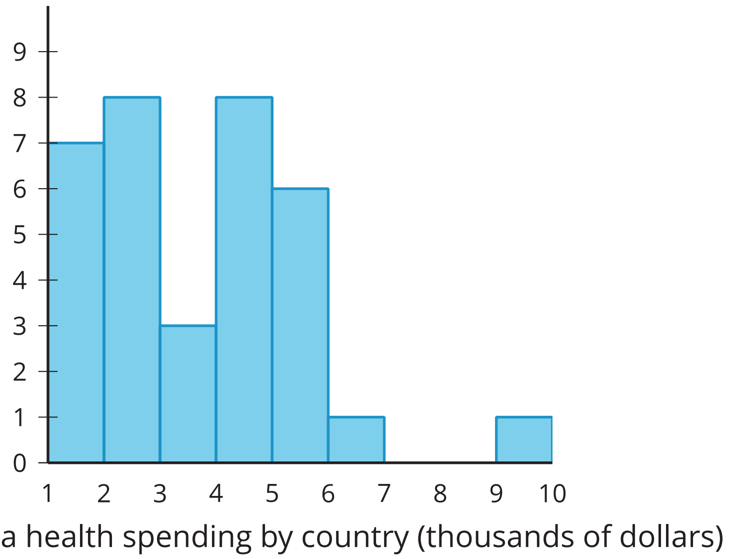
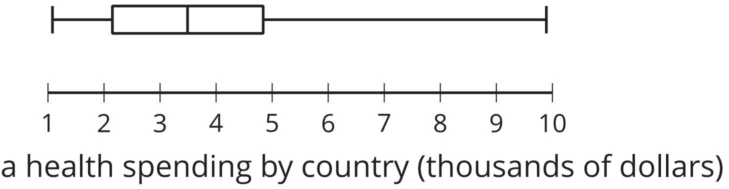
Lesson 14: Outliers

- Let’s investigate outliers and how to deal with them.

14.1: Health Care Spending

The histogram and box plot show the average amount of money, in thousands of dollars, spent on each person in the country (per capita spending) for health care in 34 countries.





1. One value in the set is an **outlier**. Which one is it? What is its approximate value?

2. By one rule for deciding, a value is an outlier if it is more than 1.5 times the IQR greater than Q3. Show on the box plot whether or not your value meets this definition of outlier.

CC BY 2019 by Illustrative Mathematics®

14.2: Investigating Outliers

Here is the data set used to create the histogram and box plot from the warm-up.

1.0803	1.0875	1.4663	1.7978	1.9702	1.9770	1.9890	2.1011
2.1495	2.2230	2.5443	2.7288	2.7344	2.8223	2.8348	3.2484
3.3912	3.5896	4.0334	4.1925	4.3763	4.5193	4.6004	4.7081
4.7528	4.8398	5.2050	5.2273	5.3854	5.4875	5.5284	5.5506
6.6475	9.8923						

1. Use technology to find the mean, standard deviation, and five-number summary.

2. The maximum value in this data set represents the spending for the United States. Should the per capita health spending for the United States be considered an outlier? Explain your reasoning.

3. Although outliers should not be removed without considering their cause, it is important to see how influential outliers can be for various statistics. Remove the value for the United States from the data set.

   a. Use technology to calculate the new mean, standard deviation, and five-number summary.

   b. How do the mean, standard deviation, median, and interquartile range of the data set with the outlier removed compare to the same summary statistics of the original data set?

CC BY 2019 by Illustrative Mathematics®

14.3: Origins of Outliers

1. The number of property crime (such as theft) reports is collected for 50 colleges in California. Some summary statistics are given:

15	17	27	31	33	39	39	45
46	48	49	51	52	59	72	72
75	77	77	83	86	88	91	99
103	112	136	139	145	145	175	193
198	213	230	256	258	260	288	289
337	344	418	424	442	464	555	593
699	768						

- mean: 191.1 reports
- minimum: 15 reports
- Q1: 52 reports
- median: 107.5 reports
- Q3: 260 reports
- maximum: 768 reports

a. Are any of the values outliers? Explain or show your reasoning.

b. If there are any outliers, why do you think they might exist? Should they be included in an analysis of the data?

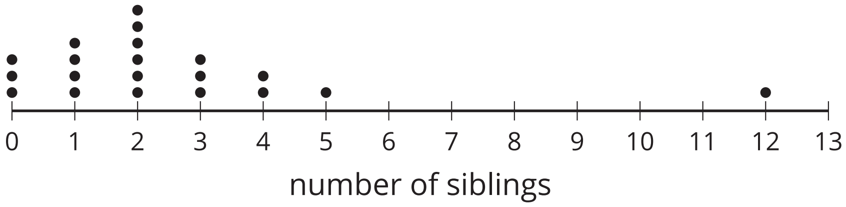
CC BY 2019 by Illustrative Mathematics®

2. The situations described here each have an outlier. For each situation, how would you determine if it is appropriate to keep or remove the outlier when analyzing the data? Discuss your reasoning with your partner.

   a. A number cube has sides labelled 1–6. After rolling 15 times, Tyler records his data:

      1, 1, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 5, 6, 20

   b. The dot plot represents the distribution of the number of siblings reported by a group of 20 people.



   c. In a science class, 11 groups of students are synthesizing biodiesel. At the end of the experiment, each group recorded the mass in grams of the biodiesel they synthesized. The masses of biodiesel are

      0, 1.245, 1.292, 1.375, 1.383, 1.412, 1.435, 1.471, 1.482, 1.501, 1.532

CC BY 2019 by Illustrative Mathematics®

Are you ready for more?

Look back at some of the numerical data you and your classmates collected in the first lesson of this unit.

1. Are any of the values outliers? Explain or show your reasoning.

2. If there are any outliers, why do you think they might exist? Should they be included in an analysis of the data?

Lesson 14 Summary

In statistics, an **outlier** is a data value that is unusual in that it differs quite a bit from the other values in the data set.

Outliers occur in data sets for a variety of reasons including, but not limited to:

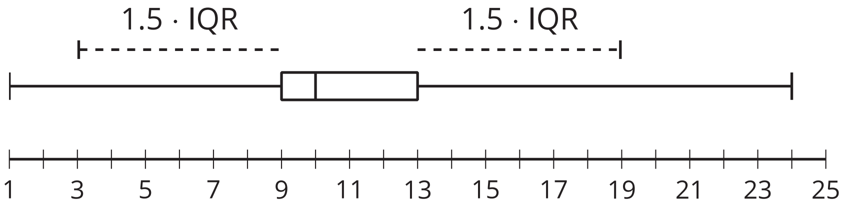
- errors in the data that result from the data collection or data entry process
- results in the data that represent unusual values that occur in the population

Outliers can reveal cases worth studying in detail or errors in the data collection process. In general, they should be included in any analysis done with the data.

A value is an outlier if it is

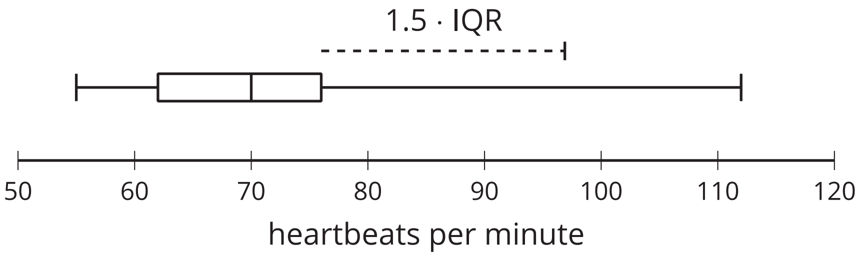
- more than 1.5 times the interquartile range greater than Q3 (if $x > \text{Q3} + 1.5 \cdot \text{IQR}$)
- more than 1.5 times the interquartile range less than Q1 (if $x < \text{Q1} - 1.5 \cdot \text{IQR}$)

In this box plot, the minimum and maximum are at least two outliers.



CC BY 2019 by Illustrative Mathematics®

It is important to identify the source of outliers because outliers can impact measures of center and variability in significant ways. The box plot displays the resting heart rate, in beats per minute (bpm), of 50 athletes taken five minutes after a workout.



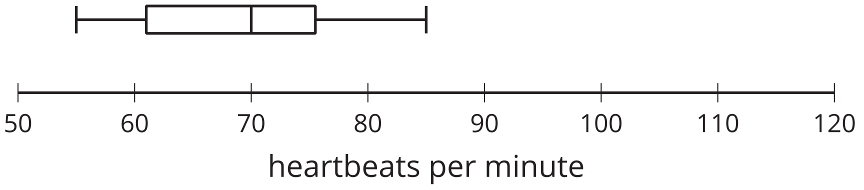
Some summary statistics include:

- mean: 69.78 bpm
- standard deviation: 10.71 bpm
- minimum: 55 bpm
- Q1: 62 bpm
- median: 70 bpm
- Q3: 76 bpm
- maximum: 112 bpm

It appears that the maximum value of 112 bpm may be an outlier. Since the interquartile range is 14 bpm ($76 - 62 = 14$) and $\text{Q3} + 1.5 \cdot \text{IQR} = 97$, we should label the maximum value as an outlier. Searching through the actual data set, it could be confirmed that this is the only outlier.

After reviewing the data collection process, it is discovered that the athlete with the heart rate measurement of 112 bpm was taken one minute after a workout instead of five minutes after. The outlier should be deleted from the data set because it was not obtained under the right conditions.

Once the outlier is removed, the box plot and summary statistics are:



- mean: 68.92 bpm
- standard deviation: 8.9 bpm
- minimum: 55 bpm
- Q1: 61 bpm
- median: 70 bpm
- Q3: 75.5 bpm
- maximum: 85 bpm

CC BY 2019 by Illustrative Mathematics®

The mean decreased by 0.86 bpm and the median remained the same. The standard deviation decreased by 1.81 bpm which is about 17% of its previous value. Based on the standard deviation, the data set with the outlier removed shows much less variability than the original data set containing the outlier. Since the mean and standard deviation use all of the numerical values, removing one very large data point can affect these statistics in important ways.

The median remained the same after the removal of the outlier and the IQR increased slightly. These measures of center and variability are much more resistant to change than the mean and standard deviation. The median and IQR measure the middle of the data based on the number of values rather than the actual numerical values themselves, so the loss of a single value will not often have a great effect on these statistics.

The source of any possible errors should always be investigated. If the measurement of 112 beats per minute was found to be taken under the right conditions and merely included an athlete whose heart rate did not slow as much as the other athletes, it should not be deleted so that the data reflect the actual measurements. If the situation cannot be revisited to determine the source of the outlier, it should not be removed. To avoid tampering with the data and to report accurate results, data values should not be deleted unless they can be confirmed to be an error in the data collection or data entry process.

CC BY 2019 by Illustrative Mathematics®