CCSS: 6.SP.2, 6.SP.5.a, 6.SP.5.b, 6.SP.5.d

Lesson 31: Describing Data Sets

To solve a statistical question, you need to follow the following steps.

Step 1: Identify whether the question requires statistics.

As covered in Lesson 28, questions may or may not require statistics to be solved. If the answer to the question will not include variability in the data, then statistics may not be required.

For example, "How tall is Robert?" is not a statistical question because there is no variability. The question, "How tall are the students in Robert's class?" *is* a statistical question because there will be variability in the heights of the students.

Step 2: Identify the variable.

For any statistical question, the quantity that is being questioned is a variable. You need to determine how this variable will be measured. You must also determine which units of measurement will be used to describe the variable.

For example, the question, "How tall are the students in Robert's class?" can be answered using a tape measure. If the tape measure uses the U.S. system of measurement, the units can be feet or inches.

Step 3: Collect the data for the data set.

Using the measurement system determined in Step 2, collect the data. There are many methods for collecting data. It is important to consider that the collection must answer the statistical question directly.

Consider the following question: "What is the average amount of rain that falls in the park each day?" The collection method must be accurate and consistent to answer the question. For example, a container must be located in a place that measures the rainfall accurately. Placing the container below a tree could influence the results.

A data set will be more accurate if more values are collected for the set. For example, you could collect rainfall every day for a week and answer the question. However, it would be more accurate to measure the rainfall every day for a year. The larger data set would provide a better solution to the statistical question.

CCSS: 6.SP.2, 6.SP.5.a, 6.SP.5.b, 6.SP.5.d

Step 4: Calculate the measures of the data set.

There are several measures that can be used to summarize a data set. Those measures include the mean, median, mode, range, interquartile range, and mean absolute deviation. Each of these measures represents a different way to interpret the data.

The mean, median, and mode show the **measures of center** of the data. Measures of center show a way that the average of the data can be represented. The range, interquartile range, and mean absolute deviation are **measures of variability**. Measures of variability show how the data varies within the set. Measures of center and variability both represent the data set in different ways.

Step 5: Display the data.

There are many ways to display data. For example, you can use a dot plot, a box plot, or a histogram. Each type of display has its own benefits. For example, a dot plot can easily show the most common values in a data set. A box plot can show the range, median, and quartiles of a data set. A histogram shows continuous data using intervals. You can also represent the same data set using multiple displays.

Step 6: Analyze the data.

Once the data from the data set are displayed, you can analyze the data. Look for patterns or deviations in the data. For example, if the display is a histogram, what is the shape of the histogram? Is the histogram left- or right-skewed? Does the data show a bell-shaped distribution of values? Is there anything unusual that stands out in the data?

Consider how the data were collected. Determine whether the context of the data collection affected the results.

Example

Follow the steps to answer the following question.

What are the ages of the customers at a G-rated movie?

Step 1: Identify whether the question requires statistics.

It is unlikely that everyone in the movie theater is the same age. Therefore, this question will likely include variability in the data. It is a statistical question.

Step 2: Identify the variable.

The information that you will need to find is the ages of the customers at the movie. The unit of measurement to use for this data is a year. Knowing the exact age in terms of days, weeks, or months is probably not necessary.

The simplest way to measure the data is to simply ask the customers for their ages. No additional tools are needed for measurement.

Step 3: Collect the data for the data set.

The data can be collected from different places. For example, the person collecting the data could ask the customers for their ages outside the ticket window, by the entrance, at the snack bar, or inside the theater. A survey conducted outside the ticket window may ask people on the street who will not be customers of the G-rated movie.

A person who is trying to collect the data may not be able to ask every customer for his or her age. The more customers who participate in the survey, the more accurate the results will be. The following table shows the ages of the 32 customers entering the G-rated movie.

Ages of Customers at Movie (in years)							
5	8	28	30	6	11	33	6
52	3	39	8	34	12	7	42
33	45	11	4	6	44	68	12
8	6	40	27	35	10	5	37

CCSS: 6.SP.2, 6.SP.5.a, 6.SP.5.b, 6.SP.5.d

Step 4: Calculate the measures of the data set.

Before any of the measures can be calculated, put the ages in order from least to greatest.

3, 4, 5, 5, 6, 6, 6, 6, 7, 8, 8, 8, 10, 11, 11, 12, 12, 27, 28, 30, 33, 33, 34, 35, 37, 39, 40, 42, 44, 45, 52, 68

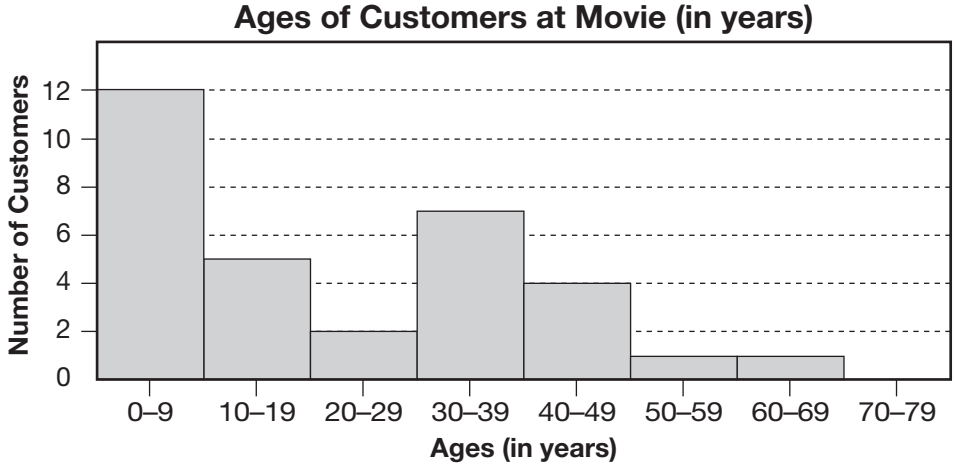
The sum of the values is 715. There are 32 values. The mean age is $\frac{715}{32}$, or about 22 years old. There is an even number of values in the set. The two middle numbers are both 12. Because $\frac{12 + 12}{2} = 12$, the median age is 12. The age 6 appears more often than any other value in the data set, so the mode is 6 years old. The range is $68 - 3$, or 65.

The first quartile is the median of the lower half of values, 6.5. The third quartile is the median of the upper half of values, 36. Because $36 - 6.5 = 29.5$, the interquartile range is 29.5.

To find the mean absolute deviation, find the absolute value of the difference between each value and 22. The differences are 19, 18, 17, 17, 16, 16, 16, 16, 15, 14, 14, 14, 12, 11, 11, **10, 10**, 5, 6, 8, 11, 11, 12, 13, 15, 17, 18, 20, 22, 23, 30, and 46. The mean absolute derivation is the mean (average) of those numbers, which is equal to about 15 years.

Step 5: Display the data from the data set.

A histogram can display the different age groups using intervals.



Step 6: Analyze the data.

Because the majority of the data appears on the left side, the histogram is right-skewed. The tallest bar is for 0–9, which means that more customers are between 0 and 9 than in any other age group. This is to be expected for a kids' movie. The shape of the histogram does not show that the data go down exactly with increasing age, however. After the initial peak at 0–9, there is another peak at 30–39, which probably represents the parents of the children.