SUMMARY OF VIDEO

Statistics is all about data. It is easy to get overwhelmed by an avalanche of numbers if we don’t figure out good ways to organize it.

One of the best places to start is with a picture. You’ve seen charts similar to the ones in Figure 2.1 before – bar charts, pie charts, and dotplots – in this case, all ways to visualize the weight of newborn babies. Visualizing data like this can be a good first step toward organizing it and understanding it.

Figure 2.1. Graphic displays of weights of newborns.

In addition to the overall pattern displayed by the charts in Figure 2.1, the charts provide a framework to contextualize a particular baby’s birth weight relative to the rest of the data. In other words, the charts can help us decide whether the baby was small, in the middle of the pack, or large compared to the other babies.
There are a variety of ways to visualize data and many real world datasets to work with. Let’s step into the Army’s boots to see the data it collected to help outfit each and every soldier with the right size uniform and gear. Soldiers’ measurements have changed over the years – over time, soldiers’ sizes have both increased and become more variable.

To better assess the outfitting needs of the soldiers, the Army periodically embarks on a measurement project in which many measurements – foot length, shoulder width, head size, and so forth – are taken on a large random sample of soldiers. With a better sense of the most frequently-found dimensions, the Army knows which sizes of uniforms to keep well-stocked, and which sizes are rare enough that it’s cheaper to custom order them. As an illustration, here are the foot lengths (cm) of thirty soldiers:

27.2  26.9  26.6  
28.0  26.8  26.1  
26.2  27.3  27.6  
25.7  29.0  26.5  
32.8  28.8  26.9  
25.0  26.7  24.6  
26.3  26.8  27.0  
28.0  27.3  26.5  
27.4  25.0  26.6  
25.8  27.0  25.9

When you see a bunch of unorganized numbers, it is hard to determine whether or not there are any important patterns. But if we organize these numbers into a stemplot, we can get a better sense of how widely foot size varied. First, using technology or a calculator, we can sort the foot sizes in order from smallest to largest. The sorted data already give us a little better sense of soldiers’ foot sizes. The smallest is 24.6 centimeters and the largest is 32.8 centimeters.

Next, we separate each measurement into a stem (the first digits) and a leaf (the final digit). The stems are lined up vertically and then the leaves are filled in opposite the appropriate stems. Always include all possible stems in your data range, even those that don’t have leaves to go with them. The final step is to organize the leaves in numerical order. The result is the stemplot in Figure 2.2.
Figure 2.2. Stemplot of soldiers’ foot lengths.

Displayed as a stemplot, we can see the overall pattern of our data: 26- centimeter values are the most common, and values on either side of that single peak are less common. A stemplot also lets you see at a glance how spread out the distribution is. The data points range from the smallest at 24.6 centimeters to the largest at 32.8 centimeters. Check out the overall shape – it looks pretty symmetric, except for the value of 32.8. An individual measurement like this one that falls outside the overall pattern of the data is called an outlier.

Next, let’s consider another dataset where a stemplot can help us visualize the numbers – fuel economy information (city mpg) on Toyota’s 2012 vehicle line. The data have been organized into the stemplot in Figure 2.3. This time, the stems have been arranged from highest at the top to lowest at the bottom. (Note: the 5|1 at the top is for 51 mpg.)

Figure 2.3 Stemplot for 2012 Toyota’s city mpg.

Take a look at the overall pattern of the stemplot (Figure 2.3). Most of the mpgs are clustered at the lower end of the plot. We can expand the stem to change the resolution of the picture.
We break each stem into two, so the low digit leaves 0, 1, 2, 3, 4 are on a different stem than the high digit leaves 5, 6, 7, 8, 9. The expanded plot appears in Figure 2.4.

Figure 2.4. Stemplot with expanded stem.

Notice that we have outliers again, but this time an explanation is obvious. The high numbers are due to the super fuel-efficient hybrid vehicles Toyota makes.

Stemplots can be used to compare two different datasets as well. Say we wanted to compare Toyota’s 2012 numbers with those from their 1984 line. We can make a back-to-back stemplot to see how mileage numbers have changed over the decades.

Figure 2.5. Comparing Toyota’s 1984 line with its 2012 line.
What is interesting is that in 2012 Toyota had more vehicles way down at the low end, and a few more up at the high end. These extremes are easy to explain when you think about what you see on the roads – modern car buyers are interested in not-so-efficient SUVs and trucks as well as uber-efficient hybrids.

So you can see how stemplots help to tease meaning out of the disorder of raw data. They are useful for visualizing the shape of your data's distribution, and figuring out how frequently particular data classes pop up in your sea of numbers.

Later videos will show other ways to display data.
A. Be able to differentiate between measurement or count data and categorical data.

B. Understand that the distribution of a variable shows what values the variable takes on and how often.

C. When presented with unorganized raw data, begin by making a graphical display of the data.

D. Be able to construct a stemplot to display the distribution of a variable for small datasets.

E. Be able to describe a graphical display such as a stemplot by first describing the overall pattern and then deviations from that pattern. Be able to identify outliers as important deviations from the overall pattern.

F. In terms of the overall shape of a distribution, recognize when it is roughly symmetric and approximate the center of the distribution.
This unit on stemplots is the first in a series of units to focus on graphical representation of quantitative data. The emphasis in this unit is not simply to construct a stemplot but to use the plot to interpret the data’s story.

The clearest picture of the distribution of values of a variable is just that – a picture. A stemplot (or stem-and-leaf plot) is a simple kind of graph that is constructed using the numbers themselves. Here’s an example of head sizes in inches of 30 male soldiers. The head size was measured by putting a tape measure around each soldier’s forehead.

23.0 22.2 21.7 22.0 22.3 22.6
22.7 21.5 22.7 24.9 20.8 23.3
24.2 23.5 23.9 23.4 20.8 21.5
23.0 24.0 22.7 22.6 23.9 21.8
23.1 21.9 21.0 22.4 23.5 22.5

To make a stemplot of these measurements, we first separate each observation into a stem, which is the first digit or digits, and a leaf, the final digit. The stems can consist of any number of digits, but the leaves generally have only a single digit. For the head circumference data, the measurements range from about 21 inches to about 25 inches and are measured to tenths of an inch. We’ll take the whole inches as stems and the tenths as leaves.

First arrange the stems in order with a vertical line to their right as shown in Figure 2.6.

Figure 2.6. Setting up the stem of the stemplot.
Next, go through the list of observations, putting each leaf on the proper stem. The first soldier’s head size was 23.0 inches, so we put leaf 0 on stem 23. The second head size is 22.2, so we put leaf 2 on stem 22. When we are finished, we have the display in Figure 2.7.

| 20 | 88 |
| 21 | 755890 |
| 22 | 2036777645 |
| 23 | 035940915 |
| 24 | 920 |

*Figure 2.7. Stemplot with unordered leaves.*

As a final step, arrange the leaves in order from smallest to largest. Figure 2.8 shows the completed stemplot. (This final step is unnecessary if technology is used to order the data from smallest to largest.)

| 20 | 88 |
| 21 | 055789 |
| 22 | 0234566777 |
| 23 | 00134599 |
| 24 | 029 |

*Figure 2.8. The completed stemplot.*

If there are too many stems with no leaves or only one leaf, it often helps to truncate the numbers and then to make a stemplot of the truncated numbers. (Truncation is faster than rounding.) If the leaves are crowded onto too few stems, expand the stem. For example, each stem can be split into two, one for leaf digits 0, 1, 2, 3, 4 and the other for leaf digits 5, 6, 7, 8, 9. (Or split each stem into five, using leaf digits 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9 for the five stems.)

Splitting stems can often reveal new information, as was the case of the fuel economy of Toyota’s 2012 vehicles that was shown in Figures 2.3 and 2.4. Hence, don’t be afraid to experiment with different stems or truncation to see what additional information might be learned. Finally, placing stemplots back-to-back is a good way to compare two datasets. (See Figure 2.5.)
Making the stemplot isn’t the end in itself. It is a tool to help unlock the data’s story. For example, the completed stemplot in Figure 2.8 gives a picture of the distribution of soldiers’ head sizes. From the stemplot, we learn that the smallest head size was 20.8 and the largest was 24.9. The shape of the distribution is mound shaped (one peak). Although the two sides of the stemplot would not line up exactly if we folded the plot along the 22 stem, they come pretty close. So, we can say that this distribution is roughly symmetric. A middle value is somewhere in the 22-inch range (in other words, somewhere between 22.0 and 22.7).

The art of looking at stemplots intelligently is as important as the skill of making them. In looking at any distribution, always look first for the overall pattern of the distribution and then for any striking deviations from that pattern. In sizing up the overall pattern, look for and try to describe the following:

- center and spread
- one peak or several
- a regular shape, such as symmetric

For now, identify a center by looking at the stemplot and selecting a number that appears to best measure the middle of the distribution. (In later units, we will cover specific measures of center such as the mean and median).
KEY TERMS

A **variable** describes some characteristic of interest that can vary in value. Some variables are **categorical** (soldiers’ gender – male or female). Others are **quantitative** (soldiers’ head circumference or foot length).

The **distribution** of a variable describes the possible values the variable takes and how often it takes these values. Stemplots are one way to graph the distribution of a quantitative variable.

**Shape, center, and spread** describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **unimodal** (single peak) or **symmetric** (one side is the mirror image of the other).

**Outliers** are data values that lie outside the overall pattern of the distribution. Always look for gaps in the data and outliers and try to explain them.

A **stemplot** (or **stem-and-leaf plot**) is a useful tool for conveying the shape of relatively small datasets and identifying outliers. It consists of two columns, one for the stems and the other for the leaves (often separated by a vertical line).
Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. List some of the variables that were taken on soldiers for the sizing data bank.

2. What was the overall shape of the distribution of soldiers' foot lengths? About where was the center of the distribution?

3. What variable was used to measure fuel economy on Toyota's line of vehicles?

4. Focus on the stemplot of fuel economy for Toyota's 2012 line. What new information became evident (or more clear) when the stem was expanded?

5. What was learned from back-to-back stemplots about the change in fuel economy in Toyota's vehicle line from 1984 to 2012?
UNIT ACTIVITY:
USING STEMPLOTS TO ANALYZE DATA

Part I: Answer the questions on the following survey (or a survey distributed by your instructor). After completing the survey, hand it in to your instructor so that the data can be combined into a class dataset.

Survey Questionnaire

1. How long (in seconds) did you wait while your instructor was getting ready for this activity?
2. How much money in coins are you carrying with you right now?
3. To the nearest inch, how tall are you?
4. How long (in minutes) do you study, on average, for an exam?
5. On a typical day, how many minutes do you exercise?
6. Circle your gender: Male Female

Return your answers to your instructor.

Part II: Answer the following questions based on the class data.

1. Make stemplots for the data from questions 1 – 5. Describe the key features of each of your stemplots. (Don’t be afraid to experiment with expanding stems or truncating data values.)
2. How do estimates of waiting time compare for males and females? Make back-to-back stemplots of the data from question 1 for males and females.
3. Do males or females tend to carry more change? Make back-to-back stemplots of the data from question 2 for the males and females in this class. Analyze the results.
EXERCISES

1. The video mentioned the fact that soldiers are bulking up along with the rest of America. Even so, soldiers are expected to be physically fit. Suggest several quantitative variables that you might use to measure fitness. Keep in mind that each of your variables must produce a number that represents fitness.

2. Below are the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees, 1920 – 1934.

   54  59  35  41  46
   25  47  60  54  46
   49  46  41  34  22

   a. Make a stemplot of the home run data. Then use your stemplot to answer questions (b) and (c).
   
   b. Describe the shape of the distribution. Is it roughly symmetric or not? Is it unimodal (single peak) or multimodal (more than one peak)?
   
   c. What is the center (this is the number of home runs the Babe hit in a typical year)?
   
   d. Ruth’s record of 60 home runs in 1927 stood for more than 30 years. Is 60 an observation that falls outside the pattern of the other observations and hence could be considered an outlier?

3. The SAT is a standardized test for college admissions that is widely used in the United States. Table 2.1 contains the average score for each state on the Critical Reading, Math, and Writing sections of the SAT for 2010-11.

   (See table on next page...)
Table 2.1: Average SAT Scores by State 2010-2011.

a. Make a stemplot of the 51 average SAT Critical Reading scores.

b. Based on your stemplot in (a), describe the overall shape of the distribution of SAT Critical Reading scores. Approximate the center. Are there any outliers?

c. Make a back-to-back stemplot of the SAT Math and SAT Writing scores.

d. Based on your stemplot in (c) compare the distributions of the SAT Math and SAT Writing scores. Compare shape, center, and spread.

4. A local television station gathered data on the ages of viewers of ACTION, a program aimed at a young audience. The ages in years as reported by the rating service were as follows:

35.3  17.0  23.7  6.4  5.6  12.1  50.4  14.7  10.5
55.5  23.7  33.4  11.2  22.7  20.4  12.6  9.8  14.8
10.1  65.2  52.3  9.8  16.2  19.7  18.6  24.7  120.0
15.3  48.6  26.3  21.4  12.1  17.3  60.9  6.2  13.1
31.5  20.9  16.6  8.1  30.9  42.0  50.9  27.7

Make a stemplot of the age distribution. As a first step, truncate the number by discarding the digit after the decimal point. Then, describe the main features of the distribution.
Review Questions

1. Return to Table 2.1 with data from SAT scores (Exercise 3).

a. Make a stemplot of the percentage of graduates from each state who took the SAT. Use an expanded stem that shows increments of 5. (In other words, each stem value should be repeated twice.) Leave room on the left side to add a second stemplot.

b. Describe the overall shape of the distribution. Identify any gaps in the data and possible outliers. What does this tell you about the percentage of students who take the SATs from various states?

c. The percentage of students who took the SAT in each state in 1990 is given below. Transform your stemplot from (a) into a back-to-back stemplot so that the 1990 percentages can be compared to the 2010/2011 percentages. Describe the similarities and differences between the percentages for the two years.

<table>
<thead>
<tr>
<th>8</th>
<th>42</th>
<th>25</th>
<th>6</th>
<th>45</th>
<th>28</th>
<th>74</th>
<th>58</th>
<th>68</th>
<th>44</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>52</td>
<td>17</td>
<td>16</td>
<td>54</td>
<td>5</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>60</td>
</tr>
<tr>
<td>59</td>
<td>72</td>
<td>12</td>
<td>14</td>
<td>4</td>
<td>12</td>
<td>20</td>
<td>10</td>
<td>24</td>
<td>67</td>
</tr>
<tr>
<td>69</td>
<td>12</td>
<td>70</td>
<td>55</td>
<td>6</td>
<td>22</td>
<td>9</td>
<td>49</td>
<td>64</td>
<td>62</td>
</tr>
<tr>
<td>54</td>
<td>5</td>
<td>12</td>
<td>42</td>
<td>5</td>
<td>62</td>
<td>58</td>
<td>44</td>
<td>15</td>
<td>11</td>
</tr>
</tbody>
</table>

2. In determining the sizes of combat boots needed to outfit today’s soldiers, the Army measures the width of soldiers’ feet in addition to the length. Below are the width measurements (in millimeters) from a random sample of 40 male soldiers. Make a stemplot of these data. Expand the stem using increments of 2 (in other words, each stem value will be repeated 5 times). Based on your stemplot, what recommendations would you make to the Army about which boot widths need to be stocked and which widths are sufficiently rare that boots for that size foot should be specially ordered?

<table>
<thead>
<tr>
<th>103</th>
<th>97</th>
<th>110</th>
<th>90</th>
<th>111</th>
<th>105</th>
<th>110</th>
<th>95</th>
<th>91</th>
<th>97</th>
</tr>
</thead>
<tbody>
<tr>
<td>104</td>
<td>103</td>
<td>108</td>
<td>98</td>
<td>102</td>
<td>100</td>
<td>107</td>
<td>92</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>95</td>
<td>111</td>
<td>107</td>
<td>113</td>
<td>98</td>
<td>91</td>
<td>98</td>
<td>93</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>110</td>
<td>108</td>
<td>102</td>
<td>103</td>
<td>101</td>
<td>94</td>
<td>110</td>
<td>104</td>
<td>102</td>
<td>119</td>
</tr>
</tbody>
</table>

For question 3, you will need access to Stemplots from the Interactive Tools menu.
3. Samples of 20 6-year-old boys and 20 6-year-old girls were selected from children who were participating in the Infant Growth Study. One purpose of the study was to look at factors leading to childhood obesity. Table 2.2 gives the body mass index (BMI) for each of the children in the two samples.

<table>
<thead>
<tr>
<th>BMI, 6-Year-Old Boys</th>
<th>BMI, 6-Year-Old Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.8</td>
<td>26.9</td>
</tr>
<tr>
<td>14.0</td>
<td>14.7</td>
</tr>
<tr>
<td>17.5</td>
<td>25.2</td>
</tr>
<tr>
<td>13.2</td>
<td>15.8</td>
</tr>
<tr>
<td>13.5</td>
<td>15.1</td>
</tr>
<tr>
<td>13.7</td>
<td>14.4</td>
</tr>
<tr>
<td>14.4</td>
<td>15.8</td>
</tr>
<tr>
<td>15.5</td>
<td>14.2</td>
</tr>
<tr>
<td>14.5</td>
<td>13.4</td>
</tr>
<tr>
<td>24.5</td>
<td>15.3</td>
</tr>
<tr>
<td>15.4</td>
<td>14.0</td>
</tr>
<tr>
<td>14.8</td>
<td>14.4</td>
</tr>
<tr>
<td>20.6</td>
<td>14.7</td>
</tr>
<tr>
<td>17.6</td>
<td>16.3</td>
</tr>
<tr>
<td>17.3</td>
<td>14.9</td>
</tr>
<tr>
<td>15.7</td>
<td>15.9</td>
</tr>
<tr>
<td>16.0</td>
<td>15.9</td>
</tr>
<tr>
<td>15.5</td>
<td>16.5</td>
</tr>
<tr>
<td>16.3</td>
<td>14.1</td>
</tr>
<tr>
<td>15.7</td>
<td>15.5</td>
</tr>
</tbody>
</table>

*Table 2.2. BMI for 6-year-old boys and girls.*

a. Use the Stemplots tool from the Interactive Tools menu to construct back-to-back stemplots for boys’ and girls’ BMIs. Here’s how:

- Launch Stemplots and select calculation mode.
- Enter the boys’ BMIs into the box for Dataset #1.
- Click Back-to-back Datasets.
- Enter the girls’ BMIs into the box for Dataset #2.
- Click Generate Stemplot.

Copy the output from the Stemplots interactive tool.

b. Describe the data on boys’ BMI. Begin with a description of the overall pattern and then describe deviations from that pattern.
c. Repeat (b) for the girls’ data.

d. Compare the distribution of 6-year-old girls’ BMI to the distribution of 6-year-old boys’ BMI.
SUMMARY OF VIDEO

Many people are afraid of getting hit by lightning. And while getting hit by lightning is against the odds, it is not against all odds. Hundreds of people are struck by lightning every year in the U.S. What’s more, fires started by lightning strikes cause hundreds of millions of dollars of property damage. Meteorologist Raul Lopez and his associates began collecting detailed data on lightning strikes back in the 1980s and soon were overwhelmed by the vast amount of data. In one year, they collected three-quarters of a million flashes in a small area of Colorado. They decided to focus on when lightning strikes occurred. The data on the times of the first lightning strike needed to be organized, summarized, and displayed graphically. One of the statistical tools that Raul Lopez turned to was the graphic display called a histogram. For example, data on the percent of first lightning flashes for each hour of the day is displayed in the histogram in Figure 3.1.

Figure 3.1. Histogram of the time of the first lightning strike.
Before the histogram could be constructed, each day was broken into hours (horizontal axis), the number of first flashes in each hour was counted, and then the counts were converted to percentages (vertical axis). So, in this histogram, each bar represents one hour, and its height is the percentage of days in which the first lightning flash fell in that hour. This histogram has two very striking features. First, it is roughly symmetric about the tallest bar, which represents the percentage of first flashes between 11 a.m. and noon. The second rather surprising feature is how tightly the time of first strike clusters around the center bar, with a range from 10 a.m. to 1 p.m. accounting for most of the days’ first strikes. And there are no first strikes at night. This pattern helped explain how lightning storms form in this area. This region is mountainous and winds from the eastern plains carry warm moist air. When the wind hits the mountains it is forced upward where it meets and mixes with colder air higher in the atmosphere forming clouds. And this turns out to be a regular daily occurrence during the Colorado summer.

Lopez and his colleagues next looked at the time of day when the maximum number of lightning flashes occurred. (See Figure 3.2.) They found a similar pattern, with a peak showing that most flashes occur between 4 p.m. and 5 p.m.

![Figure 3.2. Histogram of the time of maximum flash rate.](image)

But there is one big difference from the first flash histogram in Figure 3.1. On a few days the maximum was in the early hours of the morning. Data points like these, which stand out from the overall pattern of the distribution, are called outliers. Outliers are often the most intriguing features of a histogram. Outliers should always be investigated and, if possible, explained.
The explanation that Lopez and his colleagues came up with was that they occur on days when larger weather systems, specifically very strong winds from fast moving weather fronts, overpower the local effect.

Data collection on Colorado lightning has continued since the pioneering work of Raul Lopez and his colleagues. Figure 3.3 shows a histogram produced from more recent data showing the number of people injured or killed by lightning strikes in the last 30 years. It shows the same clustering pattern as Raul Lopez’s histograms, but interestingly, the peak time for getting struck by lightning is around 2 p.m., about midway between the peaks of the first strike and maximum activity histograms.

![Figure 3.3. Histogram of time when people were struck by lightning.](image)

When constructing histograms it is very important to choose the best class size – that is, the choice of the interval widths for the horizontal axis. Lopez chose one hour for his data, and it works well. But suppose we turn our attention to a different context, the weekday traffic density on a portion of the Massachusetts Turnpike. First, we look at a histogram with class intervals of three hours. (See Figure 3.4.)
Figure 3.4. Histogram of traffic density in three-hour intervals.

The histogram in Figure 3.4 is not terribly informative. Next, we changed the interval width to one hour, which was better. However, using one-half hour widths as shown in Figure 3.5 is even better. Now, the increased traffic density during morning rush hour and evening rush hour is clearly visible in the pattern of two peaks.

Figure 3.5. Histogram of traffic density in half-hour intervals.
But what if we went even finer-grained and used 5-minute intervals? Take a look at Figure 3.6. Now the peaks begin disappearing again back into the numbers and the histogram becomes less informative.

![Figure 3.6. Histogram of traffic density in 5-minute intervals.](image)

Figure 3.6. Histogram of traffic density in 5-minute intervals.

So, we have seen how histograms can literally show at a glance the essence of a whole lot of numbers. Here is one last example. Figure 3.7 shows a histogram of the weekly wages of workers in the U.S. in the year 1992.

![Figure 3.7. Histogram of weekly wages (1992).](image)

Figure 3.7. Histogram of weekly wages (1992).

Notice how strikingly it is skewed, with most people earning around $450 per week. As you go out to what is called the tail of the distribution (to the right), the salaries get bigger, but the
percent of people earning those salaries gets smaller. Statisticians say a distribution like this is skewed to the right, because the right side of the histogram extends much further out than the left side. Now look at the histogram in Figure 3.8 of the same variable, weekly wages, but for the year 2011.

![Histogram of weekly wages (2011)](image)

*Figure 3.8. Histogram of weekly wages (2011).*

Now, the skew has become much more pronounced, and the tail has grown much longer. Suddenly our little discourse on histograms could become highly political!
A. Understand that the distribution of a variable consists of what values the variable takes and how often. (This is a repeat of an objective from Unit 2, Stemplots.)

B. Be able to construct a histogram to display the distribution of a variable for moderate amounts of data (say, data sets with fewer than 200 observations).

C. Understand that class intervals should be of equal width; choose appropriate class widths to effectively reveal informative patterns in the data.

D. Understand that the vertical axis of the histogram may be scaled for frequency, proportion, or percentage. The choice of vertical scaling for any data set does not affect the important features revealed by a histogram.

E. Be able to describe a graphical display of data by first describing the overall pattern and then deviations from that pattern. Describe the shape of the overall pattern and identify any gaps in data and potential outliers.

F. Recognize rough symmetry and clear skewness in the overall pattern of a distribution.
Rows and rows of data provide little information. For example, below are thickness measurements, in millimeters, from a sample of 25 polished wafers used in the manufacture of microchips. Notice that it is difficult to extract much information from staring at these numbers. The numbers need to be organized, summarized, and displayed graphically in order to unlock the information they contain.

0.402  0.496  0.533  0.387  0.384  
0.528  0.411  0.367  0.462  0.499  
0.539  0.546  0.425  0.457  0.586  
0.558  0.588  0.425  0.437  0.479  
0.427  0.485  0.443  0.441  0.658

A frequency distribution is one method of organizing and summarizing data in a table. The basic idea behind a frequency distribution is to set up categories (class intervals), classify data values into the categories, and then determine the frequency with which data values are placed into each category. The steps below outline the process of making a frequency distribution table.

**Creating a frequency distribution table**

**Step 1:** Identify an interval that is wide enough to contain all the data.

**Step 2:** Subdivide the interval identified in Step 1 into class intervals of equal width. The class intervals will serve as the categories.

**Step 3:** Set up a table with three columns for the following: class interval, tally, and frequency. (The tally column can be removed in the final table.)

**Step 4:** To complete the table, determine the frequency with which data values fall into each class interval.

Convention: Any data value that falls on a class interval boundary is placed in the class interval to the right. If the data value is a maximum, it is generally put in the interval that contains the maximum at its right endpoint.
Now, we apply Steps 1 – 4 to make a frequency distribution table for the thickness measurements.

**Step 1:** In this case the smallest data value is 0.367 mm and the largest is 0.698 mm. We choose the interval from 0.3 mm to 0.7 mm, which contains all the thickness measurements.

**Step 2:** The total width of the interval from 0.3 to 0.7 is 0.4. Dividing this interval into eight class intervals works out nicely – each class interval will have width 0.05.

**Step 3:** We have set up Table 3.1 to have three columns, which we have labeled Thickness, Tally, and Frequency. We have entered the endpoints of the eight class intervals into the Thickness column.

<table>
<thead>
<tr>
<th>Thickness (mm)</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.30 – 0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.35 – 0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.40 – 0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.45 – 0.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50 – 0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55 – 0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.60 – 0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.65 – 0.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 3.1: Setting up a frequency distribution table.*

**Step 4:** The easiest way to determine the frequencies is to draw a tally line for each data value that falls into a particular class interval. When drawing tally lines, keep the following in mind:

- As you draw tally lines, instead of drawing a fifth tally line, cross out the previous four.
- If a data value falls on the boundary of a class interval, record it in the interval with the larger values.

Once a tally line has been drawn for each data value, count the number of tally lines corresponding to each class interval and record that number in the frequency column as shown in Table 3.2.
Table 3.2: A completed frequency distribution table.

The frequency distribution in Table 3.2 reveals more information about the data than a quick look at the 25 numbers. For example, from the frequency distribution, we learn that more measurements fall in the interval 0.40 – 0.45 than in any of the other class intervals. Also, we learn there is a gap in the data – no data values fall between 0.60 and 0.65.

Although a frequency distribution table is a useful tool for extracting information from data, a histogram can often convey the same information more effectively. Next, we outline how to construct a histogram from a frequency distribution.

### Creating a histogram from a frequency distribution

**Step 1:** Draw a set of axes. On the horizontal axis, mark the boundaries of the class intervals. On the vertical axis, set up a scale appropriate for the frequencies. (Later this scale can be changed to proportion or percent.)

**Step 2:** Label the horizontal axis with the name of the variable being measured and the units.

**Step 3:** Over each interval, draw a rectangle with the interval as its base. The height of the rectangle should match the frequency of data contained in that interval.

Next, we apply Steps 1 – 3 for creating a histogram to the frequency distribution in Table 3.2. Figure 3.9 shows the results.
Particularly if you are comparing histograms from samples with a different number of data values, it is useful to replace the frequency scale on the vertical axis with the proportion or percent.

**Calculating Proportions and Percents**

- To calculate a proportion, divide the frequency by the sample size.
- To convert a proportion into a percent, multiply the proportion by 100%.

In describing a histogram, we first look for the overall pattern of the distribution. In sizing up the overall pattern, look for the following:

  - center and spread;
  - one peak or several (unimodal or multimodal);
  - a regular shape, such as symmetric or skewed.

In the case of the histogram in Figure 3.9, the overall pattern is single-peaked (or unimodal) and skewed to the right. Next, we look for any striking deviations from that pattern. An important kind of deviation from an overall pattern is an outlier, an individual observation.
that lies clearly outside the overall pattern. Once identified, outliers should be investigated. Sometimes they are errors in the data and sometimes they have interesting stories related to the data. For Figure 3.9, there is a gap between 0.6 and 0.65 and there is one data value between 0.65 and 0.70, which might be an outlier.
A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.

*Figure 3.10. Shapes of histograms.*
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. The video opens by describing a study of lightning strikes in Colorado. What variable does the first histogram display?

2. In this lightning histogram, what does the horizontal scale represent? What does the vertical scale represent?

3. Was the overall shape of this histogram symmetric, skewed, or neither?

4. Why were a few values in the second lightning histogram called outliers?

5. When you choose the classes for a histogram, what property must the classes have if the histogram is to be correct?

6. What happens to a histogram if you use too many classes? What happens if you use too few?
UNIT ACTIVITY:
WAFER THICKNESS

What do automobiles, singing Barbie dolls, cell phones and computers have in common? To a worker in the semiconductor industry, the answer is obvious – they all use microchips, tiny electronic circuits etched on chips of silicon (or some other semiconductor material).

Manufacturing microchips is a complex process. It begins with cylinders of silicon, called ingots, which are 6 to 16 inches in diameter. The ingots are sliced into thin wafers, which are then polished. (See Figure 3.11.) The polished wafers are imprinted with microscopic patterns of circuits, which are etched out with acids and replaced with conductors (such as aluminum or copper). Once completed, the wafers are cut into individual chips. (See Figure 3.12.)

In order to remain competitive in a global market, American companies must process microchips correctly and repeatedly with almost perfect consistency. The only way to accomplish this is to measure and control all of the highly complex processes used to manufacture microchips. These companies rely on statistical techniques to ensure quality control at critical points in the processing. It is simply too costly to wait until the end and then reject defective chips.

Figure 3.11: Silicon ingots and polished wafers.

Figure 3.12: The grid pattern shows individual microchips on a wafer. (Credit: Peellden)
One critical stage in the manufacture of microchips is the grinding and polishing processes used to produce polished wafers. The wafers need to be consistent in thickness, not warped or bowed, and free of surface imperfections. The focus in this activity will be on adjusting controls in order to produce polished wafers that are consistently close to 0.5 mm in thickness.

The Wafer Thickness tool found in the Interactive Tools menu allows you to set three controls that adjust the grinding and polishing processes. Each control has three levels. After setting the controls, you can take a sample of polished wafers and measure their thicknesses.

1. Set all three controls to 1. Select a sample size of 10 and select Real Time mode. Then press the "Collect Sample Data" button. Watch as the sample wafers are measured. A graphic display (called a histogram) is formed in real time as data become available.

   a. Describe what happens to the graphic display each time a new data value is added to the table. In other words, how is the histogram constructed?

   b. Describe the shape or features of the histogram. Here are some questions to consider when describing the shape of the data:

      • Is the histogram roughly symmetric about some center?

      • Is there one interval that contains more data than other intervals?

      • Are there any gaps between bars? In other words, are there intervals that do not contain data?

      • In what interval did the smallest data value fall? In what interval did the largest data value fall?

      • If you had to summarize the location (or “center”) of these data with one number, what number would you choose? How did you choose this number?

      • Do you think the controls are properly set to produce wafers of consistent 0.5 mm thickness?

2. a. Would another sample of 10 wafers manufactured under the same control settings as your first sample behave exactly as the first sample? To find out leave all settings as they were in Question 1 and click the “Collect Sample Data” button.

   b. Answer Question 1b for the new sample.
3. Leave all settings as they are. Click the “Jump To Results” button. The sample size is now set at 25. Collect two more samples by clicking the “Collect Sample Data” button twice. Then click the “Compare to Previous” button. What characteristics do the two histograms have in common? How do the two histograms differ?

The manufacturer wants wafers that are 0.5 mm thick. However, it is not possible to grind and polish wafers so that every wafer has a thickness of exactly 0.5 mm. There will always be some variability in thickness. Hence, the problem is to determine the control settings that produce wafers that are consistently close to 0.5 mm in thickness. For the remainder of this activity, use the “Jump To Results” mode for data collection and select 50 for the sample size. (You could use sample size 25, but you may find that using the larger sample size gives better results.)

4. Your first task is to determine how each control affects the thickness of a sample of wafers. In other words, you should answer the following questions:

   • How do the settings of Control 1 affect wafer thickness?
   • How do the settings of Control 2 affect wafer thickness?
   • How do the settings of Control 3 affect wafer thickness?

a. You will need to be systematic in how you change the controls so that you can determine how each control affects wafer thickness. Describe the strategy you will use to collect data that will allow you to answer the question about the controls. (You may need to collect more than one sample from each set of control settings before you are able to see changes in the data.)

b. Carry out the strategy you have outlined in (a). Describe what affect Controls 1, 2, and 3 have on wafer thickness. Print (or draw) some histograms that support your conclusions.

5. What control settings would you recommend in order to produce wafers that are consistently close to 0.5 mm in thickness?

   Explain why you chose the settings that you did.
Table 3.3 is needed for Exercises 1 – 3.

<table>
<thead>
<tr>
<th>State</th>
<th>Total</th>
<th>65 and older</th>
<th>Percent 65 and older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>4,780</td>
<td>658</td>
<td>13.80%</td>
</tr>
<tr>
<td>Alaska</td>
<td>710</td>
<td>55</td>
<td>7.70%</td>
</tr>
<tr>
<td>Arizona</td>
<td>6,392</td>
<td>882</td>
<td>13.80%</td>
</tr>
<tr>
<td>Arkansas</td>
<td>2,916</td>
<td>420</td>
<td>14.40%</td>
</tr>
<tr>
<td>California</td>
<td>37,254</td>
<td>4,247</td>
<td>11.40%</td>
</tr>
<tr>
<td>Colorado</td>
<td>5,029</td>
<td>550</td>
<td>10.90%</td>
</tr>
<tr>
<td>Connecticut</td>
<td>3,574</td>
<td>507</td>
<td>14.20%</td>
</tr>
<tr>
<td>Delaware</td>
<td>898</td>
<td>129</td>
<td>14.40%</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>602</td>
<td>69</td>
<td>11.50%</td>
</tr>
<tr>
<td>Florida</td>
<td>18,801</td>
<td>3,260</td>
<td>17.30%</td>
</tr>
<tr>
<td>Georgia</td>
<td>9,688</td>
<td>1,032</td>
<td>10.70%</td>
</tr>
<tr>
<td>Hawaii</td>
<td>1,360</td>
<td>195</td>
<td>14.30%</td>
</tr>
<tr>
<td>Idaho</td>
<td>1,568</td>
<td>195</td>
<td>12.40%</td>
</tr>
<tr>
<td>Illinois</td>
<td>12,831</td>
<td>1,609</td>
<td>12.50%</td>
</tr>
<tr>
<td>Indiana</td>
<td>6,484</td>
<td>841</td>
<td>13.00%</td>
</tr>
<tr>
<td>Iowa</td>
<td>3,046</td>
<td>453</td>
<td>14.90%</td>
</tr>
<tr>
<td>Kansas</td>
<td>2,853</td>
<td>376</td>
<td>13.20%</td>
</tr>
<tr>
<td>Kentucky</td>
<td>4,339</td>
<td>578</td>
<td>13.30%</td>
</tr>
<tr>
<td>Louisiana</td>
<td>4,533</td>
<td>558</td>
<td>12.30%</td>
</tr>
<tr>
<td>Maine</td>
<td>1,328</td>
<td>211</td>
<td>15.90%</td>
</tr>
<tr>
<td>Maryland</td>
<td>5,774</td>
<td>708</td>
<td>12.30%</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>6,548</td>
<td>903</td>
<td>13.80%</td>
</tr>
<tr>
<td>Michigan</td>
<td>9,884</td>
<td>1,362</td>
<td>13.80%</td>
</tr>
<tr>
<td>Minnesota</td>
<td>5,304</td>
<td>683</td>
<td>12.90%</td>
</tr>
<tr>
<td>Mississippi</td>
<td>2,967</td>
<td>380</td>
<td>12.80%</td>
</tr>
<tr>
<td>Missouri</td>
<td>5,989</td>
<td>838</td>
<td>14.00%</td>
</tr>
<tr>
<td>Montana</td>
<td>989</td>
<td>147</td>
<td>14.90%</td>
</tr>
<tr>
<td>Nebraska</td>
<td>1,826</td>
<td>247</td>
<td>13.50%</td>
</tr>
<tr>
<td>Nevada</td>
<td>2,701</td>
<td>324</td>
<td>12.00%</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>1,316</td>
<td>178</td>
<td>13.50%</td>
</tr>
<tr>
<td>New Jersey</td>
<td>8,792</td>
<td>1,186</td>
<td>13.50%</td>
</tr>
<tr>
<td>New Mexico</td>
<td>2,059</td>
<td>272</td>
<td>13.20%</td>
</tr>
<tr>
<td>New York</td>
<td>19,378</td>
<td>2,618</td>
<td>13.50%</td>
</tr>
<tr>
<td>North Carolina</td>
<td>9,535</td>
<td>1,234</td>
<td>12.90%</td>
</tr>
<tr>
<td>North Dakota</td>
<td>673</td>
<td>97</td>
<td>14.40%</td>
</tr>
<tr>
<td>Ohio</td>
<td>11,537</td>
<td>1,622</td>
<td>14.10%</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>3,751</td>
<td>507</td>
<td>13.50%</td>
</tr>
<tr>
<td>Oregon</td>
<td>3,831</td>
<td>534</td>
<td>13.90%</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12,702</td>
<td>1,959</td>
<td>15.40%</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>1,053</td>
<td>152</td>
<td>14.40%</td>
</tr>
<tr>
<td>South Carolina</td>
<td>4,625</td>
<td>362</td>
<td>7.80%</td>
</tr>
<tr>
<td>South Dakota</td>
<td>814</td>
<td>117</td>
<td>14.40%</td>
</tr>
<tr>
<td>Tennessee</td>
<td>6,346</td>
<td>853</td>
<td>13.40%</td>
</tr>
<tr>
<td>Texas</td>
<td>25,146</td>
<td>2,602</td>
<td>10.30%</td>
</tr>
<tr>
<td>Utah</td>
<td>2,764</td>
<td>249</td>
<td>9.00%</td>
</tr>
<tr>
<td>Vermont</td>
<td>626</td>
<td>91</td>
<td>14.50%</td>
</tr>
<tr>
<td>Washington</td>
<td>6,725</td>
<td>828</td>
<td>12.30%</td>
</tr>
<tr>
<td>Virginia</td>
<td>8,001</td>
<td>977</td>
<td>12.20%</td>
</tr>
<tr>
<td>West Virginia</td>
<td>1,853</td>
<td>297</td>
<td>16.00%</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>5,687</td>
<td>777</td>
<td>13.70%</td>
</tr>
<tr>
<td>Wyoming</td>
<td>564</td>
<td>70</td>
<td>12.40%</td>
</tr>
</tbody>
</table>

Table 3.3. Count (in Thousands) of people over 65 by State and the District of Columbia in 2010.

1. How many people in your state are at least 65 years old? The answer varies from state to state. Table 3.3 gives the data for all 50 states and the District of Columbia for the year 2010.

a. Make a histogram for these data. Use class intervals of width 500,000.

b. Darken the bar in which your state’s data value would fall. Does your state tend to have more or fewer residents 65 and older than the other states, or would you say that your state is close to typical?

c. Describe the overall shape of the distribution of age 65 and older. Identify any gaps in the distribution and potential outliers.

d. Redraw the histogram this time using class intervals of 1,000 thousand. What information is now hidden using this size of class intervals?
2. You would expect highly populated states to have higher numbers of residents over 65 than less populated states. But would the percentage of people 65 and over still be higher?

   a. Make a histogram of the percentage of people over 65 in each state. Choose interval widths of 1%. Darken the bar in which your state’s percentage would fall. Does your state tend to have a higher or lower percentage of residents 65 and older than the other states, or would you say that your state is close to typical?

   b. Describe the overall shape of the distribution of percentages. Then identify any gaps in the distribution and potential outliers.

3. Finally, we consider the total population of the states.

   a. Make a histogram of the total population of the states. Choose a class interval width that shows key features of the distribution.

   b. Write a brief description of the most important features of the distribution of total number of state residents. Is the distribution roughly symmetric, clearly skewed, or neither? What states are unusual in their population sizes?

4. In a laboratory experiment, students were asked to estimate the breaking strength of wooden stakes. The dimensions of the stakes, measured in inches, were 8 × 1.5 × 1.5. From the experiment students found the load in pounds needed to break the stakes in a sample of 20 stakes. The class data, measurements of the breaking strength in hundreds of pounds, appear below.

   166  161  115  120  159  
   165  155  151  163  160  
   156  164  118  152  168  
   144  166  164  161  160

   a. Even though the wooden stakes were nearly identical, did the breaking strengths vary? Explain.

   b. Make a histogram of these data. Use class intervals of width 5.

   c. Which class interval(s) contained the most data?
d. Modify your histogram in (b) so that the scale on the vertical axis is the percent of the stakes whose breaking strength is in each class interval. How does the shape of your modified histogram compare to your histogram in (b)?

e. Write a short paragraph describing key features of the distribution of breaking strengths.
Table 3.4, needed for questions 1 and 2, consists of a list of the top 100 major league baseball players ranked according to career batting average. (Notice that because of ties for the 100th place, there are actually 104 players on this list.)

The table contains the following information for each player: number of career years, the last year for which data were collected, career batting average, and career number of home runs.

*(See full table on next page...)*
<table>
<thead>
<tr>
<th>First</th>
<th>Last</th>
<th>Career Years</th>
<th>Last Career Year</th>
<th>Career Batting Avg</th>
<th>Career Home Runs</th>
<th>First</th>
<th>Last</th>
<th>Career Years</th>
<th>Last Career Year</th>
<th>Career Batting Avg</th>
<th>Career Home Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cap Anson</td>
<td>1897</td>
<td>0.331</td>
<td>97</td>
<td>Willie Keeler</td>
<td>1901</td>
<td>0.341</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luke Appling</td>
<td>1950</td>
<td>0.310</td>
<td>45</td>
<td>Joe Kelly</td>
<td>1908</td>
<td>0.317</td>
<td>65</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earl Averill</td>
<td>1941</td>
<td>0.318</td>
<td>238</td>
<td>Chuck Klein</td>
<td>1944</td>
<td>0.320</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ginger Beaumont</td>
<td>1910</td>
<td>0.311</td>
<td>39</td>
<td>Nap Lajoie</td>
<td>1916</td>
<td>0.339</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wade Boggs</td>
<td>1999</td>
<td>0.328</td>
<td>118</td>
<td>Henry Larkin</td>
<td>1893</td>
<td>0.310</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jim Bottomley</td>
<td>1937</td>
<td>0.310</td>
<td>219</td>
<td>Freddie Lindstrom</td>
<td>1936</td>
<td>0.311</td>
<td>103</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dan Brouthers</td>
<td>1904</td>
<td>0.342</td>
<td>106</td>
<td>Denny Lyons</td>
<td>1897</td>
<td>0.318</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pete Browning</td>
<td>1894</td>
<td>0.349</td>
<td>46</td>
<td>Heinie Manush</td>
<td>1939</td>
<td>0.330</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jessie Burkett</td>
<td>1905</td>
<td>0.338</td>
<td>75</td>
<td>Edgar Martinez</td>
<td>2004</td>
<td>0.312</td>
<td>309</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miguel Cabrera</td>
<td>2012</td>
<td>0.318</td>
<td>312</td>
<td>Joe Mauer</td>
<td>2012</td>
<td>0.322</td>
<td>93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rod Carew</td>
<td>1985</td>
<td>0.326</td>
<td>92</td>
<td>Barney McCosky</td>
<td>1953</td>
<td>0.312</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fred Clarke</td>
<td>1915</td>
<td>0.312</td>
<td>67</td>
<td>John McGraw</td>
<td>1906</td>
<td>0.334</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roberto Clemente</td>
<td>1972</td>
<td>0.317</td>
<td>240</td>
<td>Joe Medwick</td>
<td>1948</td>
<td>0.324</td>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ty Cobb</td>
<td>1929</td>
<td>0.366</td>
<td>117</td>
<td>Irish Meusel</td>
<td>1927</td>
<td>0.310</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mickey Cochrane</td>
<td>1937</td>
<td>0.320</td>
<td>119</td>
<td>Bing Miller</td>
<td>1936</td>
<td>0.311</td>
<td>116</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eddie Collins</td>
<td>1930</td>
<td>0.333</td>
<td>47</td>
<td>Dale Mitchell</td>
<td>1956</td>
<td>0.312</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earle Combs</td>
<td>1935</td>
<td>0.325</td>
<td>58</td>
<td>Johnny Mize</td>
<td>1953</td>
<td>0.312</td>
<td>359</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roger Connor</td>
<td>1897</td>
<td>0.317</td>
<td>138</td>
<td>Stan Musial</td>
<td>1963</td>
<td>0.331</td>
<td>475</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiki Cuyler</td>
<td>1938</td>
<td>0.321</td>
<td>128</td>
<td>Tip O'Neill</td>
<td>1892</td>
<td>0.334</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ed Delahanty</td>
<td>1903</td>
<td>0.346</td>
<td>101</td>
<td>Jim O'Rourke</td>
<td>1904</td>
<td>0.310</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bill Dickey</td>
<td>1946</td>
<td>0.313</td>
<td>202</td>
<td>Kirby Puckett</td>
<td>1995</td>
<td>0.318</td>
<td>207</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joe DiMaggio</td>
<td>1951</td>
<td>0.325</td>
<td>361</td>
<td>Albert Pujols</td>
<td>2012</td>
<td>0.325</td>
<td>474</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mike Donlin</td>
<td>1914</td>
<td>0.333</td>
<td>51</td>
<td>Rip Radcliff</td>
<td>1943</td>
<td>0.311</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hugh Duffy</td>
<td>1906</td>
<td>0.324</td>
<td>106</td>
<td>Manny Ramirez</td>
<td>2011</td>
<td>0.312</td>
<td>555</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bibb Falk</td>
<td>1931</td>
<td>0.314</td>
<td>69</td>
<td>Sam Rice</td>
<td>1934</td>
<td>0.322</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elmer Flick</td>
<td>1910</td>
<td>0.313</td>
<td>48</td>
<td>Jackie Robinson</td>
<td>1956</td>
<td>0.311</td>
<td>137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob Fothergill</td>
<td>1933</td>
<td>0.325</td>
<td>36</td>
<td>Edd Roush</td>
<td>1931</td>
<td>0.323</td>
<td>68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jack Fournier</td>
<td>1927</td>
<td>0.327</td>
<td>136</td>
<td>Babe Ruth</td>
<td>1935</td>
<td>0.342</td>
<td>714</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jimmie Foxx</td>
<td>1945</td>
<td>0.325</td>
<td>534</td>
<td>Joe Sewell</td>
<td>1933</td>
<td>0.312</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frankie Frisch</td>
<td>1937</td>
<td>0.316</td>
<td>105</td>
<td>Al Simmons</td>
<td>1944</td>
<td>0.334</td>
<td>307</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nomar Garciaparra</td>
<td>2009</td>
<td>0.313</td>
<td>229</td>
<td>George Sisler</td>
<td>1930</td>
<td>0.340</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lou Gehrig</td>
<td>1937</td>
<td>0.340</td>
<td>493</td>
<td>Elmer Smith</td>
<td>1901</td>
<td>0.312</td>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charlie Gehringer</td>
<td>1942</td>
<td>0.320</td>
<td>184</td>
<td>Tris Speaker</td>
<td>1928</td>
<td>0.345</td>
<td>117</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goose Goslin</td>
<td>1935</td>
<td>0.316</td>
<td>248</td>
<td>Riggs Stephenson</td>
<td>1934</td>
<td>0.336</td>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hank Greenberg</td>
<td>1947</td>
<td>0.314</td>
<td>331</td>
<td>Ichiro Suzuki</td>
<td>2012</td>
<td>0.322</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vladimir Guerrero</td>
<td>2011</td>
<td>0.318</td>
<td>449</td>
<td>Bill Terry</td>
<td>1936</td>
<td>0.341</td>
<td>154</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tony Gwynn</td>
<td>2001</td>
<td>0.338</td>
<td>135</td>
<td>Sam Thompson</td>
<td>1906</td>
<td>0.331</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chick Hafey</td>
<td>1937</td>
<td>0.317</td>
<td>164</td>
<td>Mike Tieman</td>
<td>1899</td>
<td>0.311</td>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Billy Hamilton</td>
<td>1901</td>
<td>0.344</td>
<td>40</td>
<td>Cecil Travis</td>
<td>1947</td>
<td>0.314</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harry Hellmann</td>
<td>1932</td>
<td>0.342</td>
<td>183</td>
<td>Pie Traynor</td>
<td>1937</td>
<td>0.320</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Todd Helton</td>
<td>1912</td>
<td>0.320</td>
<td>354</td>
<td>George Van Haltren</td>
<td>1903</td>
<td>0.316</td>
<td>69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Babe Herman</td>
<td>1945</td>
<td>0.324</td>
<td>181</td>
<td>Arky Vaughan</td>
<td>1948</td>
<td>0.318</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matt Hollday</td>
<td>1912</td>
<td>0.314</td>
<td>227</td>
<td>Bobby Veach</td>
<td>1925</td>
<td>0.310</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rogers Hornsby</td>
<td>1937</td>
<td>0.359</td>
<td>301</td>
<td>Honus Wagner</td>
<td>1917</td>
<td>0.327</td>
<td>101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baby Doll</td>
<td>1927</td>
<td>0.311</td>
<td>83</td>
<td>Larry Walker</td>
<td>2005</td>
<td>0.313</td>
<td>383</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joe Jackson</td>
<td>1920</td>
<td>0.356</td>
<td>54</td>
<td>Paul Waner</td>
<td>1945</td>
<td>0.333</td>
<td>113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hughie Jennings</td>
<td>1918</td>
<td>0.311</td>
<td>18</td>
<td>Lloyd Waner</td>
<td>1945</td>
<td>0.316</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Derek Jeter</td>
<td>2005</td>
<td>0.313</td>
<td>254</td>
<td>Zack Wheat</td>
<td>1927</td>
<td>0.317</td>
<td>132</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Willie Keeler</td>
<td>1910</td>
<td>0.341</td>
<td>33</td>
<td>Ted Williams</td>
<td>1960</td>
<td>0.344</td>
<td>521</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joe Kelly</td>
<td>1908</td>
<td>0.317</td>
<td>65</td>
<td>Ken Williams</td>
<td>1929</td>
<td>0.319</td>
<td>196</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chuck Klein</td>
<td>1944</td>
<td>0.320</td>
<td>300</td>
<td>Tuffy Wright</td>
<td>1949</td>
<td>0.311</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nap Lajoie</td>
<td>1916</td>
<td>0.339</td>
<td>83</td>
<td>Ross Youngs</td>
<td>1926</td>
<td>0.322</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.4: Top 100 career batting averages in baseball (at the end of the 2012 season).
1. Make two histograms for the career home run data. For the first histogram, use class intervals of size 100 and in the second, use class intervals of size 50. Describe the overall shape of the data based on each of your histograms. Also identify any potential outliers. Explain what new information you can obtain from the second histogram that was not visible in the first.

2. a. Make two histograms for Career Years. Use the following class intervals:


b. Did any of the career years fall on a boundary of a class interval? If so, how did you classify those data values?

c. Describe the overall shape of each of the two histograms. In particular would you describe the shape as symmetric or skewed? Would you characterize the shape as unimodal (one peak), bimodal (two peaks), or multimodal? Did changing the class intervals affect the shape of the distribution?

3. The duration of 40 phone calls (in minutes) for technical support is given below.

   12.0  3.3  0.5  48.7  16.7  1.2  14.8  8.2  9.0  5.7
   11.5  17.5  3.2  20.8  7.3  8.0  0.2  51.2  3.3  5.2
   12.3  24.5  13.3  7.7  13.5  4.3  13.7  10.7  18.8  15.7
   3.2  38.7  16.2  23.3  9.7  4.7  6.5  0.5  45.1  5.3

a. Make a copy of Table 3.5 and then complete the frequency distribution table for the call duration data.

(See table on next page...
Table 3.5. Frequency distribution table for duration of phone calls.

<table>
<thead>
<tr>
<th>Duration (minutes)</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 – 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 – 18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 – 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 – 36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36 – 42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42 – 48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48 – 54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. What percentage of phone calls lasted less than 12 minutes?

c. What percentage of calls lasted a half hour or more?

d. Represent the frequency distribution with a histogram. Use a percent scale on the vertical axis.

e. Describe the shape of the distribution. Are there any gaps in the data? Outliers?
SUMMARY OF VIDEO

One number most people pay a lot of attention to is the one on their paycheck! Today’s workforce is pretty evenly split between men and women, but is the salary distribution for women the same as for men? The histograms in Figure 4.1 show the weekly wages for Americans in 2011, separated by gender.

Figure 4.1. Histograms comparing men’s and women’s wages.

Both histograms are skewed to the right with most people making moderate salaries while a few make much more. For comparison’s sake, it would help to numerically describe the centers of these distributions. A statistic called the median splits the distribution in half as shown in Figure 4.2 – half the wages lie above it, and half below. The median wage for men in 2011 was $865. The median wage for women was only $692, about 80% of what men make.
Simply using the median, we have identified a real disparity in wages, but it is much harder to figure out why it exists. Some of the difference can be accounted for by differences in education, age, and years in the workforce. Another reason for the earnings gap is that women tend to be concentrated in lower-paying jobs – but that begs the question: Are these jobs worth less in some sense? Or are these jobs lower paid because they are primarily held by women? This is the central issue in the debate over comparable worth – the idea that men and women should be paid equally, not only for the same job but for different jobs of equal worth.

Back in 1988 the city of Colorado Springs, Colorado, was at the forefront of this debate. As part of its normal operation, the city government evaluated all municipal jobs with criteria like working conditions, necessary skills, and accountability required. Each job got a numerical rank. It turned out that many clerical jobs, which are mostly filled by women, scored the same number of points as operations and maintenance jobs, which are mostly filled by men. However, the median wage for men’s jobs was always higher than the corresponding median wage for the women even though these jobs were judged to be exactly equal in requirements and responsibility. A group of clerical workers used this evidence to pressure the city for a more equitable pay structure. The numbers were hard to argue with and the clerical workers won. The city agreed to equalize the median salaries for jobs of comparable worth. And the plan had a benefit for the city as well – the relatively high turnover rate for jobs held by women decreased.
Colorado Springs relied on the median statistic to identify the inequality in men's and women's salaries. Next, we take a look at how to calculate this measure of center. Below are the weekly salaries from a small hypothetical company that has 19 employees. The salaries have been arranged in order from the lowest, $290 for an entry-level receptionist, up to the highest, $2,000 for the president.

290 350 400 400 450 450 450 500 500 500
550 550 650 750 800 1200 1300 1500 2000

The median represents a typical wage. To calculate the median, determine the number of observations, \( n \). In this case, we have 19 salaries and so, \( n = 19 \). The location of the median is at \( (n + 1)/2 \), or \( (19 + 1)/2 = 10 \). Count up 10 spots from the bottom (or down 10 spots from the top) and read off the median: $500. Since we had an odd number of paychecks, it is easy to count up 10 places to the middle number. But what if we had an even number of observations to deal with? Suppose we add a paycheck of $550 as shown below.

290 350 400 400 450 450 450 500 500 500
550 550 550 650 750 800 1200 1300 1500 2000

With the additional paycheck, \( n = 20 \). Now, we count up \( (20 + 1)/2 \), or 10.5 spaces. That puts us right in between the two middle values of 500 and 550. So, the median is actually halfway between those two salaries, $525.

The median is not the only measure for center. Another way to measure the center of a distribution of values is by taking the average. Statisticians call this number the mean, which is denoted by \( \bar{x} \). It is calculated by adding up all the values and dividing by the number of values:

\[
\bar{x} = \frac{\sum x}{n}
\]

If we return to our original 19 paychecks, we find the mean as follows:

\[
\bar{x} = \frac{290 + 350 + \ldots + 1500 + 2000}{19} = \frac{13,590}{19} \approx 715.26
\]
Notice that the mean, which is about $715, is higher than the median of $500. You can think of the mean as the balancing point of all the values. It is the value of the pivot point shown in Figure 4.3 that balances all the observations.

![Figure 4.3. The mean as the balancing point.](image)

The mean is influenced much more by the one high salary going to the president of the company. The median, on the other hand, is what statisticians call resistant. The median doesn't depend on what the values are out there at the extremes of our distribution. If the president doubled his salary while everyone else stayed at the same wage, the mean would bump up to $820.53, or around $821. But our median would stay at $500.

The shape of a distribution can give you some hints about the relationship between the mean and median. For a fairly symmetric distribution, such as the one shown in Figure 4.4, the mean and median are roughly the same.
If the distribution is skewed to the right, like the scores on the difficult exam pictured in Figure 4.5, the mean is larger than the median.

Likewise, if the distribution is skewed to the left, like the scores on one easy exam shown in Figure 4.6, the mean is smaller than the median. Remember the mean is influenced by values at the extremes and the median is not.
Figure 4.6. Median larger than mean.
A. Understand that graphical descriptions of data are more meaningful when supplemented with numerical measures of center.

B. Know that the median (midpoint or typical value) and mean (arithmetic average) are common measures of center (or location) for a distribution. Sometimes the mode is also used as a measure of center.

C. Be able to calculate the median, mean, and mode of a small data set.

D. Know that the mean and median should be close in symmetric distributions and that the mean is pulled toward the long tail of a skewed distribution. Know that the mean is a non-resistant measure of center because it is strongly influenced by extreme observations and that the median is a resistant measure of center.

E. Be able to choose an appropriate measure of center in practice.
Before describing data numerically, we always begin with a graph such as a stemplot or a histogram. The graph shows us the overall pattern of the data and any striking deviations such as outliers. The next step is to give a numerical description of some important aspects of the data. The focus of this unit is on numerical descriptions of the center or location of a distribution. The median, mean, and mode are three numerical measures that use different ideas of “center.” We begin with the median.

The median is the midpoint of a distribution, the value with half the observations lying below it and half above. Instructions for calculating this midpoint number are given below.

### Calculating the Median of $n$ Observations

**Step 1:** Arrange the observations from smallest to largest.

**Step 2:** Determine the location of the median: $(n + 1)/2$.

**Step 3:** Find the median in the ordered list from Step 1:

- If $n$ is odd, count up $(n + 1)/2$ spots in the ordered list and select this value. The median will be the middle number in the ordered list.

- If $n$ is even, count up the number of spots on either side of $(n + 1)/2$ and average these two values. This median will be the average of the two middle numbers.

The median is easy to calculate once the data are ordered from smallest to largest. However, if the data set is large, use software to sort the data from largest to smallest. Another approach to ordering the data would be to make a stemplot. As an example, consider the 22 exam scores listed below.

40 41 50 68 69 72 76 79 79 80 82 85 86 87 88 88 90 91 92 93 96 98

The exam scores have already been ordered from smallest to largest. Notice that repeat scores are included in the list. For example, two people scored 88 and so, the score of 88 appears twice on this list.
Now, we compute the location of the median:

\[ \frac{n+1}{2} = \frac{22+1}{2} = 11.5 \]

Start at 40 and count up 11 and 12 positions. Exam scores 82 and 85 are in the 11th and 12th position. The median is the average of these two numbers:

\[ \text{median} = \frac{82 + 85}{2} = 83.5 \]

Next, we discuss the mean as a measure of center. The **mean** is the average value. It is the balance point of the distribution (See Figure 4.3.). If the observations are from a sample of \( x \) values, we often use the notation \( \bar{x} \) to represent the mean.

Here’s how to calculate the mean:

**Calculating the Mean**

For \( n \) observations of \( x \) values:

\[
\bar{x} = \frac{\text{sum of the observations}}{\text{number of observations}} = \frac{\sum x}{n}
\]

Returning to our example of 22 exam scores, we calculate the mean as follows:

\[
\bar{x} = \frac{40 + 41 + \ldots + 96 + 98}{22} = \frac{1730}{22} \approx 78.6
\]

If we had started with a graphic display of the exam scores as shown in Figure 4.7, we should have expected a mean that was less than the median. The histogram is skewed to the left, with a few exam scores in the left tail of the distribution. The median is unaffected by these scores, but they drag the mean down.
Lastly, there is one other measure that is sometimes used as a measure of center, and that is the mode. The **mode** is the most frequent observation. In our list of exam scores, there are two scores that appear twice in the list, 79 and 88. Since both of these scores are tied for occurring most frequently, the mode is not unique – instead there are two modes.

We have discussed three measures of center or location, the median, mean, and mode. How do you decide which is best for a given situation? In choosing an appropriate measure of center, start with a graphic display of the data. Consider the overall shape of the data and deviations from that shape before deciding whether to use the mean or median to summarize the location of the data. Keep in mind that the median is a **resistant** measure of center, which is not influenced by a few extreme data values whereas a few extreme outliers can pull the mean in the direction of the extreme values.

For roughly symmetric distributions the mean and median will be close in value. For highly skewed data, or data with extreme outliers, the median is generally the better choice for a measure of the center or location of the data. For data sets with multiple peaks, the modes may give a better indication of location.

---

**Figure 4.7. Histogram of exam scores.**

[Histogram image]
KEY TERMS

The **median** gives the midpoint of a set of data – it separates the upper half of the data from the lower half. To calculate the median, order the data from smallest to largest and count up \((n + 1)/2\) places in the ordered list.

The **mean** is the arithmetic average or balance point of a set of data. To calculate the mean, sum the data and divide by the number of data:

\[
\bar{x} = \frac{\sum x}{n}
\]

The **mode** is the data value that occurs most frequently.

A **resistant measure** of some aspect of a distribution (such as its center) is relatively unaffected by a small subset of extreme data values.
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What variable is examined in comparing men and women workers at the beginning of the video?

2. Would you describe the shape of the distribution of men’s weekly wages as symmetric, skewed to the left or skewed to the right?

3. What is the most important difference between the distributions of weekly wages for men and for women?

4. Would a few very large incomes pull the mean of a group of incomes up, down, or leave the mean unaffected?

5. Would a few very large incomes pull the median of a group of incomes up, down, or leave the median unaffected?
UNIT ACTIVITY:
MEAN, MEDIAN AND DISTRIBUTION SHAPE

This activity will provide an opportunity to practice computing the mean and the median. In addition, the activity will emphasize the relevance of a distribution’s shape to the relationship between the mean and median. You will need to access Stemplots from the Interactive Tools menu.

1. Work in Quiz mode. Set the number of observations to 10 and the maximum to 100. Assume that the Stemplots tool is generating 10 hypothetical test scores.

   a. Make a copy of the stemplot. Based only on the shape of the stemplot, which do you think is larger, the mean or the median? Justify your choice.

   b. Calculate the mean and median. Show your calculations. Submit your answers to make sure they are correct. (If not, revise your answers and re-submit.) Now that you have done the calculations, was your answer to (a) correct?

2. Repeat question 1 with a new sample of 10 test scores.

3. Use the Stemplots tool to generate 17 final exam scores. The final exam is worth 150 points; so, set the Maximum Observation Value to 150.

   a. Make a copy of the stemplot.

   b. Based only on the shape of the stemplot, which do you think is larger, the mean or the median? Justify your choice.

   c. Calculate the mean and median. Use the Stemplots tool to check that your calculations are correct. Was your answer to (b) correct?
4. Repeat question 3 with a new sample of 17 final exam scores.

5. Below are 70 exam scores from a very difficult exam given to a large class.

   64   78   67   35   74   73   69   66   36   69
   74   38   72   79   36   46   77   69   39   38
   63   32   36   80   35   35   36   39   35   35
   67   73   58   43   64   64   69   69   37
   50   63   36   39   74   36   35   60   62   65
   69   69   35   34   49   67   65   61   33   36
   36   37   36   36   65   69   40   72   69   66

   a. Work in calculation mode. Enter the exam scores into the Stemplots tool. Use the interactive tool to make the stemplot. Describe the shape of the plot.

   b. Determine the median, mean, and mode(s) for the exam scores.

   c. Based on the plot, which gives a better description of the location of these data, the median, mean, or mode(s)? Explain.

6. Below are 30 exam scores from a statistics exam.

   90   76   78   76   75   74   85   74   65   78
   75   60   75   76   75   78   70   75   65   85
   72   74   70   76   72   80   80   72   78   74

   a. Work in calculation mode. Enter the exam scores into the Stemplots tool. Use the interactive tool to make the stemplot. Describe the shape of the plot.

   b. Determine the median, mean, and mode(s) for the exam scores.

   c. Based on the plot, which gives a better description of the location of these data, the median, mean, or mode(s)? Explain.
1. Here are the starting salaries, in thousands of dollars, offered to the 20 students who earned degrees in computer science in 2011 at a university.

<table>
<thead>
<tr>
<th>63</th>
<th>56</th>
<th>66</th>
<th>77</th>
<th>50</th>
<th>53</th>
<th>78</th>
<th>55</th>
<th>90</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>69</td>
<td>59</td>
<td>76</td>
<td>48</td>
<td>54</td>
<td>49</td>
<td>68</td>
<td>51</td>
<td>50</td>
</tr>
</tbody>
</table>

a. Make a graph to describe the distribution and write a brief description of its important features.

b. Find the median salary.

c. Find the mean salary.

d. Find the mode of the salaries.

e. Is the mean about the same as the median or not? What feature of the distribution explains the difference between the mean and the median? Is the mode a good measure of the center for these data?

2. Each month, the Commerce Department reports the “average” price of new single-family homes. For August 2012, the two “averages” reported were $256,900 and $295,300. Which of these numbers was the mean price and which was the median price? Explain your answer.

3. In 1961 New York Yankee outfielder Roger Maris held the major league record for home runs in a single season, with 61 home runs. That record held for 37 years. Here are Maris’s home run totals for his 10 years in the American League.

| 13, 23, 26, 16, 33, 61, 28, 39, 14, 8 |

a. Find the mean number of home runs that Maris hit in a year, both with and without his record 61. How does removing the record number of home runs affect his mean number of runs?
b. Find the median number of home runs that Maris hit in a year, both with and without his record 61. How does removing the record number of home runs affect his median number of runs?

c. If you had to choose between the mean and median to describe Maris’s home run hitting pattern, which would you use?

4. Refer to Table 3.3 (Unit 3). This table gives the number and percentage of residents 65 and older in each state and the District of Columbia.

a. Unit 3, Exercise 1(a) asked you to draw a histogram of the numbers of residents 65 and older. (If you haven’t already done so, draw the histogram.) Compute the mean and median of these data. Which measure of location, the mean or the median, better describes the location of the numbers of residents 65 and older? Justify your choice based on a histogram of these data.

b. Unit 3, Exercise 2(a) asked you to draw a histogram of the percentage of residents 65 and older. (If you haven’t already done so, draw the histogram.) Compute the mean and median of these data. Which measure of location, the mean or the median, better describes the location of the percentage of residents 65 and older? Justify your choice based on a histogram of the percentages.
1. A man in a nursing home has his pulse taken every day. His pulse readings (beats per minute) over a one-month period appear below.

\[
72 \quad 56 \quad 56 \quad 68 \quad 78 \quad 72 \quad 70 \quad 70 \quad 60 \quad 72 \quad 68 \quad 74 \\
76 \quad 64 \quad 70 \quad 62 \quad 74 \quad 70 \quad 72 \quad 74 \quad 72 \quad 78 \quad 76 \quad 74 \\
72 \quad 68 \quad 70 \quad 72 \quad 68 \quad 74 \quad 70
\]

a. Make a stemplot of the pulse data. Break the stem into 5 (for digits 01, 23, 45, 67, and 89).

b. Determine the mean, median and mode for these data. Be sure to include units in your answers.

c. Based on these data, which measure (or measures) from (b) do you think best describes the man’s typical pulse rate? Explain your reasoning.

2. Eating fish contaminated with mercury can cause serious health problems. Mercury contamination from historic gold mining operations is fairly common in sediments of rivers, lakes, and reservoirs today. A study was conducted on Lake Natoma in California to determine if the mercury concentration in fish in the lake exceeded guidelines for safe human consumption. A sample of 83 largemouth bass was collected and the concentration of mercury from sample tissue was measured. Mercury concentration is measured in micrograms of mercury per gram or \( \mu g/g \). The histogram in Figure 4.8 presents results from this study.

![Figure 4.8. Histogram of mercury concentration in fish.](image)

Figure 4.8. Histogram of mercury concentration in fish.
a. The primary objective of the study was to determine if mercury concentrations in fish tissue exceeded safety guidelines for human consumption. The U.S. Environmental Protection Agency (USEPA) human health criterion for methylmercury in fish is 0.30 μg/g. Approximately how many of the fish in the sample had mercury concentrations below the level set by the EPA (and hence were considered safe for human consumption)?

b. Approximately what percentage of the sample had mercury concentrations higher than the level set by the EPA? Show how you arrived at your answer.

c. Would the mean mercury concentration be larger, smaller, or about the same as the median mercury concentration? Explain.

3. A student often orders french fries at a local fast-food place. She keeps track of the number of french fries in each small bag she buys. Here are her counts:

   42, 47, 49, 58, 43, 47, 44, 38, 38, 28, 55, 40, 46
   54, 45, 45, 51, 35, 46, 37, 46, 40, 43, 49, 37

   a. Calculate the mean and median for these data. Show how you computed these values.

   b. Make a stemplot of the distribution. Describe the overall shape of the distribution. Are there any outliers?

   c. Do you prefer the mean or the median as a brief description of the center of this distribution? Why?
SUMMARY OF VIDEO

Hot dogs are an American icon – and we eat billions of them every year. It seems like there is a hot dog to fit just about every taste out there… all beef, some pork, turkey, skinless, even tofu for the vegetarian hot dog lover. Not all hot dogs are created equal though, at least in terms of calories. The calorie count varies quite a bit depending on the type of hot dog and also from brand to brand of a given type. The video gives us an inside view at Vallid Labs in Agawam, Massachusetts, and the calorie counting techniques they use to find the number of calories in particular hot dogs. After turning a hot dog into mush and then applying a series of treatments with acids and bases, distillations and titrations, the number of grams of fat, protein, and carbohydrates per hot dog is determined. Then using the information from the Nutrition Facts in Figure 5.1, we can determine a hot dog’s calories.

**Nutrition Facts**

<table>
<thead>
<tr>
<th>Component</th>
<th>1g = Calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat</td>
<td>9</td>
</tr>
<tr>
<td>Protein</td>
<td>4</td>
</tr>
<tr>
<td>Carbohydrate</td>
<td>4</td>
</tr>
</tbody>
</table>

*Figure 5.1. Determining the calories in a hot dog.*

Despite similar appearances, hot dogs vary widely in nutritional content. For those calorie-conscious among us, statistics can help suggest the healthiest choice. We start with the calorie counts (listed in order from smallest to largest) of 20 different brands of all-beef hot dogs.

110 110 130 130 140 150 160 160 170 170
175 180 180 180 190 190 190 200 210 230
You can see that all-beef hot dogs range from 110 calories in the lowest brand to 230 calories in the highest. One way we can describe this distribution numerically is with the median – the number of calories in a typical beef hot dog. Since we have 20 brands, the location of the median is \((10 + 1)/2\), which is 10.5. So, we find the median by averaging the 10th and 11th calorie count in the ordered list:

\[
\text{median} = \frac{170 + 175}{2} = 172.5 \text{ Calories}
\]

So, your typical beef hot dog has 172.5 calories.

Next, we add some more numbers to our analysis of the spread of the calorie counts of beef hot dogs. We know the minimum is 110 calories and the maximum is 230 calories. However, we also need some information about the numbers in between. For that we can determine the quartiles. These are values one-quarter and three-quarters up the ordered list of calories. The first quartile – also known as \(Q_1\) – has 25% of the ordered observations at or below it. It is the median of the lower half of the data:

110 110 130 130 140 150 160 160 170 170

The median of this 10 number set is between the fifth and sixth data value: \(Q_1 = 145\). Now, we turn our attention to the third quartile – also known as \(Q_3\) – which has 75% of the ordered observations at or below it. To find \(Q_3\), take the median of the upper half of the data:

175 180 180 180 190 190 190 200 210 230

The median of these 10 upper numbers gives us \(Q_3 = 190\).
So, with our minimum, first quartile, median, third quartile, and maximum, we have what is called the five-number summary, which we’ve summarized in Figure 5.2.

![Five-number summary of all-beef hot dog calories](image)

**Figure 5.2. The five-number summary of all-beef hot dog calories.**

The five-number summary gives us a nice snapshot of both the center and spread of the data. The median marks the center. The first and third quartiles contain between them the middle half of the data. We can measure the spread of the inner 50% of the data by the interquartile range (IQR):

$$IQR = Q_3 - Q_1$$

The two extremes show how far out the data extends. We can measure that spread using the range:

$$\text{range} = \text{maximum} - \text{minimum}$$

In statistics, the best description of data often combines the precision of numbers with the clarity of pictures. A boxplot (or box-and-whisker plot) is a graphic display of the five-number summary. Figure 5.3 shows a boxplot of the all-beef hot dog calories. The box spans the first and third quartiles, the median is marked inside the box, and whiskers extend out to the extremes.
Boxplots don’t show a distribution in detail the same way that a stemplot or histogram would, but boxplots can be a great way to make a quick side-to-side comparison of a few distributions. Take a look at Figure 5.4 comparing the calories of beef, poultry, and veggie hot dogs.

Notice that the median of the poultry hot dogs is below the minimum for the beef hot dogs. So, half of the brands of poultry hot dogs have fewer calories than the lowest calorie brand of all-beef hot dogs. Now, check out the boxplot for the vegetarian hot dog. Notice that at least one
veggie brand has more calories than three quarters of the beef hot dogs!

So, now we can add boxplots to the list of ways we can graphically represent data – and, as shown by the hot dogs, this method allows for easy comparisons between groups.
STUDENT LEARNING OBJECTIVES

A. Recognize that a basic numerical description of a distribution requires both a measure of center and a measure of spread.

B. Use the quartiles and the extremes to provide information about the unequal spread in the two sides of a skewed distribution.

C. Use the $1.5 \times \text{IQR}$ rule to identify outliers.

D. Be able to calculate the quartiles and give the five-number summary of a data set of moderate size (say $n \leq 100$).

E. Understand that boxplots provide less detail than stemplots or histograms but are especially useful for comparing several distributions.
The topic of this unit is the **five-number summary** and its associated graph, the **box-and-whisker plot** or **boxplot**. The five-number summary of a set of data consists of the minimum, **first quartile**, median, **third quartile** and maximum. You already know how to calculate the minimum, median, and maximum. In this overview, we will provide algorithms for calculating the quartiles. It should be noted, however, that there are several different algorithms for calculating the quartiles. So, check with your textbook or software to see how it calculates quartiles.

First, we discuss a rationale for the five-number summary for describing a data set. The five-number summary provides information on both the center of a distribution and its spread. The median is a useful measure of the *center* of a set of observations. The median is the midpoint, the point with half of the data at or below it and half above. However, the median alone is not an adequate description of a set of data. For example, it is not enough to know that the median number of candies in bags of candy is 60 pieces. It is quite a different story if (1) some bags have as few as 40 and others have as many as 75 compared to (2) some bags have as few as 55 and others have as many as 65. To quantify these two situations, we’ll need information about the *spread* or *variability* of the data.

Because the median is the “halfway” point in a data set, one way to show spread is by giving the two quartiles along with the median. The first quartile is the one-quarter point in the data: one-fourth of the data values are at or below the first quartile and three-quarters above. The third quartile is the three-quarters point, with three-quarters of the data at or below it. The two quartiles capture the middle half of the data between them. So, the distances from the median out to the quartiles and between the quartiles show how spread out the data are, or at least how spread out the middle 50% of the data are. The distances between the first and third quartiles, $Q_3 - Q_1$, is called the **interquartile range** or **IQR**.

To calculate the quartiles, first locate the median in an ordered data list. The median divides the ordered data into a lower half and an upper half.

- If there is an odd number of data values, the median is the middle data value in the ordered list. Omit this value when forming the lower half and upper half of the ordered data.

- If there is an even number of observations, the median is between the middle two data values. So, the ordered data can be divided into a lower half and upper half about the median.
The first quartile, \( Q_1 \), is the median of the lower half of the ordered data and the third quartile, \( Q_3 \), is the median of the upper half of the ordered data.

So far, we have discussed using the median to describe the center of a distribution and the interquartile range to describe the spread of the middle half of the data. We can add information about how far the data are spread by giving the distances from the median out to the minimum and maximum data values and between the minimum and maximum. The distance between the minimum and maximum, maximum – minimum, is called the range.

Now, we work through an example. Grades from an exam are displayed in the stemplot in Figure 5.5. We use the stemplot to order the test scores from smallest to largest.

```
4   01
5
6   89
7   2699
8   0256788
9   01236
```

*Figure 5.5. Stemplot of test scores.*

Since \( n = 20 \), the median is at the \((20 + 1)/2\) or 10.5 position, midway between 82 and 85; hence, the median = 83.5. The median divides the data into a lower half and upper half:

| Lower half: | 40 | 41 | 68 | 69 | 72 | 76 | 79 | 79 | 80 | 82 |
| Upper half: | 85 | 86 | 87 | 88 | 88 | 90 | 91 | 92 | 93 | 96 |

The median of the lower half is at the \((10 + 1)/2\) or 5.5 position, midway between 72 and 76; so, \( Q_1 = 74 \). The median of the upper half is midway between 88 and 90; so, \( Q_3 = 89 \).
Here's our five-number summary of the exam grades:

\[
\text{minimum} = 40, \quad Q_1 = 74, \quad \text{median} = 83.5, \quad Q_3 = 89, \quad \text{maximum} = 96
\]

We can use the median, 83.5, as a measure of center for the test scores. The spread of the middle 50% of the test scores is given by the interquartile range, $\text{IQR} = 89 - 74 = 15$. The spread as measured from the smallest test score to the largest is given by the range $= 96 - 40 = 56$. Notice that the overall spread of the test scores is more than three times the spread of the middle 50% of the test scores.

In its basic form, a boxplot (or box-and-whisker plot) is a graphical display of the five-number summary. It can be drawn either vertically or horizontally depending on your preference. Once you have the five-number summary, it takes only three steps to draw a basic boxplot as outlined below.

**Constructing a Basic Boxplot**

The instructions below are for horizontal boxplots but easily can be adapted for vertical boxplots.

**Step 1:** Draw a number line. Add a scale that begins at or below the minimum and ends at or above the maximum.

**Step 2:** Directly above the number line, draw a rectangular box that extends from $Q_1$ to $Q_3$. Divide the box with a vertical line at the median.

**Step 3:** Draw two whiskers: one from the middle left side of the box to the minimum and the other from the middle right side of the box to the maximum.

Figure 5.6 shows the result of applying these steps to create a basic boxplot from the five-number summary for the test scores.
Figure 5.6. Basic boxplot of exam grades.

\[ Q_1 = 74, \text{median} = 83.5, Q_3 = 89, \text{maximum} = 96 \]

Each part of the boxplot – the left whisker, the box from \( Q_1 \) to the median, the box from the median to \( Q_3 \), and the right whisker – represents the spread of one quarter of the data. So, for example, because the box from \( Q_1 \) to the median is longer than the box from the median to \( Q_3 \), we know that the second quarter of the test scores are more spread out than the third quarter of the test scores.

Notice also the long left whisker that extends from \( Q_1 = 74 \) all the way down to the minimum test score of 40. We don’t know if that long whisker is the result of a single low grade, an outlier, or if the pattern of the lower quarter of the test scores spreads out over the interval from 40 to 74. A modified boxplot, which separates out the outliers and adjusts the lengths of the whiskers so that they are unaffected by outliers, will help us sort out this issue. Here are the steps needed to convert a basic boxplot into a modified boxplot (the generally preferred plot).
Constructing a Modified Boxplot

**Step 1:** After making a basic boxplot, remove the whiskers.

**Step 2:** Compute the IQR = $Q_3 - Q_1$; compute a step = $1.5 \times IQR$.

**Step 3:** Calculate the inner fences (one step on either side of the box ends):

$$Q_1 - 1 \text{ step and } Q_3 - 1 \text{ step}.$$  

Calculate the outer fences (two steps on either side of the box ends):

$$Q_1 - 2 \text{ steps and } Q_3 + 2 \text{ steps}.$$  

**Step 4:** Identify the mild outliers. Use an asterisk (*) to plot any data values that lie between the two fences. Identify the extreme outliers. Use another symbol, such as an open circle, to plot any data values that are more extreme than the outer fences.

**Step 5:** Attach a whisker from the left end of the box to the smallest data value that is not an outlier. Then attach a whisker from the right end of the box to the largest data value that is not an outlier.

Next, we convert the basic boxplot from Figure 5.6 into the modified boxplot shown in Figure 5.7. We begin by removing the whiskers from the basic boxplot. Then we calculate the inner and outer fences as follows:

$$IQR = 89 - 74 = 15$$  
$$\text{Step} = 1.5 \times IQR = 1.5(15) = 22.5$$  
$$\text{Inner fences: } Q_1 - 1 \text{ step; } Q_3 + 1 \text{ step: } 74 - 22.5 = 51.5; 89 + 22.5 = 111.5$$  
$$\text{Outer fences: } Q_1 - 2 \text{ steps; } Q_3 + 2 \text{ steps: } 74 - 2(22.5) = 29; 89 + 2(22.5) = 134$$

Two test scores, 40 and 41, fall between the lower inner fence and lower outer fence and hence are classified as mild outliers. Mark each of their locations with an asterisk. (There are no extreme outliers.) Attach the left end of the box at $Q_1$ to 68, the smallest test score that is not an outlier. Redraw the original right whisker (since all test scores were smaller than the upper fences).

The completed modified boxplot appears in Figure 5.7.
Figure 5.7. Modified boxplot of test scores.

Notice that in the modified boxplot, the length of the lower whisker is about the same as the upper whisker, which indicates that with outliers removed the lower quarter of the test scores had about the same spread as the upper quarter of the test scores. The long left whisker in the basic boxplot was due to two students whose grades were outliers.
A **five-number summary** of a set of data consists of the following:

minimum, first quartile \((Q_1)\), median, third quartile \((Q_3)\), maximum.

The **first quartile**, \(Q_1\), is the one-quarter point in an ordered set of data. To compute \(Q_1\), calculate the median of the lower half of the ordered data. The **third quartile**, \(Q_3\), is the three-quarter point in an ordered set of data. To compute \(Q_3\), calculate the median of the upper half of the ordered data.

A basic **boxplot** (or **box-and-whisker plot**) is a graphical representation of the five-number summary. A modified boxplot indicates outliers and adjusts the whiskers.

The **interquartile range** or **IQR** measures the spread of the middle half of the data:

\[ IQR = Q_3 - Q_1 \]

The **range** measures the spread of the data from its extremes:

\[ \text{range} = \text{maximum} - \text{minimum} \]
THE VIDEO: BOXPLOTS

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What *variable* is used to compare different brands of hot dogs?

2. What name do we give to the value for which one-quarter of the data values falls at or below it?

3. What numbers make up a five-number summary?

4. How do you calculate the interquartile range?

5. Boxplots show that poultry hot dogs as a group differ from all-beef hot dogs. Compare the distribution of calories between the two types of hot dogs.
Unit Activity:
Using Boxplots to Analyze Data

Return to the class survey data collected for Unit 2’s Activity. The questions are restated at the end of this activity so that you don’t have to refer back to Unit 2.

Analysis of the Survey

1. Make modified boxplots for the data from questions 1 – 5. Describe the key features of each of your boxplots.

2. How do estimates of waiting time compare for males and females? Make comparative modified boxplots of the data from survey question 1 for males and females.

3. Do males or females spend more time studying for exams? Make comparative modified boxplots of the average time spent studying for an exam for males and females. Interpret your graphs in the context of study times.

4. Make comparative modified boxplots for the amount of time that males and females exercise on a typical day.

Survey Questions (Unit 2, Activity)

1. How long (in seconds) did you wait while your instructor was getting ready for this activity?

2. How much money in coins are you carrying with you right now?

3. To the nearest inch, how tall are you?

4. How long (in minutes) do you study, on average, for an exam?

5. On a typical day, how many minutes do you exercise?

6. Circle your gender: Male Female
EXERCISES

1. A consumer testing laboratory measured the calories per hot dog in 20 brands of beef hot dogs. Here are the results:

   186  181  176  149  184  190  158  139  175  148  
   152  111  141  153  190  157  131  149  135  132

   a. Find the five-number summary of this distribution. Explain how you arrived at your answer.

   b. Compute the range and interquartile range. Explain what these numbers tell you about the variability in calories in different brands of all-beef hot dogs.

   c. Would a beef hot dog with 175 calories be in the top quarter of the data? Support your answer.

2. Return to the data on all-beef hot dog calories from exercise 1.

   a. Draw a basic boxplot for the calories per hot dog.

   b. In which quarter – the first, second, third, or fourth – are the data most concentrated? Explain how you can answer this question based on the boxplot from (a).

   c. In which quarter – the first, second, third, or fourth – is the data most spread out? Explain how you can answer this question based on the boxplot from (a).

   d. If a data value is more than $1.5 \times \text{IQR}$ below the first quartile or more than $1.5 \times \text{IQR}$ above the third quartile, it is considered an outlier. Should any of the calorie counts for the beef hot dogs be classified as outliers? Explain.

3. Make a stemplot of the calories in the sample of beef hot dogs from exercise 1. What do you learn from the stemplot that you could not learn from the boxplot?
4. The calories for 20 brands of veggie dogs are given below. (Notice these data have been ordered from smallest to largest.)

   40  45  45  45  50  50  55  57  60  60  
   70  80  80  81  90  95  100  100  110  190

   a. Make a five-number summary of the veggie dog calories.

   b. Make a modified boxplot for the veggie dog data. Use asterisks (*) to indicate any mild outliers and open circles to indicate any extreme outliers. (Leave room to add another graph to this graphic display.)

   c. Add a modified boxplot for the beef hot dog calories next to your display in (b). This will allow you to compare the calorie distributions of the two types of hot dogs.

   d. Based on your displays in (c), compare the distributions of calories for beef dogs and veggie dogs.

5. Refer to the baseball data from Table 3.4, Unit 3. Focus on the variable career home runs. (Keep in mind there are 104 players listed because of ties in career batting averages.)

   a. Order the number of career home runs from smallest to largest. (You may want to use software such as Excel or your graphing calculator to do the ordering. Otherwise, try using a stemplot to help you order these data.)

   b. Create a five-number summary of the career number of home runs.

   c. Make a modified boxplot of the career number of home runs. Mark mild outliers with asterisks (*) and extreme outliers with open circles. Show your calculations for the fences. Write the names of the players above each of the outliers.

   d. Would you describe the shape of the distribution as symmetric, skewed to the right, or skewed to the left? Justify your choice.
Review Questions

Recall that Table 2.1 in Unit 2 gives data on state average SAT scores and the percent of high school graduates in each state taking the SATs. Refer to Table 2.1 as needed for questions 1 and 2.

1. The state average SAT Critical Reading scores, ordered from smallest to largest, appear below.

| 469 | 469 | 479 | 479 | 482 | 485 | 485 | 487 | 489 | 493 | 493 | 493 | 494 |
| 495 | 495 | 499 | 499 | 509 | 512 | 513 | 514 | 515 | 515 | 517 | 520 | 523 |
| 523 | 539 | 539 | 542 | 546 | 548 | 555 | 563 | 564 | 568 | 570 | 571 | 572 |
| 575 | 576 | 580 | 583 | 584 | 585 | 586 | 590 | 592 | 593 | 596 | 599 |

a. Determine a five-number summary of the state average SAT Critical Reading scores.

b. Does California (average score 499) fall in the top half of the states in the SAT Critical Reading score? Does it fall above the bottom quarter? Support your answer.

c. Roughly what percentage of the states have scores higher than Wyoming’s 572? How many states would that be?

d. Make a basic boxplot of the states’ average SAT Critical Reading scores. Which quarter of the data, the first, second, third or fourth, shows the most amount of spread?

2. The states’ average SAT Math scores, ordered from smallest to largest, appear below.

| 457 | 469 | 487 | 489 | 490 | 490 | 493 | 496 | 499 | 500 | 501 | 501 | 501 |
| 502 | 502 | 508 | 509 | 511 | 513 | 515 | 516 | 518 | 521 | 523 | 525 | 527 |
| 529 | 537 | 539 | 541 | 541 | 543 | 545 | 550 | 559 | 565 | 568 | 569 | 570 |
| 572 | 573 | 591 | 591 | 591 | 593 | 602 | 604 | 606 | 608 | 612 | 617 |

a. Give the five-number summary of the 51 state average SAT Math scores.

b. Make boxplots to compare the distribution of the Critical Reading and Math scores. (In order to make comparisons, the boxplots must be on the same scale and positioned so that comparisons are easily made.)
c. Write a brief description comparing the distributions. Include in your descriptions comparisons of both center and spread.

3. The stemplot video in Unit 2 included data on the fuel economy information on Toyota's 2012 vehicle line. A stemplot of the city miles per gallon (mpg) data appears below.

```
1 3 4
1 67799
2 0112
2 5678
3 0
3
4 34
4
5 1
5
```

a. Make a five-number summary of the mpg data.

b. Make a basic boxplot of the mpg data.

c. Based on the stemplot, how many of the data values are potential outliers?

d. Make a modified boxplot of the mpg data. Show the calculations for the fences. Based on your plot, how many data values were identified as outliers?
SUMMARY OF VIDEO

The video begins with a tale of two cities, Portland, Oregon, and Montreal, Quebec. The average monthly precipitation in these two cities is similar – Portland’s average precipitation is 3.32 inches per month and Montreal’s is 3.4 inches per month. Not much difference there! However, the average (or mean) monthly precipitation for these two cities is not the whole story. As can be seen in Figure 6.1, Montreal’s precipitation is relatively consistent from month to month, while Portland’s is far more variable – more rain in the winter months and little rain in the summer months.

![Figure 6.1. Monthly precipitation for Montreal and Portland.](image)

We’ve hit on a case where a measure of center does not provide all the information we need. To express the climate difference between these two cities, we need a measure of how much spread or variability there is in month-to-month precipitation.

Next, the video switches to two locations in California – the sites of two Wahoo’s Fish Taco restaurants, one located at Manhattan Beach and the other in the South Coast Plaza. Wing Lam and his two brothers founded their first Wahoo’s as a single taco shop on the beach. Now they have more than 50 Wahoo’s Fish Taco restaurants in their chain. Knowing what to expect,
based on how busy each month has been in the past, lets managers plan inventory orders and staff schedules appropriate to each location.

The mean sales per four-week period for the South Coast Plaza and Manhattan Beach locations are around $130,000 and $97,000, respectively. But that’s not all that differs between these two locations. South Coast Plaza is a great store because it is indoors in a shopping mall, and therefore, sales are relatively unaffected by the weather. On the other hand, the Manhattan Beach restaurant is on the beach and so the weather has greater impact on its sales. Back-to-back stemplots in Figure 6.2 show the sales from these two restaurants over four-week periods.

![Back-to-back stemplots of sales.](image)

Because the South Coast Plaza restaurant is indoors, its sales are pretty stable. A stemplot of the South Coast Plaza restaurant data is roughly symmetric, with one outlier at $177,000 (for December). On the other hand, the sales data for the Manhattan Beach location are more variable than the South Coast Plaza location. The range for the Manhattan Beach location is over $100,000 compared to a range of $70,000 for the South Coast Plaza location. Boxplots in Figure 6.3 based on these sales figures make the variation obvious. Check out the interquartile ranges, represented by the widths of the boxes. The interquartile range for the Manhattan Beach location is wider than for the South Coast Plaza location.
Another way to quantify the spread is to measure how far observations are from their mean. This is the idea behind measures of spread called the variance and standard deviation. To find the variance of the Manhattan Beach and South Coast Plaza data, we use the following formula

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

The standard deviation is the square root of the variance. The standard deviations for the Manhattan Beach and South Coast Plaza Wahoo's are $31,075 and $17,000, respectively. This difference in variability makes sense given the locations of these two Wahoo's. At the beach, sales are affected by weather – warm, sunny weather brings people to the beach along with good sales, while cold rainy days keep people away resulting in poor sales. However, at the indoor mall location, Wahoo’s sales are fairly steady all year long.
STUDENT LEARNING OBJECTIVES

A. Know that the sample standard deviation, s, is the measure of spread most commonly used when the mean, $\bar{x}$, is used as the measure of center.

B. Be able to calculate the standard deviation s from the formula for small data sets (say $n \leq 10$).

C. Know the basic properties of the standard deviation:
   - $s \geq 0$, and $s = 0$ only when all data values are identical.
   - s increases as the spread about $\bar{x}$ increases.
   - s, like $\bar{x}$, is strongly influenced by outliers.

D. Know that the standard deviation is most useful for symmetric distributions and, in particular, for normal distributions.

E. Know that adding the same constant a to all the observations increases the value of $\bar{x}$ by a. However, adding the same constant a to all the observations does not change the value of s. That’s because adding a constant a to all data values shifts the location of the data but does not affect its spread.

F. Know that multiplying all data values by a constant amount $k$ changes $\bar{x}$ and s by a factor of $k$. 
Although the variance and standard deviation are the most common measures of spread or variability in statistical practice, they are tedious to calculate from their formulas and somewhat difficult to interpret. Their common use arises from the fact that the standard deviation is the natural measure of spread for normal distributions, in which data tend to be mound-shaped and symmetric. (Normal distributions will be covered in Units 7 – 9.) For describing distributions of data, the five-number summary is generally more useful than $\bar{x}$ and $s$ particularly for distributions that are not roughly mound-shaped and symmetric.

If we decide to use the mean $\bar{x}$ to describe the center of a set of data, then it makes sense to use the deviations from the mean, $x - \bar{x}$, as the basis for a measure of spread. Clearly, if all the deviations from the mean are small, then all data values lie close to $\bar{x}$ and there is little spread. The only problem is that some of the deviations from the mean are positive and some are negative. If you sum the deviations from the mean, $\sum(x - \bar{x})$, the result is always exactly zero. However, the sum of the squared deviations will be positive and so the “average squared deviation from the mean” makes sense as a measure of spread. This is the idea behind the variance. The standard deviation is the square root of the variance.

The variance is an average of the squared deviations from the mean:

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}, \text{ where } n \text{ is the sample size}$$

You are probably wondering why we divide by $n - 1$ in computing the average rather than by $n$. Here is an explanation. Because the sum of the deviations, $\sum(x - \bar{x})$, is always zero, the last deviation can be found once we know the first $n - 1$. That means that only $n - 1$ of the squared deviations can vary freely, and so, the average is found by dividing by $n - 1$.

Because the variance involves squaring the deviations, it does not have the same units of measurement as the original data values. For example, lengths measured in centimeters have a variance measured in squared centimeters. Taking the square root remedies this, so that the standard deviation $s = \sqrt{s^2}$ measures dispersion about the mean in the original scale.
Calculations of \( s \) or \( s^2 \) using the formula can be tedious even for relatively small data sets. As an example, we calculate the mean and standard deviation for the sample of five data values given below.

\[
90 \quad 40 \quad 85 \quad 69 \quad 79
\]

First, we calculate the sample mean as follows:

\[
\bar{x} = \frac{90 + 40 + 85 + 69 + 79}{5} = \frac{363}{5} = 72.6
\]

Now, we are ready to calculate the variance. Here are the steps:

1. For each data value, \( x \), we calculate its deviation from the mean, \( x - 72.6 \), and enter that value into the second column of Table 6.1.
2. Next, we square each of these deviations and enter them into the third column.
3. Then we sum the entries in the third column.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( x - 72.6 )</th>
<th>((x - 72.6)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>17.4</td>
<td>302.76</td>
</tr>
<tr>
<td>40</td>
<td>-32.6</td>
<td>1062.76</td>
</tr>
<tr>
<td>85</td>
<td>12.4</td>
<td>153.76</td>
</tr>
<tr>
<td>69</td>
<td>-3.6</td>
<td>12.96</td>
</tr>
<tr>
<td>79</td>
<td>6.4</td>
<td>40.96</td>
</tr>
</tbody>
</table>

\[
\text{Sum} = 1573.2
\]

Table 6.1. Calculating the Variance.

For the final step, we divide the sum by \( 5 - 1 \), or 4:

\[
s^2 = \frac{1573.2}{5 - 1} = 393.3
\]

To compute the standard deviation, we take the square root of the variance:

\[
s = \sqrt{393.3} \approx 19.8
\]

Although calculating the standard deviation from the formula is reasonable when the sample size is small, for larger data sets it is better to use graphing calculators or software (such as Minitab, SPSS, and Excel). We conclude this overview with a list of properties of the standard deviation.
• The standard deviation, \( s \), measures spread about the mean, and should be used only when the mean is chosen as the measure of center.

• \( s = 0 \) only when there is no spread, that is, when all observations have the same value. Otherwise \( s > 0 \), and \( s \) increases as the observations move farther apart (and hence farther from the mean).

• Like the mean, \( s \) is strongly affected by a few extreme observations. In fact, the use of squared deviations renders \( s \) even more sensitive to a few extreme observations than is the mean.

• Adding a constant value to each data value shifts the data affecting its location but not its spread. Therefore, \( s \) does not change.

• Multiplying each data value by a constant does affect the spread of data. If \( s \) is the standard deviation of a data set, and the data are modified by multiplying each data value by a constant \( k \), then the standard deviation of the modified data set is \( k \cdot s \).
Given a data set, one measure of center is the mean, $\bar{x}$. One way to judge the spread of the data is to look at the **deviations from the mean**, $x - \bar{x}$.

The **variance** is a measure of variability that is based on the square of the deviations from the mean. The formula for computing variance is:

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Because the units for variance are the square of the units for the original data, we generally take the square root of the variance, which gives us the standard deviation:

$$s = \sqrt{s^2}$$
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. In comparing monthly precipitation for Portland, Oregon, and Montreal, Canada, why was comparing the mean monthly precipitation rates insufficient?

2. Why don’t we measure spread about the mean by simply averaging \( x - \bar{x} \), the deviations of individual data values from their mean?

3. What did the standard deviation of four-week sales data tell you about the two Wahoo’s Taco locations, Manhattan Beach and South Coast Plaza?

4. Can the standard deviation of a set of observations be \( s = -1.5 \)? Explain.
UNIT ACTIVITY:
VISUALIZING STANDARD DEVIATION

The standard deviation is a measure of spread that is based on the deviations from the mean. For some insight into deviations from the mean, we start with the following data set: 6, 6, 2, 8, 3. We calculate the mean of these data:

$$\bar{x} = \frac{6 + 6 + 2 + 8 + 3}{5} = 5.$$  

A dotplot of the 5 data values is shown in Figure 6.4. A vertical line has been drawn at the mean, $\bar{x} = 5$. The horizontal line segments represent the deviations of each data value from the mean. The lowest horizontal line segment represents the deviation of the first data value, 6, from the mean; this deviation has length 1, since $6 - 5 = 1$.

![Figure 6.4. Dotplot with horizontal line segments.](image)

1. a. List the deviations from the mean for each of the 5 data values.
   
b. What is the sum of the five deviations from the mean listed in (a)?
   
c. Some of the deviations from the mean are positive and some are negative. To keep the deviations from cancelling each other out, square each of the deviations listed in (a). What is the sum of the square of the deviations?
   
d. To find the variance, $s^2$, divide your answer to (c) by $n - 1$, which in this case is 4. What is the variance? The standard deviation is the square root of the variance. What is the standard deviation? (Round to two decimals.)
2. Consider two more data sets:

Data Set X: 2, 4, 3, 4, 5, 3
Data Set Y: 3, 2, 3, 10, 5, 4

a. Calculate the means for Data Sets X and Y.

b. Draw dotplots of the two data sets – use the same scale for both plots. Draw a vertical line at the mean and then represent the deviations from the mean as horizontal line segments.

c. Based on your plots in (b), which data set do you think will have the larger standard deviation? Explain your reasoning.

d. Calculate the standard deviations for Data Sets x and Y. (Round answers to two decimals.) Do your calculations confirm your answer to (c)?

Figures 6.5 – 6.9 show histograms of five data sets. Your task in questions 3 and 4 will be to determine which data sets have larger standard deviations based on histograms of the data.

Figure 6.5. Histogram of Data Set A.
Figure 6.6. Histogram of Data Set B.

Figure 6.7. Histogram of Data Set C.

Figure 6.8. Histogram of Data Set D.
3. Data Sets A – E displayed graphically in Figures 6.5 – 6.9 all have means of 2.5. So, the mean does not provide any information that could be used to distinguish one data set from another. In parts (a) – (c), determine from the histograms which of the two data sets has the larger standard deviation, or if the standard deviations are about the same. In each case, give a justification of your answer.

a. Data Set A and Data Set B.

b. Data Set C and Data Set D.

c. Data Set D and Data Set E.

4. The standard deviations of the Data Sets A – E are given below in random order. Match each standard deviation with its data set.

\[
1.589 \quad 1.026 \quad 1.124 \quad 1.026 \quad 1.451
\]
EXERCISES

1. SAT Math scores in recent years have had means around 490 and standard deviations around 100. What is the variance of SAT Math scores?

2. Six ninth-grade students and six 12th-grade students were asked: How many movies have you seen this month? Here are their responses.

   Ninth-grade students:  5, 1, 2, 5, 3, 8
   12th-grade students:  4, 2, 0, 2, 3, 1

   a. Calculate the mean, variance, and standard deviation of each of these data sets. Which is more spread out, the ninth-grade or 12th-grade data set?
   b. Make a graph of both data sets. Which of these data sets appears more spread out? Does your answer agree with your conclusion in part (a)?

3. Suppose we add 2 to each of the numbers in the ninth-grade data in question 2. That modification produces the following data: 7, 3, 4, 7, 5, 10.

   a. Find the mean and the standard deviation of the modified data.
   b. Compare your answers from (a) with the mean and standard deviation from the original ninth-graders’ data in question 2. How did adding 2 to each data value change the mean? How did it change the standard deviation?
   c. Without doing the calculations, guess what will happen to the mean and standard deviation of the 12th-graders’ data from question 2 if we add 10 to each data value.

4. Return to the SAT data from Table 2.1, Unit 2.

   a. Use technology to calculate the standard deviation for the critical reading, mathematics, and writing SAT scores.
   b. Based on the standard deviations, which of the three SAT exams has the largest spread over the 50 states and the District of Columbia?
REVIEW QUESTIONS

1. a. If two distributions have exactly the same mean and standard deviation, must their histograms look exactly alike? Explain.

b. If two distributions have the same five-number summary, must their histograms be identical? Explain.

2. The Army wants to describe the distribution of head sizes of its soldiers in order to plan orders of helmets. Here are the head sizes in inches of 30 male soldiers, which were obtained by putting a tape measure around each soldier’s forehead.

   23.0  22.2  21.7  22.0  22.3  22.6
   22.7  21.5  22.7  24.9  20.8  23.3
   24.2  23.5  23.9  23.4  20.8  21.5
   23.0  24.0  22.7  22.6  23.9  21.8
   23.1  21.9  21.0  22.4  23.5  22.5

   a. Give a graphical description of these data. Is there any aspect of the distribution that would discourage the use of $\bar{x}$ and $s$ to measure center and spread?

   b. Find $\bar{x}$ and $s$ for these data. Be sure to include the units in which these numbers are measured.

   c. What percentage of the data is within one standard deviation of the mean?

3. a. The head-size data in question 2 was measured in inches. There are 2.54 centimeters per inch. Change the head-size data from inches to centimeters by multiplying each data value by 2.54.

   b. Calculate the standard deviation of the head-size data from (a).

   c. How is the standard deviation from (a), for head sizes measured in centimeters, related to the standard deviation in 2(b), for head sizes measured in inches?
4. Two types of wire, a 12 ½-gauge low-carbon wire and a thinner 14-gauge high-tensile wire, both used in barbed-wire fencing, are tested to see which wire is stronger. Data on the breaking strengths, in pounds, of both types of wire are given below.

<table>
<thead>
<tr>
<th>12 ½-gauge, low-carbon wire:</th>
</tr>
</thead>
<tbody>
<tr>
<td>455 455 495 490 410 470 475</td>
</tr>
<tr>
<td>480 445 435 405 450 500 435</td>
</tr>
<tr>
<td>430 460 480 480 445 435 405</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>14-gauge, high-tensile wire:</th>
</tr>
</thead>
<tbody>
<tr>
<td>780 780 775 770 780 780 790</td>
</tr>
<tr>
<td>780 785 800 775 760 770 800</td>
</tr>
<tr>
<td>800 780 790 790 785 770 790</td>
</tr>
</tbody>
</table>

a. Make comparative boxplots for the breaking strengths of the types of wire. Would you describe the shapes of the distributions as symmetric or skewed?

b. Determine the mean breaking strengths for the two types of wire. Which type of wire has the larger mean breaking strength?

c. The two types of wires have different properties that affect the consistency of the wire's breaking strength. Find the standard deviation of breaking strengths for each type of wire. Which wire has the more variable breaking strength?
SUMMARY OF VIDEO

Histograms of completely unrelated data often exhibit similar shapes. To focus on the overall shape of a distribution and to avoid being distracted by the irregularities about that shape, statisticians often draw smooth curves through histograms. For example, histograms of weights of Gala apples from an orchard (Figure 7.1) and of SAT Math scores from entering students at a state university (Figure 7.2) have similar shapes – both are mound-shaped and roughly symmetric. The bell-shaped curves, called normal curves, drawn over the histograms in Figures 7.1 and 7.2, summarize the overall patterns in each of these data sets. Because normal curves are symmetric, the mean $\mu$ and median are the same point, at the line of symmetry for the curve.

![Figure 7.1 Histogram of apple weight.](image1)  ![Figure 7.2. Histogram of SAT Math scores.](image2)

Many distributions in the natural world exhibit this normal-curve shape including arrival patterns of migratory birds as they pass through Manomet Center for Conservation Sciences each spring and fall. The normal curves below show the patterns of migration data for the Blackpoll Warbler and Eastern Towhee separated by a period of 32 years.
Figure 7.3. Normal curves of arrival times for (a) Blackpoll Warbler and (b) Eastern Towhee in Years 1 and 33.

In both cases, the curves for Year 1 are taller, with more area underneath, than the curves for Year 33. This is because the vertical axis uses a frequency scale and the bird population passing through Manomet declined from Year 1 to Year 33. Notice that for the Blackpoll Warbler (a), the mean arrival time is the same for Years 1 and 33. However, the first arrival date is later for Year 33 than it is for Year 1. For the Eastern Towhee (b), the mean arrival time has shifted earlier but the first arrival time appears roughly the same in both years. The conclusion? Something is affecting migration as a whole – most likely climate change.

To better understand the changes in patterns for the Eastern Towhee and Blackpoll Warbler data, we convert the normal curves in Figure 7.3 to normal density curves in Figure 7.4 by changing the scaling on the $y$-axis from frequency to relative frequency or proportion. This makes the area under each curve equal to 1 (representing 100% of the data).

Figure 7.4. Normal density curves of arrival times for (a) Eastern Towhee and (b) Blackpoll Warbler.
Notice that for the Eastern Towhee the first arrivals are happening at about the same time in both years. What is different between the two years is the proportion (or percentage) of birds that have arrived by a particular time after the first arrivals. For example, in Year 33 half of the birds had arrived within 48 days of the spring equinox, while in Year 1 only 23% of the birds had arrived by this time (See Figure 7.5.).

**Figure 7.5. Proportion of Eastern Towhee that have arrived by day 48.**

The density curves for the Blackpoll Warblers show a consistent mean at 67 in both years. So half the birds had arrived by day 67 in both Year 1 and Year 33. If we compare the proportion (or percentage) of birds that had arrived by day 56, we find that a proportion of 0.10 or 10% of the birds had arrived by this date in Year 1 whereas in Year 33 the proportion is 0.04 or only 4% (See Figure 7.6.).

**Figure 7.6. Proportion of Blackpoll Warblers that have arrived by day 56.**
So, the percentage of birds that used to arrive by day 56 is more than double what it is now. Notice the later first arrival for Year 33 compared to Year 1. The only thing causing the observed later first arrival is that fewer birds are migrating – making it tougher for researchers to spot the rarer birds. The scientists at Manomet say it is important to take into account population sizes when looking at migration times especially if the only data available are those easily-influenced first arrival dates.
STUDENT LEARNING OBJECTIVES

A. Understand that the overall shape of a distribution of a large number of observations can be summarized by a smooth curve called a density curve.

B. Know that an area under a density curve over an interval represents the proportion of data that falls in that interval.

C. Recognize the characteristic bell-shapes of normal curves. Locate the mean and standard deviation on a normal density curve by eye.

D. Understand how changing the mean and standard deviation affects a normal density curve.
   • Know that changing the mean of a normal density curve shifts the curve along the horizontal axis without changing its shape.
   • Know that increasing the standard deviation produces a flatter and wider bell-shaped curve and that decreasing the standard deviation produces a taller and narrower curve.
NORMAL CURVES

Normal curves can be convenient summaries of data whose histograms are mound-shaped and roughly symmetric. The idea is to replace a histogram with a smooth curve that describes the overall shape in an idealized way. Because area under a histogram describes what part of the data lies in any region, the same is true of the curve. Density curves have areas under their curves of exactly 1, representing a proportion of the data of 1, or 100% of the data. The area under the curve above any interval on the x-axis represents the proportion of all of the data that lie in that region.

Normal curves are a particularly important class of density curves. Many sets of real data (but by no means all) are approximately normal. Normal curves can be described exactly by an equation, but we will not do this. Instead, we emphasize the characteristic symmetric, bell-like shape. Unlike distributions in general, any normal curve is exactly described by giving its mean and its standard deviation. The usual notation for the mean of a normal curve is $\mu$ (the Greek letter mu), and its standard deviation is $\sigma$ (the Greek letter sigma). Notice that we use different notation for the mean and standard deviation of a normal distribution than we used for the sample mean, $\bar{x}$, and sample standard deviation, s, which could be calculated from data. A density curve is an idealized description of the overall pattern, not a detailed description of the specific data. We often use normal curves to describe the distribution of a large population, such as the weights of apples or the wingspans of birds. The mean weight $\bar{x}$ of a specific sample of apples or birds generally does not exactly equal $\mu$, the mean weight of the population of all apples or all birds. So, a separate notation is needed.

We can find both the mean $\mu$ and standard deviation $\sigma$ by eye on a normal curve. The mean $\mu$ is the center of symmetry for the curve. To find $\sigma$, start at the center of the curve and run a pencil outward. At first, the curve bends downward, falling ever more steeply; but then the curve, while still falling, begins to level off and then bends upward. The point where the curvature changes from ever steeper to ever less steep is one standard deviation away from the mean. Figure 7.7 shows a normal curve on which both $\mu$ and $\sigma$ are marked.
As mentioned in the previous paragraph, the mean is the center of symmetry of a normal density curve. So, changing the mean simply shifts the curve along the horizontal axis without changing its shape. Take, for example, Figure 7.8, which shows hypothetical normal curves describing weights of two types of apples. The standard deviations of the two curves are the same, but their means are different. The mean of the solid curve is at 130 grams and the mean of the dashed curve is at 160 grams.

The standard deviation controls the spread of a distribution. So, changing $\sigma$ does change the
shape. For the normal curves in Figure 7.9, the standard deviation for the dashed and solid curves are 10 and 20, respectively. Notice that the normal curve with the smaller standard deviation, $\sigma = 10$, is taller and exhibits less spread than the normal curve with the larger standard deviation, $\sigma = 20$.

![Figure 7.9. Two normal curves with different standard deviations.](image-url)
KEY TERMS

A normal density curve is a bell-shaped curve. A density curve is scaled so that the area under the curve is 1. The center line of the normal density curve is at the mean $\mu$. The change of curvature in the bell-shaped curve occurs at $\mu - \sigma$ and $\mu + \sigma$.

A normal distribution is described by a normal density curve. Any particular normal distribution is completely specified by its mean $\mu$ and standard deviation $\sigma$. 
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. Describe the characteristic shape of a normal curve.

2. How can you spot the mean of a normal curve?

3. If one normal curve is low and spread out and another is tall and skinny, which curve has the larger standard deviation?

4. Focus on the distribution of arrival times for the Eastern Towhee for Years 1 and 33. Has the mean arrival date in Year 33 increased, decreased or remained the same as the mean in Year 1?

5. The mean of the arrival times for the Blackpoll Warbler passing through Manomet in Years 1 and 33 is roughly the same. In Year 33 has the percentage of birds that have arrived by day 56 increased, decreased or remained the same as what it was in Year 1?
UNIT ACTIVITY:
EXPLORING NORMAL DENSITY CURVES

Figure 7.10. Two normal density curves with different standard deviations.

1. Figure 7.10 shows two normal density curves. The standard normal density curve is the solid curve. It represents the normal distribution with mean \( \mu = 0 \) and standard deviation \( \sigma = 1 \). A vertical line has been drawn at \( \mu = 0 \), which marks the curve’s line of symmetry.

a. Between what two values on the horizontal axis would you expect nearly all data from the standard normal distribution to fall?

b. Between what two values would you expect nearly all of the lower 50% of standard normal data to fall?

2. a. The dashed curve in Figure 7.10 is a normal density curve with the same mean as the standard normal density curve but a different standard deviation. Is the standard deviation larger or smaller than 1? How can you tell from the graphs of the two density curves?

b. To estimate the standard deviation for the dashed normal curve in Figure 7.10, start at the top of the curve and follow the curve down along its right side. Above what value on the horizontal axis does the curve switch from bending downward to bending upward? (In other
words, at what point does the curve go from falling ever more steeply to falling less and less steeply?) Use this point to estimate the standard deviation for the dashed normal density curve.

Figure 7.11. Two normal density curves with different means.

3. Figure 7.11 shows two normal density curves, the standard normal density curve (solid curve) and another normal density curve (dashed curve) with the same standard deviation but a different mean.

a. What is the mean of the normal distribution represented by the dashed curve? How did you determine your answer from the graph?

b. Between what two values would you expect nearly all the data from the normal distribution shown by the dashed curve to fall? How are these numbers related to the curve’s mean and standard deviation?
4. Figures 7.12 (a – d) show four normal density curves. Match the density curves with each of the following means and standard deviations. Explain how you were able to match each density curve to its mean and standard deviation.

i. $\mu = 4, \; \sigma = 1$

ii. $\mu = 15, \; \sigma = 3$

iii. $\mu = 4, \; \sigma = 0.5$

iv. $\mu = 5, \; \sigma = 2$
EXERCISES

1. Height of 4-year-old boys is approximately normally distributed with mean $\mu = 40$ inches and standard deviation $\sigma = 1.5$ inches. An unscaled normal curve appears in Figure 7.13. Sketch a copy of this curve. On the horizontal axis, mark the location of the mean. Then mark points on the horizontal axis that are one standard deviation on either side of the mean.

![Height of 4-Year-Old Boys (inches)](image)

*Figure 7.13. Normal density curve for height of boys.*

2. IQ scores are normally distributed with mean 100 and standard deviation 15.

![IQ scores](image)

*Figure 7.14. Normal density curve for IQ scores.*
a. Sketch the graph in Figure 7.14. Then on the horizontal axis mark the mean, 100, and one standard deviation on either side of the mean. Label the horizontal axis as IQ Test Scores.

b. Scores from 90 to 110 represent normal or average intelligence. Scores above 120 represent very superior intelligence to genius. Shade the area under your normal curve from part (a) that represents the proportion of people with normal (or average) intelligence. Then shade, in a different color or on a copy of your graph from (a), the area that represents the proportion of people with very superior intelligence or genius intelligence.

c. Based on your shaded areas in (b), are there a higher proportion of people with normal (or average) intelligence or with very superior or genius intelligence? Explain.

3. Polished silicon wafers are one step in the production of microchips. No matter how carefully the slicing and polishing equipment is adjusted, there is some variability in the thickness of the polished wafers. Periodically samples of polished wafers are selected and wafer thickness is measured. Data on thicknesses (mm) of polished wafers from two samples each of size 50 appears below. The data from each sample have been sorted from smallest to largest.

Sample 1

0.372 0.379 0.399 0.400 0.407 0.408 0.412 0.416 0.417
0.418 0.418 0.424 0.430 0.434 0.434 0.434 0.438 0.442
0.453 0.453 0.455 0.456 0.462 0.464 0.466 0.466 0.472
0.473 0.475 0.477 0.487 0.489 0.493 0.495 0.498 0.499
0.510 0.511 0.511 0.518 0.521 0.526 0.529 0.531 0.535
0.536 0.552 0.553 0.562 0.590

Sample 2

0.362 0.366 0.394 0.396 0.400 0.409 0.411 0.412 0.414
0.419 0.420 0.420 0.422 0.423 0.425 0.431 0.438 0.449
0.450 0.458 0.460 0.467 0.471 0.471 0.478 0.483 0.483
0.486 0.489 0.496 0.500 0.508 0.530 0.534 0.537 0.548
0.549 0.583 0.586 0.592 0.593 0.593 0.609 0.615 0.615
0.630 0.638 0.661 0.662 0.666

a. Create a histogram for the data from Sample 1. Use class intervals that start at 0.35 and have width 0.05.
b. Does it seem reasonable that a normal curve could represent the thicknesses of polished wafers produced under these control settings? If so, superimpose a normal curve over your histogram from part (a). If not, superimpose a smooth curve of a different shape that summarizes the pattern in the histogram.

c. The target thickness is 0.5 mm. Does the balance point (mean) of your smoothed curve from part (b) appear to be at 0.5 mm, smaller than 0.5 mm or larger than 0.5 mm?

4. Repeat question 3, this time using the data from Sample 2.
1. Two graphs of normal density curves are shown in Figure 7.15.

![Figure 7.15. Two normal density curves.](image)

a. Which normal density curve, the solid curve or the dashed curve, represents the distribution with the larger mean $\mu$? Explain how you can tell from the graph.

b. Which normal density curve represents a distribution with the larger standard deviation $\sigma$? Explain how you can tell from the graph.
2. Figure 7.16 shows a density curve of a distribution that is not normal.

![Triangular density curve](image)

*Figure 7.16. A triangular density curve.*

a. What is the mean of this distribution? How can you tell from the graph?

b. The area under the density curve over a particular interval gives the proportion of data that fall in that interval. What proportion of data from this distribution would fall below 1.5? Explain your calculations.

c. What proportion of data would fall within 0.5 units of the mean? Explain how you arrived at your answer. Draw a diagram that helps explain how to get to your answer. Explain how your diagram works.

3. The distribution of the height of 6-year-old girls is approximately normal with mean $\mu = 115$ cm and standard deviation $\sigma = 4$ cm. An unscaled normal curve appears in Figure 7.17.
Figure 7.17. Normal density curve for heights of girls.

a. Sketch a copy of Figure 7.17, and then mark the mean on the horizontal axis. Mark the points on the curve one standard deviation on either side of the mean. Now, mark off the axis in centimeters so that your curve is the normal curve representing the height of 6-year-old girls.

b. For normal data, roughly 68% of the data falls within one standard deviation of the mean. Shade the area that represents this percentage.

c. Approximately what percentage of 6-year-old girls is shorter than one standard deviation from the mean? Shade the area under the normal curve that represents this percentage.

4. a. Figure 7.18 shows a histogram of data on an individual’s pulse rates (beats/minute). Based on the shape of the histogram, do you think this individual’s pulse rate follows a normal distribution? If yes, sketch the histogram and draw a normal curve over it. If no, explain why not.

See figure on next page...
b. Figure 7.19 shows a histogram of femur bone lengths. Based on the shape of the histogram, do you think femur bone lengths are normally distributed? If yes, sketch the histogram and draw a normal curve over it. If no, explain why not.

Figure 7.19. Histogram of femur bone length.

c. Figure 7.20 shows a histogram of Body Mass Index (BMI) of 6-year-olds. Based on the shape of the histogram, do you think 6-year-olds’ BMIs are normally distributed? If yes, sketch the histogram and draw a normal curve over it. If no, explain why not.
Figure 7.20. Histogram of BMIs from 6-year-olds.
SUMMARY OF VIDEO

In this video, we continue the discussion of normal curves that was begun in Unit 7. Recall that a normal curve is bell-shaped and completely characterized by its mean, \( \mu \), and standard deviation, \( \sigma \).

Figure 8.1. Normal curve specified by its mean and standard deviation.

Given the mean and standard deviation of a normal curve, we’d like to approximate the proportion of data that falls within certain intervals. The Empirical Rule or 68-95-99.7% Rule can give us a good starting point. This rule tells us that around 68% of the data will fall within one standard deviation of the mean; around 95% will fall within two standard deviations of the mean; and 99.7% will fall within three standard deviations of the mean. The standard deviation \( \sigma \) is a natural yardstick for any measurements that follow a normal distribution. For example, the distribution of the height of American women can be described by a normal curve with mean \( \mu = 63.8 \) inches and standard deviation \( \sigma = 4.2 \) inches. To illustrate how the standard deviation works as a yardstick, begin with a normal curve centered at 63.8. Now, using our
4.2-inch standard deviation yardstick, we measure one standard deviation on either side of the mean, giving 59.6 inches and 68.0 inches. It turns out that about 68% of the total area under the normal curve is over the interval from 59.6 to 68.0 as shown in Figure 8.2. This means that roughly 68% of women's heights are between 59.6 inches and 68.0 inches.

![Figure 8.2. Measuring one standard deviation from the mean.](image)

Now, measure out two standard deviations from the mean, from 55.4 inches to 72.2 inches. According to the Empirical Rule, roughly 95% of women's heights will fall within this interval (See Figure 8.3.).

![Figure 8.3. Measuring two standard deviations from the mean.](image)
And finally, measure out three standard deviations from the mean, from 51.2 inches to 76.4 inches. About 99.7% of all women's heights will fall within this interval. That leaves only 0.15% of women who are shorter than 51.2 inches and 0.15% of women who are taller than 76.4 inches.

Next, we take a trip to a meeting of the Boston Beanstalks, a social club for tall people — women must be at least 5 feet 10 inches (70 inches) and men at least 6 feet 2 inches (74 inches). To see how far out (in terms of height) the Boston Beanstalks' women are, we convert the entry height of 70 inches into a standard unit called a z-score by subtracting the mean of 63.8 and then dividing the result by the standard deviation of 4.2:

$$z = \frac{70 - 63.8}{4.2} = 1.48$$

This tells us that in order for a woman to be a member of the Boston Beanstalks Club, her height must be at least 1.48 standard deviations above the mean. Heights of men can be described by a normal curve with $\mu = 69.4$ inches and $\sigma = 4.7$ inches. Converting the entry height for males into a z-score gives:

$$z = \frac{74 - 69.4}{4.7} = 0.98$$

Changing normal data into z-scores transforms the data into standard normal data. The standard normal curve has mean $\mu = 0$ and standard deviation $\sigma = 1$. Using the standard normal distribution, Figure 8.4 compares our z-scores for men's and women's entry height into the Boston Beanstalks. In order for men to join, they only need to be about one standard deviation above the men's mean height. The height requirements for women are more stringent than for men. Women need to be a half standard deviation further from the mean than their male counterparts.
Based on the 68-95-99.7% Rule, approximately 16% of men are tall enough to join the club (leaving roughly 84% of the men who are too short to join). But what about women? Again using the 68-95-99.7% Rule, we know that the percent of women tall enough to join is somewhere between 2.5% and 16%. In order to get a more accurate answer, we can turn to z-tables for the standard normal distribution. A z-table tells us how much of the distribution falls below any z-value. Here’s how to use the table. For a z-score of 1.48, we follow the z column in Figure 8.5 down to 1.4 and then move right until we are in the .08 column. We can now read off the proportion of women who are too short to join the Beanstalks, 0.9306 or 93.06%. That means that only about 7% of women are tall enough to join the Beanstalks compared to around 16% of men.
We can also use z-scores to compute the percent of data that falls in an interval between two values. For example, suppose we want to know the proportion of American women who are taller than our host Pardis Sabeti, who is 64.5 inches tall, but still not tall enough to make it into the Boston Beanstalks. The z-score for a height of 64.5 inches is 0.16. Recall the z-score for women’s entry into the Boston Beanstalks is 1.48. So, we need the area under the standard normal curve between 0.16 and 1.48, which is shown in Figure 8.6. We know the proportion of women shorter than the cutoff for the Beanstalks is 0.9306. Using the table in Figure 8.5, we find that the proportion of women shorter than Dr. Sabeti is 0.5636. We get the desired proportion by subtracting these two proportions:

\[
0.9306 - 0.5636 = 0.3670
\]
So, approximately 36.7% of American women are taller than Pardis Sabeti but too short to be accepted by the Boston Beanstalks.

Figure 8.6. Computing the proportion between z-scores of 0.16 and 1.48.
STUDENT LEARNING OBJECTIVES

A. Be able to use the Empirical Rule (68-95-99.7% Rule) to approximate the proportions of normal data falling in certain intervals.

B. Understand that standardizing (by subtracting the mean and dividing by the standard deviation) allows us to compare observations from different normal distributions.

C. Know that in order to use a standard normal table to do calculations involving normal distributions, we must first standardize measurements.

D. Be able to use the standard normal table (z-table) to find the proportion of observations below any value of $z$.

E. Be able to combine standardization with use of the z-table to find the proportion above, below, or between given values in any normal distribution.

F. Be able to use software such as Excel, Minitab, or SPSS, or graphing calculators to perform the calculations in D and E without having to convert to z-scores.
Normal density curves are bell-shaped curves, which have been scaled so that the area under the curve is 1. A normal density curve is completely determined by its mean, \( \mu \), and standard deviation \( \sigma \). Figure 8.7 shows two normal density curves, the solid curve has mean \( \mu = 5 \) and standard deviation \( \sigma = 3 \) and the dashed curve has mean \( \mu = 8 \) and standard deviation \( \sigma = 2 \). Vertical line segments mark the means of the two normal distributions and horizontal line segments mark the standard deviations.

![Figure 8.7. Comparing two normal density curves.](image)

The standard deviation acts as a natural yardstick for any measurement that follows a normal distribution. We can summarize use of the standard deviation as a yardstick with the Empirical Rule, also known as the 68-95-99.7% Rule.

**Empirical Rule or 68-95-99.7% Rule**

In any normal distribution with mean \( \mu \) and standard deviation \( \sigma \):

- Approximately 68% of the data fall within one standard deviation of the mean.
- Approximately 95% of the data fall within two standard deviations of the mean.
- Approximately 99.7% of the data fall within three standard deviations of the mean.
Figure 8.8 below shows the percentage of normal data falling within one, two, and three standard deviations from the mean.

![Figure 8.8. Graphic of 68-95-99.7% Rule.](image)

We can apply the Empirical Rule to the normal distribution shown by the solid curve in Figure 8.7. Since this curve has mean 5 and standard deviation 3, the Empirical Rule tells us that roughly 68% of data from this distribution will fall between 2 and 8, roughly 95% of the data will fall between -1 and 11 and nearly all (99.7%) of the data will fall between -4 and 14. If, instead, we focus on the dashed curve in Figure 8.7, we get the same percentages if we form intervals one, two, and three standard deviations on either side of the mean.

The Empirical Rule suggests that the difference between two normal curves is only a matter of scaling. All normal distributions are the same when we measure how many standard deviations an observation $x$ lies away from the mean, which we calculate as follows:

$$z = \frac{x - \mu}{\sigma}$$

The standardized value for $x$, calculated by the formula above, is called its z-score. Observations from different normal distributions are best compared by comparing their standardized values, or z-scores. The z-score states how many standard deviations the original observation falls away from the mean and in which direction. Observations larger than the mean have positive z-scores, while observations smaller than the mean have negative z-scores.
Now, return to the dashed normal curve in Figure 8.7. When we transform the numbers on the horizontal scale into a z-score scale, we get the standard normal curve, with mean $\mu = 0$ and standard deviation $\sigma = 1$, as shown in Figure 8.9. If we switch to the solid curve in Figure 8.7 and transform its horizontal scale into a z-score scale (subtract 5 and divide the result by 3), we would also wind up with the standard normal curve. So, any two normal curves can be changed into the standard normal curve by transforming the scale on the horizontal axis into a z-score scale.

![Figure 8.9. Transforming the horizontal scale to a z-score scale.](image_url)

Converting to standardized values allows us to find proportions that we can't get from the Empirical Rule. For example, suppose we want to know the percentage of data from a normal distribution with mean $\mu = 8$ and standard deviation $\sigma = 2$ (Figure 8.7, original scale) that falls below $x = 9$. We can use the Empirical Rule to learn that the percentage is between 50% and 84%, but that is not a very accurate estimate. Instead, we convert $x = 9$ into a z-score:

$$z = \frac{9 - 8}{2} = 0.5$$

Now, we find the proportion of standard normal data that falls below 0.5. To find this proportion, we use a standard normal table, which gives the proportion of data that falls below any value for $z$. Using the portion of a z-table shown Figure 8.5, we look down the z-column to locate 0.5 and then move to the right under the .00 column. Our answer is 0.6915, or around 69.15%.
We can use the same technique to find the percentage of data that fall between two values such as $x = 9$ and $x = 13$. This proportion is represented by the shaded area in Figure 8.10. First, we need to convert $x = 13$ into a $z$-score:

$$z = \frac{13 - 8}{2} = 2.5$$

To find the proportion of standard normal data that falls below $z = 2.5$, we consult a $z$-table, which gives 0.9938. We already have determined that the proportion of standard normal data that falls below $z = 0.5$ is 0.6915. All that is needed to find the shaded gray area in Figure 8.10 is to subtract these two proportions:

$$\text{proportion} = 0.9938 - 0.6915 = 0.3023$$

So, we find that just over 30% of observations from a normal distribution with $\mu = 8$ and $\sigma = 2$ will fall between $x = 9$ and $x = 13$.

*Figure 8.10. Finding the area under the curve from $x = 9$ to $x = 13$.*
The Empirical Rule or 68-95-99.7% Rule gives the approximate percentage of data that fall within one standard deviation (68%), two standard deviations (95%), and three standard deviations (99.7%) of the mean. This rule should be applied only when the data are approximately normal.

An observation $x$ from a normal distribution with mean $\mu$ and standard deviation $\sigma$ can be transformed into a standardized value called a z-score as follows:

$$z = \frac{x - \mu}{\sigma}$$

A standard normal distribution is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. 

---

**KEY TERMS**
Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is another name for the Empirical rule?

2. How tall must a woman be to join the Boston Beanstalks Club?

3. How do you calculate a $z$-score?

4. Based on $z$-scores, are eligibility requirements to join the Boston Beanstalks more difficult to meet for men or for women?
UNIT ACTIVITY:
USING AREA TO ESTIMATE STANDARD NORMAL PROPORTIONS

Figure 8.11. The standard normal density curve.

1. Estimate the number of rectangles that fit under the standard normal density curve in Figure 8.11. The area under any density curve is 1, which represents a proportion of all the data or 100% of the data.

2. Using the rectangles, you will estimate the proportion of data from a standard normal distribution that falls below 0. (You should already know the answer. So, this question will help you check the process.)

   a. Shade the area under the curve that lies to the left of \( z = 0 \) on the normal curve below. Count the number of rectangles contained in the shaded region.
b. To find the proportion of data that falls below $z = 0$, divide the number of rectangles you got in (a) by the total number of rectangles counted for question 1. What is your estimate for this proportion?

3. Next, using the rectangles, you will estimate the proportion of data that falls below $z = -1$.

   a. Shade the area under the curve that lies to the left of $z = -1$ on the normal curve below. Count the number of rectangles contained in the shaded region.

   ![Normal Curve Diagram]

   b. To find the proportion of data that falls below $z = -1$, divide the number of rectangles you got in (a) by the total number of rectangles from question 1. What is your estimate for this proportion?

4. Repeat question 3 using $z = -2$. Show your calculations.

   a.

   ![Normal Curve Diagram]

   b. Proportion = ________________
5. Repeat question 3 using \( z = 1 \).

a.

\[
\begin{array}{c|cccccccccc}
\hline
z & -3.0 & -2.5 & -2.0 & -1.5 & -1.0 & -0.5 & 0.0 & 0.5 & 1.0 & 1.5 \\
\hline
P(z) & 0.00 & 0.05 & 0.10 & 0.15 & 0.20 & 0.25 & 0.30 & 0.35 & 0.40 \\
\hline
\end{array}
\]

b. Proportion = ______________

6. Repeat question 3 with \( z = 2 \).

a.

\[
\begin{array}{c|cccccccccc}
\hline
z & -3.0 & -2.5 & -2.0 & -1.5 & -1.0 & -0.5 & 0.0 & 0.5 & 1.0 & 1.5 \\
\hline
P(z) & 0.00 & 0.05 & 0.10 & 0.15 & 0.20 & 0.25 & 0.30 & 0.35 & 0.40 \\
\hline
\end{array}
\]

b. Proportion = ______________

7. Now it is time to compare your estimates with the proportions given in the standard normal table (z-table). Given any value of \( z \), the standard normal table will tell you the proportion of data that falls at or below the given value.

a. Use a standard normal table to complete the entries in column two of Table 8.1.

b. Enter your estimates from questions 2 – 6 into column three of Table 8.1.
Table 8.1. Comparisons of proportions from a z-table and estimates.

c. How close were your estimates of the area under the standard normal curve to the actual proportions from the z-table?
EXERCISES

1. The distribution of heights of young women (18 – 24 years old) is approximately normal with mean \( \mu = 65.5 \) inches and standard deviation \( \sigma = 2.5 \) inches.

a. Between what two values do the heights of the central 95% of young women lie? Draw a normal curve and shade this area under the normal curve.

b. Based on the 68-95-99.7% Rule, explain why only about 2.5% of young women are more than two standard deviations taller than the mean.

c. What is the z-score for a 20-year-old woman who is 6 feet tall? Interpret this value.

2. There are two national college-entrance examinations, the SAT and the American College Testing program (ACT). Scores on individual SAT exams are approximately normal with mean 500 and standard deviation 100. Scores on the ACT exams are approximately normal with mean 18 and standard deviation 6. (These are the distributions for a large reference population of students used to set standards that remain the same from year to year.)

a. What percent of all SAT scores are above 600? Explain how you arrived at your answer.

b. Julie’s SAT Math score is 630. John’s ACT Math score is 22. Find the standardized scores for both Julie and John. Assuming that both tests measure the same kind of ability, who has the higher score?

3. Do you think that the distribution of the selling price of houses in a city is close to normal? Why or why not? Would you be willing to apply the 68-95-99.7% Rule to house prices? Explain.

4. Find the percentage of observations from a standard normal distribution that satisfy the following.

a. \( z < -1 \)

b. \( z \geq 2.25 \)

c. \(-1 < z < 2.25 \)
1. A study investigated the effect of car speed on accident severity. As part of the study, the vehicle speed at impact was extracted from a sample of 6000 accident reports of fatal automobile accidents. Analysis revealed that vehicle speed at impact could be described by a normal distribution with mean $\mu = 44$ mph and standard deviation $\sigma = 14$ mph.

a. Draw a normal curve that represents the speed at impact for fatal accidents. Be sure to add appropriate scaling to the horizontal axis. (For example, use the information on mean and standard deviation to mark the center of the curve and indicate its spread.)

Use the Empirical Rule to answer parts (b) – (d).

b. Approximately what proportion of vehicle speeds were between 30 mph and 58 mph?

c. Approximately what proportion of vehicle speeds were less than 30 mph?

d. Approximately what proportion of vehicle speeds exceeded 72 mph?

2. A teacher gives two different statistics tests, but one is harder than the other. Scores on test A have mean 78 and standard deviation 6, and scores on test B have mean 65 and standard deviation 9. Carrie scored 79 on test B and Pat scored 85 on test A. Who had the higher standardized score?

3. The Army finds that the head sizes (forehead circumference) of soldiers vary according to the Normal distribution with mean $\mu = 22.8$ inches and standard deviation $\sigma = 1.1$ inches.

a. What proportion of soldiers have head size at least 21 inches? Explain how to use a standard normal table to answer this question. Then verify your answer using technology.

b. What proportion of soldiers have head size between 21 inches and 23 inches. Explain how to use a standard normal table to answer this question. Then verify your answer using technology.

4. Assume that in recent years the arrival time (in days since the spring equinox) for Blackpoll Warblers at Manomet Center for Conservation Sciences follows a normal distribution with mean $\mu = 67$ and standard deviation $\sigma = 5.5$.

a. Sketch a normal curve that represents the arrival time of Blackpoll Warblers passing
through Manomet. Add an appropriate scale to the horizontal axis that uses the mean and standard deviation.

b. What percentage of Blackpoll Warblers passing through Manomet Center arrived before day 60?

c. What percentage of Blackpoll Warblers arrived after day 70?

d. What percentage of Blackpoll Warblers arrived between day 60 and day 70?
SUMMARY OF VIDEO

Examples of normal distributions can be found in all sorts of settings. Remember normal curves are symmetric and bell-shaped. But in the real world, data can be a lot messier than the idealized examples you might find in a textbook. At times it can be difficult to eyeball whether or not data are normally distributed. Down the road, it will be important to feel confident in our assumption that a particular data distribution is, in fact, normal because that is a prerequisite for some more advanced statistical techniques.

We begin our study into methods of sorting out normal data from data that is not normal at Pete & Gerry’s organic egg farm. Plenty of data get collected at the egg farm: for example, how much water and feed the birds are eating, how they are growing, how many and what sizes of eggs they are laying, and how production is running on the packaging line. Are any of the data distributions they see normal?

Let’s start by taking a look at the weights of 7-week-old hens in one flock. Pete & Gerry’s Jesse Laflamme explains that they get them as day-old chicks and they are essentially the same size. But as they grow, Pete & Gerry’s test the young hens by taking a random sample of around 100 pullets. They weigh the hens and then graph the weights from the sample. They expect the graph to have the shape of a normal curve with a target weight of where they expect the middle of that curve to be. Figure 9.1 shows a histogram of the weights from one sample.
Figure 9.1. Weights of 7-week-old hens.

The histogram for this flock does appear to be normal with the one peak in the middle at around 550 grams. The normal curve drawn over the histogram in Figure 9.2 seems to be a pretty good fit.

Figure 9.2. Overlaying a normal curve.

It’s important to consider the class size when we are eyeballing a histogram to see if the data are normal. Sometimes, changing the class size can really change the way the histogram looks and what once appeared perfectly bell-shaped now looks quite different. The histogram of hen weights in Figure 9.3 looks less like normal data than the histogram of the same data in Figure 9.2.
A second option as we assess normality is to use the same hen weights to construct a boxplot. Boxplots can act as another graphical display test to see if our data are normally distributed. If a distribution is normal, we would expect to see the box containing the middle 50% of the data to be pretty tightly grouped in the center of the distribution, with longer whiskers indicating the increased spread of the upper and lower quarters of the data. In Figure 9.4, take a look at how a truly normal distribution translates into a boxplot and compare it with our hens. The weight distribution does appear to be approximately normal, with the whiskers each longer than the Q1 to median distance and the median to Q3 distance.

Going beyond eyeballing these kinds of graphic displays, there is another more precise way to check whether a distribution is approximately normal. Statisticians use software to construct what is known as a normal quantile plot. The basic idea is to compare the ordered data values you have with values you would expect from a standard normal distribution. If your data are normally distributed, the normal-quantile-plot points will fall close to a straight line.
Since a computer will do the work for you, it is less important to understand the steps taken to construct the normal quantile plot than it is to know how to interpret it. Figure 9.5 shows the normal quantile plot for the hen weights. Our observed weights are on the x-axis and the expected values on the y-axis. The pattern of dots in the plot lies close to a straight line. So we can conclude that our data come from an approximate normal distribution.

![Normal quantile plot for hen weights.](image)

**Figure 9.5. Normal quantile plot for hen weights.**

Next, we try out these tests with another size range – egg weights. In the wild, we would expect the size of one bird species’ eggs to be normally distributed. Figure 9.6 shows a histogram of the egg weights for a day’s worth of eggs from Pete & Gerry’s.

![Histogram of egg weights.](image)

**Figure 9.6. Egg weights from a week’s worth of data from Pete & Gerry’s.**

It is a little challenging to decide if Pete & Gerry’s egg size distribution looks normal. Could this be an example of messy real-world numbers? Or has the farm’s careful control over hatching and breeding had an influence on the size range? To find out, we investigate using the normal quantile plot for these data, which is shown in Figure 9.7.
The middle of the plot looks like what you would expect from normal data (roughly a straight-line pattern). But the lower tail of the distribution does not look at all linear. That lower tail shows us that we have more eggs at the lower weight range than we would expect to see if the distribution were actually normal. We can conclude that the egg size distribution at Pete & Gerry’s is not normal, at least on the day that these data were collected. That is logical if you understand that the size of the eggs a chicken lays increases over her lifetime. The egg business has seasonal cycles.

For instance, sales increase for the year-end holidays when people are stocking up for their baking needs. Pete & Gerry’s tries to prepare for those cycles by knowing the age of the hens they will need laying to meet that heightened demand for the most desirable egg sizes, large and extra large. On the day these data were recorded, months before the peak season when the demand for large eggs is highest, there were more younger flocks laying smaller eggs.

As you get more familiar with normal quantile plots, you might start to recognize predictable patterns for various non-normal distributions. For instance, Figure 9.8 shows a histogram that is skewed to the right, indicating that there were a majority of very young birds laying smaller eggs. The normal quantile plot in this case, shown in Figure 9.9, is curved with a concave down pattern.
Figure 9.8. Histogram skewed to right. Figure 9.9. Normal quantile plot concave down.

On the other extreme, Figure 9.10 shows how things might look for a weight distribution of eggs laid by much older chickens. The histogram would be skewed to the left. The normal quantile plot is again curved but this time it is concave up.

Figure 9.10. Histogram skewed left.

So, don’t count your chickens before they hatch. Or, in this case, don’t count your eggs before they are laid, at least when it comes to assuming data are normal!
STUDENT LEARNING OBJECTIVES

A. Be familiar with the characteristic shape of histograms and boxplots of normal data.

B. Know how to calculate percentiles from a normal distribution.

C. Understand how normal $n$-quantiles are used to divide the area under a normal curve into $n$-equal areas.

D. Understand the basic construction of a normal quantile plot.

E. Use a normal quantile plot to assess whether data are from a normal distribution.
Normal distributions are a very important class of statistical distributions and certainly the most common distributions you will encounter in this course. Data produced by many natural processes can be described as normally distributed. Some examples include human bone lengths, people’s IQ scores, bird weights, and flower lengths. Later in this course, you will encounter statistical procedures that are based on an assumption of normality. So, how do we decide whether it is reasonable to assume data come from a normal distribution?

Not all mound-shaped, symmetric data turn out to be normally distributed. Furthermore, histograms and boxplots of normal data don’t always fit the pattern we expect. For example, take a look at the histogram and boxplot in Figures 9.12 and 9.13.

![Figure 9.12. Histogram of normal data.](image1)

![Figure 9.13. Boxplot of normal data.](image2)

The histogram looks more ragged than the characteristic mound-shaped, symmetric histogram we might expect from normal data. From the boxplot, the inner 50% of the data do not appear symmetric about the median, again a characteristic we would expect from normal data. So, we might conclude that these data do not come from a normal distribution – but we would be wrong! Particularly when the sample size is small, here \( n = 20 \), histograms and boxplots may not provide sufficient information to decide whether or not data are normally distributed. So, we need another tool to differentiate normal data from non-normal data – a normal quantile plot. This plot compares the ordered data with what we would expect from perfectly normal data.

Generally, we use technology to create normal quantile plots, which are difficult and tedious to construct by hand. However, we will discuss a simplified version of normal quantile plots. But first, we need some background on percentiles and quantiles.
Finding Percentiles

In Unit 8, Normal Calculations, we discussed how to use a z-table to find the proportion of standard normal data that would fall below a specific value of z. That was done by finding the area under the standard normal curve that lies to the left of that specific z-value. For example, if we use the partial z-table that appears in Figure 8.5 of Unit 8, we learn that the area to the left of \( z = 1.25 \) is 0.8944. So, we know that around 89.44% of standard normal data falls below 1.25. (See Figure 9.14.)

![Standard Normal Density Curve](image)

Figure 9.14. Finding the area that lies to the left of 1.25.

For percentiles, we work this process in reverse – we start with the area or percentage, and then determine the corresponding value of z, which we call a **percentile**. For example, finding the 50\(^{th}\) percentile of a standard normal distribution is easy – that’s the value of z for which the area under the density curve to its left is 0.50: \( z = 0 \). To find the 70th percentile of a standard normal distribution, we need to find the z-value such that the area under the curve to the left of that z-value is 0.70. Table 9.1 shows a portion of the z-table. Start in the body of the table (which contains the areas) and find numbers as close to 0.70 as possible: 0.6985 and 0.7019. Then read off the corresponding z-values: 0.52 and 0.53. So, we know the 70th percentile is between \( z = 0.52 \) and \( z = 0.53 \).

<table>
<thead>
<tr>
<th>( z )</th>
<th>0</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5000</td>
<td>0.5040</td>
<td>0.5080</td>
<td>0.5120</td>
<td>0.5160</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5398</td>
<td>0.5438</td>
<td>0.5478</td>
<td>0.5517</td>
<td>0.5557</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5793</td>
<td>0.5832</td>
<td>0.5871</td>
<td>0.5910</td>
<td>0.5948</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6179</td>
<td>0.6217</td>
<td>0.6255</td>
<td>0.6293</td>
<td>0.6331</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6554</td>
<td>0.6591</td>
<td>0.6628</td>
<td>0.6664</td>
<td>0.6700</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6915</td>
<td>0.6950</td>
<td><strong>0.6985</strong></td>
<td><strong>0.7019</strong></td>
<td>0.7054</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7257</td>
<td>0.7291</td>
<td>0.7324</td>
<td>0.7357</td>
<td>0.7389</td>
</tr>
</tbody>
</table>

*Table 9.1. Partial z-table.*
Using technology to find percentiles is easier and more accurate than using the z-table. Just use the inverse function of what you used to find the proportions (areas). (For example, in Excel use Norm.Dist to find the area to the left of $z = 1.25$ and use Norm.Inv to find the 70th percentile, the $z$-value associated with area 0.70.) Figure 9.15 shows the result using Minitab's Probability Distribution Plot, which gives $z = 0.5244$ as the 70th percentile of the standard normal distribution.

![Standard Normal Density Curve](image)

*Figure 9.15. Specify area 0.7 and read off percentile $z = 0.5244$.*

**Finding Normal Quantiles**

Quantiles are points of a distribution taken at regular percentile intervals. Suppose we want to find the 10-quantiles, or deciles, of the standard normal distribution. That means we need to determine 9 $z$-values that divide the $x$-axis into 10 intervals such that the areas over consecutive intervals are equal. In this case, the area is 0.10. So, to find the deciles, we need to find the 10th, 20th, 30th, . . . 90th percentiles, which are:

-1.282  -0.842  -0.524  -0.253  0.000  0.253  0.524  0.842  1.282

Figure 9.16 shows the 10-quartiles dividing the area under the standard normal curve into equal areas.
A normal quantile plot is a graphical method for assessing whether data come from a normal distribution. The plot compares the ordered data with what would be expected of perfectly normal data, which in a simplified version of the plot will be the normal quantiles. If the data are normally distributed, then the dots on a normal quantile plot will fall close to a straight line. We’ll test this out with the following data:

4.45  5.76  5.81  5.34  4.68  5.15  6.25  4.88  4.19

Next, we construct a normal quantile plot in order to decide whether these data come from a normal distribution. Here are the instructions for making a simplified version of a normal quantile plot.

### Constructing a Simplified Normal Quantile Plot

**Step 1:** Given a sample of $n$ data values, order the data from smallest to largest.

**Step 2.** Find the $(n+1)$-quantiles. This gives $n$ $z$-values that divide the area under the normal density curve into $n + 1$ equal areas of size $1/(n+1)$.

**Step 3.** Plot the quantiles versus the ordered data values.
We begin construction of the normal quantile plot by ordering the data from smallest to largest:

\[
4.19 \quad 4.45 \quad 4.68 \quad 4.88 \quad 5.15 \quad 5.34 \quad 5.76 \quad 5.81 \quad 6.25
\]

Since there are nine data values, we find the \((9 + 1)\)-quantiles of a standard normal distribution, which we have already done (See Figure 9.16.). The final step is to plot the normal quantiles versus the ordered data. In other words, plot \((4.19, -1.282), \ldots, (6.25, 1.282)\), which gives us the plot in Figure 9.17.

![Figure 9.17. Simplified normal quantile plot.](image)

Now, compare Figure 9.17 with the normal quantile plot in Figure 9.18, which was constructed using the statistical software Minitab. The patterns of the dots in the two plots are similar even though the scaling on the vertical axes differs. Minitab adds a line to help us see whether the pattern is linear. In addition, Minitab adds some curved bands that help us decide when the points are straying too far from the line compared to what is expected of normal data.

![Figure 9.18. Normal quantile plot using Minitab.](image)
Based on the normal quantile plot (either Figure 9.17 or Figure 9.18), it seems reasonable to assume these data are normally distributed.

For data that are not normal, the normal quantile plot can help us determine how the data deviate from normality. For example, normal quantile plots can tell us that data are skewed. In Figures 9.8 – 9.11, we see that if data are skewed to the right, then normal quantile plots are concave down, and if data are skewed to the left, then normal quantile plots are concave up. Instead, suppose data come from the uniform distribution pictured in Figure 9.19, which compares a uniform density curve to a normal density curve with the same mean.

![Figure 9.19. Comparing a normal density curve to a uniform density curve.](image)

Data from this uniform distribution will be fairly evenly spread out between 0 and 1. There are no values out in the tails (data values less than 0 or greater than 1) as you would expect from normal data. So, how does the lack of data in the tails show up in a normal quantile plot? Take a look at Figure 9.20, which is a normal quantile plot for 20 data values from this uniform distribution.

![Figure 9.20. Normal quantile plot of data from a uniform distribution.](image)
In the middle of the plot, the dots stay close to the line and hence are similar to what we would expect from normal data. However, the data at the extremes, the smallest and largest data values, are associated with curved patterns in the normal quantile plot. Consider the largest data value, marked with the double circle. The $x$-value associated with this point is 0.98. In order for this point to lie on the line, that $x$-value would need to be about 1.2. So, the largest uniform data values are too small compared to what we would expect from normal data values, which explains the pattern of an upward bend at the right end of the graph. On the other extreme, the smallest uniform data values are too large compared to what we would expect from normal data values, which explains the downward bend at the left end of the graph.
KEY TERMS

A percentile of a distribution is a value such that a certain percentage of observations from the distribution fall at or below that value. For example, a 16th percentile of a standard normal distribution is a value such that 16% of the area under the standard normal curve falls at or below that value. According to the 68-95-99.7% Rule, that value should be around $z = -1$, as shown in Figure 9.21.

Figure 9.21. The 16th percentile of a standard normal distribution.

Quantiles are points of a distribution taken at regular percentile intervals. For example, the 4-quantiles, or quartiles, are the 25th, 50th, and 75th percentiles. The quartiles for a standard normal distribution are shown in Figure 9.22.

Figure 9.22. Quartiles of a standard normal distribution.
A normal quantile plot (also known as a normal probability plot) is a graphical method for assessing whether data come from a normal distribution. The plot compares the ordered data with what would be expected of perfectly normal data. A fairly linear pattern in a normal quantile plot suggests that it is reasonable to assume that the data come from a normal distribution.
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is the shape of a normal distribution curve?

2. What characteristics do you expect to see in a histogram of normal data?

3. What characteristics do you expect to see in a boxplot of normal data?

4. What pattern in a normal quantile plot tells you that data come from a normal distribution?

5. Suppose a normal quantile plot has a curved, concave down pattern. Would you expect a histogram of the data to be symmetric, skewed to the right, or skewed to the left?
UNIT ACTIVITY:
WORKING WITH NORMAL QUANTILE PLOTS

In questions 1 – 4 you are given a histogram and normal quantile plot for each of four datasets. Compare the histograms to the normal quantile plots for each dataset. These activity questions should give you a better idea of how to interpret patterns in normal quantile plots. For the rest of the activity, you will construct normal quantile plots to assess whether data distributions are normal.

1. Figure 9.23 shows a histogram and normal quantile plot for 20 data values collected on a variable $u$.

![Histogram and normal quantile plot](image)

*Figure 9.23. Histogram (a) and normal quantile plot (b) for 20 data values.*

a. Describe the shape of the histogram.

b. Does the pattern of the dots in the normal quantile plot look fairly straight? If not, describe the general overall pattern of the dots, or if the overall pattern is straight, describe departures from the straight-line pattern.

c. Do you think it is reasonable to assume these data come from a normal distribution? Explain.

2. Figure 9.24 shows a histogram and normal quantile plot for 20 data values collected on a variable $v$. Repeat question 1 for variable $v$. 

Unit 9: Checking Assumptions of Normality | Student Guide | Page 18
Figure 9.24. Histogram (a) and normal quantile plot (b) for 20 data values.

3. Figure 9.25 shows a histogram and normal quantile plot for 20 data values collected on a variable \( w \). Repeat question 1 for variable \( w \).

Figure 9.25. Histogram (a) and normal quantile plot (b) for 20 data values.

4. Figure 9.26 shows a histogram and normal quantile plot for 20 data values collected on a variable \( x \). Repeat question 1 for variable \( x \).
Figure 9.26. Histogram (a) and normal quantile plot (b) for 20 data values.

Next, you will decide whether it is reasonable to assume that data collected from two studies are normally distributed.

5. The length in minutes of 20 calls made to a call center are given below:
   
   25  27  135  36  4  83  39  34  233
   33  241  186  246  53  30  47  32  118  1

   Construct a normal quantile plot for these data. Based on your plot, is it reasonable to assume that call lengths are normally distributed? Explain.

6. The average daily calories consumed by a sample of men are as follows:
   
   2716  2754  2484  2995  2635  2296  2741  3262  2572
   3371  2778  3041  3045  2888  2908  3457  3109  2977

   Construct a normal quantile plot for these data. Based on your plot, is it reasonable to assume that the average daily calories consumed by men are normally distributed? Explain.

7. Collect some data on two or more quantitative variables. The data can be from your class, from an online source (such as sports statistics), or from another source. Make histograms and normal quantile plots for each dataset. Decide which of your data could be described as approximately normally distributed.
1. Find the following percentiles for a standard normal distribution.
   a. 5th percentile.
   b. 10th percentile.
   c. 90th percentile.
   d. 95th percentile.
   e. What is the relationship between the 5th and 95th percentile? What is the relationship between the 10th and 90th percentile?

2. For data from a normal distribution, \( \bar{x} \) and \( s \) are appropriate measures of center and spread, respectively. However, for non-normal data, it is better to use a five-number summary to describe center and spread. In Unit 6, you were asked to compute \( \bar{x} \) and \( s \) for data on soldiers’ head size.

\[
\begin{array}{cccccccc}
23.0 & 22.2 & 21.7 & 22.0 & 22.3 & 22.6 \\
22.7 & 21.5 & 22.7 & 24.9 & 20.8 & 23.3 \\
24.2 & 23.5 & 23.9 & 23.4 & 20.8 & 21.5 \\
23.0 & 24.0 & 22.7 & 22.6 & 23.9 & 21.8 \\
23.1 & 21.9 & 21.0 & 22.4 & 23.5 & 22.5 \\
\end{array}
\]

a. Draw a histogram of these data. Then sketch a normal curve over the histogram. Based on your histogram, does it appear reasonable that these data come from a normal distribution? Explain.

b. Represent these data with a boxplot. Describe what features of the boxplot would lead you to believe that these data are normally distributed.

c. Not all data that is mound-shaped and roughly symmetric is normal. Make a normal quantile plot of these data. Based on your plot, is it reasonable to assume these data come from a normal distribution? Explain.
3. Three normal quantile plots are displayed in Figures 9.27 – 9.29. In parts (a) – (c) match the histogram of 50 data values to the normal quantile plot.

![Normal Quantile Plot #1](image1)

**Figure 9.27. Normal Quantile Plot #1.**

![Normal Quantile Plot #2](image2)

**Figure 9.28. Normal Quantile Plot #2.**

![Normal Quantile Plot #3](image3)

**Figure 9.29. Normal Quantile Plot #3.**
a. Match the histogram of data on variable $v$ (Figure 9.30) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.

![Histogram of data values for variable v.](image)

Figure 9.30. Histogram of data values for variable $v$.

b. Match the histogram of data on variable $w$ (Figure 9.31) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.

![Histogram of data values for w.](image)

Figure 9.31. Histogram of data values for $w$.

c. Match the histogram of data on variable $x$ (Figure 9.32) with one of the normal quantile plots in Figures 9.27 – 9.29. Justify your choice.
Figure 9.32. Histogram of data on variable x.
**Review Questions**

1. The 5-quartiles are called quintiles.
   
a. What proportion of standard normal data falls between two consecutive quintiles?

b. What are the quintiles for a standard normal distribution?

2. IQ scores follow a normal distribution with mean $\mu$ and standard deviation $\sigma$. For each of the following questions, find the percentile. Then draw a graph of the normal distribution of IQ scores and shade the area associated with the percentile.
   
a. What is the 25$^{th}$ percentile of IQ scores?

b. What is the 50$^{th}$ percentile of IQ scores?

c. What is the 75$^{th}$ percentile of IQ scores?

d. How are the 25$^{th}$ and 75$^{th}$ percentiles related?

The aging population is of interest to states because of costs connected with caring for older citizens. Table 9.2 gives the population in thousands of residents 85 or older in each state and the District of Columbia. But it isn’t just the population size of older residents that is problematic. More important is the percentage of a state’s residents who are older. Questions 3 and 4 are based on the data in Table 9.2, which contains information on the population size and percentage of state residents who are 85 years old or older.

<table>
<thead>
<tr>
<th>State</th>
<th>Population 85 and Over (thousands)</th>
<th>Percent</th>
<th>State</th>
<th>Population 85 and Over (thousands)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>76</td>
<td>1.58%</td>
<td>Montana</td>
<td>20</td>
<td>2.02%</td>
</tr>
<tr>
<td>Alaska</td>
<td>5</td>
<td>0.66%</td>
<td>Nebraska</td>
<td>39</td>
<td>2.15%</td>
</tr>
<tr>
<td>Arizona</td>
<td>103</td>
<td>1.62%</td>
<td>Nevada</td>
<td>30</td>
<td>1.12%</td>
</tr>
<tr>
<td>Arkansas</td>
<td>51</td>
<td>1.76%</td>
<td>New Hampshire</td>
<td>25</td>
<td>1.88%</td>
</tr>
<tr>
<td>California</td>
<td>601</td>
<td>1.61%</td>
<td>New Jersey</td>
<td>180</td>
<td>2.04%</td>
</tr>
<tr>
<td>Colorado</td>
<td>70</td>
<td>1.38%</td>
<td>New Mexico</td>
<td>32</td>
<td>1.55%</td>
</tr>
<tr>
<td>Connecticut</td>
<td>85</td>
<td>2.38%</td>
<td>New York</td>
<td>391</td>
<td>2.02%</td>
</tr>
<tr>
<td>Delaware</td>
<td>16</td>
<td>1.75%</td>
<td>North Carolina</td>
<td>147</td>
<td>1.55%</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>10</td>
<td>1.71%</td>
<td>North Dakota</td>
<td>17</td>
<td>2.48%</td>
</tr>
<tr>
<td>Florida</td>
<td>434</td>
<td>2.31%</td>
<td>Ohio</td>
<td>230</td>
<td>2.00%</td>
</tr>
<tr>
<td>Georgia</td>
<td>114</td>
<td>1.17%</td>
<td>Oklahoma</td>
<td>62</td>
<td>1.65%</td>
</tr>
<tr>
<td>Hawaii</td>
<td>30</td>
<td>2.22%</td>
<td>Oregon</td>
<td>78</td>
<td>2.03%</td>
</tr>
<tr>
<td>Idaho</td>
<td>25</td>
<td>1.61%</td>
<td>Pennsylvania</td>
<td>306</td>
<td>2.41%</td>
</tr>
<tr>
<td>Illinois</td>
<td>235</td>
<td>1.83%</td>
<td>Rhode Island</td>
<td>27</td>
<td>2.54%</td>
</tr>
<tr>
<td>Indiana</td>
<td>115</td>
<td>1.78%</td>
<td>South Carolina</td>
<td>71</td>
<td>1.53%</td>
</tr>
<tr>
<td>Iowa</td>
<td>75</td>
<td>2.45%</td>
<td>South Dakota</td>
<td>19</td>
<td>2.36%</td>
</tr>
</tbody>
</table>
3. a. Make a histogram of the state population sizes of residents 85 or older. (The data are in Table 9.2). Describe the shape of the histogram. Is it reasonable to assume that the state populations of residents 85 or older are normally distributed?

b. Based on your histogram in (a), do you think a normal quantile plot for these data would be concave up, concave down, or mostly linear? Explain.

c. Make a normal quantile plot for these data. Based on your normal quantile plot, is it reasonable to assume that the state populations of residents 85 or older are approximately normally distributed? Explain.

4. Return to the data in Table 9.2. The focus here is on the percentage of state residents who are 85 or older.

a. Make a histogram of the percent data. Describe the shape of the histogram. Is it reasonable to assume that the state percentages of residents 85 years old or over are approximately normally distributed?

b. Based on your histogram in (a), do you think a normal quantile plot for these data would be concave up, concave down, or mostly linear? Explain.

c. Make a normal quantile plot for the percentage data. Based on your normal quantile plot, is it reasonable to assume that the state percentages of residents 85 years old or older are normally distributed? Explain.

<table>
<thead>
<tr>
<th>State</th>
<th>Population 85 and Over (thousands)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kansas</td>
<td>59</td>
<td>2.08%</td>
</tr>
<tr>
<td>Kentucky</td>
<td>69</td>
<td>1.59%</td>
</tr>
<tr>
<td>Louisiana</td>
<td>66</td>
<td>1.45%</td>
</tr>
<tr>
<td>Maine</td>
<td>29</td>
<td>2.19%</td>
</tr>
<tr>
<td>Maryland</td>
<td>98</td>
<td>1.70%</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>145</td>
<td>2.22%</td>
</tr>
<tr>
<td>Michigan</td>
<td>192</td>
<td>1.94%</td>
</tr>
<tr>
<td>Minnesota</td>
<td>107</td>
<td>2.01%</td>
</tr>
<tr>
<td>Mississippi</td>
<td>44</td>
<td>1.49%</td>
</tr>
<tr>
<td>Missouri</td>
<td>114</td>
<td>1.90%</td>
</tr>
<tr>
<td>Tennessee</td>
<td>100</td>
<td>1.57%</td>
</tr>
<tr>
<td>Texas</td>
<td>305</td>
<td>1.21%</td>
</tr>
<tr>
<td>Utah</td>
<td>31</td>
<td>1.12%</td>
</tr>
<tr>
<td>Vermont</td>
<td>13</td>
<td>2.05%</td>
</tr>
<tr>
<td>Virginia</td>
<td>122</td>
<td>1.53%</td>
</tr>
<tr>
<td>Washington</td>
<td>117</td>
<td>1.74%</td>
</tr>
<tr>
<td>West Virginia</td>
<td>36</td>
<td>1.94%</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>119</td>
<td>2.08%</td>
</tr>
<tr>
<td>Wyoming</td>
<td>9</td>
<td>1.53%</td>
</tr>
</tbody>
</table>
| Table 9.2. Population and percentage of state residents 85 and over.
SUMMARY OF VIDEO

The video opens with views of manatees, large gentle sea creatures that live along the coast of Florida. Each year, a number of manatees are killed by powerboat propellers. From 1977 to 2011 the number of yearly manatee deaths appeared to be rising, but then so were the number of powerboat registrations. One avenue of investigation into manatee deaths is to look at the data for clues to the relationship between the number of powerboats registered in Florida in a given year and the number of manatees killed by powerboats. These are both quantitative variables; that is, they are measured in meaningful numerical units for which arithmetic operations make sense.

Powerboat registrations help explain manatee deaths, so powerboat registrations is the explanatory variable and manatees killed by powerboats is the response variable. The scatterplot in Figure 10.1, with the explanatory variable on the horizontal axis and the response variable on the vertical axis, shows the relationship between the two variables.

![Figure 10.1. Scatterplot of manatee deaths and powerboat registrations.](image)

The relationship shown in the scatterplot confirms our suspicions: as powerboat registrations rose, so did manatee fatalities. Our scatterplot shows positive association because as one variable increased, the other also tended to increase. An example of negative association
is shown in Figure 10.2, a graph of hypothetical data on the time to make a pie (response variable) versus numbers of tries in making this type of pie (explanatory variable). As the number of pies produced increased, the time required decreased because we get more efficient in making this type of pie.

![Figure 10.2. Scatterplot of preparation times and pies produced.](image)

When looking at a scatterplot, think about the overall pattern, how strong it is, and its direction. The manatee scatterplot has an overall pattern that is linear, as is shown in Figure 10.3; the points lie roughly in a straight line. The points stay pretty close to that line—hence, the relationship is strong. We’ve already noted that the relationship is positive and so is the slope of the line.

![Figure 10.3. Linear relationship between the variables.](image)
Keep in mind that not all relationships are linear. For example, the relationship shown in Figure 10.4 shows a curved pattern.

![Figure 10.4. Relationship with a curved pattern.](image)

Although a scatterplot shows the nature of a relationship between two variables, it doesn't prove that one variable causes the changes in the other. Even so, identifying a relationship between powerboat registrations and manatee deaths provided sufficient evidence for the state of Florida to consider additional ways to protect the manatees from those deadly propellers.
STUDENT LEARNING OBJECTIVES

A. Know how to make a scatterplot of quantitative bivariate data.

B. Recognize when there is an explanatory/response distinction and put the explanatory variable on the horizontal axis in making a scatterplot.

C. Recognize patterns in a scatterplot, especially positive or negative association and a linear (straight line) pattern.

D. Recognize outliers in a scatterplot and understand that these points should be investigated.
A bivariate data set consists of measurements or observations on two variables from the same individual or subject. When the two variables are quantitative, such as the height and weight of a group of children, we can display the data in a scatterplot by plotting the ordered pairs of data values.

In plotting bivariate quantitative data, we need to decide which variable to put on the horizontal axis and which to put on the vertical axis. In many cases, it is correct to put either variable on the horizontal axis. But if one variable, the explanatory variable, explains or causes changes in the other variable, the response variable, the explanatory variable is put on the horizontal axis. In plotting a variable against time, time is always on the horizontal axis.

A fundamental strategy of data analysis is: Make a graph of your data, then look for an overall pattern and for striking deviations from that pattern. In applying this strategy to relationships between two quantitative variables, the basic graph is the scatterplot. Look for the direction and shape of the pattern and identify any deviations from the overall pattern. Positive association and negative association describe the direction. Linear or curved (quadratic, exponential, and so on) describe the shape. Not all plots have either a clear direction or a clear shape; hence, a simple description may not be possible in all cases. Outliers are the most noticeable kind of deviation from the overall pattern.
KEY TERMS

For **bivariate data** measurements or observations are recorded on two attributes for each individual or subject under study. For example, in the video segment on manatees, the two attributes are number of manatee deaths and powerboat registrations and the subject is the year. For **multivariate data** measurements or observations are recorded on two or more attributes for each individual or subject under study.

A **response variable** measures an outcome of a study. The response variable is always plotted on the vertical axis of a scatterplot. An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is always plotted on the horizontal axis of a scatterplot.

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values of one tend to accompany below-average values of the other. In a scatterplot a positive association would appear as a pattern of dots in the lower left to the upper right. Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa. In a scatterplot a negative association would appear as a pattern of dots in the upper left to the lower right. (See Figure 10.5.)

![Figure 10.5. Scatterplots illustrating positive and negative association.](image)

A scatterplot has **linear form** when the dots appear to be randomly scattered on either side of a straight line. However, sometimes the data form a curved pattern. In that case, we say the
scatterplot has **nonlinear form**. Figure 10.6 shows two scatterplots, one with linear form and one with nonlinear form.

![Figure 10.6. Scatterplots illustrating linear and nonlinear form.](image)

*Figure 10.6. Scatterplots illustrating linear and nonlinear form.*
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is a manatee?

2. What does a scatterplot show about the relationship between the number of powerboats registered in Florida and the number of manatees killed by powerboats?

3. Why is the number of boats plotted on the horizontal axis of this scatterplot?

4. What trend would you expect to see in a scatterplot of two variables that have a negative association?
UNIT ACTIVITY:
RELATIONSHIPS BETWEEN PARENT AND STUDENT HEIGHTS

This activity is based on class data. You will need to measure your height. If you are male, measure your father’s height; if you are female, measure your mother’s height.

Record the height data from the class in Tables 10.1 and 10.2, which follow this activity.

1. Make a scatterplot of female student’s height versus mother’s height. (Mother’s height goes on the horizontal axis.)

2. Make a scatterplot of male student’s height versus father’s height. (Father’s height goes on the horizontal axis.)

3. Compare the patterns of the two scatterplots from questions 1 and 2.

4. Next make a scatterplot in which you can visualize three variables: student height, parent height, and student gender. Use parent’s height for the horizontal axis and student’s height for the vertical axis. Use different symbols (or different colors) for data from males and females.

5. Write a brief description about what can be learned from the scatterplot drawn for question 4.
### HEIGHT DATA FOR FEMALE STUDENTS

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Class (sophomore, junior, etc.)</th>
<th>Student’s Height (inches)</th>
<th>Mother’s Height (inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 10.1. Female students and mothers’ heights*
# HEIGHT DATA FOR MALE STUDENTS

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Class (sophomore, junior, etc.)</th>
<th>Student’s Height (inches)</th>
<th>Father’s Height (inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 10.2. Male students and fathers’ heights.*
Exercises

1. In each of the following situations, tell whether you would be interested simply in exploring the relationship between the two variables or whether you would want to view one of the variables as an explanatory variable and the other as a response variable. In the latter case, state which variable is the explanatory variable and which is the response variable.

   a. The amount of time spent studying for a statistics exam and the grade on the exam.

   b. The weight and height of a person.

   c. The amount of yearly rainfall and the yield of a crop.

   d. Hand length and foot length of a person.

2. A study was conducted on mercury (Hg) concentrations in fish taken from Lake Natoma in California. The researchers were concerned that mercury concentration levels in sample fish tissue might differ depending on the lab testing the fish. Fish tissue samples from 10 largemouth bass were sent to two labs, Columbia Environmental Research Center (CERC) and University of California, Davis (UC Davis), for inter-laboratory comparison. The data appear in Table 10.3. Mercury concentration is measured in micrograms of mercury per gram of fish tissue (dry weight).

<table>
<thead>
<tr>
<th>CERC</th>
<th>UC Davis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hg (µg/g dry wt.)</td>
<td>Hg (µg/g dry wt.)</td>
</tr>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>2.76</td>
<td>2.67</td>
</tr>
<tr>
<td>2.31</td>
<td>2.35</td>
</tr>
<tr>
<td>1.75</td>
<td>1.63</td>
</tr>
<tr>
<td>1.27</td>
<td>1.19</td>
</tr>
<tr>
<td>3.66</td>
<td>3.54</td>
</tr>
<tr>
<td>0.88</td>
<td>0.91</td>
</tr>
<tr>
<td>1.10</td>
<td>1.04</td>
</tr>
<tr>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td>2.00</td>
<td>2.05</td>
</tr>
<tr>
<td>3.24</td>
<td>3.13</td>
</tr>
</tbody>
</table>

Table 10.3. Mercury concentration in fish as determined by two labs.
a. Make a scatterplot of the data in Table 10.3.

b. Is the relationship between mercury concentrations determined by the two labs an example of positive association or negative association or neither. Explain.

c. Does the relationship have linear or nonlinear form? Explain.

3. Satellites are one of the many tools used for predicting flash floods, heavy rainfall, and large amounts of snow. Geostationary Operational Environmental Satellites (GOES) collect data on cloud top brightness temperatures (measured in degrees Kelvin (°K)). It turns out that colder cloud temperatures are associated with higher and thicker clouds, which could be associated with heavier precipitation. Data consisting of cloud top temperature measured by a GOES satellite and rainfall rate measured by ground radar appear in Table 10.4. Because ground radar can be limited by location and obstructions, having an alternative for predicting the rainfall rates can be useful.

<table>
<thead>
<tr>
<th>Temperature (°K)</th>
<th>Radar Rain Rate (mm/h)</th>
<th>Temperature (°K)</th>
<th>Radar Rain Rate (mm/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>195</td>
<td>150</td>
<td>203</td>
<td>44</td>
</tr>
<tr>
<td>196</td>
<td>150</td>
<td>204</td>
<td>39</td>
</tr>
<tr>
<td>197</td>
<td>150</td>
<td>205</td>
<td>39</td>
</tr>
<tr>
<td>198</td>
<td>118</td>
<td>206</td>
<td>35</td>
</tr>
<tr>
<td>199</td>
<td>109</td>
<td>207</td>
<td>38</td>
</tr>
<tr>
<td>200</td>
<td>95</td>
<td>208</td>
<td>31</td>
</tr>
<tr>
<td>201</td>
<td>63</td>
<td>209</td>
<td>20</td>
</tr>
<tr>
<td>202</td>
<td>66</td>
<td>210</td>
<td>24</td>
</tr>
</tbody>
</table>

*Table 10.4. Sixteen data pairs of (temperature, rain rate) data.*

a. In this situation, which variable is the explanatory variable and which is the response variable? Explain your choice.

b. Make a scatterplot of the data in Table 10.4.

c. Is the association between the variables positive or negative? Does the pattern of the dots in your scatterplot appear roughly linear? Explain.
4. The video discussed the linkage between powerboat registrations in Florida and manatee deaths. Table 10.5 contains data from 1977 to 2011 on the number of manatee deaths and the number of powerboat registrations in Florida. A time-series graph is a scatterplot where time is on the horizontal axis and the variable being measured is on the vertical axis.

a. Make a time-series graph of yearly manatee deaths. Would you describe the pattern as linear or nonlinear? Explain.

b. Make a time-series graph of the yearly number of powerboat registrations in Florida. Would you describe the pattern as linear or nonlinear? Explain.

<table>
<thead>
<tr>
<th>Year</th>
<th>Powerboat Registrations (thousands)</th>
<th>Manatee Deaths</th>
<th>Year</th>
<th>Powerboat Registrations (thousands)</th>
<th>Manatee Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>1977</td>
<td>447</td>
<td>13</td>
<td>1995</td>
<td>713</td>
<td>42</td>
</tr>
<tr>
<td>1978</td>
<td>460</td>
<td>21</td>
<td>1996</td>
<td>732</td>
<td>60</td>
</tr>
<tr>
<td>1979</td>
<td>481</td>
<td>24</td>
<td>1997</td>
<td>755</td>
<td>54</td>
</tr>
<tr>
<td>1980</td>
<td>498</td>
<td>16</td>
<td>1998</td>
<td>809</td>
<td>66</td>
</tr>
<tr>
<td>1981</td>
<td>513</td>
<td>24</td>
<td>1999</td>
<td>830</td>
<td>82</td>
</tr>
<tr>
<td>1982</td>
<td>512</td>
<td>20</td>
<td>2000</td>
<td>880</td>
<td>78</td>
</tr>
<tr>
<td>1983</td>
<td>526</td>
<td>15</td>
<td>2001</td>
<td>944</td>
<td>81</td>
</tr>
<tr>
<td>1984</td>
<td>559</td>
<td>34</td>
<td>2002</td>
<td>962</td>
<td>95</td>
</tr>
<tr>
<td>1985</td>
<td>585</td>
<td>33</td>
<td>2003</td>
<td>978</td>
<td>73</td>
</tr>
<tr>
<td>1986</td>
<td>614</td>
<td>33</td>
<td>2004</td>
<td>983</td>
<td>69</td>
</tr>
<tr>
<td>1987</td>
<td>645</td>
<td>39</td>
<td>2005</td>
<td>1010</td>
<td>79</td>
</tr>
<tr>
<td>1988</td>
<td>675</td>
<td>43</td>
<td>2006</td>
<td>1024</td>
<td>92</td>
</tr>
<tr>
<td>1989</td>
<td>711</td>
<td>50</td>
<td>2007</td>
<td>1027</td>
<td>73</td>
</tr>
<tr>
<td>1990</td>
<td>719</td>
<td>47</td>
<td>2008</td>
<td>1010</td>
<td>90</td>
</tr>
<tr>
<td>1991</td>
<td>681</td>
<td>53</td>
<td>2009</td>
<td>982</td>
<td>97</td>
</tr>
<tr>
<td>1992</td>
<td>679</td>
<td>38</td>
<td>2010</td>
<td>942</td>
<td>83</td>
</tr>
<tr>
<td>1993</td>
<td>678</td>
<td>35</td>
<td>2011</td>
<td>922</td>
<td>88</td>
</tr>
<tr>
<td>1994</td>
<td>696</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10.5. Manatee data for years 1977 – 2011.
Table 10.6. NBA attendance and ticket prices 2012/13 Season

1. Table 10.6 gives the average attendance and average ticket price for each of the 30 teams in the National Basketball Association in the 2012-2013 season. We want to investigate the relationship between ticket prices and attendance.

   a. Give some reasons why you might expect a positive association between ticket prices and attendance. Then give some reasons why you might expect a negative association.

   b. Make a scatterplot of the data to display the relationship between price and attendance. Explain your choice about what graph to make. Describe the relationship in words. Are there any outliers?
2. Table 10.7 contains data on the number of doctors in the United States (in thousands) for the 30-year period from 1970 to 1999.

<table>
<thead>
<tr>
<th>Year</th>
<th>( x ), Years Since 1970</th>
<th>Total Number of Physicians (thousands)</th>
<th>Number of Women Physicians (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970</td>
<td>1</td>
<td>330</td>
<td>25</td>
</tr>
<tr>
<td>1975</td>
<td>5</td>
<td>390</td>
<td>36</td>
</tr>
<tr>
<td>1980</td>
<td>10</td>
<td>470</td>
<td>55</td>
</tr>
<tr>
<td>1985</td>
<td>15</td>
<td>550</td>
<td>80</td>
</tr>
<tr>
<td>1990</td>
<td>20</td>
<td>620</td>
<td>100</td>
</tr>
<tr>
<td>1995</td>
<td>25</td>
<td>720</td>
<td>150</td>
</tr>
<tr>
<td>1999</td>
<td>30</td>
<td>800</td>
<td>190</td>
</tr>
</tbody>
</table>

*Table 10.7. Number of physicians by year.*

a. Notice that the values for \( x \) in the second column, the number of years since 1970, are missing. Complete this column.

b. Make a scatterplot of the total number of physicians versus \( x \). (We are using \( x \) as the explanatory variable.) Leave room to extend the horizontal axis out to 40 (for the year 2010).

c. Does your scatterplot from (b) have linear form? If not, describe the nonlinear nature of the scatterplot.

d. Make a scatterplot of the number of women physicians versus \( x \). Leave room to extend the horizontal axis out to 40 (for the year 2010).

e. Look at your scatterplot from (d). Does your scatterplot from (d) have linear form? If not, describe the nonlinear nature of the scatterplot.

f. In 2010 (\( x = 40 \)), the total number of physicians was 850,000 and the total number of female physicians was 250,000. Add this information to your scatterplots in (b) and (d). Does the added data point fit the overall pattern of the original data or deviate from the overall pattern? Explain.
SUMMARY OF VIDEO

Scatterplots are a great way to visualize the relationship between two quantitative variables. For example, the scatterplot of temperatures and coral reef growth in Figure 11.1 shows that as temperatures go up, new coral growth goes down.

![Figure 11.1. Scatterplot of coral growth versus ocean temperature.](image)

Another example is the scatterplot in Figure 11.2 of height versus femur bone length in humans, which exhibits a positive linear relationship. In this case, the femur length is the explanatory variable and hence, is on the horizontal axis. The response variable is height, which is on the vertical axis. Since the dots in the scatterplot appear to form a linear pattern, a line has been added to the scatterplot. Statisticians call the line that describes how the response variable changes with the explanatory variable a regression line.
Any line can be described by an equation of the form $y = a + bx$, where $a$ is the $y$-intercept and $b$ is the slope of the line. Recall the $y$-intercept is the $y$-value corresponding to $x = 0$ and the slope measures how much $y$ changes when $x$ increases by one unit. Sizing up the data points, we can just eyeball where the line should fall, but there is a statistical technique to figure out how best to fit a line to data. Once we have that line we can use it to make predictions. That's how a forensic scientist can estimate an unidentified crime victim's height just by using a femur bone measurement from incomplete skeletal remains.

The Colorado Climate Center uses regression lines for less sinister scenarios – to forecast the state’s seasonal water supply. Farmers, city planners, and businesses, all need to know how much water is going to be available each year so they can plan accordingly. Climatologist Nolan Doesken introduces an important question for Colorado: How can we predict the water supply we’re going to have as far ahead of time as possible? To answer this question, researchers have developed a model based on two types of data: the amount of winter snowpack in the high elevations and the resulting volume of water that flows out of the mountains throughout the summer. During the winter, Colorado’s Natural Resources Conservation Service heads into the Rockies to collect data on the snowpack. Later, when the snowpack starts to melt, data related to the volume of water runoff are collected. Figure 11.3 shows a scatterplot of the data collected over many years.
Figure 11.3. Scatterplot of annual water runoff versus snow water equivalent.

From the scatterplot in Figure 11.3, there appears to be a pretty strong positive linear relationship between the two variables and we have drawn a line to summarize that relationship. Of course, in the real world, all the points don’t fall exactly on a line. So, we need a technique to determine the regression line that in some way minimizes the vertical distances of our data points from the line.

To get a better idea of how to fit a line to data, we zoom in on three of the data points from our Colorado water data. (See Figure 11.4.)

Figure 11.4. Zooming in on three data points and visualizing residuals.

The vertical distance of a data point from the line, together with a positive sign if the point lies above the line and a negative sign if the point lies below the line, is called a residual. As we shift the line trying to find the one that best fits these three points (see Figure 11.5.), some
residuals get smaller while others get larger. So, we try to find the sweet spot where we’ve got them as a group as small as possible.

Figure 11.5. Residuals corresponding to different lines.

Statisticians use a method called least squares to find the “best-fitting” line. Since some of the residuals are positive and some are negative, we square them, which makes them all positive. If you add up the squared residuals, the bigger the sum, the more the line misses the points. So, we want to make that sum as small as possible. Software can compute the equation of the least-squares regression line – the line with the smallest sum of the squared residuals.

Back to the Colorado water data. Figure 11.6 shows both the graph of the least-squares regression line and its equation.

Figure 11.6. Graph of the least-squares regression line.

Note that in Figure 11.6 we have used the notation \( \hat{y} \) (pronounced y-hat) to emphasize that we are talking about the predicted value of \( y \), not a measured value from our data set. The slope of 1,941 predicts that for every one-inch increase in snowpack, the runoff increases by
1,941 acre-feet. The \( y \)-intercept is at -7,920. This value seems to say that if the snowpack was 0, the runoff would be -7,920 acre feet. Obviously that doesn’t make sense, and it is a good reminder that you can’t extrapolate from the regression line too far outside the range of the observed data. Keeping that limitation in mind, though, the regression line can be very useful for Colorado water users.

Now, we are ready to use the least-squares line to make a prediction. If you know that this winter the Rockies saw 30 inches of snowpack, you can look at the line in Figure 11.7 to predict how much water is going to flow into the system in the spring.

\[
\hat{y} = -7,920 + 1,941(30)
\]

\[
\hat{y} = 50,310
\]

**Figure 11.7. Using the least-squares equation to make a prediction.**

The regression line works well to predict the Colorado water supply because the relationship between snowpack and runoff is linear. If the relationship you are trying to describe has a curved pattern, a *straight* regression line won’t be a good fit. One way to assess how well a regression line fits the data is to make a residual plot, a plot of the residuals versus the explanatory variable. If the dots in the residual plot appear randomly scattered with no strong pattern, then the regression line has nicely captured the pattern in the data and a linear model is a good choice to describe it. For example, take a look at the residual plot for the Colorado data in Figure 11.8. The dots appear randomly scattered, with roughly equal numbers of dots above and below the horizontal axis.
What if your scatterplot had a curved pattern such as the one in Figure 11.9(a), made by data on alligator weight versus alligator length? If you fit a least-squares line to these data, then the residual plot will be curved as shown in Figure 11.9(b). You wouldn’t want to use the equation of a line to make predictions about alligators! Instead you need to find an equation that describes the curved shape of the data.

Figure 11.9. (a) Curved relationship between alligator weight and alligator length.  
(b) Residual plot for fitting a line to alligator data.
STUDENT LEARNING OBJECTIVES

A. Be able to predict the value of the response variable corresponding to a given value of the explanatory variable from a graph of a line.

B. Understand the least-squares criterion for determining the line of "best fit" to data for the purpose of predicting $y$ from $x$.

C. Know how to calculate the equation of the least-squares line for a small set of bivariate data (say $n < 10$). Know how to use software (such as Minitab or Excel) or a graphing calculator to calculate the equation of the least-squares line.

D. Understand the effect on the least-squares line of influential points that are extreme in the $x$-direction.

E. Know how to check the adequacy of a linear model by (1) viewing a scatterplot of the data to see if the pattern is linear and by (2) making a residual plot.

F. Know the difference between extrapolation and interpolation. Know when to beware of extrapolation.
In this unit, we discuss a common method for fitting a line to data that show a linear pattern but whose points do not fall exactly on a line. We begin with a simple example that involves four data points: (2,1), (3,5), (4,3), and (5,2). Figure 11.10 shows a plot of these points and a line that might be a good fit because it goes through two of the four points.

The line in Figure 11.10 can be described by the equation $y = x - 1$. But is this the best line to describe the pattern in these data? To measure how far off the line is from a point, we calculate the residual (also called the residual error) associated with each point:

$$\text{residual} = \text{actual } y\text{-value from data} - \text{predicted } y\text{-value from the line}$$

For example, the residual corresponding to (2,1) is 0. When $x = 2$, the $y$-value from the data point, $y = 1$, is the same as the predicted $y$-value from the equation, $\hat{y} = 2 - 1 = 1$. The residual corresponding to the point (4,3) is also zero, because this data point lies on the line. However, the residuals corresponding to (3,5) and (5,2) are 3 and -2, respectively. The residuals are represented graphically in Figure 11.10 by the dashed vertical lines. Notice that when a data point lies above the line, its residual is positive and when a data point lies below the line its residual is negative.

The least-squares method is the most common means of fitting a “best” line to data on an explanatory variable, $x$, and a response variable, $y$. This method fits a line that has the smallest
sum of squares of residual (errors), or SSE for short. We express the formula for the least-squares regression line as \( y = a + bx \), where

\[
b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}
\]

\[
a = \bar{y} - bx
\]

The formulas above are a bit complicated. So, to calculate the equation of the least-squares line for the data in Figure 11.10, we have set up Table 11.1. In the first two columns, we list the data values for \( x \) and \( y \). Then, we calculate the mean for the \( x \)'s and \( y \)'s: \( \bar{x} = 3.5 \) and \( \bar{y} = 2.75 \). In columns three and four we subtract \( \bar{x} \) and \( \bar{y} \) from the \( x \)'s and \( y \)'s, respectively. In columns five and six, we form the products needed for the numerator and denominator of the formula for \( b \). Finally, the sum of the entries in columns five and six gives us the numerator and denominator of \( b \), respectively.

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x - 3.5 )</th>
<th>( y - 2.75 )</th>
<th>( (x - 3.5)(y - 2.5) )</th>
<th>( (x - 3.5)^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>-1.5</td>
<td>-1.75</td>
<td>2.625</td>
<td>2.25</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>-0.5</td>
<td>2.25</td>
<td>-1.125</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.5</td>
<td>-0.75</td>
<td>-1.125</td>
<td>2.25</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td>0.5</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

\( b = 0.5 / 5 = 0.1 \)

\( a = 2.75 - (0.1)(3.5) = 2.4 \)

The equation of the least-squares line can be expressed as \( y = 0.1x + 2.4 \). Figure 11.11 shows the graph of the least-squares line added to Figure 11.10. Unlike our first line, the least-squares line does not contain any of the data points. However, the SSE for the least-squares line, the sum of the squared lengths of the vertical line segments, will be less than the SSE for the line \( y = x - 1 \). (For confirmation, work through question 5 in the Review Questions.)
Although you now know how to compute the equation for the least-squares line, it is much faster to use technology for the calculations particularly for larger data sets. Statistical software, spreadsheet software, and graphing calculators all have built-in linear regression capabilities that will report the equation of the least-squares line.

Regression assumes that we want to predict the response variable $y$ given values of the explanatory variable $x$. The distinction between an explanatory and a response variable is essential; reversing the roles of the two variables produces a different regression line. (See question 6 in the Unit Activity.) In Figure 11.3, the response variable, $y$, was the annual water runoff (in acre-feet) in Colorado and the explanatory variable, $x$, was the snowpack (inches). We can use the equation of the least-squares line, $\hat{y} = -7,920 + 1,941x$ to make predictions. For example, suppose the snowpack one year turned out to be $x = 40$ inches. Then the predicted water flow for the next spring would be $\hat{y} = -7,920 + 1,941(40) = 69,720$ acre-feet.

Of course, it only makes sense to use the equation of the regression line to make predictions if the line adequately describes the pattern of the data. One way to check is with a residual plot, a plot of the residuals versus the explanatory variable. Figure 11.8 shows a residual plot for the Colorado runoff-snowpack data. A good residual plot looks like a bunch of dots randomly thrown onto the graph – there should be no strong pattern, and roughly half the points should lie above the horizontal axis and half below. The residual plot in Figure 11.8 meets both of these criteria. If, on the other hand, the residual plot shows a strongly curved pattern then the least-squares line fails to adequately describe the pattern in the data. Using the least-squares equation to make predictions in this case, could lead to some very inaccurate predictions. What is needed is a search for another model, one that would better describe the curved pattern of the data.
We conclude with two warnings. The first is that the least-squares regression line can be strongly influenced by one or more extreme points. Points that lie far from the other data in the $x$-direction (as might be the case if data appear in clusters) are particularly dangerous. In this case, a residual plot will most likely exhibit some strong patterns indicating that the model is not adequate to describe the pattern in the data. The second warning is to beware of extrapolation – the use of a fitted line for prediction outside the range of $x$-values in the data. For example, return to the problem of predicting Colorado water runoff. Suppose in one very dry year the snowpack measured 2 inches, much lower than anything observed in the data. Using the least-squares equation to predict the spring water runoff gives:

$$y = -7920 + 1941(2) = -4038 \text{ acre feet}.$$

Clearly, the volume of water runoff can’t be negative. Contrast this situation with the previous prediction of 69,720 acre feet for a snowpack of 40 inches. That prediction made sense and is an example of interpolation, because a 40-inch snowpack lies within the range of observed values for snowpack.
A residual (or residual error) is the vertical deviation of a data point from the regression model:

\[ \text{residual} = \text{observed } y - \text{predicted } y. \]

The least-squares regression line is the line that makes the sum of the squares of the residual errors, \( \text{SSE} \), as small as possible. The equation of the least-squares line is \( \hat{y} = a + bx \), where the \( y \)-intercept, \( a \), and slope, \( b \), are calculated from \( n \) data values on an explanatory variable \( x \) and a response value \( y \). To calculate the values of \( a \) and \( b \) without the use of software, use the following formulas:

\[
b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}
\]

\[
a = \bar{y} - b\bar{x}
\]

Because the formula for computing the least-squares regression line will fit a line to any data set on two variables \( x \) and \( y \), it is important to assess the adequacy of the least-squares model for describing the pattern in the data. One way to judge the adequacy of a linear model is to make a residual plot – a plot of the residuals versus the explanatory variable. If the dots in the residual plot appear randomly scattered with roughly half above and half below the \( x \)-axis, then the linear model is adequate to describe the pattern in the data. Otherwise, look for a new model.

An observation is influential if removing it would greatly change the position of the regression line. Points that are separated in the \( x \)-direction from the other observations, such as in a cluster pattern, are often influential.

Interpolation is the use of the regression equation to predict \( y \) for values of \( x \) that lie inside the range of values in the data used to fit the line. Extrapolation is the use of the regression equation to predict \( y \) for values of \( x \) outside the range of values in the data used to fit the line. Models don’t hold forever, hence extrapolation can be risky.
THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. How is the snowpack during wintertime in the Colorado Mountains measured?

2. What is a residual?

3. How does the least-squares method decide which line best fits the points in a scatterplot?

4. How can a particular year’s data on the snowpack be used to predict the amount of water running downstream in the spring?

5. The video showed two examples of residual plots. What does a residual plot tell you if the dots in the plot appear to be randomly scattered? What if the dots appear to form a strong curved pattern instead?
UNIT ACTIVITY:
THE RELATIONSHIP BETWEEN FOREARM LENGTH AND FOOT LENGTH

In determining normal proportions in human bodies, doctors look at the relationships between the lengths of various body parts. Artists are also interested in these relationships. Knowing such relationships helps them draw human figures that appear appropriately proportioned. On a less serious note, this activity was inspired by a piece of trivia from Julia Roberts’ character in the 1990 movie Pretty Woman, in which she states – your forearm length is the same as your foot length. To check the validity of this assertion, we collected data on forearm and foot lengths of 26 college students enrolled in an introductory statistics course. These data appear in Table 11.2.

<table>
<thead>
<tr>
<th>Forearm Length (inches), $x$</th>
<th>Foot Length (inches), $y$</th>
<th>Forearm Length (inches), $x$</th>
<th>Foot Length (inches), $y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.00</td>
<td>9.50</td>
<td>8.75</td>
<td>9.00</td>
</tr>
<tr>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>10.50</td>
</tr>
<tr>
<td>10.00</td>
<td>9.50</td>
<td>8.50</td>
<td>11.00</td>
</tr>
<tr>
<td>10.00</td>
<td>10.00</td>
<td>10.25</td>
<td>11.50</td>
</tr>
<tr>
<td>11.50</td>
<td>12.50</td>
<td>10.25</td>
<td>11.25</td>
</tr>
<tr>
<td>9.00</td>
<td>11.50</td>
<td>8.50</td>
<td>9.00</td>
</tr>
<tr>
<td>8.50</td>
<td>9.00</td>
<td>9.25</td>
<td>10.50</td>
</tr>
<tr>
<td>6.75</td>
<td>9.25</td>
<td>10.50</td>
<td>10.50</td>
</tr>
<tr>
<td>10.00</td>
<td>10.00</td>
<td>8.25</td>
<td>8.50</td>
</tr>
<tr>
<td>8.25</td>
<td>8.25</td>
<td>9.00</td>
<td>10.00</td>
</tr>
<tr>
<td>8.25</td>
<td>9.50</td>
<td>7.00</td>
<td>8.75</td>
</tr>
<tr>
<td>9.00</td>
<td>9.50</td>
<td>9.50</td>
<td>8.75</td>
</tr>
<tr>
<td>8.00</td>
<td>9.50</td>
<td>9.75</td>
<td>10.00</td>
</tr>
</tbody>
</table>

Table 11.2: Data on forearm and foot length.

Since students objected to removing their shoes in order to measure their feet, forearm length is the explanatory variable and foot length is the response variable. That way, after determining a relationship between these two variables, we can use it to predict students’ foot lengths from their forearm lengths (and they won’t have to take off their shoes!).
1. a. Make a scatterplot of foot length, $y$, versus forearm length, $x$.

b. Based on your scatterplot, does it appear that students with longer forearms tend to have bigger feet? Explain how you can tell from the scatterplot.

2. Use technology to determine the equation of the least-squares regression line. Superimpose a graph of this line on your scatterplot. Does the line appear to do a good job of summarizing the pattern of the points in the scatterplot?

3. Use the equation of the least-squares line to predict the foot length of a person with a 10.5 inch forearm.

4. Find the residual corresponding to the first data point (10, 9.50). Show your calculations.

5. Before using the least-squares line to make predictions (as we did in question 3), we first should have checked on the adequacy of the linear model. Figure 11.12 shows a residual plot. Based on the residual plot, is the least-squares line adequate to describe the pattern in these data? Explain.

![Residual Plot](image)

*Figure 11.12. Residual plot.*
6. Danny got his explanatory and response variables mixed up. He fit a least-squares line to forearm length, \( x \), versus foot length, \( y \), and got the following equation:

\[ x = 2.865 + 0.6332y \]

Sarah, his partner, told him not to worry but to solve for \( y \) and he would get the correct equation for the least-squares line. Follow Sarah’s advice and solve for \( y \) in terms of \( x \). Do your results agree with the equation you got for question 2? What does this tell you about Sarah’s strategy for fixing her partner’s mistake?

7. Linda fit a line to the data by finding the equation of a line that contains (8,9.5) and (10.5,10.5). The equation for her line is \( y = 0.4x + 6.3 \). What can you say about the sum of squared residuals, SSE, for her line compared to the SSE for the least-squares line? Explain.

Extension to question 7: Find the SSE for the least-squares line and for Linda’s line. Which line had the smaller SSE?
The data on the average sea surface temperature and coral reef growth shown in Figure 11.1 appear below.

<table>
<thead>
<tr>
<th>Temperature (°C), $x$</th>
<th>29.7</th>
<th>29.9</th>
<th>30.2</th>
<th>30.2</th>
<th>30.5</th>
<th>30.7</th>
<th>30.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coral Growth (mm), $y$</td>
<td>2.63</td>
<td>2.58</td>
<td>2.60</td>
<td>2.48</td>
<td>2.26</td>
<td>2.38</td>
<td>2.26</td>
</tr>
</tbody>
</table>

*Table 11.3. Temperature and coral growth data.*

1. a. Make a scatterplot of coral reef growth versus the average sea surface temperature. Describe the pattern. Are there any outliers (data points that appear to deviate from the overall pattern)?

b. Computer software gives the equation of the least-squares regression line as

\[ \hat{y} = 12.3 - 0.325x \]

Add a graph of this line to your scatterplot in (a).

2. Return to the coral reef data in Table 11.3. Next, you will verify the equation in exercise 1(b) using the formulas given in the Content Overview.

a. Determine $\bar{x}$ and $\bar{y}$.

b. Make a copy of Table 11.4 and complete the entries. Record the sum of the last two columns in the shaded cells.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$(x - \bar{x})$</th>
<th>$(y - \bar{y})$</th>
<th>$(x - \bar{x})(y - \bar{y})$</th>
<th>$(x - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.7</td>
<td>2.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29.9</td>
<td>2.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.2</td>
<td>2.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.2</td>
<td>2.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.5</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.7</td>
<td>2.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30.9</td>
<td>2.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 11.4. Table used to compute $b$.*

c. Using the information from (b), show the calculations for the slope, $b$, and $y$-intercept, $a$, of the least-squares line.
3. a. Figure 11.13 shows a residual plot corresponding to the least-squares regression line determined in exercise 2. Based on this plot is the least-squares line adequate to describe the pattern in the data? Explain.

![Residual Plot](image)

Figure 11.13. Residual plot.

b. Regardless of your answer to (a), assume that the least-squares line is adequate to describe the pattern in the data from Table 11.2. Use the equation given in 1(b) to predict the coral growth if the average sea surface temperature rises to 40°.

4. Satellites are one of the many tools used for predicting flash floods, heavy rainfall, and large amounts of snow. Geostationary Operational Environmental Satellites (GOES) collect data on cloud top brightness temperatures (measured in degrees Kelvin (°K)). It turns out that colder cloud temperatures are associated with higher and thicker clouds, which could be associated with heavier precipitation. Data consisting of cloud top temperature measured by a GOES satellite and rainfall rate measured by ground radar appear in Table 11.5. Because ground radar can be limited by location and obstructions, having an alternative for predicting the rainfall rates can be useful.
Table 11.5. Sixteen data pairs of (temperature, rain rate) data.

<table>
<thead>
<tr>
<th>Temperature (°K)</th>
<th>Radar Rain Rate (mm/h)</th>
<th>Temperature (°K)</th>
<th>Radar Rain Rate (mm/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>195</td>
<td>150</td>
<td>203</td>
<td>44</td>
</tr>
<tr>
<td>196</td>
<td>150</td>
<td>204</td>
<td>39</td>
</tr>
<tr>
<td>197</td>
<td>150</td>
<td>205</td>
<td>39</td>
</tr>
<tr>
<td>198</td>
<td>118</td>
<td>206</td>
<td>35</td>
</tr>
<tr>
<td>199</td>
<td>109</td>
<td>207</td>
<td>38</td>
</tr>
<tr>
<td>200</td>
<td>95</td>
<td>208</td>
<td>31</td>
</tr>
<tr>
<td>201</td>
<td>63</td>
<td>209</td>
<td>20</td>
</tr>
<tr>
<td>202</td>
<td>66</td>
<td>210</td>
<td>24</td>
</tr>
</tbody>
</table>

a. You first encountered these data in Unit 10, exercise 3, where you were asked to make a scatterplot of these data. Since we want to use temperature to predict rain rate, temperature is the explanatory variable. Fit a least-squares line to rain rate versus temperature data. Sketch a scatterplot of the data and the line. Report the equation of your line.

b. Use your equation from (a) to predict the rain rate when the cloud temperature is 220 °K. Does your answer make sense? Explain.

c. Make a residual plot. Do you think a straight-line model is adequate to describe the pattern in these data? Explain why or why not.
1. A random sample of femur bone lengths (mm) and heights (cm) from 20 males appears in Table 11.6. These data are from the Forensic Anthropology Data Bank at the University of Tennessee.

<table>
<thead>
<tr>
<th>Femur Length (mm)</th>
<th>Height (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>447</td>
<td>168</td>
</tr>
<tr>
<td>444</td>
<td>168</td>
</tr>
<tr>
<td>470</td>
<td>175</td>
</tr>
<tr>
<td>459</td>
<td>170</td>
</tr>
<tr>
<td>482</td>
<td>178</td>
</tr>
<tr>
<td>520</td>
<td>191</td>
</tr>
<tr>
<td>464</td>
<td>175</td>
</tr>
<tr>
<td>470</td>
<td>172</td>
</tr>
<tr>
<td>482</td>
<td>182</td>
</tr>
<tr>
<td>462</td>
<td>178</td>
</tr>
<tr>
<td>522</td>
<td>193</td>
</tr>
<tr>
<td>461</td>
<td>171</td>
</tr>
<tr>
<td>422</td>
<td>160</td>
</tr>
<tr>
<td>520</td>
<td>185</td>
</tr>
<tr>
<td>476</td>
<td>180</td>
</tr>
<tr>
<td>508</td>
<td>183</td>
</tr>
<tr>
<td>477</td>
<td>173</td>
</tr>
<tr>
<td>504</td>
<td>175</td>
</tr>
<tr>
<td>547</td>
<td>189</td>
</tr>
<tr>
<td>508</td>
<td>198</td>
</tr>
</tbody>
</table>

*Table 11.6. Data on femur bone length and height.*

a. We would like to predict the height of a male given the length of his femur bone. Which variable is the explanatory variable and which is the response variable? Explain.

b. Enter the data into columns of computer software (or calculator lists). Make a scatterplot of the data. Describe the pattern in your scatterplot. Does the pattern appear linear or nonlinear? Is the association between the variables positive or negative?

c. Fit a least-squares line to the data and report its equation. Overlay a graph of the line on your scatterplot from (b).
d. Do there appear to be any outliers? If so, identify the point(s). Do you think these are mistakes or real data values? Explain.

2. Return to your work from question 1.

a. Interpret the slope and y-intercept in the context of these data. Do these quantities make sense in the given context?

b. Two femur bones presumed to be from two men are measured and their lengths differ by 5 mm. Use the least-squares regression equation to predict the difference in heights between the two men.

c. Predict the height of a male whose femur length is 475 mm. Is this a reasonable height for a man? (Convert your answer to feet and inches. Recall there are 2.54 centimeters per inch.)

d. The femur length of a boy measures 250 mm. Predict the height of the child. Explain why this prediction might not be trustworthy.

3. Table 11.7 contains data on mercury concentration in tissue samples from 20 largemouth bass taken from Lake Natoma in California. Only fish of legal/edible size were used in this study.

<table>
<thead>
<tr>
<th>Total Length (mm)</th>
<th>Mercury Concentration (µg/g wet wt.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>341</td>
<td>0.515</td>
</tr>
<tr>
<td>353</td>
<td>0.268</td>
</tr>
<tr>
<td>387</td>
<td>0.450</td>
</tr>
<tr>
<td>375</td>
<td>0.516</td>
</tr>
<tr>
<td>389</td>
<td>0.342</td>
</tr>
<tr>
<td>395</td>
<td>0.495</td>
</tr>
<tr>
<td>407</td>
<td>0.604</td>
</tr>
<tr>
<td>415</td>
<td>0.695</td>
</tr>
<tr>
<td>425</td>
<td>0.577</td>
</tr>
<tr>
<td>446</td>
<td>0.692</td>
</tr>
<tr>
<td>490</td>
<td>0.807</td>
</tr>
<tr>
<td>315</td>
<td>0.320</td>
</tr>
<tr>
<td>360</td>
<td>0.332</td>
</tr>
<tr>
<td>385</td>
<td>0.584</td>
</tr>
<tr>
<td>390</td>
<td>0.580</td>
</tr>
</tbody>
</table>
Table 11.7. Fish length and mercury concentration in fish tissue samples.

<table>
<thead>
<tr>
<th>Length (mm)</th>
<th>Mercury Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>410</td>
<td>0.722</td>
</tr>
<tr>
<td>425</td>
<td>0.550</td>
</tr>
<tr>
<td>480</td>
<td>0.923</td>
</tr>
<tr>
<td>448</td>
<td>0.653</td>
</tr>
<tr>
<td>460</td>
<td>0.755</td>
</tr>
</tbody>
</table>

a. We want to be able to predict mercury concentration from fish length. Which variable is the explanatory variable and which is the response variable?

b. Fit a least-squares line to the data from Table 11.7. Report its equation. (Round the slope and y-intercept to four decimals.) Also, show a scatterplot of the data and the least-squares line.

c. Make a residual plot. Based on your plot, is the least-squares model adequate to describe the overall pattern in the data? Explain.

d. Interpret the slope of the least-squares line in the context of this problem. Does the interpretation of slope make sense in the given context? Explain why or why not.

e. Interpret the y-intercept of the least-squares line in the context of this problem. Does the interpretation of the y-intercept make sense in the given context? Explain why or why not.

4. Return to the data in exercise 3, Table 11.7. Use your answer to 3(b) to make the following predictions:

a. Predict the mercury concentration in a largemouth bass that is 430 mm in length. Is this prediction an example of interpolation or extrapolation? Explain.

b. Predict the mercury concentration in a largemouth bass that is 90 mm in length, which is below the legal/edible size. Is this an example of interpolation or extrapolation? Explain.

5. In the Content Overview, we fit two lines to the data in Table 11.8. Graphs of these lines appear in Figure 11.11.
Table 11.8. Data from Figure 11.10.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

a. The equation of the first line was $y = x - 1$. Calculate the sum of the squares of the residuals, SSE, for this line.

b. The equation of the least-squares line is $y = 0.1x + 2.4$. Calculate the SSE for the least-squares line.

c. Which line had the smaller SSE? Why should we not be surprised by this result?