Identifying Outliers

Task

Students were asked to report how far (in miles) they each live from school. The following distances were recorded.

	Student	Distance
1	Zane	0.4
2	Jackson	0.5
3	Benjamin	1.0
4	Bethany	1.0
5	Joe	2.4
6	Noelle	2.7
7	Tianye	3.0
8	Anthony	3.2
9	Amanda	4.1
10	Michaela	4.2
11	Miranda	4.4

12	Joseph	5.0
13	John	7.5

1. Summary statistics for the distances are given below.

Min	Q1	Median	Q3	Max	Mean
0.4	1.0	3.0	4.3	7.5	3.03

Construct a box plot for the distances and describe the main features of the distribution.

2. John currently lives furthest from school. Would you consider his distance from school (7.5 miles) to be ”unusual"? Explain.

3. John’s family is considering moving to a house that is 10 miles from school. How will this move affect the summary statistics? Change John’s distance from 7.5 miles to 10 miles and complete the table based on this new data.

Min	Q1	Median	Q3	Max	Mean

Has the mean increased, decreased or remained unchanged? Has the median increased, decreased or remained unchanged? Explain how John’s move has affected these measures of center.

4. Construct a boxplot of the distances after John's move. Would you consider John's new distance from school (10 miles) to be "unusual" now? Explain.

5. A data point can be considered an "outlier" if it is more than 1.5 times the IQR above Q3 or more than 1.5 times the IQR below Q1. Using this description of an outlier, was John's distance from school considered an outlier before the move? How about after the move? Support your answer with appropriate work.

6. Zane lives closest to school. Using the description of an outlier given in question 5, is his distance considered an outlier?

IM Commentary

In high school, students interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers). The standard deviation is introduced as a measure of variability, and students calculate and interpret the standard deviation in context. The purpose of this task is to develop students understanding how extreme data points affect measures of center and how to use the mathematical definition to classify data points as "outliers" (S-ID.3).

It is important for students to understand how extreme values affect the locations of the mean and median. In the discussion of questions 2 and 4, students are asked to give their opinion about John's distance from school. In questions 5 and 6, students are introduced a rule for identifying outliers and see how to apply it to the data set consisting of the distances students live from school.

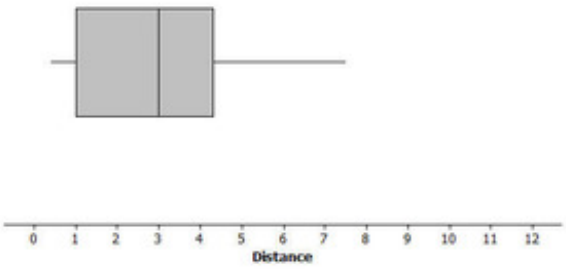
Other discussion points include the following.

- Discussion should move students from giving their judgment about calling John's distance an outlier to using one possible rule for identifying outliers.
- Students should understand how to compare extreme points to the outlier boundaries to classify points as outliers.
- Students should recognize how John's move affects the mean but not the median.
- It is possible to have multiple outliers.
- It is possible to have outliers on both extremes within the same distribution.
- Any skewness in a distribution tends to pull the mean in the direction of the skew.
- If a distribution is highly skewed, it may be inappropriate to use the mean as the measure of center as the mean may not be near many "typical" values in the

distribution.For example, suppose Bill Gates (one of the wealthiest men in the world) was meeting with a group of 10 teachers whose salaries are all less than $100,000. Although the mean annual salary for all of these people would be in the millions, no one individual makes anywhere near that amount. It would be more appropriate to use the median salary in this case.

Solution

1.



0.4 = 7.1of the students living between 1.0 and 4.3 miles from school. Skewness is present, as evidenced by the longer whisker on the right representing the upper 25% of the distances.

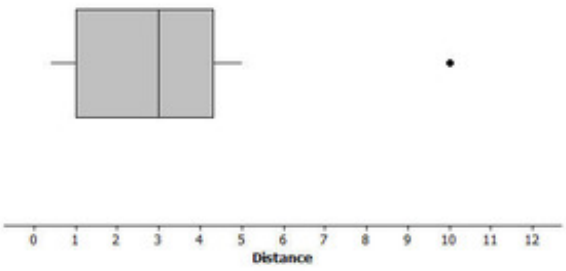
2. Some students may say yes and others may say no. Which ever answer the student gives should be justified. A student who thinks John's distance is unusual may justify it by saying that the John's data point is at the far right and that the right whisker is relatively long. A student who says his distance is not unusual may justify it by saying that the right whisker is not "extremely" long or by saying that the shortest distance (Zane's) is not unusual and John's distance from the median is not much larger than Zane's distance from the median. The purpose of this question is to motivate the need for a more objective definition of an outlier.

3.

Min	Q1	Median	Q3	Max	Mean
0.4	1.0	3.0	4.3	10.0	3.22

The mean has increased and the median is unchanged. The mean has increased because it is calculated by taking the sum of all of the distances and dividing by 13. Since John's distance has increased, the sum of all distances will increase and will cause the mean to also increase. Since the median is simply the middle value in a sorted list of distances, it remains unchanged, as John is still the student who lives furthest from school.

4.



Most students will agree that John is now an "outlier". They should be using the length of the right whisker relative to the rest of the boxplot to justify their choice.

5. The upper boundary for outliers is
Q3 = 1.5(IQR) = 4.3 + 1.5(4.3 - 1.0) = 4.3 + 4.95 = 9.25

Before the move, John's distance is not considered an outlier since 7.0 is less than 9.25. After the move, his distance in greater than 9.25, and now considered an outlier.

Note: The value Q3 + 1.5(IQR) is often referred to as the "upper fence". When asked why he decided on the scalar 1.5, John Tukey replied, "1.0 was too small and 2.0 was too big."

6. The lower boundary for outliers is
Q3 + 1.5(IQR) = 0.4 - 1.5(4.3 - 1.0) = 0.4 - 4.95 = -4.55

Since Zane's distance is greater than this "lower fence", his distance is not considered an outlier.