In this lesson you will

- create and interpret box plots for sets of data
- use the interquartile range (IQR) to identify potential outliers and graph them on a modified box plot
- identify the shape of data as skewed or symmetric

To make it easier to understand and interpret a large set of data, you can display the values in a graph and compute numerical measures, or statistics, that summarize the data.

The mean, median, and mode are statistics that give an indication of the typical value in a data set. You probably learned about these measures of central tendency in earlier math classes. Review these measures by working through the examples and exercises in the lesson Refreshing Your Skills for Chapter 2.

A good description of a data set includes a measure of central tendency along with information about the shape and spread of the data. A box plot is a useful tool for showing the shape and spread of the data. Work through Example A of Lesson 2.1 to review how to find the five-number summary of a data set and create a box plot.

When analyzing data, it is important to know how the data were collected. If the backpack data were collected only from students leaving a particular class, the data might be biased, or unfair. Data from a simple random sample of all the students would be more representative of the backpack weights for all students. Read the text after Example A in your book to better understand these terms and to review the five-number summary values.

The box plot below shows the backpack data, with the five-number summary values highlighted and explained.

The left edge of the box is the first quartile, $Q_1$, which is the median of the data values below the median.

The right edge of the box is the third quartile, $Q_3$, which is the median of the data values above the median.

The minimum, $Q_1$, median, $Q_3$, and maximum are collectively called the five-number summary.

The segments extending from the “box” are called “whiskers.” Notice how long the right whisker is. In the backpack data, the heaviest backpack weighs 33 lb, which is far more than the next largest weight of 20 lb. Extreme values in a data set are called outliers. In a modified box plot, any points that are more than 1.5 times the IQR from the ends of the box are plotted as separate points. To help you better understand outliers and modified box plots, work through Example B in your book.

(continued)
Lesson 2.1 • Box Plots (continued)

Statisticians use the word shape to describe how data are distributed relative to the position of the measure of central tendency. Symmetric data are balanced, or nearly balanced, at the center. Skewed data are spread out more on one side of the center than on the other.

This box plot shows a symmetric data set.

Skewed right implies that the data are spread more to the right of the center than to the left.

This data set is skewed left.

Investigation: Pulse Rates

The investigation in your book involves collecting resting and exercise pulse rates from all the students in your class. If you aren’t able to collect real data, use the data below.

Resting rates: 68, 76, 84, 80, 76, 72, 60, 68, 68, 68, 80, 68, 68, 64, 64, 72, 76, 72, 68, 56, 88, 80, 76, 68, 56, 64, 60, 92, 72, 84, 72


To find the five-number summary for the resting rates, start by ordering them.

56, 56, 60, 60, 64, 64, 64, 68, 68, 68, 68, 68, 68, 68, 72, 72, 72, 72, 76, 80, 80, 80, 80, 84, 84, 84, 88, 92

The minimum value for these sample data is 56, and the maximum is 92. There are 30 values in the data set, so the median is the mean of the 15th and 16th values, which is 72. The first quartile is the median of the lower 15 values, which is 68. The third quartile is the median of the upper 15 values, which is 80.

So, the five-number summary for the sample resting-rate data is 56, 68, 72, 80, 92. Using the same process, you can determine that the five-number summary for the sample exercise-rate data is 121, 133, 140, 153, 164. Use the five-number summaries to construct box plots. Using a scale from 50 to 170 allows you to display both sets of data on the same axis.

Here are a few conclusions you can draw from the box plots. Try to come up with at least three more conclusions on your own.

- The minimum exercise rate is almost 30 beats per minute more than the maximum resting rate.
- Both sets of data are skewed right, with the rates above the median more spread out than the rates below the median.
- The median exercise rate is almost double the median resting rate.

Can you draw any conclusions about a larger population? If you collected data from students in your class, then the data might be representative of students your age. However, it’s possible that your class might have a larger percentage of athletes than the general population of students your age, or have other special characteristics that would make the data less representative. What other factors might influence whether you can generalize from your class data to a larger population?
Measures of Spread

In this lesson you will

- find measures of spread for a data set
- find and interpret the standard deviation of a data set

If you did Exercise 7 in Lesson 2.1, you may recall the following data. Here are the scores on assignments for two students, listed from low to high.

Connie’s scores: 82, 82, 84, 84, 85, 85, 86
Oscar’s scores: 72, 76, 76, 84, 90, 94, 96

Both data sets have a mean of 84 and a median of 84. However, the data sets are very different. In this lesson you will learn some ways to describe the variability, or spread, of a data set.

The box plots of the data show that Oscar’s scores are more spread out relative to the median, whereas Connie’s scores are clustered closer to the median.

The box plots show the spread of data relative to the median, but sometimes you may want to look at the spread of data relative to the mean. Read page 93 of your book to learn how to calculate deviations, which are the signed differences between data values and the mean.

In the investigation you'll explore ways to describe the variability of results of an experiment.

Investigation: A Good Design

In this investigation you will conduct several trials of an experiment. If your experiment is well designed and you perform it in a consistent way, you should get a similar result each time.

Conduct one of the two experiments described in your book. Here we will use sample data for the Rubber-Band Launch Experiment, but you should follow along using your own data.

Sample rubber-band data (cm): 182.2, 135.9, 187.6, 162.5, 150.0, 186.5, 180.0

Step 1 Calculate the mean distance for your trials and then calculate the deviations. For the sample data, the mean is 169.2 cm. To calculate the deviations, subtract the mean from each value in your data set. For the sample data, the deviations are 13.0, −33.3, 18.4, −6.7, −19.2, 17.3, and 10.8, respectively.

Step 2 These values give you an idea of how controlled the setup of your experiment was. If most of the differences are close to 0, it means you were able to perform the trials in a consistent way, obtaining a similar result each time.
Lesson 2.2 • Measures of Spread (continued)

Try to find a way to calculate a single value that tells how consistent your results are. One possibility is to find the average of the absolute values of the differences. For the sample data, this value is 17.0, indicating that, on average, each value is 17 cm from the mean.

**Step 3** The **standard deviation** is a measure of the spread of data away from the mean. Figure out how to use your calculator to find the standard deviation of a data set (see Calculator Note 2A). Then, enter the results of your experiment and calculate the standard deviation. For the sample data, the standard deviation is 20.2 cm.

**Step 4** Before reading the following text, try to find a formula or procedure for finding the standard deviation by hand. (*Hint: Standard deviation involves both squaring and taking a square root.*)

To calculate the standard deviation, follow these steps:

1. Square each deviation from Step 1. For the sample data, the results are 169, 1108.9, 338.56, 44.89, 368.64, 299.29, and 116.64.
2. Find the sum of the squared deviations. For the sample data, the sum is 2445.91.
3. Divide the sum of the squared deviations by one less than the number of data values. This value is called the **variance** of the data. For the sample data, the variance is 407.65.
4. Take the square root of the result from the previous step. The result is the standard deviation. For the sample data, the standard deviation is 20.2 cm.

In statistics, the mean is often referred to by the symbol $\bar{x}$, pronounced “x bar.” Another symbol, $\Sigma$ (capital sigma), is used to express the sum of a set of data values. For example, $\sum_{i=1}^{5} x_i$ means $x_1 + x_2 + x_3 + x_4 + x_5$, where $x_1$, $x_2$, $x_3$, $x_4$, and $x_5$ are the individual data values. This is called **sigma notation**, or **summation notation**. You can use these symbols to summarize the steps for finding standard deviation into a single formula:

$$ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} $$

where $x_i$ represents the individual data values, $n$ is the number of values, and $\bar{x}$ is the mean.

**Step 5** Repeat the experiment and collect another set of seven or eight trials. As you work, be as careful and consistent as possible. Calculate the standard deviation of the results. How does the standard deviation for the second set of trials compare with the standard deviation for the first set? Did you perform the trials in a more consistent way the second time?

Read the text on pages 95–97 in your book and make sure you understand it. Then practice finding the standard deviation and identifying outliers by working through the example in your book.
In this lesson you will

- construct and interpret histograms
- find the percentile rank of a data value
- apply all the statistics and graphs you have learned to analyze a set of data

A histogram shows how numerical data are distributed over different intervals, giving you a vivid picture of clusters and gaps in the data set. The columns of a histogram, called bins, indicate how many data values fall within a given interval.

The level of detail shown in a histogram depends on the bin width. Both of the histograms below show the backpack data from Lesson 2.1 of your book.

The bin width for the plot on the left is 6, and the bin width for the plot on the right is 3. Each bin includes the left endpoint value but not the right endpoint value. So, for example, the second bin of the left histogram includes backpacks that weigh 6 pounds but not backpacks that weigh 12 pounds. A 12-pound backpack would be counted in the third bin.

The right histogram gives more detail about the distribution than the left. For example, from the left histogram, you can see that there are four backpacks weighing less than 6 pounds. The right histogram shows that each of these four backpacks weighs at least 3 pounds.

**EXAMPLE**

Consider the histograms above.

a. What is the total number of backpacks represented by the histograms?

b. Describe how each histogram shows clusters and gaps in the data.

c. Use the left histogram to determine the interval that includes the median weight. Now, use the right histogram to determine the interval that includes the median. Which histogram gives you a more accurate estimate of the median?

d. What percentage of the backpacks weigh less than 9 pounds?
Lesson 2.3 • Histograms and Percentile Ranks (continued)

**Solution**

a. Use either histogram and add up the number of backpacks in each bin. Using the left histogram, you get $4 + 19 + 5 + 1 + 1 = 30$. So, 30 backpacks are represented. (Check this answer using the right histogram.)

b. The left histogram shows that the most common backpack weights are between 6 and 12 pounds, that all but two backpacks weigh less than 18 pounds, and that no backpack weighs between 24 and 30 pounds. The right histogram shows that the most common backpack weights are between 9 and 12 pounds, that all but seven of the backpacks weigh between 3 and 12 pounds, and that no backpacks weigh between 21 and 33 pounds.

c. There are a total of 30 values, so the median is between the 15th and 16th values. In the left histogram, the median value is in the second bin, which includes values greater than or equal to 6 pounds and strictly less than 12 pounds. This can be expressed as $6 \text{ lb} \leq \text{weight} < 12 \text{ lb}$. In the right histogram, the median is in the fourth bin, which includes values such that $9 \text{ lb} \leq \text{weight} < 12 \text{ lb}$. The right histogram gives the more accurate estimate.

d. Use the right histogram because it has a bin with an endpoint value of 9. Add the bin frequencies to the left of 9. You get $4 + 8 = 12$. So, $\frac{12}{30}$ or 40%, of the backpacks weigh less than 9 pounds.

Now, work through Example A in your book. Make sure you understand how to create a histogram on your calculator (see Calculator Note 2C).

The **percentile rank** of a data value gives the percentage of values that are below that value. In part d of the previous example, 40% of the backpacks weigh less than 9 pounds, so a backpack that weighs 9 pounds has a percentile rank of 40. Work through Example B in your book to practice working with percentiles and standard deviations.

**Investigation: Eating on the Run**

The investigation in your book asks you to conduct a statistical analysis about the nutritional value of fast-food items. Read the investigation carefully and then conduct your statistical analysis.

If you choose to analyze the saturated fat content of the food items, you might consider the following statistics and graphs.

<table>
<thead>
<tr>
<th>Saturated fat (g)</th>
<th>All sandwiches</th>
<th>Burgers</th>
<th>Chicken</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>≈ 11.9</td>
<td>≈ 17.6</td>
<td>≈ 6.2</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>8</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td>≈ 9.2</td>
<td>≈ 9.8</td>
<td>≈ 2.8</td>
</tr>
</tbody>
</table>