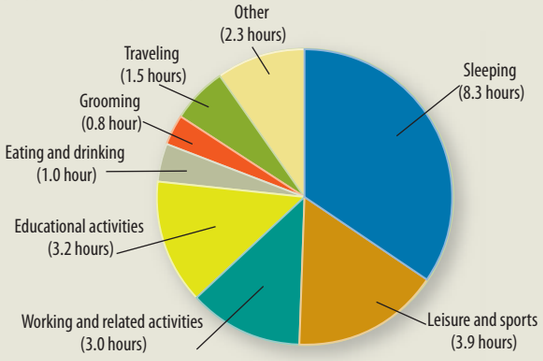
Descriptive Analysis and Presentation of Single-Variable Data

Ever wonder if your typical day measures up to that of other college students? If you look at the graph below, think about what you do during the day and how much time you spend doing those activities. Does your day break up into the categories shown below? Or do you have an extra category or two? On the average, how does the amount of time you spend compare? Perhaps you have different categories. You may wish that you too could average 8.3 hours of sleep!

TIME USE ON AN AVERAGE WEEKDAY FOR FULL-TIME UNIVERSITY AND COLLEGE STUDENTS



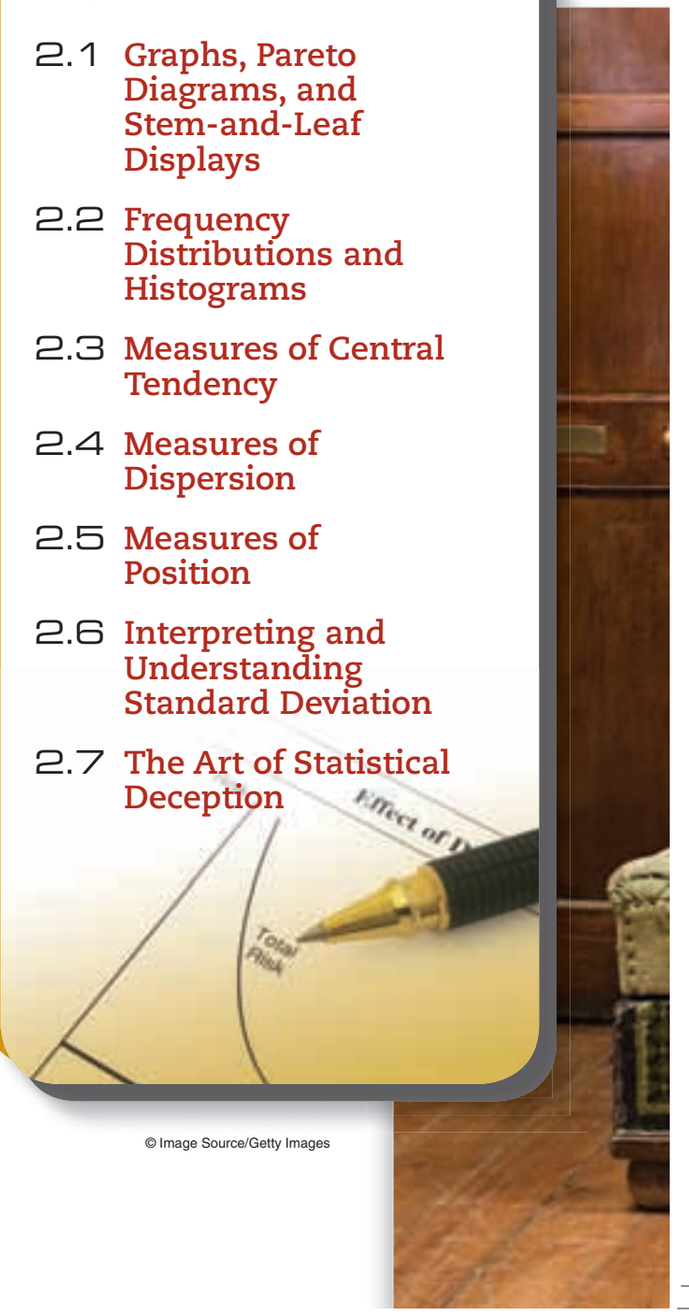
Note: Data include individuals, ages 15 to 49, who were enrolled full time at a university or college. Data include non-holiday weekends and are averages for 2003–07.

SOURCE: Bureau of Labor Statistics

Can you imagine all this information written out in sentences? Graphical displays can truly be worth a thousand words. This one pie chart summarizes "Time Use" information from a 2003–2007 American Time Use Survey (ATUS) of over 50,000 Americans. ATUS is a federally administered, continuous survey on time use in the United States sponsored by the Bureau of Labor Statistics and conducted by the U.S. Census Bureau. Since it is a cross-sectional survey, this graph included only the full-time college students who participated.

Now that you know the source and see the overall sample size, you may feel that these data portray a relatively accurate picture of a college student's day. Or do some of the numbers seem off to you? The 0.8 hour per day grooming may have a gender difference

objectives



© Image Source/Getty Images

and strike you as inaccurate. As you work through Chapter 2, you will begin to learn how to organize and summarize data into graphical displays and numerical statistics in order to accurately and appropriately describe data.

2.1 Graphs, Pareto Diagrams, and Stem-and-Leaf Displays

ONCE THE SAMPLE DATA HAVE BEEN COLLECTED, WE MUST "GET ACQUAINTED" WITH THEM.

CHAPTER 2 STAT

One of the most helpful ways to become acquainted with the data is to use an initial exploratory data analysis technique that will result in a pictorial representation of the data. The display will visually reveal patterns of behavior of the variable being studied. There are several graphic (pictorial) ways to describe data. The type of data and the idea to be presented determine which method is used.

NOTE: There is no single correct answer when constructing a graphic display. The analyst's judgment and the circumstances surrounding the problem play a major role in the development of the graph.

Qualitative Data

Graphs can be used to summarize qualitative, or attribute, or categorical, data. **Pie charts (circle graphs)** show the amount of data that belong to each category as a proportional part of a circle. **Bar graphs** show the amount of data that belong to each category as a proportionally sized rectangular area. Any graphic representation used, regardless of type, needs to be completely self-explanatory. That includes a descriptive, meaningful title and proper identification of the quantities and variables involved. To appreciate the differences between these two types of graphical representations, let's compare them by using the same data set to create one of each.

To get a better sense of what's involved in graphing qualitative data, let's consider an example about surgeries at a hospital. Table 2.1 lists the number of cases of each type of operation performed at General Hospital last year.

*Table 2.1 Operations Performed at General Hospital Last Year

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23
Total	498

* Tables marked with an asterisk have data sets online at cengagebrain.com.

The data in Table 2.1 are displayed on a pie chart in Figure 2.1, with each type of operation represented by a relative proportion of a circle, found by dividing the number of cases by the total sample size—namely, 498. The proportions are then reported as percentages (for example, 25% is $\frac{1}{4}$ of the circle). Figure 2.2 displays the same "type of operation" data but in the form of a bar graph. Bar graphs of attribute data should be drawn with a space between bars of equal width.

When the bar graph is presented in the form of a *Pareto diagram*, it presents additional and very helpful information. That's because in a **Pareto diagram** the bars are arranged from the most numerous category to the least

© Pali Rao/iStockphoto.com

Pie charts (circle graphs) Graphs that show the amount of data belonging to each category as a proportional part of a circle.

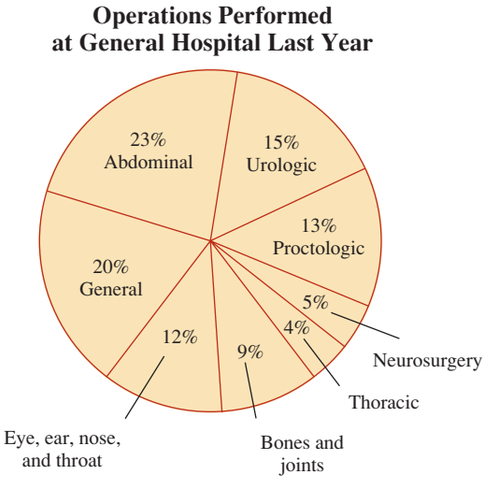
Bar graphs Graphs that show the amount of data belonging to each category as a proportionally sized rectangular area.

Pareto diagram A bar graph with the bars arranged from the most numerous category to the least numerous category. It includes a line graph displaying the cumulative percentages and counts for the bars.

numerous category. A Pareto diagram also includes a line graph displaying the cumulative percentages and counts for the bars. The Pareto diagram is popular in quality-control applications. A Pareto diagram of types of defects will show the ones that have the greatest effect on the defective rate in order of effect. It is then easy to see which defects should be targeted in order to most effectively lower the defective rate.

Pareto diagrams can also be useful in evaluating crime statistics. The FBI reported the number of hate crimes by category for 2003 (www.fbi.gov). The Pareto diagram in Figure 2.3 shows the 8,715 categorized hate crimes, their percentages, and cumulative percentages.

Figure 2.1 Pie Chart



Quantitative Data

One major reason for constructing a graph of quantitative data is to display its **distribution,** or the pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable. Two popular methods for displaying distribution of qualitative data are the **dotplot** and the **stem-and-leaf** display.

Distribution The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.

Figure 2.2 Bar Graph

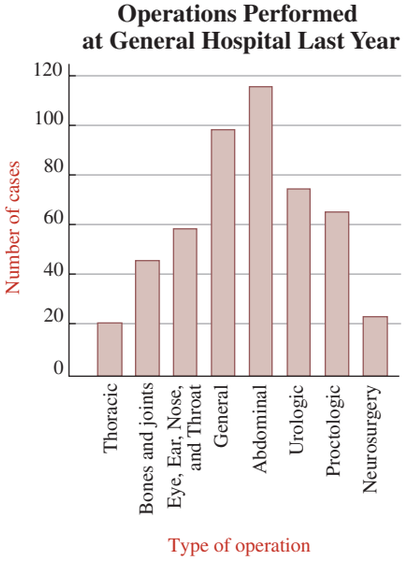
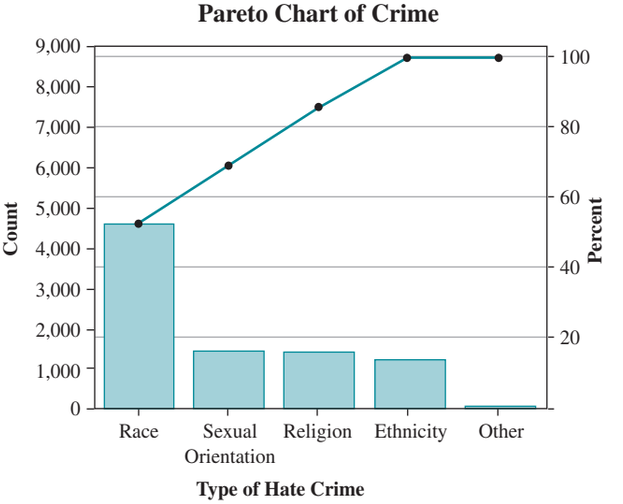


Figure 2.3 Pareto Diagram



Count	4,574	1,430	1,426	1,236	49
Percent	52.5	16.4	16.4	14.2	0.6
Cum %	52.5	68.9	85.3	99.4	100.0

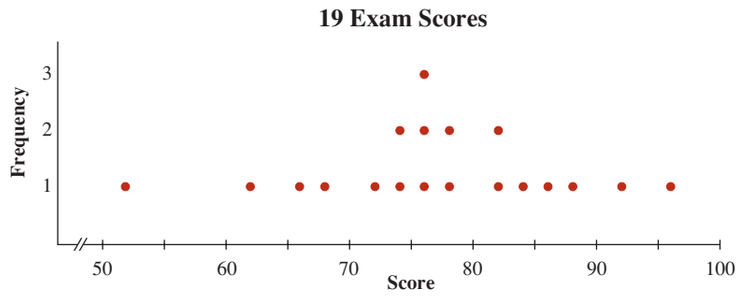
Dotplot

One of the simplest graphs used to display a distribution is the **dotplot display.** The dotplot displays the data of a sample by representing each data value with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale. The dotplot display is a convenient technique to use as you first begin to analyze the data. It results in a picture of the data and it sorts the data into numerical order. (To *sort* data is to list the data in rank order according to numerical value.) Table 2.2 provides a sample of 19 exam grades randomly selected from a large class. Notice how the data in Figure 2.4 are bunched near the center and more spread out near the extremes.

*Table 2.2 Sample of 19 Exam Grades

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

Figure 2.4 Dotplot



© Tracy Hebden/iStockphoto.com

Stem-and-Leaf Display

In recent years a technique known as the **stem-and-leaf display** has become very popular for summarizing numerical data. It is a combination of a graphic technique and a sorting technique. These displays are simple to create and use, and they are well suited to computer applications. A stem-and-leaf display presents the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The leading digit(s) becomes the stem, and the trailing digit(s) becomes the leaf. The stems are located along the main axis, and a leaf for each data value is located so as to display the distribution of the data.

Dotplot display Displays the data of a sample by representing each data with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale.

Stem-and-leaf display A display of the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The leading digit(s) becomes the stem, and the trailing digit(s) becomes the leaf. The stems are located along the main axis, and a leaf for each data value is located so as to display the distribution of the data.

Let's construct a stem-and-leaf display for the 19 exam scores from Table 2.2. At a quick glance we see that there are scores in the 50s, 60s, 70s, 80s, and 90s. Let's use the first digit of each score as the stem and the second digit as the leaf. Typically, the display is constructed vertically. We draw a vertical line and place the stems, in order, to the left of the line.

```
5 |
6 |
7 |
8 |
9 |
```

Next we place each leaf on its stem. This is done by placing the trailing digit on the right side of the vertical line opposite its corresponding leading digit. Our first data value is 76; 7 is the stem and 6 is the leaf. Thus, we place a 6 opposite the 7 stem:

```
7 | 6
```

The next data value is 74, so a leaf of 4 is placed on the 7 stem next to the 6.

```
7 | 6 4
```

The next data value is 82, so a leaf of 2 is placed on the 8 stem.

```
7 | 6 4
8 | 2
```

We continue until each of the other 16 leaves is placed on the display. Figure 2.5A shows the resulting stem-and-leaf display; Figure 2.5B shows the completed stem-and-leaf display after the leaves have been ordered.

From Figure 2.5B, we see that the grades are centered around the 70s. In this case, all scores with the same tens digit were placed on the same branch, but this may not always be desired. Suppose we reconstruct the display; this time instead of grouping ten possible values on each stem, let's group the values so that only five possible values could fall on each stem. Do you notice a difference in the appearance of Figure 2.6? The general shape is approximately symmetrical about the high 70s. Our information is a little more refined, but basically we see the same distribution.

It is fairly typical of many variables to display a distribution that is concentrated (mounded) about a central value and then in some manner dispersed in one or both directions. Often a graphic display reveals something that the analyst may or may not have anticipated. The example that follows demonstrates what generally occurs when two populations are sampled together.

Figure 2.5A Unfinished Stem-and-Leaf Display

19 Exam Scores

5	2
6	6 8 2
7	6 4 6 8 2 6 8 4
8	2 6 4 2 8
9	6 2

Figure 2.5B Final Stem-and-Leaf Display

19 Exam Scores

5	2
6	2 6 8
7	2 4 4 6 6 6 8 8
8	2 2 4 6 8
9	2 6

Figure 2.6 Stem-and-Leaf Display

19 Exam Scores

(50–54) 5	2
(55–59) 5	
(60–64) 6	2
(65–69) 6	6 8
(70–74) 7	2 4 4
(75–79) 7	6 6 6 8 8
(80–84) 8	2 2 4
(85–89) 8	6 8
(90–94) 9	2
(95–99) 9	6

Overlapping Distributions

Let's examine overlapping distributions by considering a random sample of 50 college students. Their weights were obtained from their medical records. The resulting data are listed in Table 2.3. Notice that the weights range from 98 to 215 pounds. Let's group the weights on stems of ten units using the hundreds and the tens digits as stems and the units digit as the leaf (see Figure 2.7). The leaves have been arranged in numerical order.

*Table 2.3 Weights of 50 College Students

Student	1	2	3	4	5	6	7	8	9	10
Male/Female	F	M	F	M	M	F	F	M	M	F
Weight	98	150	108	158	162	112	118	167	170	120
Student	11	12	13	14	15	16	17	18	19	20
Male/Female	M	M	M	F	F	M	F	M	M	F
Weight	177	186	191	128	135	195	137	205	190	120
Student	21	22	23	24	25	26	27	28	29	30
Male/Female	M	M	F	M	F	F	M	M	M	M
Weight	188	176	118	168	115	115	162	157	154	148
Student	31	32	33	34	35	36	37	38	39	40
Male/Female	F	M	M	F	M	F	M	F	M	M
Weight	101	143	145	108	155	110	154	116	161	165
Student	41	42	43	44	45	46	47	48	49	50
Male/Female	F	M	F	M	M	F	F	M	M	M
Weight	142	184	120	170	195	132	129	215	176	183

Figure 2.7 Stem-and-Leaf Display

Weights of
50 College Students (lb)
Stem-and-Leaf of WEIGHT
***N* = 50 Leaf Unit = 1.0**

9	8
10	1 8 8
11	0 2 5 5 6 8 8
12	0 0 0 8 9
13	2 5 7
14	2 3 5 8
15	0 4 4 5 7 8
16	1 2 2 5 7 8
17	0 0 6 6 7
18	3 4 6 8
19	0 1 5 5
20	5
21	5

Figure 2.8 "Back-to-Back" Stem-and-Leaf Display

Weights of 50 College Students (lb)

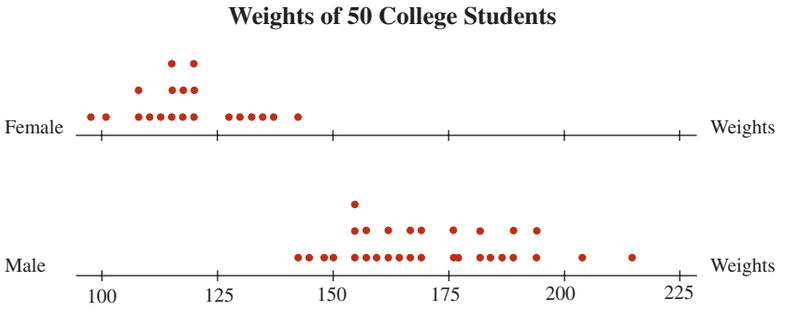
Female		Male
8	09	
1 8 8	10	
0 2 5 5 6 8 8	11	
0 0 0 8 9	12	
2 5 7	13	
2	14	3 5 8
	15	0 4 4 5 7 8
	16	1 2 2 5 7 8
	17	0 0 6 6 7
	18	3 4 6 8
	19	0 1 5 5
	20	5
	21	5

Close inspection of Figure 2.7 suggests that two overlapping distributions may be involved. That is exactly what we have: a distribution of female weights and a distribution of male weights. Figure 2.8 shows a "back-to-back" stem-and-leaf display of this set of data and makes it obvious that two distinct distributions are involved.

Figure 2.9, a "side-by-side" dotplot (same scale) of the same 50 weight data, shows the same distinction between the two subsets.

Based on the information shown in Figures 2.8 and 2.9, and on what we know about people's weight, it seems reasonable to conclude that female college students weigh less than male college students. Situations involving more than one set of data are discussed further in Chapter 3.

Figure 2.9 Dotplots with Common Scale



2.2 Frequency Distributions and Histograms

LISTS OF LARGE SETS OF DATA DO NOT PRESENT MUCH OF A PICTURE.

Sometimes we want to condense the data into a more manageable form. This can be accomplished by creating a **frequency distribution** that pairs the values of a variable with their frequency. Frequency distributions are often expressed in chart form.

To demonstrate the concept of a frequency distribution, let's use this set of data:

3	2	2	3	2	4	4	1	2	2
4	3	2	0	2	2	1	3	3	1

If we let x represent the variable, then we can use a frequency distribution to represent this set of data by listing the x values with their frequencies. For example, the value 1 occurs in the sample three times; therefore, the frequency for $x = 1$ is 3. The complete set of data is shown in the frequency distribution in Table 2.4.

Frequency distribution A listing, often expressed in chart form, that pairs values of a variable with their frequency.

Frequency The number of times the value x occurs in the sample.

Table 2.4 Ungrouped Frequency Distribution

x	f
0	1
1	3
2	8
3	5
4	3

The **frequency** f is the number of times the value x occurs in the sample. Table 2.4 is an **ungrouped frequency distribution**—"ungrouped" because each value of x in the distribution stands alone. When a large set of data has many different x values instead of a few repeated values, as in the previous example, we can group the values into a set of classes and construct a **grouped frequency distribution.** The stem-and-leaf display in Figure 2.5B (page 27) shows, in picture form, a grouped frequency distribution. Each stem represents a class. The number of leaves on each stem is the same as the frequency for that same *class* (sometimes called a *bin*). The data represented in Figure 2.5B are listed as a grouped frequency distribution in Table 2.5.

Table 2.5 Grouped Frequency Distribution

		Class	Frequency
50 or more to less than 60	⟶	$50 \le x < 60$	1
60 or more to less than 70	⟶	$60 \le x < 70$	3
70 or more to less than 80	⟶	$70 \le x < 80$	8
80 or more to less than 90	⟶	$80 \le x < 90$	5
90 or more to less than 100	⟶	$90 \le x < 100$	2
			19

The stem-and-leaf process can be used to construct a frequency distribution; however, the stem representation is not compatible with all *class widths*. For example, class widths of 3, 4, and 7 are awkward to use. Thus, sometimes it is advantageous to have a separate procedure for constructing a grouped frequency distribution.

Constructing Grouped Frequency Distribution

To illustrate this grouping (or classifying) procedure, let's use a sample of 50 final exam scores taken from last semester's elementary statistics class. Table 2.6 lists the 50 scores.

*Table 2.6 Statistics Exam Scores

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

PROCEDURE

1. Identify the high score ($H = 98$) and the low score ($L = 39$), and find the range:

   $$\text{range} = H - L = 98 - 39 = 59$$

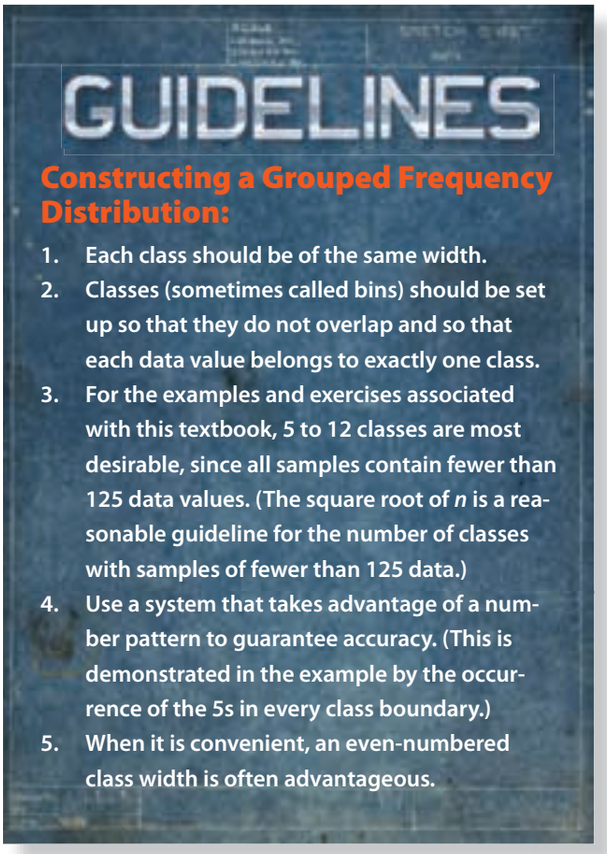
2. Select a number of classes ($m = 7$) and a class width ($c = 10$) so that the product ($mc = 70$) is a bit larger than the range (range = 59).

3. Pick a starting point. This starting point should be a little smaller than the lowest score L. Suppose we start at 35; counting from there by tens (the class width), we get 35, 45, 55, 65, ... , 95, 105. These are called the **class boundaries.** The classes for the data in Table 2.6 are:

35 or more to less than 45	⟶	$35 \le x < 45$
45 or more to less than 55	⟶	$45 \le x < 55$
55 or more to less than 65	⟶	$55 \le x < 65$
65 or more to less than 75	⟶	$65 \le x < 75$
	⋮	$75 \le x < 85$
		$85 \le x < 95$
95 or more to and including 105	⟶	$95 \le x \le 105$

NOTES:

1. At a glance you can check the number pattern to determine whether the arithmetic used to form the classes was correct (35, 45, 55, ... , 105).
2. For the interval $35 \le x < 45$, 35 is the lower class boundary and 45 is the upper class boundary. Observations that fall on the lower class boundary stay in that interval; observations that fall on the upper class boundary go into the next higher interval.
3. The class width is the difference between the upper and lower class boundaries.
4. Many combinations of class widths, numbers of classes, and starting points are possible when classifying data. There is no one best choice. Try a few different combinations, and use good judgment to decide on the one to use.



Once the classes are set up, we need to sort the data into those classes. The method used to sort will depend on the current format of the data: If the data are ranked, the frequencies can be counted; if the data are not ranked, we will **tally** the data to find the frequency

Table 2.7 Standard Chart for Frequency Distribution

Class Number	Class Tallies	Boundaries	Frequency
1	||	$35 \le x < 45$	2
2	||	$45 \le x < 55$	2
3	||||| ||	$55 \le x < 65$	7
4	||||| ||||| |||	$65 \le x < 75$	13
5	||||| ||||| |	$75 \le x < 85$	11
6	||||| ||||| |	$85 \le x < 95$	11
7	||||	$95 \le x \le 105$	4
			50

Notes: 𝍸 𝍸

1. If the data have been ranked (list form, dotplot, or stem-and-leaf), tallying is unnecessary; just count the data that belong to each class.
2. If the data are not ranked, be careful as you tally.
3. The frequency f for each class is the number of pieces of data that belong in that class.
4. The sum of the frequencies should equal the number of pieces of data n ($n = \Sigma f$). This summation serves as a good check.

numbers. When classifying data, it helps to use a standard chart (see Table 2.7).

Now you can see why it is helpful to have an even class width. An odd class width would have resulted in a class midpoint with an extra digit. (For example, the class 45–54 is 9 wide and the class midpoint is 49.5.)

Each class needs a single numerical value to represent all the data values that fall into that class. The **class midpoint** (sometimes called the *class mark*) is the numerical value that is exactly in the middle of each class. It is found by adding the class boundaries and dividing by 2. Table 2.8 shows an additional column for the class midpoint, x. As a check of your arithmetic, successive class midpoints should be a class width apart, which is 10 in this example (40, 50, 60, ... , 100 is a recognizable pattern).

Table 2.8 Frequency Distribution with Class Midpoints

Class Number	Class Boundaries	Frequency f	Class Midpoints x
1	$35 \le x < 45$	2	40
2	$45 \le x < 55$	2	50
3	$55 \le x < 65$	7	60
4	$65 \le x < 75$	13	70
5	$75 \le x < 85$	11	80
6	$85 \le x < 95$	11	90
7	$95 \le x \le 105$	4	100
		50	

When we classify data into classes, we lose some information. Only when we have all the raw data do we know the exact values that were actually observed for each class. For example, we put a 47 and a 50 into class 2, with class boundaries of 45 and 55. Once they are placed in the class, their values are lost to us and we use the class midpoint, 50, as their representative value.

Histograms

One way statisticians visually represent frequency counts of a quantitative variable is to use a bar graph called a **histogram.** A histogram is made up of three components:

1. A title, which identifies the population or sample of concern.
2. A vertical scale, which identifies the frequencies in the various classes.
3. A horizontal scale, which identifies the variable x. Values for the class boundaries or class midpoints may be labeled along the x-axis. Use whichever method of labeling the axis best presents the variable.

Class midpoint (class mark) The numerical value that is exactly in the middle of each class.

Histogram A bar graph that represents a frequency distribution of a quantitative variable.

The frequency distribution from Table 2.8 appears in histogram form in Figure 2.10.

Sometimes the **relative frequency** of a value is important. The relative frequency is a proportional measure of the frequency for an occurrence. It is found by dividing the class frequency by the total number of observations. Relative frequency can be expressed as a common fraction, in decimal form, or as a percentage. In our example about the exam scores, the frequency associated with the third class (55–65) is 7. The relative frequency for the third class is $\frac{7}{50}$, or 0.14, or 14%. Relative frequencies are often useful in a presentation because nearly everybody understands fractional parts when they are expressed as percentages. Relative frequencies are particularly useful when comparing the frequency distributions of two different size sets of data. Figure 2.11 is a **relative frequency histogram** of the sample of the 50 final exam scores from Table 2.8.

A stem-and-leaf display contains all the information needed to create a histogram, for example, Figure 2.5B (page 27). In Figure 2.12A the stem-and-leaf has been rotated 90° and labels have been added to show its relationship to a histogram. Figure 2.12B shows the same set of data as a completed histogram.

Histograms are valuable tools. For example, the histogram of a sample should have a distribution shape very similar to that of the population from which the sample was drawn. If the reader of a histogram is at all

Figure 2.10 Frequency Histogram

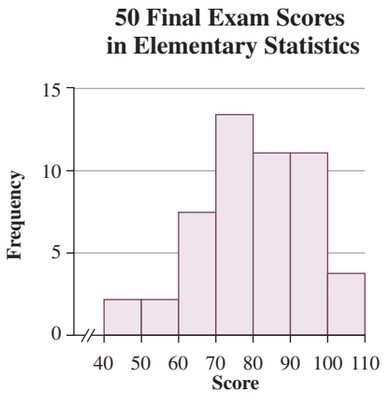


Figure 2.11 Relative Frequency Histogram

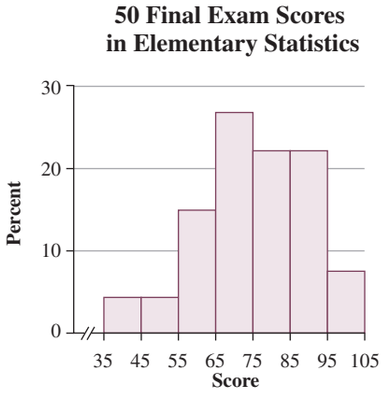


Figure 2.12A Modified Stem-and-Leaf Display

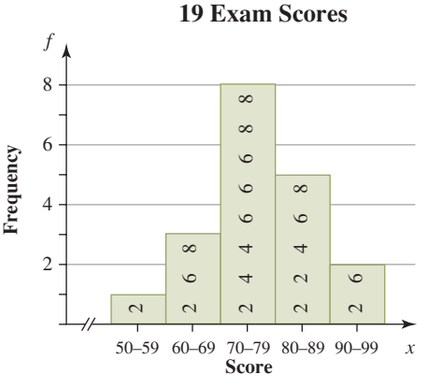


Figure 2.12B Histogram

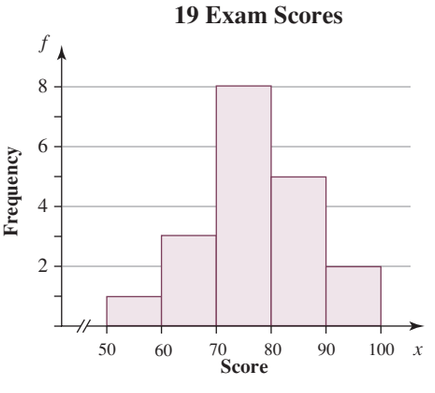
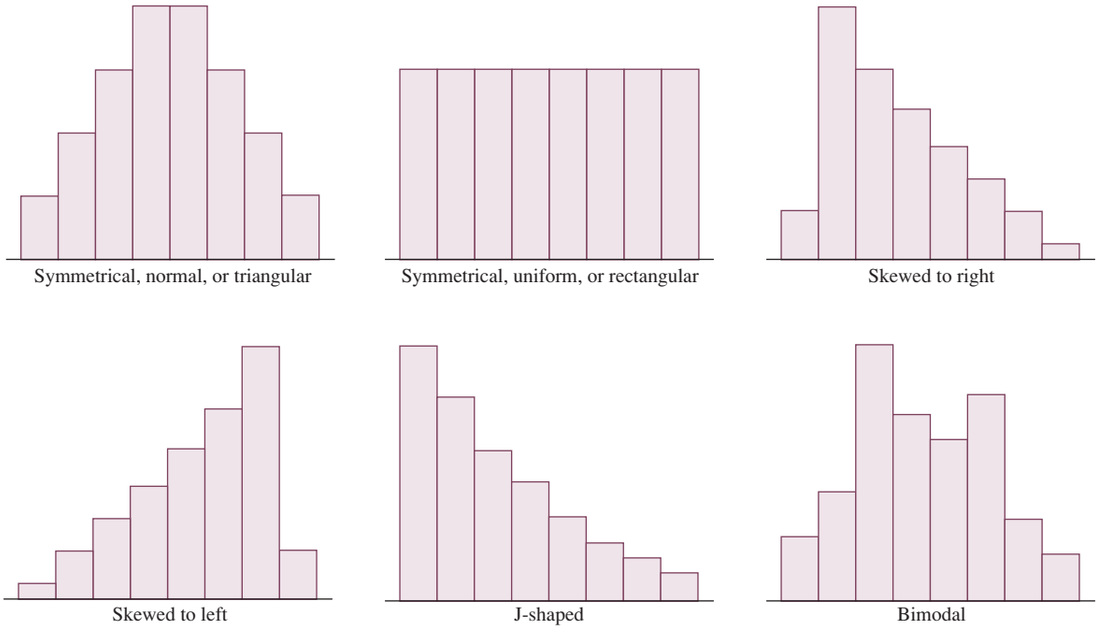


Figure 2.13 Shapes of Histograms



familiar with the variable involved, he or she will usually be able to interpret several important facts. Figure 2.13 presents histograms with descriptive labels resulting from their geometric shape.

Briefly, the terms used to describe histograms are as follows:

Symmetrical: Both sides of this distribution are identical (halves are mirror images).

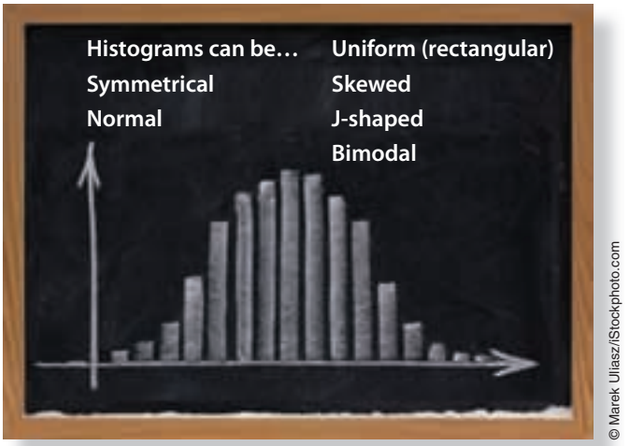
Normal (triangular): A symmetrical distribution is mounded up about the mean and becomes sparse at the extremes. (Additional properties are discussed later.)

Uniform (rectangular): Every value appears with equal frequency.

Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.

J-shaped: There is no tail on the side of the class with the highest frequency.

Bimodal: The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled. (See Figure 2.7, page 28.)



NOTES:

1. The mode is the value of the data that occurs with the greatest frequency. (Mode will be discussed in Objective 2.3)
2. The modal class is the class with the highest frequency.
3. A bimodal distribution has two high-frequency classes separated by classes with lower frequencies. It is not necessary for the two high frequencies to be the same.

Cumulative Frequency Distribution and Ogives

Another way to express a frequency distribution is to use a **cumulative frequency distribution** to pair cumulative frequencies with values of the variable.

The cumulative frequency for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values. Table 2.9 shows the cumulative frequency distribution from Table 2.8.

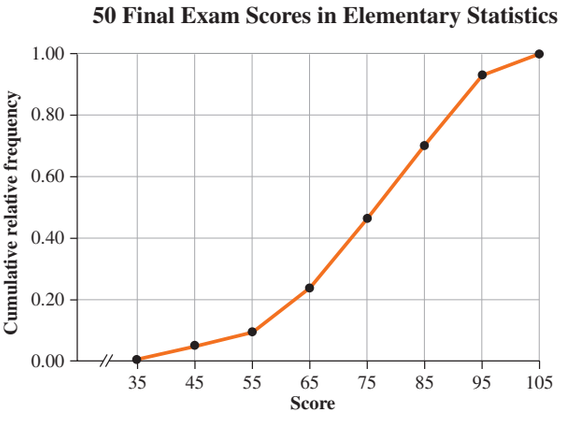
Table 2.9 Using Frequency Distribution to Form a Cumulative Frequency Distribution

Class Number	Class Boundaries	Frequency f	Cumulative Frequency
1	$35 \le x < 45$	2	2 (2)
2	$45 \le x < 55$	2	4 (2 + 2)
3	$55 \le x < 65$	7	11 (7 + 4)
4	$65 \le x < 75$	13	24 (13 + 11)
5	$75 \le x < 85$	11	35 (11 + 24)
6	$85 \le x < 95$	11	46 (11 + 35)
7	$95 \le x \le 105$	4	50 (4 + 46)

The same information can be presented by using a *cumulative relative frequency distribution* (see Table 2.10). This combines the cumulative frequency and the relative frequency ideas.

Cumulative distributions can be displayed graphically using an ogive. Whereas a histogram is a bar graph, an **ogive** is a line graph of a cumulative frequency or cumulative relative frequency distribution. An ogive has the following three components:

Figure 2.14 Ogive



1. A title, which identifies the population or sample.
2. A vertical scale, which identifies either the cumulative frequencies or the cumulative relative frequencies. (Figure 2.14 shows an ogive with cumulative relative frequencies.)
3. A horizontal scale, which identifies the upper class boundaries. Until the upper boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an ogive is always based on the upper class boundaries.

The ogive can be used to make percentage statements about numerical data much like a Pareto diagram does for attribute data. For example, suppose we want to know what percent of the final exam scores were not passing if scores of 65 or greater are considered passing.

Table 2.10 Cumulative Relative Frequency Distribution

Class Number	Class Boundaries	Cumulative Relative Frequency	*Cumulative frequencies are for the interval 35 up to the upper boundary of that class.*
1	$35 \le x < 45$	2/50, or 0.04	⟵ *from 35 up to less than 45*
2	$45 \le x < 55$	4/50, or 0.08	⟵ *from 35 up to less than 55*
3	$55 \le x < 65$	11/50, or 0.22	⟵ *from 35 up to less than 65*
4	$65 \le x < 75$	24/50, or 0.48	
5	$75 \le x < 85$	35/50, or 0.70	
6	$85 \le x < 95$	46/50, or 0.92	
7	$95 \le x < 105$	50/50, or 1.00	⟵ *from 35 up to less than 105*

Cumulative frequency distribution A frequency distribution that pairs cumulative frequencies with values of the variable.

Ogive (pronounced ō´jīv) A line graph of a cumulative frequency or cumulative relative frequency distribution.

Following vertically from 65 on the horizontal scale to the ogive line and reading from the vertical scale, approximately 22% of the final exam scores were not passing grades.

NOTE: Every ogive starts on the left with a relative frequency of zero at the lower class boundary of the first class and ends on the right with a cumulative relative frequency of 100% at the upper class boundary of the last class.

2.3 Measures of Central Tendency

MEASURES OF CENTRAL TENDENCY ARE NUMERICAL VALUES THAT LOCATE, IN SOME SENSE, THE CENTER OF A SET OF DATA.

The term *average* is often associated with all measures of central tendency, including the mean, median, mode, and midrange.

Finding the Mean

The **mean,** also called the **arithmetic mean,** is the average with which you are probably most familiar. The sample mean is represented by $\bar{x}$ (read "x-bar" or "sample mean"). The mean is found by adding all the values of the variable x (this sum of x values is symbolized Σx) and dividing the sum by the number of these values, n (the "sample size"). We express this in formula form as

$$\text{sample mean:} \quad x\text{-bar} = \frac{\text{sum of all } x}{\text{number of } x}$$

$$\bar{x} = \frac{\Sigma x}{n} \tag{2.1}$$

NOTE: The population mean, μ (lowercase mu, Greek alphabet), is the mean of all x values for the entire population.

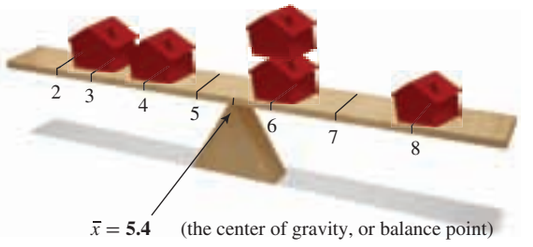
FYI: *The mean is the middle point by weight.*

Let's work on finding the mean using a set of data consisting of the five values 6, 3, 8, 6, and 4. To find the mean, we'll first use formula (2.1). Doing that, we find

$$\bar{x} = \frac{\Sigma x}{n} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5.4$$

A physical representation of the mean can be constructed by thinking of a number line balanced on a fulcrum. A weight is placed on a number on the line corresponding to each number in the sample of our example above. In Figure 2.15 there is one weight each on the 3, 8, and 4 and two weights on the 6, since there are two 6s in the sample. The mean is the value that balances the weights on the number line—in this case, 5.4.

Figure 2.15 Physical Representation of the Mean



Finding the Median

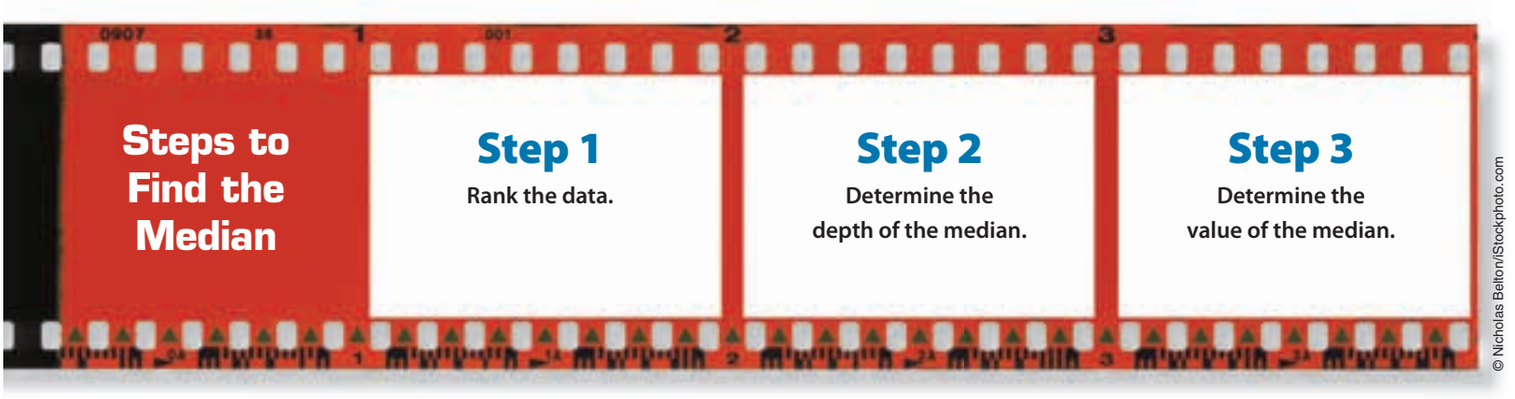
The value of the data that occupies the middle position when the data are ranked in order according to size is called the **median.** The sample median is represented by $\tilde{x}$ (read "x-tilde" or "sample median"). The population median, M (uppercase mu in the Greek alphabet), is the data value in the middle position of the entire ranked population.

Finding the median involves three basic steps. First, you need to rank the data. Then you determine the depth of the median. The **depth** (number of positions from either end), or position, of the median is determined by the formula

$$\text{depth of median} = \frac{\text{sample size} + 1}{2}$$

$$d(\tilde{x}) = \frac{n+1}{2} \tag{2.2}$$

The median's depth (or position) is found by adding the position numbers of the smallest data (1) and the largest data (n) and dividing the sum by 2 (n is the number of pieces of data). Finally, you must determine the value of the median. To do this, you count the ranked data, locating the data in the $d(\tilde{x})$th position. The median will be the same regardless of which end



of the ranked data (high or low) you count from. In fact, counting from both ends will serve as an excellent check.

The following two examples demonstrate this procedure as it applies to both odd-numbered and even-numbered sets of data.

FYI: *The value of* $d(\tilde{x})$ *is the depth of the median, NOT the value of the median* $\tilde{x}$.

Median for Odd *n*

Let's practice finding the median by first working with an odd number n. We'll find the median for the set of data {6, 3, 8, 5, 3}. First, we rank the data. In this case, the data, ranked in order of size, are 3, 3, 5, 6, and 8. Next, we'll find the depth of the median: $d(\tilde{x}) = \frac{n+1}{2} = \frac{5+1}{2} = 3$ (the "3rd" position). We can now identify the median. The median is the third number from either end in the ranked data, or $\tilde{x} = 5$.

Notice that the median essentially separates the ranked set of data into two subsets of equal size (see Figure 2.16).

As in the above example, when n is odd, the depth of the median, $d(\tilde{x})$, will always be an integer. When n is even, however, the depth of the median, $d(\tilde{x})$, will always be a half-number, as shown next.

The median is the middle point by count.

Median for Even *n*

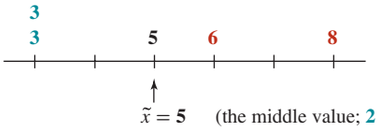
We can now compare the process we just completed with one in which we have an even number of points in our data set. Let's find the median of the sample 9, 6, 7, 9, 10, 8.

As before, we'll first rank the data by size. In this case, we have 6, 7, 8, 9, 9, and 10.

The depth of the median now is: $d(\tilde{x}) = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$ (the "3.5th" position).

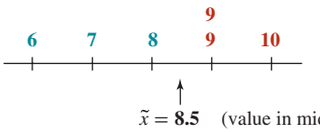
Finally, we can identify the median. The median is halfway between the third and fourth data values. To find the number halfway between any two values, add the two values together and divide the sum by 2. In this case, add the third value (8) and the fourth value (9) and then divide the sum (17) by 2. The median is $\tilde{x} = \frac{8+9}{2} = 8.5$, a number halfway between the "middle" two numbers (see Figure 2.17). Notice that the median again separates the ranked set of data into two subsets of equal size.

Figure 2.16 Median of {3, 3, 5, 6, 8}



$\tilde{x} = 5$ (the middle value; **2 data values are smaller, 2 are larger**)

Figure 2.17 Median of {6, 7, 8, 9, 9, 10}

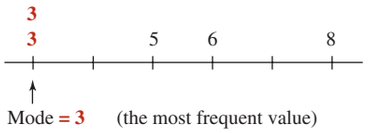


$\tilde{x} = 8.5$ (value in middle; **3 data values are smaller, 3 are larger**)

Finding the Mode

The **mode** is the value of x that occurs most frequently. In the set of data we used to find the median for odd n, {3, 3, 5, 6, 8}, the mode is 3 (see Figure 2.18).

Figure 2.18 Mode of {3, 3, 5, 6, 8}



In the sample 6, 7, 8, 9, 9, 10, the mode is 9. In this sample, only the 9 occurs more than once; in our earlier data set {6, 3, 8, 5, 3}, only the 3 occurs more than once. If two or more values in a sample are tied for the highest frequency (number of occurrences), we say there is *no mode*. For example, in the sample 3, 3, 4, 5, 5, 7, the 3 and the 5 appear an equal number of times. There is no one value that appears most often; thus, this sample has no mode.

Finding the Midrange

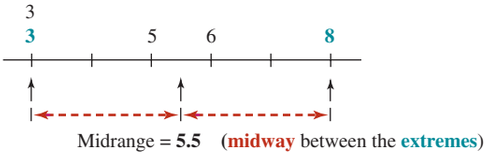
The number exactly midway between a lowest data value L and a highest data value H is called the **midrange.** To find the midrange, average the low and the high values:

$$midrange = \frac{low\ value + high\ value}{2}$$

$$midrange = \frac{L + H}{2} \qquad (2.3)$$

For the set of data {3, 3, 5, 6, 8}, $L = 3$ and $H = 8$ (see Figure 2.19).

Figure 2.19 Midrange of {3, 3, 5, 6, 8}



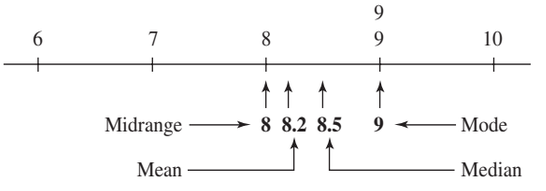
Therefore, the midrange is

$$midrange = \frac{L + H}{2} = \frac{3 + 8}{2} = 5.5$$

The four measures of central tendency represent four different methods of describing the middle. These four values may be the same, but more likely they will be different.

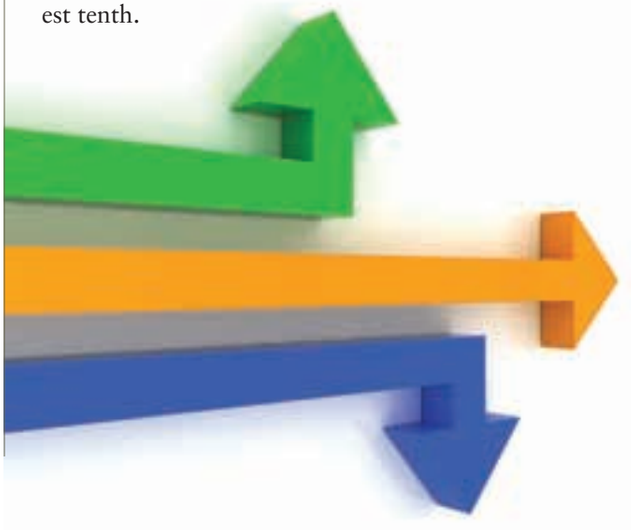
For the sample data set {6, 7, 8, 9, 9, 10}, the mean $\bar{x}$ is 8.2, the median $\tilde{x}$ is 8.5, the mode is 9, and the midrange is 8. Their relationship to one another and to the data is shown in Figure 2.20.

Figure 2.20 Measures of Central Tendency for {6, 7, 8, 9, 9, 10}



Round-Off Rule

When rounding off an answer, let's agree to keep one more decimal place in our answer than was present in the original information. To avoid round-off buildup, round off only the final answer, not the intermediate steps. That is, avoid using a rounded value to do further calculations. In our previous examples, the data were composed of whole numbers; therefore, those answers that have decimal values should be rounded to the nearest tenth.



"Average" Means Different Things

When it comes to convenience, few things can match that wonderful mathematical device called *averaging*. With an **average** you can take a fistful of figures on any subject and compute one figure that will represent the whole fistful.

But there is one thing to remember. There are several kinds of measures ordinarily known as averages, and each gives a different picture of the figures it is called on to represent. Take an example: Table 2.11 contains the annual incomes of ten families.

*Table 2.11 Annual Income of 10 Families

$54,000	$39,000	$37,500	$36,750	$35,250
$31,500	$31,500	$31,500	$31,500	$25,500

What would this group's "typical" income be? Averaging would provide the answer, so let's compute the typical income by the simpler and more frequently used kinds of averaging.

- **The arithmetic mean.** It is the most common form of average, obtained by adding items in the series and then dividing by the number of items: $35,400. The mean is representative of the series in the sense that the sum of the amounts by which the higher figures exceed the mean is exactly the same as the sum of the amounts by which the lower figures fall short of the mean.
- **The median.** As you may have observed, six families earn less than the mean, four earn more. You might very well wish to represent this varied group by the income of the family that is right smack dab in the middle of the whole bunch. The median works out to $33,375.
- **The midrange.** Another number that might be used to represent the group is the midrange, computed by calculating the figure that lies halfway between the highest and lowest incomes: $39,750.
- **The mode.** So far we've seen three kinds of averages, and not one family actually has an income matching any of them. Say you want to represent the group by stating the income that occurs most frequently. That is called a mode. $31,500 would be the modal income.

Four different averages, each valid, correct, and informative in its way. But how they differ!

Arithmetic Mean	Median	Midrange	Mode
$35,400	$33,375	$39,750	$31,500

And they would differ still more if just one family in the group were a millionaire—or one were jobless!

So there are three lessons: First, when you see or hear an average, find out which average it is. Then you'll know what kind of picture you are being given. Second, think about the figures being averaged so you can judge whether the average used is appropriate. And third, don't assume that a literal mathematical quantification is intended every time somebody says "average." It isn't. All of us often say "the average person" with no thought of implying a mean, median, or mode. All we intend to convey is the idea of other people who are in many ways a great deal like the rest of us.

SOURCE: Reprinted by permission from *Changing Times* magazine (March 1980 issue). Copyright by The Kiplinger Washington Editors.

BASIC TERMS – TAKE TWO

Mean (arithmetic mean) The mean, also called the arithmetic mean, is the average with which you are probably most familiar. The sample mean is represented by $\bar{x}$ (read "x-bar" or "sample mean"). The mean is found by adding all the values of the variable x (this sum of x values is symbolized Σx) and dividing the sum by the number of these values, n (the "sample size").

Median The value of the data that occupies the middle position when the data are ranked in order according to size. The sample median is represented by $\tilde{x}$ (read "x-tilde" or "sample median").

Mode The mode is the value of x that occurs most frequently.

Midrange The number exactly midway between the lowest-valued data L and the highest-valued data H.

Range The difference in value between the highest data value (H) and the lowest data value (L).

Deviation from the mean A deviation from the mean, $x - \bar{x}$, is the difference between the value of x and the mean $\bar{x}$.

Sample variance The sample variance, s^2, is the mean of the squared deviations.

Sample standard deviation The standard deviation of a sample, s, is the positive square root of the variance.

2.4 Measures of Dispersion

HAVING LOCATED THE "MIDDLE" WITH THE MEASURES OF CENTRAL TENDENCY, OUR SEARCH FOR INFORMATION FROM DATA SETS NOW TURNS TO THE MEASURES OF DISPERSION (SPREAD).

The measures of dispersion include the range, variance, and standard deviation. These numerical values describe the amount of spread, or variability, that is found among the data: Closely grouped data have relatively small values, and more widely spread out data have larger values. The closest possible grouping occurs when the data have no dispersion (all data are the same value); in this situation, the measure of dispersion will be zero. There is no limit to how widely spread out the data can be; therefore, measures of dispersion can be very large. The simplest measure of dispersion is **range,** which is the difference in value between the highest data value (H) and the lowest data value (L):

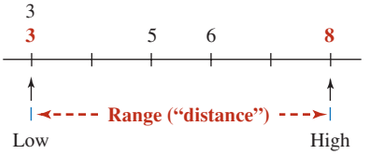
$$range = high\ value - low\ value$$

$$range = H - L \tag{2.4}$$

The sample 3, 3, 5, 6, 8 has a range of $H - L = 8 - 3 = 5$. The range of 5 tells us that these data all fall within a 5-unit interval (see Figure 2.21).

The other measures of dispersion to be studied in this chapter are measures of dispersion about the mean.

Figure 2.21 Range of {3, 3, 5, 6, 8}



To develop a measure of dispersion about the mean, let's first answer the following question: How far is each x from the mean? The difference between the value of x and the mean $\bar{x}$, or $x - \bar{x}$, is called a **deviation from the mean.** Each individual value x deviates from the mean by an amount equal to $(x - \bar{x})$. This deviation $(x - \bar{x})$ is zero when x is equal to the mean $\bar{x}$. The deviation $(x - \bar{x})$ is positive when x is larger than $\bar{x}$ and negative when x is smaller than $\bar{x}$.

When you square the deviations and take an average of those, you get something called the **sample variance,** s^2. It is calculated using $n - 1$ as the divisor:

sample variance:

$$s\text{-squared} = \frac{\text{sum of (deviations squared)}}{\text{number} - 1}$$

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} \qquad (2.5)$$

where n is the sample size—that is, the number of data values in the sample.

The variance of the sample 6, 3, 8, 5, 3 is calculated in Table 2.12 using formula (2.5).

NOTES:

1. The sum of all the x values is used to find $\bar{x}$.
2. The sum of the deviations, $\Sigma(x - \bar{x})$, is always zero, provided the exact value of $\bar{x}$ is used. Use this fact as a check in your calculations, as was done in Table 2.12 (denoted by ✓).
3. If a rounded value of $\bar{x}$ is used, then $\Sigma(x - \bar{x})$ will not always be exactly zero. It will, however, be reasonably close to zero.
4. The sum of the squared deviations is found by squaring each deviation and then adding the squared values.

To graphically demonstrate what variances of data sets are telling us, consider a second set of data: {1, 3, 5, 6, 10}. Note that the data values are more dispersed than the data in Table 2.12. Accordingly, its calculated variance is larger at $s^2 = 11.5$. An illustrative side-by-side graphical comparison of these two samples and their variances is shown in Figure 2.22.

Sample Standard Deviation

Variance is instrumental in the calculation of the **standard deviation of a sample,** s, which is the positive square root of the variance:

sample standard deviation:

$$s = \text{square root of sample variance}$$

$$s = \sqrt{s^2} \qquad (2.6)$$

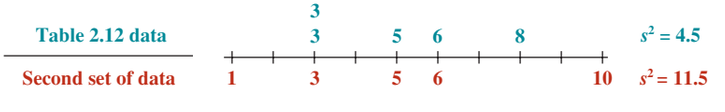
For the samples shown in Figure 2.22, the standard deviations are $\sqrt{4.5}$ or 2.1, and $\sqrt{11.5}$ or 3.4.

NOTE: The numerator for the sample variance, $\Sigma(x - \bar{x})^2$, is often called the *sum of squares for x* and

Table 2.12 Calculating Variance Using Formula (2.5)

Step 1. Find Σx	Step 2. Find $\bar{x}$	Step 3. Find each $x - \bar{x}$	Step 4. Find $\Sigma(x - \bar{x})^2$	Step 5. Find s^2
6	$\bar{x} = \frac{\Sigma x}{n}$	$6 - 5 = 1$	$(1)^2 = 1$	$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$
3		$3 - 5 = -2$	$(-2)^2 = 4$	
8		$8 - 5 = 3$	$(3)^2 = 9$	
5	$\bar{x} = \frac{25}{5}$	$5 - 5 = 0$	$(0)^2 = 0$	$s^2 = \frac{18}{4}$
3		$3 - 5 = -2$	$(-2)^2 = 4$	
$\Sigma x = 25$	$\bar{x} = 5$	$\Sigma(x - \bar{x}) = 0$ ✓	$\Sigma(x - \bar{x})^2 = 18$	$s^2 = 4.5$

Figure 2.22 Comparison of Data



symbolized by SS(x). Thus, formula (2.5) can be expressed as

$$\text{sample variance:} \quad s^2 = \frac{SS(x)}{n-1} \tag{2.7}$$

The formulas for variance can be modified into other forms for easier use in various situations.

The arithmetic becomes more complicated when the mean contains nonzero digits to the right of the decimal point. However, the **sum of squares for x,** the numerator of formula (2.5), can be rewritten so that $\bar{x}$ is not included:

$$\text{sum of squares:} \quad SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \tag{2.8}$$

Combining formulas (2.7) and (2.8) yields the "shortcut formula" for sample variance:

$$\textit{s-squared} = \frac{(\textit{sum of } x^2) - \left[\dfrac{(\textit{sum of } x)^2}{\textit{number}}\right]}{\textit{number} - 1}$$

$$\text{sample variance:} \quad s^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n-1} \tag{2.9}$$

Formulas (2.8) and (2.9) are called "shortcuts" because they bypass the calculation of $\bar{x}$. The computations for SS(x), s^2, and s using formulas (2.8), (2.9), and (2.6) are performed as shown in Table 2.13.

The unit of measure for the standard deviation is the same as the unit of measure for the data. For example, if our data are in pounds, then the standard deviation s will also be in pounds. The unit of measure for variance might then be thought of as *units squared*. In our example of pounds, this would be *pounds squared*. As you can see, the unit has very little meaning.

2.5 Measures of Position

MEASURES OF POSITION ARE USED TO DESCRIBE THE POSITION A SPECIFIC DATA VALUE POSSESSES IN RELATION TO THE REST OF THE DATA.

Table 2.13 Calculating Standard Deviation Using the Shortcut Method

Step 1. Find Σx	Step 2. Find Σx^2	Step 3. Find SS(x)	Step 4. Find s^2	Step 5. Find s
6	$6^2 = 36$	$SS(x) = \Sigma x^2 - \dfrac{(\Sigma x)^2}{n}$	$s^2 = \dfrac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n-1}$	$s = \sqrt{s^2}$
3	$3^2 = 9$			$s = \sqrt{5.7}$
8	$8^2 = 64$	$SS(x) = 138 - \dfrac{(24)^2}{5}$		$s = 2.4$
5	$5^2 = 25$		$s^2 = \dfrac{22.8}{4}$	
2	$2^2 = 4$	$SS(x) = 138 - 115.2$		
$\Sigma x = 24$	$\Sigma x^2 = 138$	$SS(x) = 22.8$	$s^2 = 5.7$	

Quartiles and *percentiles* are two of the most popular measures of position. Other measures of position include midquartiles, 5-number summaries, and standard scores, or z-scores.

Quartiles

Quartiles are values of the variable that divide the ranked data into quarters; each set of data has three quartiles. The *first quartile,* Q_1, is a number such that at most 25% of the data are smaller in value than Q_1 and at most 75% are larger. The *second quartile* is the median. The *third quartile,* Q_3, is a number such that at most 75% of the data are smaller in value than Q_3 and at most 25% are larger (see Figure 2.23).

The procedure for determining the values of the quartiles is the same as that for **percentiles,** which are the values of the variable that divide a set of ranked data into 100 equal subsets; each set of data has 99 percentiles (see Figure 2.24). The kth percentile, P_k, is a value such that at most k% of the data are smaller in value than P_k and at most $(100 - k)$% of the data are larger (see Figure 2.25).

NOTES:

1. The first quartile and the 25th percentile are the same; that is, $Q_1 = P_{25}$. Also, $Q_3 = P_{75}$.
2. The median, the second quartile, and the 50th percentile are all the same: $\tilde{x} = Q_2 = P_{50}$. Therefore, when asked to find P_{50} or Q_2, use the procedure for finding the median.

Percentiles

The procedure for determining the value of any kth percentile (or quartile) involves four basic steps as outlined in Figure 2.26.

Using the sample of 50 elementary statistics final exam scores listed in Table 2.14, find the first quartile Q_1, the 58th percentile P_{58}, and the third quartile Q_3.

Quartiles Values of the variable that divide the ranked data into quarters; each set of data has three quartiles.

Percentiles Values of the variable that divide a set of ranked data into 100 equal subsets; each set of data has 99 percentiles.

Figure 2.23 Quartiles

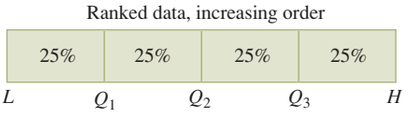


Figure 2.24 Percentiles

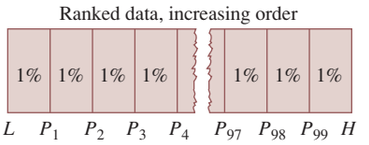


Figure 2.25 kth Percentile

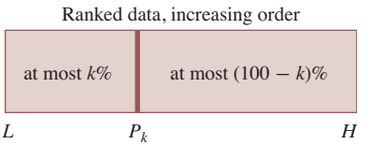


Table 2.14 Raw Scores for Elementary Statistics Exam

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

Figure 2.26 Finding P_k Procedure

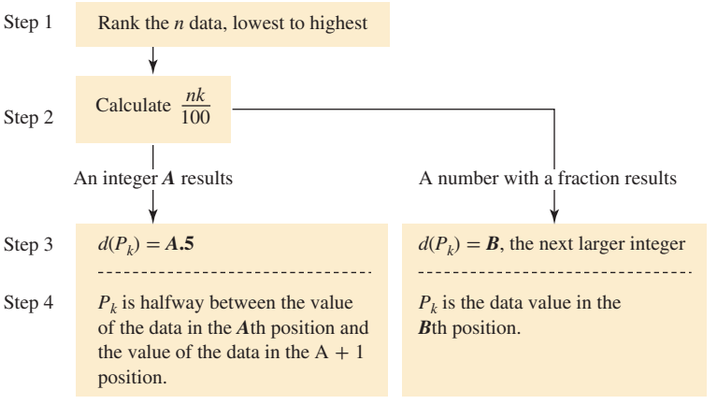


Table 2.15
Ranked Data: Exam Scores

39	64	72	78	89
44	66	72	80	90
47	67	74	82	90
50	68	74	82	91
55	68	75	83	92
58	70	76	85	94
58	70	77	86	95
60	70	77	86	95
63	72	77	88	97
64	72	78	88	98

2nd position from L
13th position from L
29th and 30th positions from L
2nd position from H
13th position from H

Figure 2.27
Final Exam Scores

Stem-and-leaf of score $N = 50$
Leaf Unit = 1.0

1	3	9
2	4	4
3	4	7
4	5	0
7	5	5 8 8
11	6	0 3 4 4
15	6	6 7 8 8
24	7	0 0 0 2 2 2 2 4 4
(7)	7	5 6 7 7 7 8 8
19	8	0 2 2 3
15	8	5 6 6 8 8 9
9	9	0 0 1 2 4
4	9	5 5 7 8

PROCEDURE

Step 1 Rank the data: A ranked list may be formulated (see Table 2.15), or a graphic display showing the ranked data may be used. The dotplot and the stem-and-leaf are handy for this purpose. The stem-and-leaf is especially helpful since it gives depth numbers counted from both extremes when it is computer generated (see Figure 2.27). Step 1 is the same for all three statistics.

Find Q_1:

Step 2 Find $\frac{nk}{100}$: $\frac{nk}{100} = \frac{(50)(25)}{100} = \mathbf{12.5}$
($n = 50$ and $k = 25$, since $Q_1 = P_{25}$.)

Step 3 Find the depth of Q_1: $d(Q_1) = \mathbf{13}$
(Since 12.5 contains a fraction, B is the next larger integer, 13.)

Step 4 Find Q_1: Q_1 is the 13th value, counting from L (see Table 2.15 or Figure 2.27), $Q_1 = \mathbf{67}$

Find P_{58}:

Step 2 Find $\frac{nk}{100}$: $\frac{nk}{100} = \frac{(50)(58)}{100} = \mathbf{29}$
($n = 50$ and $k = 58$ for P_{58}.)

Step 3 Find the depth of P_{58}: $d(P_{58}) = 29.5$
(Since $A = 29$, an integer, add 0.5 and use 29.5.)

Step 4 Find P_{58}: P_{58} is the value halfway between the values of the 29th and the 30th pieces of data, counting from L (see Table 2.15 or Figure 2.27), so

$$P_{58} = \frac{77 + 78}{2} = 77.5$$

FYI: $d(P_k)$ = *depth or location of the* k^{th} *percentile.*

Optional technique: **When k is greater than 50, subtract k from 100 and use $(100 - k)$ in place of k in Step 2. The depth is then counted from the largest data value H.**

Find Q_3 using the optional technique:

Step 2 **Find $\frac{nk}{100}$: $\frac{nk}{100} = \frac{(50)(25)}{100} = \mathbf{12.5}$**
($n = 50$ and $k = 75$, since $Q_3 = P_{75}$, and $k > 50$; use $100 - k = 100 - 75 = 25$.)

Step 3 **Find the depth of Q_3 from H: $d(Q_3) = \mathbf{13}$**

Step 4 **Find Q_3: Q_3 is the 13th value, counting from H (see Table 2.15 or Figure 2.27), $Q_3 = \mathbf{86}$**

Therefore, it can be stated that "at most 75% of the exam grades are smaller in value than 86." This is also equivalent to stating that "at most 25% of the exam grades are larger in value than 86."

Therefore, it can be stated that "at most 58% of the exam grades are smaller in value than 77.5." This is also equivalent to stating that "at most 42% of the exam grades are larger in value than 77.5."

NOTE: An ogive of these grades would graphically determine these same percentiles, without the use of formulas.

Other Measures of Position

Let's now examine three other measures of position: midquartile, 5-number summary, and standard scores.

Midquartiles

Using the fundamental calculations of quartiles, you can now calculate the measure of central tendency known as the **midquartile,** or the numerical value midway between the first quartile and the third quartile.

$$\text{midquartile} = \frac{Q_1 + Q_3}{2} \qquad (2.10)$$

So, to find the midquartile for the set of 50 exam scores given in our exam score example, you would simply add 67 to 86 and divide by 2.

$Q_1 = 67$ and $Q_3 = 86$, thus,

$$\text{midquartile} = \frac{Q_1 + Q_3}{2} = \frac{67 + 86}{2} = \mathbf{76.5}$$

The median, the midrange, and the midquartile are not necessarily the same value. Each is the middle value, but by different definitions of "middle." Figure 2.28 summarizes the relationship of these three statistics as applied to our set of 50 exam scores.

5-Number Summary

Another measure of position based on quartiles and percentiles is the **5-number summary.** Not only is the 5-number summary very effective in describing a set of data, it is easy information to obtain and is very informative to the reader.

The 5-number summary is composed of:

1. L, the smallest value in the data set,
2. Q_1, the first quartile (also called P_{25}, the 25th percentile),
3. $\tilde{x}$, the median,
4. Q_3, the third quartile (also called P_{75}, the 75th percentile), and
5. H, the largest value in the data set.

The 5-number summary for our set of 50 exam scores is

39	67	75.5	86	98
L	Q_1	$\tilde{x}$	Q_3	H

Notice that these five numerical values divide the set of data into four subsets, with one-quarter of the data in each subset. From the 5-number summary we can observe how much the data are spread out in each of the quarters. We can now define an additional measure of dispersion. The **interquartile range** is the difference between the first and third quartiles. It is the range of the middle 50% of the data. The 5-number summary makes it very easy to see the interquartile range.

The 5-number summary is even more informative when it is displayed on a diagram drawn to scale. A

Midquartile The numerical value midway between the first quartile and the third quartile.

5-number summary The presentation of 5 numbers that give a statistical summary of a data set: the smallest value in the data set, the first quartile, the median, the third quartile, and the largest value in the data set.

Interquartile range The difference between the first and third quartiles. It is the range of the middle 50% of the data.

Figure 2.28 Final Exam Scores

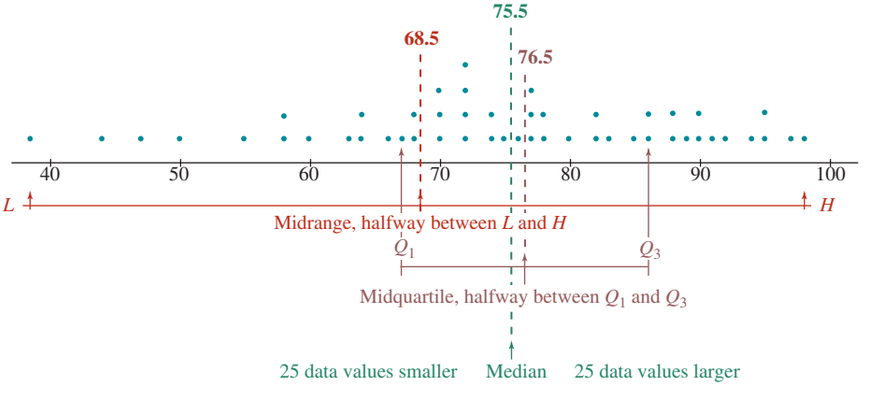
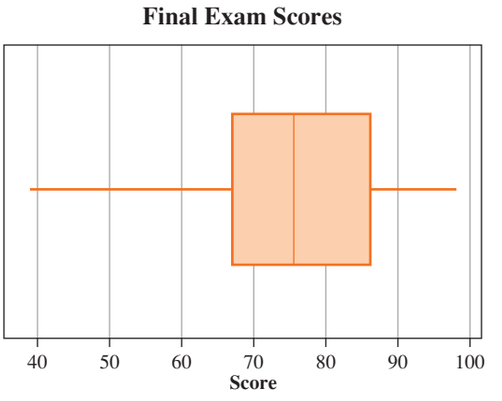


Figure 2.29 Box-and-Whiskers Display



computer-generated graphic display that accomplishes this is known as the **box-and-whiskers display.** In this graphic representation of the 5-number summary, the five numerical values (smallest, first quartile, median, third quartile, and largest) are located on a scale, either vertical or horizontal. The box is used to depict the middle half of the data that lie between the two quartiles. The whiskers are line segments used to depict the other half of the data: One line segment represents the quarter of the data that are smaller in value than the first quartile, and a second line segment represents the quarter of the data that are larger in value than the third quartile.

Figure 2.29 is a box-and-whiskers display of the 50 exam scores.

Standard Scores (*z*-scores)

So far, we've examined *general* measures of position, but sometimes it is necessary to measure the position of a *specific* value in terms of the mean and standard deviation. In those cases, the *standard score,* commonly called the *z-score,* is used. The **standard score** (or ***z*-score**) is the position a particular value of x has relative to the mean, measured in standard deviations. The z-score is found by the formula:

$$z = \frac{value - mean}{st.\ dev.} = \frac{x - \bar{x}}{s} \qquad (2.11)$$

Let's apply this formula to finding the standard scores for (a) 92 and (b) 72 with respect to a sample of exam grades that has a mean score of 74.92 and a standard deviation of 14.20.

Box-and-whiskers display A graphic representation of the 5-number summary.

Standard score or *z*-score The position a particular value of x has relative to the mean, measured in standard deviations.

SOLUTION

a. $x = 92, \bar{x} = 74.92, s = 14.20.$

Thus, $z = \frac{x - \bar{x}}{s} = \frac{92 - 74.92}{14.20} = \frac{17.08}{14.20} = \mathbf{1.20.}$

b. $x = 72, \bar{x} = 74.92, s = 14.20.$

Thus, $z = \frac{x - \bar{x}}{s} = \frac{72 - 74.92}{14.20} = \frac{-2.92}{14.20} = \mathbf{-0.21.}$

This means that the score 92 is approximately one and one-fifth standard deviations above the mean, while the score 72 is approximately one-fifth of a standard deviation below the mean.

NOTES:

1. Typically, the calculated value of z is rounded to the nearest hundredth.
2. z-scores typically range in value from approximately -3.00 to $+3.00$.

Because the z-score is a measure of relative position with respect to the mean, it can be used to help us compare two raw scores that come from separate populations. For example, suppose you want to compare a

grade you received on a test with a friend's grade on a comparable exam in her course. You received a raw score of 45 points; she got 72 points. Is her grade better? We need more information before we can draw a conclusion. Suppose the mean on the exam you took was 38 and the mean on her exam was 65. Your grades are both 7 points above the mean, but we still can't draw a definite conclusion. The standard deviation on the exam you took was 7 points, and it was 14 points on your friend's exam. This means that your score is one (1) standard deviation above the mean ($z = 1.0$), whereas your friend's grade is only one-half of a standard deviation above the mean ($z = 0.5$). Since your score has the "better" relative position, you conclude that your score is slightly better than your friend's score. (Again, this is speaking from a relative point of view.)

2.6 Interpreting and Understanding Standard Deviation

STANDARD DEVIATION IS A MEASURE OF VARIATION (DISPERSION) IN THE DATA.

It has been defined as a value calculated with the use of formulas. Even so, you may be wondering what it really is and how it relates to the data. It is a kind of yardstick by which we can compare the variability of one set of data with another. This particular "measure" can be understood further by examining two statements that tell us how the standard deviation relates to the data: the *empirical rule* and *Chebyshev's theorem*.

The Empirical Rule and Testing for Normality

The **empirical rule** states that if a variable is normally distributed, then: within one standard deviation of the mean there will be approximately 68% of the data; within two standard deviations of the mean there will be approximately 95% of the data; and within three standard deviations of the mean there will be approximately 99.7% of the data. This rule applies specifically to a *normal (bell-shaped) distribution*, but it is frequently applied as an interpretive guide to any mounded distribution.

Figure 2.30 shows the intervals of one, two, and three standard deviations about the mean of an approximately normal distribution. Usually these proportions do not occur exactly in a sample, but your observed values will be close when a large sample is drawn from a normally distributed population.

If a distribution is approximately normal, it will be nearly symmetrical and the mean will divide the distribution in half (the mean and the median are the same in a symmetrical distribution). This allows us to refine the empirical rule, as shown in Figure 2.31.

Figure 2.30 Empirical Rule

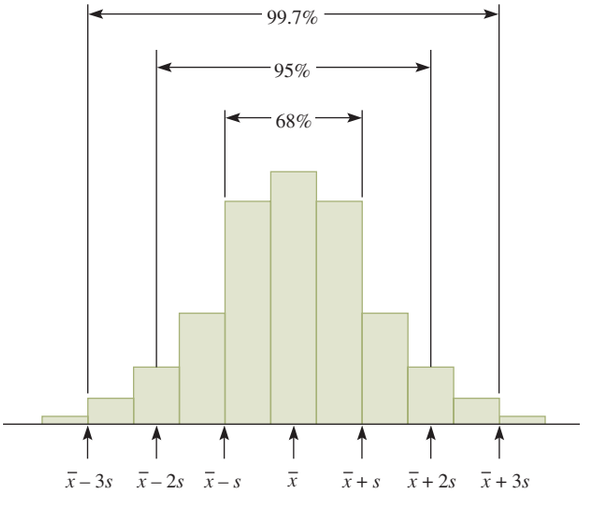
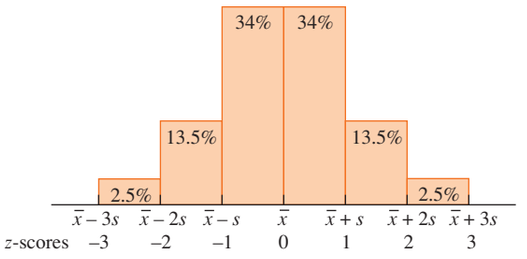


Figure 2.31 Refinement of Empirical Rule



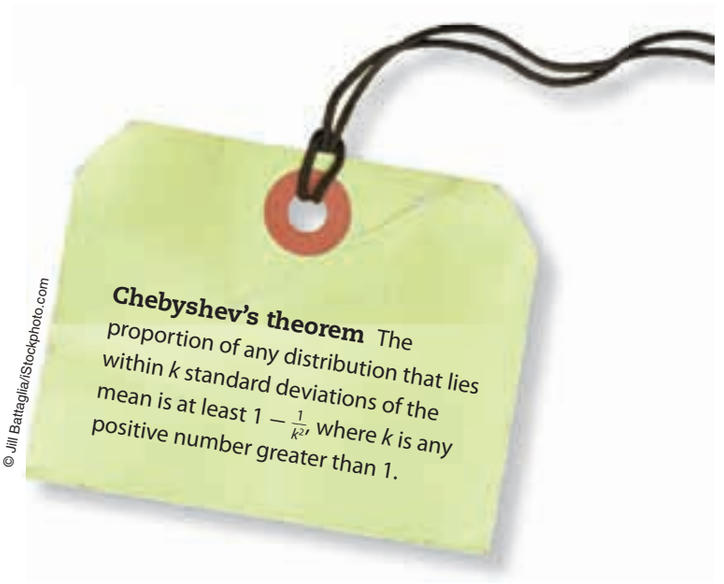
The empirical rule can be used to determine whether or not a set of data is approximately normally distributed. Let's demonstrate this application by working with the distribution of final exam scores that we have been using throughout this chapter. The mean, $\bar{x}$, was found to be 74.92, and the standard deviation, s, was 14.20. The interval from one standard deviation below the mean, $\bar{x} - s$, to one standard deviation above the mean, $\bar{x} + s$, is $74.92 - 14.20 = 60.72$ to $74.92 + 14.20 = 89.12$. This interval (60.72 to 89.12) includes 61, 62, 63, ... , 89. Upon inspection of the ranked data (Table 2.15, page 43), we see that 34 of the 50 data, or 68%, lie within one standard deviation of the mean. Furthermore, $\bar{x} - 2s = 74.92 - (2)(14.20) = 74.92 - 28.40 = 46.52$ to $\bar{x} + 2s = 74.92 + 28.40 = 103.32$ gives the interval from 46.52 to 103.32. Of the 50 data, 48, or 96%, lie within two standard deviations of the mean. All 50 data, or 100%, are included within three standard deviations of the mean (from 32.32 to 117.52). This information can be placed in a table for comparison with the values given by the empirical rule (see Table 2.16).

Table 2.16 Observed Percentages versus the Empirical Rule

Interval	Empirical Rule Percentage	Percentage Found
$\bar{x} - s$ to $\bar{x} + s$	≈68	68
$\bar{x} - 2s$ to $\bar{x} + 2s$	≈95	96
$\bar{x} - 3s$ to $\bar{x} + 3s$	≈99.7	100

The percentages found are reasonably close to those predicted by the empirical rule. By combining this evidence with the shape of the histogram, we can safely say that the final exam data are approximately normally distributed.

Another method for testing normality is to draw a probability plot using a computer or graphing calculator.



Chebyshev's Theorem

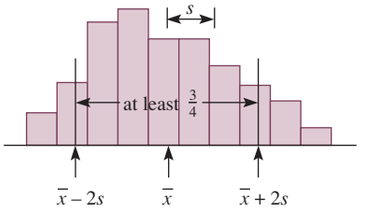
In the event that the data do not display an approximately normal distribution, **Chebyshev's theorem** gives us information about how much of the data will fall within intervals centered at the mean for all distributions. It states that the proportion of any distribution that lies within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$, where k is any positive number greater than 1. This theorem applies to all distributions of data.

This theorem says that within two standard deviations of the mean ($k = 2$), you will always find at least 75% (that is, 75% or more) of the data:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0.75, \textbf{ at least 75\%}$$

Figure 2.32 shows a mounded distribution that illustrates at least 75%.

Figure 2.32 Chebyshev's Theorem with $k = 2$

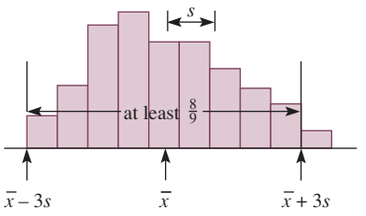


If we consider the interval enclosed by three standard deviations on either side of the mean ($k = 3$), the theorem says that we will always find at least 89% (that is, 89% or more) of the data:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 0.89, \textbf{ at least 89\%}$$

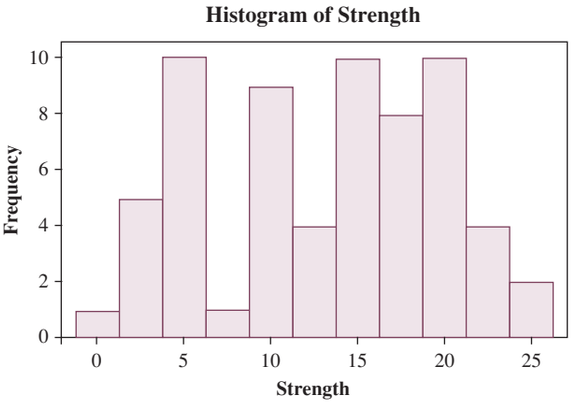
Figure 2.33 shows a mounded distribution that illustrates at least 89%.

Figure 2.33 Chebyshev's Theorem with $k = 3$



Imagine that all the third graders at Roth Elementary School were given a physical-fitness strength test. Their test results are listed below in rank order and are shown on the histogram (*data set).

1	2	2	3	3	3	4	4	4	5	5	5	5	6	6	6
8	9	9	9	9	9	9	10	10	11	12	12	12	13	14	14
14	15	15	15	15	16	16	16	17	17	17	17	18	18	18	18
19	19	19	19	20	20	20	21	21	21	22	22	22	23	24	24



Some questions of interest are: Does this distribution satisfy the empirical rule? Does Chebyshev's theorem hold true? Is this distribution approximately normal?

To answer the first two questions, we need to find the percentages of data in each of the three intervals about the mean. The mean is 13.0, and the standard deviation is 6.6.

Mean ± (k × Std. Dev.)	Interval	Percentage Found	Empirical	Chebyshev
13.0 ± (1 × 6.6)	6.4 to 19.6	36/64 = 56.3%	68%	—
13.0 ± (2 × 6.6)	−0.2 to 26.2	64/64 = 100%	95%	At least 75%
13.0 ± (3 × 6.6)	−6.8 to 32.8	64/64 = 100%	99.70%	At least 89%

It is left to you to verify the values of the mean, the standard deviation, the intervals, and the percentages.

The three percentages found (56.3, 100, and 100) do not approximate the 68, 95, and 99.7 percentages stated in the empirical rule. The two percentages found (100 and 100) do agree with Chebyshev's theorem in that they are greater than 75% and 89%. Remember, Chebyshev's theorem holds for all distributions. With the distribution seen on the histogram and the three percentages found, it is reasonable to conclude that these test results are not normally distributed.

2.7 The Art of Statistical Deception

"THERE ARE THREE KINDS OF LIES—LIES, DAMNED LIES, AND STATISTICS."

These remarkable words spoken by Benjamin Disraeli (19th-century British prime minister) represent the cynical view of statistics held by many people. Most people are on the consumer end of statistics and therefore have to "swallow" them.

Good Arithmetic, Bad Statistics

Let's explore an outright statistical lie. Suppose a small business employs eight people who earn between $300 and $350 per week. The owner of the business pays himself $1,250 per week. He reports to the general public that the average wage paid to the employees of his firm is $430 per week. That may be an example of good arithmetic, but it is bad statistics. It is a misrepresentation of the situation because only one employee, the owner, receives more than the mean salary. The public will think that most of the employees earn about $430 per week.

Graphic Deception

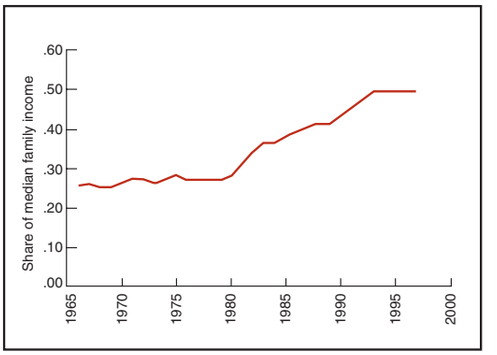
Graphic representations can be tricky and misleading. The frequency scale (which is usually the vertical axis) should start at zero in order to present a total picture. Usually, graphs that do not start at zero are used to save space. Nevertheless, this can be deceptive. Graphs in which the frequency scale starts at zero tend to emphasize the size of the numbers involved, whereas graphs that are chopped off may tend to emphasize the variation in the numbers without regard to the actual size of the numbers. The labeling of the horizontal scale can be misleading also. You need to inspect graphic presentations very carefully before you draw any conclusions from the "story being told."

Superimposed Misrepresentation

The "clever" graphic overlay from the *Ithaca Times* (December 7, 2000) has to be the worst graph ever to make a front page. The cover story, "Why does college have to cost so much?" pictures two graphs superimposed on a Cornell University campus scene. The two broken lines represent "Cornell's Tuition" and "Cornell's Ranking," with the tuition steadily increasing and the ranking staggering and falling. A very clear image is created: Students get less and pay more!

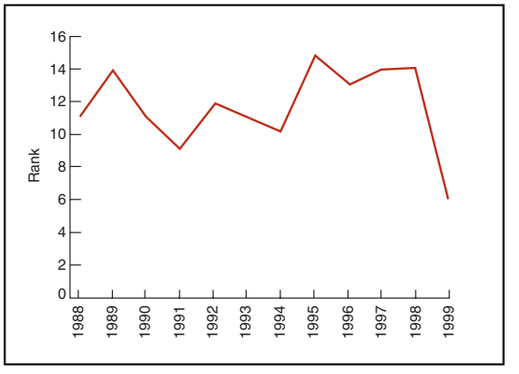
http://www.math.yorku.ca/SCS/Gallery/context.html

Now view the two complete graphs separately. Notice: (1) The graphs cover two different time periods. (2) The vertical scales differ. (3) The "best" misrepresentation comes from the impression that a "drop in rank" represents a lower quality of education. Wouldn't a rank of 6 be better than a rank of 15?



BY THE NUMBERS: OVER 35 YEARS, CORNELL'S TUITION HAS TAKEN AN INCREASINGLY LARGER SHARE OF ITS MEDIAN STUDENT FAMILY INCOME

SOURCE: http://www.math.yorku.ca/SCS/Gallery/context.html



PECKING ORDER: OVER 12 YEARS, CORNELL'S RANKING IN *US NEWS & WORLD REPORT* HAS RISEN AND FALLEN ERRATICALLY.

What it all comes down to is that statistics, like all languages, can be and is abused. In the hands of the careless, the unknowledgeable, or the unscrupulous, statistical information can be as false as "damned lies."

problems

Objective 2.1

2.1 Results from Self.com poll on "What is your top cold-weather beauty concern?" were reported in the December 2008 issue of *Self* magazine: dry skin—57%, chapped lips—25%, dull hair—10%, rough feet—8%.

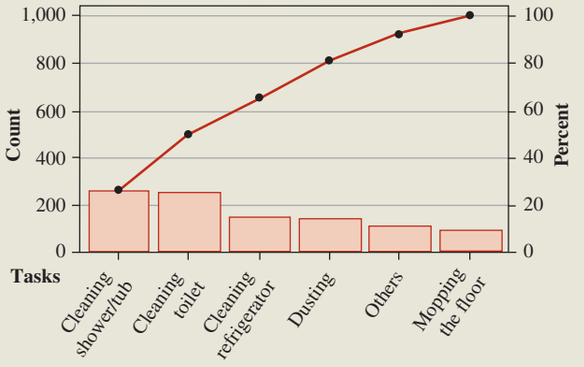
a. Construct a pie chart showing the top cold-weather beauty concerns.

b. Construct a bar graph showing the top cold-weather beauty concerns.

c. In your opinion, does the pie chart in part (a) or the bar graph in part (b) result in a better representation of the information? Explain.

2.2 Some cleaning jobs are disliked more than others. According to the July 17, 2009 *USA Today* Snapshot on a survey of women by Consumer Reports National Research Center, the cleaning tasks women dislike the most are presented in the following Pareto diagram.

Cleaning Tasks That Women Dislike the Most



Count	262	252	151	141	111	91
Percent	26.0	25.0	15.0	14.0	11.0	9.0
Cum %	26.0	51.0	66.0	80.0	91.0	100.0

a. How many total women were surveyed?

b. Verify the 15% listed for "Cleaning refrigerator."

c. Explain how the "cum % for dusting" value of 80% was obtained and what it means.

d. What three tasks would make no more than 75% of the women surveyed happy if those tasks were eliminated?

2.3 A shirt inspector at a clothing factory categorized the last 500 defects as: 67–missing button, 153–bad seam, 258–improperly sized, 22–fabric flaw. Construct a Pareto diagram for this information.

*2.4 Shown below are the heights (in inches) of the basketball players who were the first round picks by the National Basketball Association professional teams for 2009.

82	86	76	77	75	72	75	81	78	74
77	77	81	81	82	80	76	72	74	74
73	82	80	84	74	81	80	77	74	78

SOURCE: http://www.mynbadraft.com/NBA-Draft-Results/

a. Construct a dotplot of the heights of these players.

b. Use the dotplot to uncover the shortest and the tallest players.

c. What is the most common height and how many players share that height?

d. What feature of the dotplot illustrates the most common height?

*2.5 Every year *Fortune* magazine ranks America's top 100 employers. The top 15 companies to work for and their job growth are listed below.

Company	Job Growth (%)	Company	Job Growth (%)
NetApp	12	Goldman Sachs	2
Edward Jones	9	Nugget Market	22
Boston Consulting Group	10	Adobe Systems	9
Google	40	Recreational Equipment (REI)	11
Wegmans Food Markets	6	Devon Energy	11
Cisco Systems	7	Robert W. Baird	4
Genentech	5	W. L. Gore & Associates	5
Methodist Hospital System	1		

SOURCE: http://money.cnn.com

a. Construct a stem-and-leaf display of the data.

b. Based on the stem-and-leaf display, describe the distribution of percentages of growth.

2.6 Given the following stem-and-leaf display:

Stem-and-Leaf of C1 N = 16 Leaf Unit = 0.010		
1	59	7
4	60	148
(5)	61	02669
7	62	0247
3	63	58
1	64	3

a. What is the meaning of Leaf Unit = 0.010?

b. How many data values are shown on this stem-and-leaf display?

c. List the first four data values.

d. What is the column of numbers down the left-hand side of the figure?

Objective 2.2

*2.7 The U.S. Women's Olympic Soccer Team had a great year in 2008. One way to describe the players on that team is by their individual heights.

Height (inches)								
70	68	65	64	68	66	66	67	68
68	67	65	65	66	64	69	66	65

SOURCE: http://www.ussoccer.com

a. Construct an ungrouped frequency distribution for the heights.

b. Construct a frequency histogram of this distribution.

c. Prepare a relative frequency distribution for this same data.

d. What percentage of the team is at least 5 ft 6 in tall?

*2.8 A survey of 100 resort club managers on their annual salaries resulted in the following frequency distribution:

Annual Salary ($1,000s)	15–25	25–35	35–45	45–55	55–65
No. of Managers	12	37	26	19	6

a. The data value "35" belongs to which class?
b. Explain the meaning of "35–45."
c. Explain what "class width" is, give its value, and describe three ways that it can be determined.
d. Draw a frequency histogram of the annual salaries for resort club managers. Label class boundaries.

*2.9 During the Spring 2009 semester, 200 students took a statistics test from a particular instructor. The resulting grades are given in the following table.

Test Grades	Number
50–60	13
60–70	44
70–80	74
80–90	59
90–100	9
100–110	1
Total	200

a. What is the class width?
b. Draw and completely label a frequency histogram of the statistics test grades.
c. Draw and completely label a relative frequency histogram of the statistics test grades.
d. Carefully examine the two histograms in parts (b) and (c), and explain why one of them might be more useful to a student and to the instructor.

*2.10 The speeds of 55 cars were measured by a radar device on a city street:

27	23	22	38	43	24	35	26	28	18	20
25	23	22	52	31	30	41	45	29	27	43
29	28	27	25	29	28	24	37	28	29	18
26	33	25	27	25	34	32	36	22	32	33
21	23	24	18	48	23	16	38	26	21	23

a. Classify these data into a grouped frequency distribution by using class boundaries 12–18, 18–24, . . . ,48–54.
b. Find the class width.
c. For the class 24–30, find the class midpoint, the lower class boundary, and the upper class boundary.
d. Construct a frequency histogram of these data.

*2.11 A survey of 100 resort club managers on their annual salaries resulted in the following frequency distribution.

Annual Salary ($1,000s)	15–25	25–35	35–45	45–55	55–65
No. of Managers	12	37	26	19	6

a. Prepare a cumulative frequency distribution for the annual salaries.
b. Prepare a cumulative relative frequency distribution for the annual salaries.
c. Construct an ogive for the cumulative relative frequency distribution found in part (b).
d. What value bounds the cumulative relative frequency of 0.75?
e. 75% of the annual salaries are below what value? Explain the relationship between (d) and (e).

Objective 2.3

*2.12 The cost for taking your pet aboard a flight with you in the continental United States varies according to airline. The prices charged by 14 of the major U.S. airlines in June 2009 were (in dollars):

69	100	100	100	125	150	100	60	100	125	75	100	125	100

Find the mean cost for flying your pet with you.

2.13 For those 7th graders with cell phones, the number of programmed numbers in their phones were:

100	37	12	20	53	10	20	50	35	30

a. Find the mean number of programmed numbers on a 7th grader's cell phone.
b. Find the median number of programmed numbers on a 7th grader's cell phone.
c. Explain the difference in values of the mean and median.
d. Remove the most extreme value and answer (a) through (c) again.
e. Did removing the extreme value have more of an effect on the mean or median? Explain why.

2.14 The number of cars owned per apartment in a sample of tenants in a large complex is 1, 2, 1, 2, 2, 2, 1, 2, 3, 2. What is the mode?

2.15 Each year around 160 colleges compete in the American Society of Civil Engineer's National Concrete Canoe Competition. Each team must design a seaworthy canoe from concrete, a substance not known for its capacity to float. The canoes must weigh between 100 and 350 pounds. When last year's entries weighed in, the weights ranged from 138 to 349 pounds.
a. Find the midrange.
b. The information given contains 4 weight values, explain why you did use two of them in (a) and did not use the other two.

2.16. Consider the sample 2, 4, 7, 8, 9. Find the following:
a. mean, $\bar{x}$ b. median, $\tilde{x}$
c. mode d. midrange

Objective 2.4

2.17 a. The data value $x = 45$ has a deviation value of 12. Explain the meaning of this.
b. The data value $x = 84$ has a deviation value of -20. Explain the meaning of this.

2.18 All measures of variation are nonnegative in value for all sets of data.
   a. What does it mean for a value to be "nonnegative"?
   b. Describe the conditions necessary for a measure of variation to have the value zero.
   c. Describe the conditions necessary for a measure of variation to have a positive value.

2.19 Consider the sample 2, 4, 7, 8, 9. Find the following:
   a. Range
   b. Variance s^2, using formula (2.5)
   c. Standard deviation, s

2.20 Fifteen randomly selected college students were asked to state the number of hours they slept the previous night. The resulting data are 5, 6, 6, 8, 7, 7, 9, 5, 4, 8, 11, 6, 7, 8, 7. Find the following:
   a. Variance s^2, using formula (2.5)
   b. Variance s^2, using formula (2.9)
   c. Standard deviation, s

2.21 Consider the following two sets of data:

Set 1	45	80	50	45	30
Set 2	30	80	35	30	75

Both sets have the same mean, which is 50. Compare these measures for both sets: $\Sigma(x - \bar{x})$, SS(x), and range. Comment on the meaning of these comparisons relative to the distribution.

2.22 Comment on the statement: "The mean loss for customers at First State Bank (which was not insured) was $150. The standard deviation of the losses was −$125."

Objective 2.5

2.23 Refer to the table of exam scores in Table 2.15 on page 43 for the following.
   a. Using the concept of depth, describe the position of 91 in the set of 50 exam scores in two different ways.
   b. Find P_{20} and P_{35} for the exam scores in Table 2.15.
   c. Find P_{80} and P_{95} for the exam scores in Table 2.15.

*2.24 The U.S. Geological Survey collected atmospheric deposition data in the Rocky Mountains. Part of the sampling process was to determine the concentration of ammonium ions (in percentages). Here are the results from the 52 samples:

2.9	4.1	2.7	3.5	1.4	5.6	13.3	3.9	4.0
2.9	7.0	4.2	4.9	4.6	3.5	3.7	3.3	5.7
3.2	4.2	4.4	6.5	3.1	5.2	2.6	2.4	5.2
4.8	4.8	3.9	3.7	2.8	4.8	2.7	4.2	2.9
2.8	3.4	4.0	4.6	3.0	2.3	4.4	3.1	5.5
4.1	4.5	4.6	4.7	3.6	2.6	4.0		

   a. Find Q_1
   b. Find Q_2
   c. Find Q_3
   d. Find the midquartile
   e. Find P_{30}
   f. Find the 5-number summary
   g. Draw the box-and-whiskers display

2.25 An exam produced grades with a mean score of 74.2 and a standard deviation of 11.5. Find the z-score for each test score x:
   a. $x = 54$
   b. $x = 68$
   c. $x = 79$
   d. $x = 93$

2.26 A sample has a mean of 120 and a standard deviation of 20.0. Find the value of x that corresponds to each of these standard scores:
   a. $z = 0.0$
   b. $z = 1.2$
   c. $z = -1.4$
   d. $z = 2.05$

2.27 The ACT Assessment® is designed to assess high school students' general educational development and their ability to complete college-level work. The table lists the mean and standard deviation of scores attained by the 3,908,557 high school students from the 2006 to 2008 graduating classes who took the ACT exams.

2006–2008	English	Mathematics	Reading	Science	Composite
Mean	20.6	21.0	21.4	20.9	21.1
Standard deviation	6.0	5.1	6.1	4.8	4.9

SOURCE: American College Testing

Convert the following ACT test scores to z-scores for both English and Math. Compare placement between the two tests.
   a. $x = 30$
   b. $x = 23$
   c. $x = 12$
   d. Explain why the relative positions in English and Math changed for the ACT scores of 30 and 12.
   e. If Jessica had a 26 on one of the ACT exams, on which one of the exams would she have the best possible relative score? Explain why.

Objective 2.6

2.28 The empirical rule indicates that we can expect to find what proportion of the sample included between the following?
   a. $\bar{x} - s$ and $\bar{x} + s$
   b. $\bar{x} - 2s$ and $\bar{x} + 2s$
   c. $\bar{x} - 3s$ and $\bar{x} - 3s$

2.29 The mean lifetime of a certain tire is 30,000 miles and the standard deviation is 2,500 miles.
   a. If we assume the mileages are normally distributed, approximately what percentage of all such tires will last between 22,500 and 37,500 miles?
   b. If we assume nothing about the shape of the distribution, approximately what percentage of all such tires will last between 22,500 and 37,500 miles?

2.30 Using the empirical rule, determine the approximate percentage of a normal distribution that is expected to fall within the interval described.
   a. Less than the mean
   b. Greater than 1 standard deviation above the mean
   c. Less than 1 standard deviation above the mean
   d. Between 1 standard deviation below the mean and 2 standard deviations above the mean

*2.31 Each year, NCAA college football fans like to learn about the up-and-coming freshman class of players. Following are the heights (in inches) of the nation's top 100 high school football players for 2009.

73	75	71	76	74	77	74	72	73	72
74	72	74	72	72	78	73	76	75	72
77	76	73	72	76	72	73	70	75	72
71	74	77	78	74	75	71	75	71	76
70	76	72	71	74	74	71	72	76	71
75	79	78	79	74	76	76	76	75	73
74	70	74	74	75	75	75	75	76	71
74	75	74	78	72	73	71	72	73	72
74	75	77	73	77	75	77	71	72	70
74	76	71	73	76	76	79	77	74	78

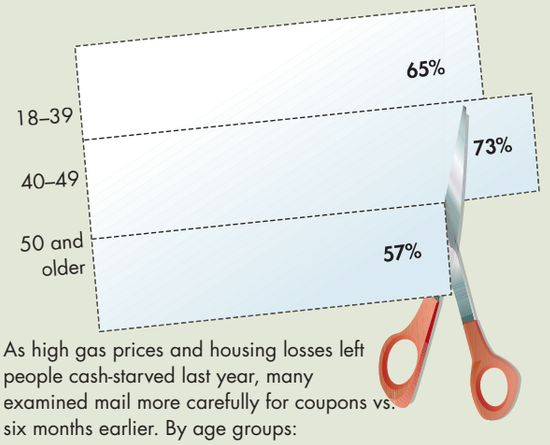
SOURCE: http://www.takkle.com/

a. Construct a histogram and one other graph of your choice that displays the distribution of heights.
b. Calculate the mean and standard deviation.
c. Sort the data into a ranked list.
d. Determine the values of $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$, and determine the percentage of data within one, two, and three standard deviations of the mean.
e. Do the percentages found in (d) agree with the empirical rule? What does this imply? Explain.
f. Do the percentages found in (d) agree with Chebyshev's theorem? What does that mean?
g. Does the graph show a distribution that agrees with your answers in part (e)? Explain.
h. Utilize one of the "testing for normality" technology instructions on your Chapter 2 Tech Card. Compare the results with your answer to part (e).

Objective 2.7

2.32

Clipping Coupons



As high gas prices and housing losses left people cash-starved last year, many examined mail more carefully for coupons vs. six months earlier. By age groups:

SOURCE: DMNews for Pitney Bowes, survey conducted online among 1,003 adults, September 9–16, 2008.

a. Is the graph a bar graph or a histogram? Explain how you determined your answer.
b. The age grouping used in the Clipping Coupons graphic does not lead to a very informative graph. Describe how the age groups might have been formed and how your suggested grouping would give additional meaning to the graph.

2.33 What kinds of financial transactions do you do online? Are you worried about your security? According to Consumer Internet Barometer, the source of a March 25, 2009, *USA Today* Snapshot titled "Security of Online Accounts," the following transactions and percent of people concerned about their online security were reported.

What	Percent
Banking	72
Paying bills	70
Buying stocks, bonds	62
Filing taxes	62

SOURCE: *USA Today* and Consumer Internet Barometer

Prepare two bar graphs to depict the percentage data. Scale the vertical axis on the first graph from 50 to 80. Scale the second graph from 0 to 100. What is your conclusion concerning how the percentages of the four responses stack up based on the two bar graphs, and what would you recommend, if anything, to improve the presentations?

***remember**

Problems marked with an asterisk have data sets available on the CourseMate for STAT2 site. Login at cengagebrain.com.