should be ignored.

Sometimes an outlier is an important piece of information that should not
be ignored. For example, if one student does much better or much worse
than the rest of the class on a test.

Outliers may not always be obvious.
Identifying outliers is then a matter of choice.

Example

Here are the marks out of 100 on an English test for students in a Grade 7 class:

21, 23, 24, 24, 27, 29, 29, 29, 32, 37, 37, 38, 39,

40, 50, 50, 51, 54, 56, 57, 58, 59, 61, 71, 80, 99

a) How many students were in the class? How do you know?

b) What is the outlier? Explain your choice.

c) Calculate the mean, median, and mode.

d) Calculate the mean, median, and mode without the outlier.
   What do you notice?

e) Should the outlier be used when reporting the average test mark? Explain.

A Solution

a) Count the number of marks to find the number of students in the class.
   There are 26 students.

b) There is only one number, 99, that is significantly different.
   The outlier is 99.
   The difference between the outlier and the nearest mark is $99 - 80 = 19$.
   This difference is much greater than that between other pairs of adjacent marks.

c) There are 26 marks. To find the mean mark, add the marks:
   $21 + 23 + 24 + 24 + 27 + 29 + 29 + 29 + 32 + 37 + 37 + 38 + 39 +$
   $40 + 50 + 50 + 51 + 54 + 56 + 57 + 58 + 59 + 61 + 71 + 80 + 99 = 1175$
   Divide the total by the number of marks, 26: $1175 \div 26 \doteq 45.2$
   The answer is written to the nearest tenth.
   The mean mark is about 45.2.
   The median mark is the mean of the 13th and 14th marks.
   The 13th mark is 39. The 14th mark is 40.
   So, the median is: $\frac{39 + 40}{2} = \frac{79}{2} = 39.5$
   The mode is the mark that occurs most often. This is 29.

d) Without the outlier, there are 25 marks and the sum of the marks is: $1175 - 99 = 1076$
   The mean is: $1076 \div 25 = 43.04$
   The median is the 13th mark: 39
   The mode is 29.
   When the outlier was removed, the mean and median decreased.
   The mode remained the same.

e) The outlier should be used when reporting the average test mark.
   To understand how the class is performing, all test marks should be included.

1. This set of data represents the waiting time, in minutes, at a fast-food restaurant:

5, 5, 5, 6, 5, 7, 0, 5, 1, 7, 7, 5, 6, 5, 5, 5, 8, 5, 0, 5, 4, 5, 2, 7, 9

> Remember to arrange the data in order before finding the median.

  a) Calculate the mean, median, and mode.
  b) Identify the outliers. Explain your choice.
  c) Calculate the mean, median, and mode without the outliers. How is each average affected when the outliers are not included?

2. Bryan recorded the time he spent on the school bus each day for one month. Here are the times, in minutes:

15, 21, 15, 15, 18, 19, 14, 20, 95, 18, 21, 14, 15, 20, 16, 14, 22, 21, 15, 19

  a) Calculate the mean, median, and mode times.
  b) Identify the outlier. How can you explain this time?
  c) Calculate the mean, median, and mode times without the outlier. How is each average affected when the outlier is not included?
  d) A classmate asks Bryan, "What is the average time you spend on the bus each day?" How should Bryan answer? Give reasons.

3. A clothing store carries pant sizes 28 to 46. A sales clerk records the sizes sold during her 6-h shift:

28, 36, 32, 32, 34, 4, 46, 44, 42, 38, 36, 36, 40, 32, 36

  a) Calculate the mean, median, and mode sizes.
  b) Is there an outlier? If so, why do you think it is an outlier?
  c) Calculate the mean, median, and mode sizes without the outlier. How is each average affected when the outlier is not included?
  d) Should the outlier be used when the sales clerk reports the average pant size sold during her shift? Explain your thinking.

4. Here are the science test marks out of 100 for the Grade 7 students in a combined-grades class:

0, 66, 65, 72, 78, 93, 82, 68, 64, 90, 65, 68

  a) Calculate the mean, median, and mode marks.
  b) Identify the outlier. How might you explain this mark?
  c) Calculate the mean, median, and mode marks without the outlier. How is each average affected when the outlier is not included?
  d) Should the outlier be used when reporting the average test mark? Explain.

5. a) Give an example of a situation in which outliers would not be used in reporting the averages. Explain why they would not be included.

b) Give an example of a situation in which outliers would be used in reporting the averages. Explain why they would be included.

6. (**Assessment Focus**) A Grade 7 class wanted to find out if a TV advertisement was true. The ad claimed that *Full of Raisins* cereal guaranteed an average of 23 raisins per cup of cereal. Each pair of students tested one box of cereal. Each box contained 20 cups of cereal. The number of raisins in each cup was counted.

a) Assume the advertisement is true. How many raisins should there be in 1 box of cereal?

b) Here are the results for the numbers of raisins in 15 boxes of cereal:
473, 485, 441, 437, 489, 471, 400, 453, 465, 413, 499, 428, 419, 477, 467

   i) Calculate the mean, median, and mode numbers of raisins.
   ii) Identify the outliers. Explain your choice.
   iii) Calculate the mean, median, and mode without the outliers. How do the outliers affect the mean?
   iv) Should the outliers be used when reporting the average number of raisins? Explain.
   v) Was the advertisement true? Justify your answer.

7. Take It Further Here is a set of data: 2, 3, 5, 5, 7, 8
An outlier has been removed.
a) Calculate the mean, median, and mode without the outlier.
b) The outlier is returned to the set.
The averages become:
Mean: 7   Median: 5   Mode: 5
What is the outlier? Show your work.

Reflect

Your friend is having difficulty recognizing outliers in a data set. What advice would you give your friend?

Applications of Averages

Focus | Understand which average best describes a set of data.

Explore

Record on the board how many siblings you have.
Use the class data.
Find the mean, the median, and the mode.
Find the range.

Reflect & Share

With a classmate, discuss which measure
best describes the average number of siblings.

Connect

A clothing store sold jeans in these sizes in one day:
28 30 28 26 30 32 28 32 26 28 34 38 36 30 34 32 30
To calculate the mean jeans size sold, add the sizes, then divide by
the number of jeans sold.

$$\text{Mean} = \frac{28 + 30 + 28 + 26 + 30 + 32 + 28 + 32 + 26 + 28 + 34 + 38 + 36 + 30 + 34 + 32 + 30}{17}$$

$$= \frac{522}{17}$$

$$\doteq 30.7$$

The mean size is approximately 30.7.

To calculate the median, order the jeans sold from least size to
greatest size. There are 17 numbers.
The middle number is the median. The middle number is the 9th.
26, 26, 28, 28, 28, 28, 30, 30, **30**, 30, 32, 32, 32, 34, 34, 36, 38
The median size is 30.

The mode is the number that occurs most often.
But, there are two numbers that occur most often.
So, there are two modes.
They are 28 and 30.
So, the mode sizes are 28 and 30.

> When there is an odd number of data,
> to find the middle number: Add 1 to
> the number of data, then divide by 2.
> This gives the position of the middle
> number. For example: $\frac{17 + 1}{2} = \frac{18}{2} = 9$;
> the middle number is the 9th.

In this situation, the mean, 30.7, is of little use.
The mean does not represent a size.

The median, 30, shows about one-half of the customers bought jeans of size 30
or smaller, and about one-half of the customers bought jeans of size 30 or larger.

The modes, 28 and 30, tell which sizes are purchased more often.
The mode is most useful to the storeowner.
He may use the mode to order extra stock of the most popular sizes.

Example

A bookstore has 15 books in its young adult section.
There are 5 different prices.
This table shows the number of books at each price.
a) Find the mean, median, and mode prices.
b) Which measure best represents the
   average price of a young adult book?
c) What is the range of the prices?

Young Adult Books	
Price ($)	Number of Books
8.99	3
9.99	5
13.99	5
32.99	1
37.99	1

A Solution

Make a list of the prices, in dollars:
8.99, 8.99, 8.99, 9.99, 9.99, 9.99, 9.99, 9.99, 13.99,
13.99, 13.99, 13.99, 13.99, 32.99, 37.99
a) Mean price:
   • Multiply each price by the total number
     of books at that price, then add the prices.
     $(8.99 \times 3) + (9.99 \times 5) + (13.99 \times 5) + 32.99 + 37.99 = 217.85$
   • Divide the total price by the total number of books: 15
     $\frac{217.85}{15} \doteq 14.52$, to two decimal places
   The mean price is approximately $14.52.
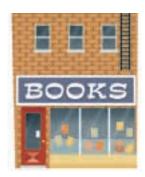
   Median price:
   There are 15 books.
   The list shows the books in order from least price
   to greatest price.
   The median price is the 8th price. The 8th price is $9.99.
   The median price is $9.99.

   Mode price:
   There are two mode prices. They are $9.99 and $13.99.

b) The mean price is not charged for any of the books.
Only two books cost more than the mean of $14.52.

There are two mode prices.
One mode, $9.99, is the same as the median price.

One-half the books cost the median price or less.
One-half cost more. So, the median price, $9.99, best represents the average price
of a young adult book at the store.

c) For the range, subtract the lowest price from the highest price:
37.99 − 8.99 = 29.00
The range of prices is $29.00.

The mean is usually the best average when no numbers in the
data set are significantly different from the other numbers.

The median is usually the best average when there are numbers
in the data set that are significantly different.

The mode is usually the best average when the data represent measures,
such as shoe sizes or clothing sizes.
A store needs to restock the sizes that sell most often.

Practice

1. The daily high temperatures for one week at Clearwater
   Harbour were: 27°C, 31°C, 23°C, 25°C, 28°C, 23°C, 28°C
   a) Find the mean, median, and mode for these data.
   b) Which average do you think best describes the daily high
      temperature at Clearwater Harbour that week? Explain.
   c) The weather channel reported the average temperature for
      Clearwater Harbour that week was 23°C. Is this correct? Explain.

2. Caitlin received these test marks in each subject.
   a) Find the mean, median, and mode mark for
      each subject.
   b) Explain what information each average gives.
   c) Which subject do you think Caitlin is best at?
      Worst at? Explain your reasoning.

Caitlin's Marks							
Math	85	69	92	55	68	75	78
Music	72	81	50	69	81	96	92
French	68	74	82	80	76	67	74

3. The table shows the tips earned by five waiters
and waitresses during two weeks in December.
   a) Calculate the mean, median, and mode tips for each week.
   b) Calculate the mean, median, and mode tips
   for the two-week period.
   c) Compare your answers in parts a and b.
   Which are the same? Which are different?
   Explain why.
   d) Explain which average best represents the tips
   earned during the two weeks.

Weekly Tips Earned ($)		
Waiter	Week 1	Week 2
James	1150	600
Kyrra	700	725
Tamara	800	775
Jacob	875	860
George	600	1165

4. A small engineering company has an owner and
5 employees. This table shows their salaries.
   a) Calculate the mean, median,
   and mode annual salaries.
   b) What is the range of the annual salaries?
   c) Which measure would you use to describe the
   average annual salary in each case? Explain.
   i) You want to attract a new employee.
   ii) You want to suggest the company does not
   pay its employees well.

Company Salaries	
Position	Annual Salary ($)
Owner	130 000
Manager	90 000
2 Engineers	50 000
Receptionist	28 000
Secretary	28 000

5. Is each conclusion correct? Explain your reasoning.
   a) The mean cost of a medium pizza is $10.
   So, the prices of three medium pizzas could be $9, $10, and $11.
   b) The number of raisins in each of 30 cookies was counted.
   The mean number of raisins was 15.
   So, in 10 cookies, there would be a total of 150 raisins.

6. **Assessment Focus** In each case, which average
do you think is most useful: the mean, median, or mode?
Justify your answer.
   a) A storeowner wants to know which sweater sizes she should order.
   Last week she sold 5 small, 15 medium, 6 large, and 2 X-large sweaters.
   b) Five of Robbie's friends said their weekly allowances are:
   $10, $13, $15, $11, and $10.
   Robbie wants to convince his parents to increase his allowance.
   c) Tina wants to know if her math mark was in the top half
   or bottom half of the class.

7. A quality control inspector randomly selects
boxes of crackers from the production line.
She measures their masses.
On one day she selects 15 boxes, and records these data:
- 6 boxes: 405 g each
- 2 boxes: 395 g each
- 4 boxes: 390 g each
- 2 boxes: 385 g each
- 1 box: 380 g
  a) Calculate the mean, median, and mode masses.
  b) What is the range of the masses?
  c) For the shipment of crackers to be acceptable,
     the average mass must be at least 398 g.
     Which average would you use to describe
     this shipment to make it acceptable? Explain.

8. Take It Further Andrew has these marks:
English 82%, French 75%, Art 78%, Science 80%
  a) What mark will Andrew need in math if he wants his mean
     mark in these 5 subjects to be each percent?
     i) 80%   **ii)** 81%   **iii)** 82%
  b) Is it possible for Andrew to get
     a mean mark of 84% or higher?
     Justify your answer.

9. Take It Further Celia received a mean mark of 80%
in her first three exams.
She then had 94% on her next exam.
Celia stated that her overall mean mark was 87%
because the mean of 80 and 94 is 87.
Is Celia's reasoning correct? Explain.

87%

Reflect

Use your answers from *Practice*. Describe a situation for each case.
  a) The mean is the best average.
  b) The median is the best average.
  c) The mode is the best average.
Justify your choices.

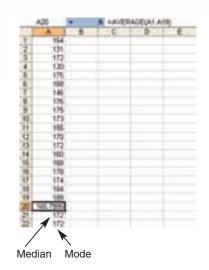# Using Spreadsheets to Investigate Averages

You can use spreadsheet software to find the mean, median, and mode of a set of data.

A spreadsheet program allows us to calculate the averages for large sets of data values quickly and efficiently.

You can also use the software to see how these averages are affected by outliers.

Here are the heights, in centimetres, of all Grade 7 students who were on the school track team: 164, 131, 172, 120, 175, 168, 146, 176, 175, 173, 155, 170, 172, 160, 168, 178, 174, 184, 189

In some spreadsheet software, the mean is referred to as the average.

Use spreadsheet software.

➤ Input the data into a column of the spreadsheet.

➤ Use the statistical functions of your software to find the mean, median, and mode. Use the Help menu if you have any difficulties.

➤ Investigate the effect of an outlier on the mean, median, and mode.
Delete 120.
What happens to the mean? Median? Mode? Explain.

➤ Suppose one member of the track team with height 155 cm is replaced by a student with height 186 cm.
How does this substitution affect the mean, median, and mode? Explain.

Median   Mode

Notice that when you add or remove data values, the averages change to reflect the adjustments.

1. Enter these data into a spreadsheet.
   They are the donations, in dollars, that were made to a Toy Wish Fund.
   5, 2, 3, 9, 10, 5, 2, 8, 7, 15, 14, 17, 28, 30, 16, 19, 4, 7, 9, 11, 25, 30,
   32, 15, 27, 18, 9, 10, 16, 22, 34, 19, 25, 18, 20, 17, 9, 10, 15, 35
   a) Find the mean, median, and mode.
   b) Add some outliers to your spreadsheet.
      State the values you added.
      How do the new mean, median, and mode compare to
      their original values? Explain.

2. Enter these data into a spreadsheet.
   They are purchases, in dollars, made by customers at a grocery store.
   55.40, 48.26, 28.31, 14.12, 88.90, 34.45, 51.02, 71.87, 105.12, 10.19,
   74.44, 29.05, 43.56, 90.66, 23.00, 60.52, 43.17, 28.49, 67.03, 16.18,
   76.05, 45.68, 22.76, 36.73, 39.92, 112.48, 81.21, 56.73, 47.19, 34.45
   a) Find the mean, median, and mode.
   b) Add some outliers to your spreadsheet.
      State the values you added.
      How do the new mean, median, and mode compare to
      their original values? Explain.

3. Enter these data into a spreadsheet.
   They are the number of ice-cream bars sold at the community centre
   each day in the month of July.
   101, 112, 127, 96, 132, 125, 116, 97, 124, 136, 123, 113, 78, 102, 118, 130,
   87, 108, 114, 99, 126, 86, 94, 117, 121, 107, 122, 119, 111, 105, 93
   Find the mean, median, and mode.
   What happens when you try to find the mode? Explain.

4. Repeat question 1 parts a and b.
   This time, enter data you find in the newspaper or on the Internet.

Reflect

List some advantages of using a spreadsheet to find the mean, median,
and mode of a set of data. What disadvantages can you think of?

Mid-Unit Review

7.1
7.2

1. Here are the heights, in centimetres, of the students in a Grade 7 class:
162, 154, 166, 159, 170, 168, 158, 162, 172, 166, 157, 170, 171, 165, 162, 170, 153, 167, 164, 169, 167, 173, 170

   a) Find the mean, median, and mode heights.

   b) What is the range of the heights?

2. The mean of five numbers is 20.
The median is 23.
What might the numbers be?
Find 2 different sets of data.

7.3

3. The cost of hotel rooms at *Stay in Comfort* range from $49 to $229 per night. Here are the rates charged, in dollars, for one particular night:
70, 75, 85, 65, 75, 90, 70, 75, 60, 80, 95, 85, 75, 20, 65, 229

   a) Calculate the mean, median, and mode costs.

   b) Identify the outliers.
   How can you explain these costs?

   c) Calculate the mean, median, and mode costs without the outliers. How is each average affected when the outliers are not included?

   d) Should the outliers be used when reporting the average cost of a hotel room? Explain.

7.4 **4.** A quality control inspector measures the masses of boxes of raisins. He wants to know if the average mass of a box of raisins is 100 g. The inspector randomly chooses boxes of raisins.
The masses, in grams, are:
99.1, 101.7, 99.8, 98.9, 100.8, 100.3, 98.3, 100.0, 97.8, 97.6, 98.5, 101.7, 100.2, 100.2, 99.4, 100.3, 98.8, 102.0, 100.3, 98.0, 99.4, 99.0, 98.1, 101.8, 99.8, 101.3, 100.5, 100.7, 98.7, 100.3, 99.3, 102.5

   a) Calculate the mean, median, and mode masses.

   b) For the shipment to be approved, the average mass of a box of raisins must be at least 100 g. Which average could someone use to describe this shipment to get it approved? Explain.

5. Is each conclusion true or false? Explain.

   a) The mode number of books read last month by students in James' class is 5. Therefore, most of the students read 5 books.

   b) A random sample of 100 people had a mean income of $35 000. Therefore, a random sample of 200 people would have a mean income of $70 000.