MATH 156

Statistics Course Materials

Department of Mathematics & Statistics
Camosun College

Contents

Chapter 5

Describing Data with Graphs

Reading for Sections 5.1 and 5.2:

The following reading is excerpted from:

> Mendenhall, Beaver, Beaver, and Ahmed. Introduction to Probability and Statistics. 3rd Canadian edition, Nelson, 2014, pages 10-17, 20-22, 757.

Section 5.1 : Variables

1.1 VARIABLES AND DATA

In Chapters 1 and 2, we will present some basic techniques in *descriptive statistics*—the branch of statistics concerned with describing sets of measurements, both *samples* and *populations.* Once you have collected a set of measurements, how can you display this set in a clear, understandable, and readable form? First, you must be able to define what is meant by measurements or "data" and to categorize the types of data that you are likely to encounter in real life. We begin by introducing some definitions—new terms in the statistical language that you need to know.

Definition A **variable** is a characteristic that changes or varies over time and/or for different individuals or objects under consideration.

For example, body temperature is a variable that changes over time within a single individual; it also varies from person to person. Religious affiliation, ethnic origin, income, height, age, and number of offspring are all variables—characteristics that vary depending on the individual chosen.

In the Introduction, we defined an *experimental unit* or an *element of the sample* as the object on which a measurement is taken. Equivalently, we could define an experimental unit as the object on which a variable is measured. When a variable is actually measured on a set of experimental units, a set of measurements or **data** result.

Definition An **experimental unit** is the individual or object on which a variable is measured. A single **measurement** or data value results when a variable is actually measured on an experimental unit.

If a measurement is generated for every experimental unit in the entire collection, the resulting data set constitutes the *population* of interest. Any smaller subset of measurements is a *sample.*

Definition A **population** is the set of all measurements of interest to the investigator.

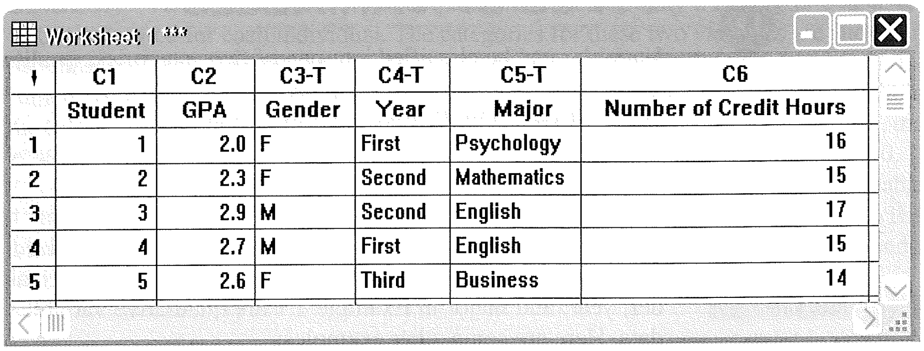
Definition A **sample** is a subset of measurements selected from the population of interest.

EXAMPLE 1.1

A set of five students is selected from all undergraduates at a large Canadian university, and measurements are entered into a spreadsheet as shown in Figure 1.1. Identify the various elements involved in generating this set of measurements.

Solution There are several *variables* in this example. The *experimental unit* on which the variables are measured is a particular undergraduate student on the campus, identified in column C1. Five variables are measured for each student: grade point average (GPA), gender, year in university, major, and current number of credit hours. Each of these characteristics varies from student to student. If we consider the GPAs of all students at this university to be the population of interest, the five GPAs in column C2 represent a *sample* from this population. If the GPA of each undergraduate student at the university had been measured, we would have generated the entire *population* of measurements for this variable.

The second variable measured on the students is gender, in column C3-T. This variable can take only one of two values—male (M) or female (F). It is not a numerically valued



↓	C1	C2	C3-T	C4-T	C5-T	C6
	Student	GPA	Gender	Year	Major	Number of Credit Hours
1	1	2.0	F	First	Psychology	16
2	2	2.3	F	Second	Mathematics	15
3	3	2.9	M	Second	English	17
4	4	2.7	M	First	English	15
5	5	2.6	F	Third	Business	14

FIGURE 1.1
Measurements on five undergraduate students

variable and hence is somewhat different from GPA. The population, if it could be enumerated, would consist of a set of Ms and Fs, one for each student at the university. Similarly, the third and fourth variables, year and major, generate nonnumerical data. Year has four categories (first, second, third, fourth), and major has one category for each undergraduate major on campus. The last variable, current number of credit hours, is numerically valued, generating a set of numbers rather than a set of qualities or characteristics.

Although we have discussed each variable individually, remember that we have measured each of these five variables on a single experimental unit: the student. Therefore, in this example, a "measurement" really consists of five observations, one for each of the five measured variables. For example, the measurement taken on student 2 produces this observation:

(2.3, F, Second, Mathematics, 15)

You can see that there is a difference between a *single* variable measured on a single experimental unit and *multiple* variables measured on a single experimental unit as in Example 1.1.

EXAMPLE 1.2

A city roads department in Ottawa would like to repair some potholes on the busiest section of a road. The best way to do the repair is when the flow of traffic is low to avoid traffic congestion and inconvenience for drivers. Practically, it would be expensive and time consuming if not impossible to monitor and record the traffic flow on every day. It was decided to record how many vehicles pass on this section of road during one particular day. Using this data information, the department decides when to repair the potholes.

Note that in this example, population constitutes all the automobiles that flow on all days on that section of the road and the sample is defined as the automobiles that passed on that section of the road on the particular day.

Definition **Univariate data** result when a single variable is measured on a single experimental unit.

Definition **Bivariate data** result when two variables are measured on a single experimental unit. **Multivariate data** result when more than two variables are measured.

If you measure the body temperatures of 148 people, the resulting data are *univariate*. In Example 1.1, five variables were measured on each student, resulting in *multivariate* data.

1.2 TYPES OF VARIABLES

Variables can be classified into one of two categories: **qualitative** or **quantitative.**

Definition **Qualitative variables** measure a quality or characteristic on each experimental unit. **Quantitative variables** measure a numerical quantity or amount on each experimental unit.

NEED A TIP?

Qualitative ⇔ "quality" or characteristic

Quantitative ⇔ "quantity" or number

Qualitative variables produce data that can be categorized according to similarities or differences in kind; hence, they are often called **categorical data.** The variables gender, year, and major in Example 1.1 are qualitative variables that produce categorical data. Here are some other examples:

- Political affiliation: Liberals, Conservatives, NDP, Green, Independent
- Taste ranking: excellent, good, fair, poor
- Colour of an M&M® candy: brown, yellow, red, orange, green, blue

Quantitative variables, often represented by the letter x, produce numerical data, such as those listed here:

- $x =$ Prime interest rate
- $x =$ Number of unregistered taxicabs in a city
- $x =$ Weight of a package ready to be shipped
- $x =$ Volume of orange juice in a glass

Notice that there is a difference in the types of numerical values that these quantitative variables can assume. The number of unregistered taxicabs, for example, can take on only the values $x = 0, 1, 2, \ldots$, whereas the weight of a package can take on any value greater than zero, or $0 < x < \infty$. To describe this difference, we define two types of quantitative variables: **discrete** and **continuous.**

Definition A **discrete variable** can assume only a finite or countable number of values. A **continuous variable** can assume the infinitely many values corresponding to the points on a line interval.

NEED A TIP?

Discrete ⇔ "listable"

Continuous ⇔ "unlistable"

The name *discrete* relates to the discrete gaps between the possible values that the variable can assume. Variables such as number of family members, number of new car sales, and number of defective tires returned for replacement are all examples of discrete variables. On the other hand, variables such as height, weight, time, distance, and volume are *continuous* because they can assume values at any point along a line interval. For any two values you pick, a third value can always be found between them!

EXAMPLE 1.3

Identify each of the following variables as qualitative or quantitative:

1. The most frequent use of your microwave oven (reheating, defrosting, warming, other)
2. The number of consumers who refuse to answer a telephone survey
3. The door chosen by a mouse in a maze experiment (A, B, or C)
4. The winning time for a horse running at the Woodbine racetrack, Toronto
5. The number of children in a fifth-grade class who are reading at or above grade level

Solution Variables 1 and 3 are both *qualitative* because only a quality or characteristic is measured for each individual. The categories for these two variables are shown in parentheses. The other three variables are *quantitative.* Variable 2, the number of consumers, is a *discrete* variable that can take on any of the values $x = 0, 1, 2, \ldots$, with a maximum value depending on the number of consumers called. Similarly, variable 5, the number of children reading at or above grade level, can take on any of the values $x = 0, 1, 2, \ldots$, with a maximum value depending on the number of children in the class. Variable 4, the winning time for a Woodbine horse, is the only *continuous* variable in the list. The winning time, if it could be measured with sufficient accuracy, could be 121 seconds, 121.5 seconds, 121.25 seconds, or any values between any two times we have listed.

NEED A TIP?
Discrete variables often involve the "number of" items in a set.

EXAMPLE 1.4

Identify each of the following quantitative variables as discrete or continuous:

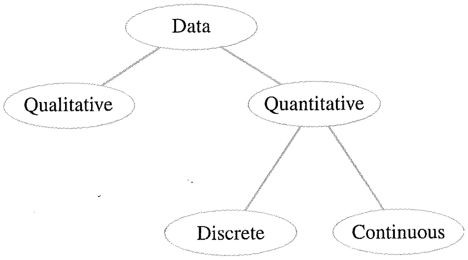
a. Average daily temperature for a small city in Quebec during a summer month
b. Number of bees on a flower
c. Driving time between Regina, Saskatchewan and Winnipeg, Manitoba
d. Number of passengers (excluding the airline staff) on a flight from Edmonton to Vancouver
e. Amount of propane gas for a BBQ cylinder filled at Costco in Montreal

Solution

a. Continuous variable: The temperature if it could be measured with a reasonable accuracy, it could be 22.5° Celsius, 26.1° Celsius, 21.9° Celsius, or any other possible values.
b. Discrete variable: It can take any value from 0 to 5 or more.
c. Continuous variable: The time could be reasonably measured, per say 6 hours 20 minutes, or 7 hours 44 minutes, or even 6 hours 11 minutes and 50 seconds, if measured in seconds too. This variable can take any possible value on a certain time interval.
d. Discrete variable: Like in part b, depending on the size of plane, it can take any value from 0 to a maximum number, depending on how many seats were sold and how many passengers showed up for this particular flight.
e. Continuous variable: A typical 13.6 kg (30 lb) cylinder holds approximately 9 kg (20 lb) of propane. This extra unfilled volume leaves some room for the liquid to expand. However, in reality, fill, if accurately measured, could be 9.12 kg, 8.93 kg, or 9.5 kg.

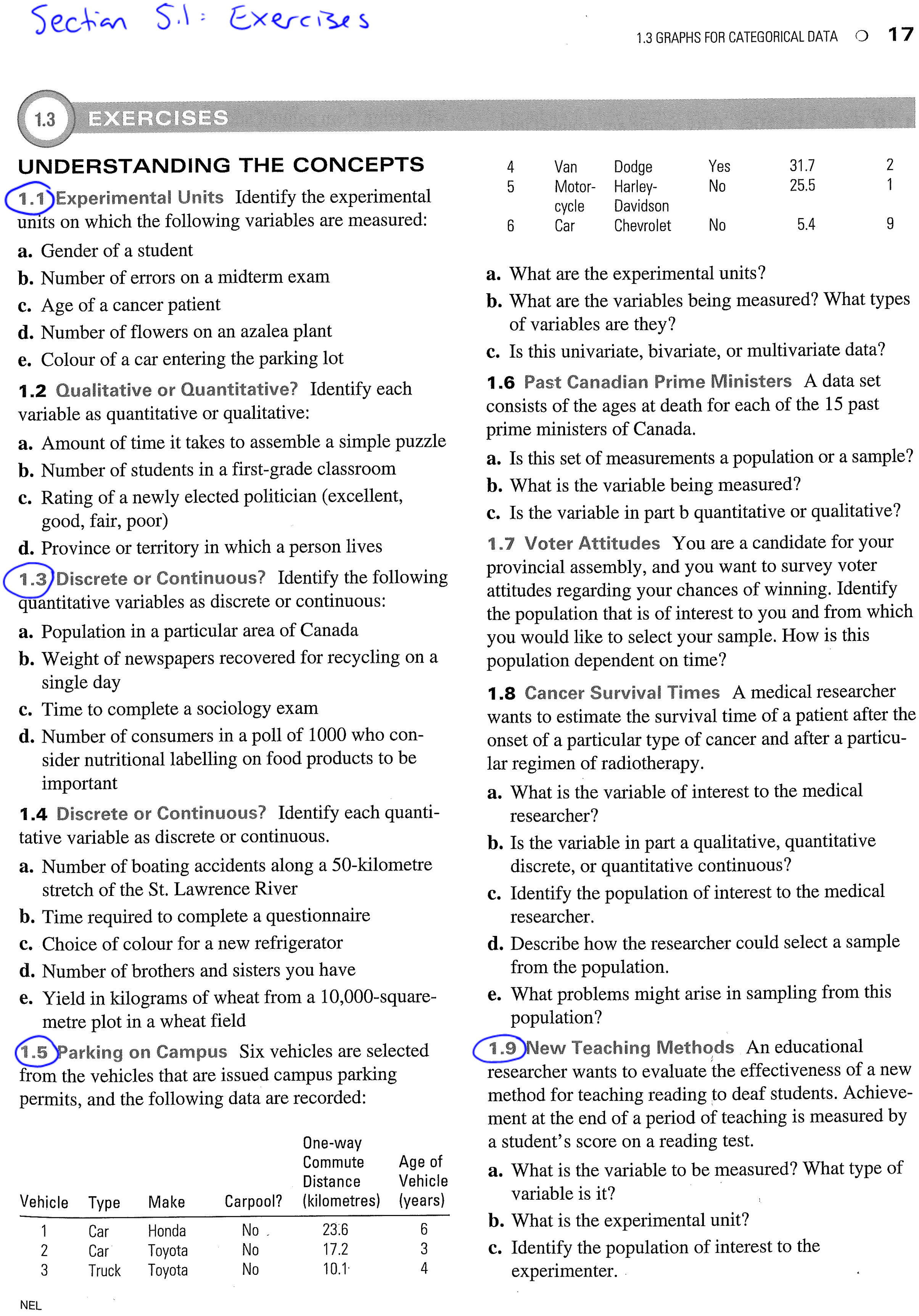
Figure 1.2 depicts the types of data we have defined. Why should you be concerned about different kinds of variables and the data that they generate? The reason is that

FIGURE 1.2
Types of data



the methods used to describe data sets depend on the type of data you have collected. For each set of data that you collect, the key will be to determine what type of data you have and how you can present them most clearly and understandably to your audience!

Section 5.1: Exercises



1.3 EXERCISES

UNDERSTANDING THE CONCEPTS

1.1 Experimental Units Identify the experimental units on which the following variables are measured:

a. Gender of a student

b. Number of errors on a midterm exam

c. Age of a cancer patient

d. Number of flowers on an azalea plant

e. Colour of a car entering the parking lot

1.2 Qualitative or Quantitative? Identify each variable as quantitative or qualitative:

a. Amount of time it takes to assemble a simple puzzle

b. Number of students in a first-grade classroom

c. Rating of a newly elected politician (excellent, good, fair, poor)

d. Province or territory in which a person lives

1.3 Discrete or Continuous? Identify the following quantitative variables as discrete or continuous:

a. Population in a particular area of Canada

b. Weight of newspapers recovered for recycling on a single day

c. Time to complete a sociology exam

d. Number of consumers in a poll of 1000 who consider nutritional labelling on food products to be important

1.4 Discrete or Continuous? Identify each quantitative variable as discrete or continuous.

a. Number of boating accidents along a 50-kilometre stretch of the St. Lawrence River

b. Time required to complete a questionnaire

c. Choice of colour for a new refrigerator

d. Number of brothers and sisters you have

e. Yield in kilograms of wheat from a 10,000-square-metre plot in a wheat field

1.5 Parking on Campus Six vehicles are selected from the vehicles that are issued campus parking permits, and the following data are recorded:

Vehicle	Type	Make	Carpool?	One-way Commute Distance (kilometres)	Age of Vehicle (years)
1	Car	Honda	No	23.6	6
2	Car	Toyota	No	17.2	3
3	Truck	Toyota	No	10.1	4
4	Van	Dodge	Yes	31.7	2
5	Motorcycle	Harley-Davidson	No	25.5	1
6	Car	Chevrolet	No	5.4	9

a. What are the experimental units?

b. What are the variables being measured? What types of variables are they?

c. Is this univariate, bivariate, or multivariate data?

1.6 Past Canadian Prime Ministers A data set consists of the ages at death for each of the 15 past prime ministers of Canada.

a. Is this set of measurements a population or a sample?

b. What is the variable being measured?

c. Is the variable in part b quantitative or qualitative?

1.7 Voter Attitudes You are a candidate for your provincial assembly, and you want to survey voter attitudes regarding your chances of winning. Identify the population that is of interest to you and from which you would like to select your sample. How is this population dependent on time?

1.8 Cancer Survival Times A medical researcher wants to estimate the survival time of a patient after the onset of a particular type of cancer and after a particular regimen of radiotherapy.

a. What is the variable of interest to the medical researcher?

b. Is the variable in part a qualitative, quantitative discrete, or quantitative continuous?

c. Identify the population of interest to the medical researcher.

d. Describe how the researcher could select a sample from the population.

e. What problems might arise in sampling from this population?

1.9 New Teaching Methods An educational researcher wants to evaluate the effectiveness of a new method for teaching reading to deaf students. Achievement at the end of a period of teaching is measured by a student's score on a reading test.

a. What is the variable to be measured? What type of variable is it?

b. What is the experimental unit?

c. Identify the population of interest to the experimenter.

Section 5.1: Answers

Answers to Selected Exercises

Chapter 1

1.1 **a.** the student **b.** the exam **c.** the patient **d.** the plant **e.** the car

1.3 **a.** discrete **b.** continuous **c.** continuous **d.** discrete

1.5 **a.** vehicles **b.** type (qualitative); make (qualitative); carpool (qualitative); distance (quantitative continuous); age (quantitative continuous) **c.** multivariate

1.7 The population is the set of voter preferences for all voters in the province. Voter preferences may change over time.

1.9 **a.** score on the reading test; quantitative **b.** the student **c.** the set of scores for all deaf students who hypothetically might take the test

1.11 **a.** a pair of jeans **b.** the province in which the jeans are produced; qualitative **e.** 8/25 **f.** Ontario **g.** The three provinces produce roughly the same numbers of jeans.

1.13 **a.** no

1.15 **a.** yes **b.** yes **c.** the bar chart

1.17 **a.** Mound-shaped distribution

b. 1.6 (1 6)

c. both 4.9 (4 9)

1.19 **a.**

```
3 | 2 3 4 5 5 5 6 6 7 9 9 9 9
4 | 0 0 2 2 3 3 3 4 4 5 8    leaf digit = 0.1, 1 2
                             represents 1.2
```

b.

```
3 | 2 3 4
3 | 5 5 5 6 6 7 9 9 9 9
4 | 0 0 2 2 3 3 3 4 4    leaf digit = 0.1, 1 2
4 | 5 8                  represents 1.2
```

Yes.

1.21 **b.** The ones digit must be the stem, and the leaf will be a zero digit.

c.

```
0 | 0 0 0 0 0
1 | 0 0 0 0 0 0 0 0 0
2 | 0 0 0 0 0 0
```

d. Yes, if the stem and leaf plot is turned 90 degrees and stretched to resemble the dotplot.

1.25 **b.** Skewed right

c. 0.72

1.27 **a.**

```
3 | 0 0 0 1 1 2 2 2 3 3 4 4
3 | 5 5 5 6 6 6 6 6 7 7 8 8 9 9 9
4 | 0 0 0 0 1 1 1 1 2 2 3 3
4 | 5 5 6 6 6 7 8 8
5 | 0 0
5 | 5
```

c. Both are very similar; however, the relative frequency histogram may be more helpful.

d. .54

e. .94

1.29 **a.**

```
0 | 2 2 3 3 3 4 4 4
0 | 5 5 6 6 6 6 7 7 7 8 8 8 8 9 9
1 | 0 0 1 1 1 1 1 1 1 2 2 2 3 3 3 4 4
1 | 6 6 7 7 8 8 8 8 9 9
2 | 1 2 3
2 | 5 8     leaf digit = 0.1, 1 2 represents 1.2
3 | 1 1
3 | 6
4 |
4 | 5
5 | 2
```

b. 0.4167

c. 0.2

1.31 **c.** Same range, no outliers.

1.33 **b.** Relatively mound-shaped, centred at 5.2.

c. Somewhat unusual.

1.35 **a.** Skewed right, with two outliers.

b. Dot plot is more informative; better display of the data shape with outliers shown.

1.37 **c.** Pareto chart seems more effective since it is very easy to compare the relative membership of the organized religions.

1.39 **a.** skewed **b.** symmetric **c.** symmetric **d.** symmetric **e.** skewed **f.** skewed

1.41 **a.** continuous **b.** continuous **c.** discrete **d.** discrete **e.** discrete

1.43

```
 7 | 8 9
 8 | 0 1 7
 9 | 0 1 2 4 4 5 6 6 6 8 8
10 | 1 7 9
11 | 2
```

1.45 **a.** skewed right

Section 5.2: Pie Charts and Bar Charts

1.3 GRAPHS FOR CATEGORICAL DATA

After the data have been collected, they can be consolidated and summarized to show the following information:

- What values of the variable have been measured
- How often each value has occurred

For this purpose, you can construct a *statistical table* that can be used to display the data graphically as a data distribution. The type of graph you choose depends on the type of variable you have measured.

When the variable of interest is *qualitative*, the statistical table is a list of the categories being considered along with a measure of how often each value occurred. You can measure "how often" in three different ways:

- The **frequency,** or number of measurements in each category
- The **relative frequency,** or proportion of measurements in each category
- The **percentage** of measurements in each category

For example, if you let n be the total number of measurements in the set, you can find the relative frequency and percentage using these relationships:

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

$$\text{Percent} = 100 \times \text{Relative frequency}$$

You will find that the sum of the frequencies is always n, the sum of the relative frequencies is 1, and the sum of the percentages is 100%.

The categories for a qualitative variable should be chosen so that

- a measurement will belong to one and only one category
- each measurement has a category to which it can be assigned

NEED a tip? NEED A TIP?

Three steps to a data distribution:
(1) raw data ⇒
(2) statistical table ⇒
(3) graph

For example, if you categorize meat products according to the type of meat used, you might use these categories: beef, chicken, seafood, pork, turkey, other. To categorize ranks of university faculty, you might use these categories: professor, associate professor, assistant professor, instructor, lecturer, other. The "other" category is included in both cases to allow for the possibility that a measurement cannot be assigned to one of the earlier categories.

Once the measurements have been categorized and summarized in a *statistical table*, you can use either a pie chart or a bar chart to display the distribution of the data. A **pie chart** is the familiar circular graph that shows how the measurements are distributed among the categories. A **bar chart** shows the same distribution of measurements in categories, with the height of the bar measuring how often a particular category was observed.

EXAMPLE 1.5

In a survey concerning public education, 400 school administrators were asked to rate the quality of education in Canada. Their responses are summarized in Table 1.1. Construct a pie chart and a bar chart for this set of data.

Solution To construct a pie chart, assign one sector of a circle to each category. The angle of each sector should be proportional to the proportion of measurements

TABLE 1.1 **Canadian Education Rating by 400 Educators**

Rating	Frequency
A	35
B	260
C	93
D	12
Total	400

(or *relative frequency*) in that category. Since a circle contains 360°, you can use this equation to find the angle:

$$\text{Angle} = \text{Relative frequency} \times 360^\circ$$

Table 1.2 shows the ratings along with the frequencies, relative frequencies, percentages, and sector angles necessary to construct the pie chart. Figure 1.3 shows the pie chart constructed from the values in the table. While pie charts use percentages to determine the relative sizes of the "pie slices," bar charts usually plot frequency against the categories. A bar chart for these data is shown in Figure 1.4.

NEED A TIP?

Proportions add to 1.
Percents add to 100.
Sector angles add to 360°.

TABLE 1.2 **Calculations for the Pie Chart in Example 1.5**

Rating	Frequency	Relative Frequency	Percent	Angle
A	35	35/400 = 0.09	9	0.09 × 360 = 32.4°
B	260	260/400 = 0.65	65	234.0°
C	93	93/400 = 0.23	23	82.8°
D	12	12/400 = 0.03	3	10.8°
Total	400	1.00	100%	360°

FIGURE 1.3
Pie chart for Example 1.5

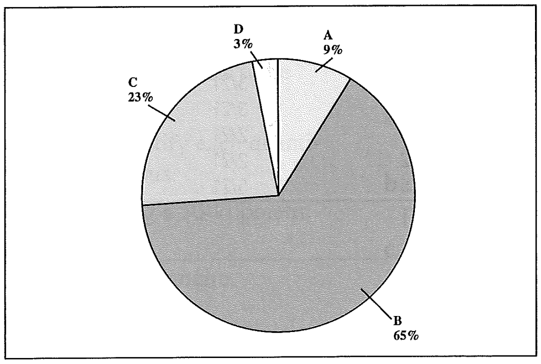
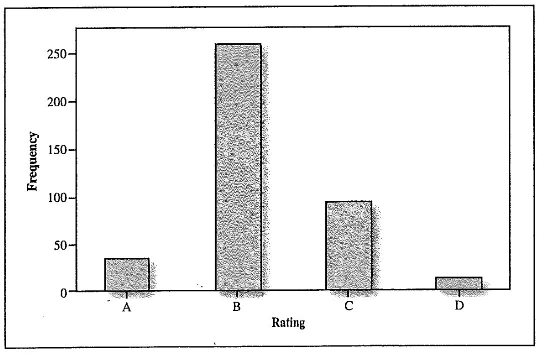


FIGURE 1.4
Bar chart for Example 1.5



The visual impact of these two graphs is somewhat different. The pie chart is used to display the relationship of the parts to the whole; the bar chart is used to emphasize the actual quantity or frequency for each category. Since the categories in this example are ordered "grades" (A, B, C, D), we would not want to rearrange the bars in the chart to change its *shape*. In a pie chart, the order of presentation is irrelevant.

EXAMPLE 1.6

A snack size bag of peanut M&M® candies contains 21 candies with the colours listed in Table 1.3. The variable "colour" is *qualitative*, so Table 1.4 lists the six categories along with a tally of the number of candies of each colour. The last three columns of Table 1.4 give the three different measures of how often each category occurred. Since the categories are colours and have no particular order, you could construct bar charts with many different *shapes* just by reordering the bars. To emphasize that brown is the most frequent colour, followed by blue, green, and orange, we order the bars from largest to smallest and generate the bar chart using *MINITAB* in Figure 1.5. A bar chart in which the bars are ordered from largest to smallest is called a **Pareto chart.**

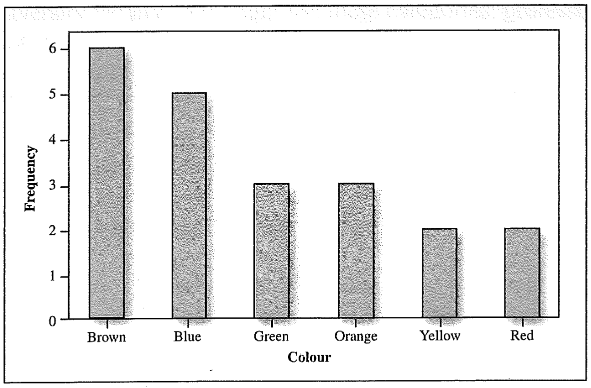
TABLE 1.3 **Raw Data: Colours of 21 Candies**

Brown	Green	Brown	Blue
Red	Red	Green	Brown
Yellow	Orange	Green	Blue
Brown	Blue	Blue	Brown
Orange	Blue	Brown	Orange
Yellow			

TABLE 1.4 **Statistical Table: M&M Data for Example 1.6**

Category	Tally	Frequency	Relative Frequency	Percent
Brown	𝍸 I	6	6/21	28
Green	III	3	3/21	14
Orange	III	3	3/21	14
Yellow	II	2	2/21	10
Red	II	2	2/21	10
Blue	𝍸	5	5/21	24
Total		21	1	100%

FIGURE 1.5
MINITAB bar chart for Example 1.6



1.4 GRAPHS FOR QUANTITATIVE DATA

Quantitative variables measure an amount or quantity on each experimental unit. If the variable can take only a finite or countable number of values, it is a *discrete* variable. A variable that can assume an infinite number of values corresponding to points on a line interval is called *continuous.*

Pie Charts and Bar Charts

Sometimes information is collected for a quantitative variable measured on different segments of the population, or for different categories of classification. For example, you might measure the average incomes for people of different age groups, different genders, or living in different geographic areas of the country. In such cases, you can use pie charts or bar charts to describe the data, using the amount measured in each category rather than the frequency of occurrence of each category. The *pie chart* displays how the total quantity is distributed among the categories, and the *bar chart* uses the height of the bar to display the amount in a particular category.

EXAMPLE 1.7

Canadian Defence Budget Expected to Rise to $20 billion by 2010 Like all federal institutions during the 1990s, the Department of Defence underwent budget cuts as part of the federal government's effort to eliminate the deficit. Consequently, the budget, which totalled $12 billion in 1993–1994, declined to $9.38 billion by 1998–1999. Since then, the Department of National Defence has received three successive budget increases totalling more than $5 billion, to be delivered between 2001–2002 and 2006–2007.

The amount of money estimated for the fiscal year 2002–2003 budget ($ millions) by the Department of National Defence, Government of Canada, is shown in Table 1.5.[2] Construct both a pie chart and a bar chart to describe the data.

Compare the two forms of presentation.

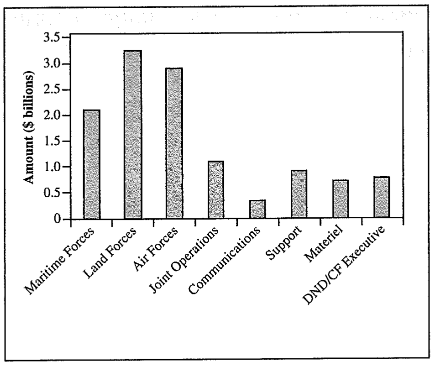
TABLE 1.5 **Expenses by Category**

Category	Amount (in dollars)
Maritime forces	2,053,210,000
Land forces	3,181,330,000
Air forces	2,828,760,000
Joint operations and civil emergency preparedness	1,086,310,000
Communications and information management	304,020,000
Support to the personnel function	860,850,000
Materiel, infrastructure, and environment support	754,080,000
DND/CF executive	786,240,000
Total	**11,834,800,000**

Source: National Defence and Canadian Forces Budget-Budget 2002-2003 http://www/collectionscanada.gc.ca/webarchives/20060327215046/http://www.forces.gc.ca/site/about/budgete.asp. Reproduced with the permission of the Minister of Public Works and Government Services, 2007.

Solution Two variables are being measured: the category of expenditure (qualitative) and the amount of the expenditure (quantitative). The bar chart in Figure 1.6 displays the categories on the horizontal axis and the amount on the vertical axis.

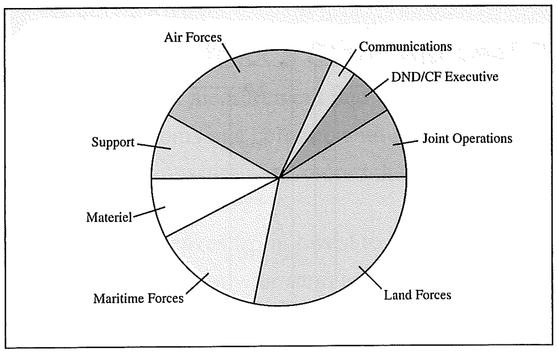
FIGURE 1.6
Bar chart for Example 1.7



For the pie graph in Figure 1.7, each "pie slice" represents the proportion of the total expenditure ($11,834,800,000). For example, for the Air Forces category, the angle of the sector is

$$\frac{2,828,760,000}{11,834,800,000} \times 360° = 86.04°$$

FIGURE 1.7
Pie chart for Example 1.7



Both graphs show that the largest amounts of money were spent on land forces. Since there is no inherent order to the categories, you are free to rearrange the bars or sectors of the graphs in any way you like. The *shape* of the bar chart has no bearing on its interpretation.

Line Charts

When a quantitative variable is recorded over time at equally spaced intervals (such as daily, weekly, monthly, quarterly, or yearly), the data set forms a **time series.** Time series data are most effectively presented on a **line chart** with time as the horizontal axis. The idea is to try to discern a pattern or **trend** that will likely continue into the future, and then to use that pattern to make accurate predictions for the immediate future.

EXAMPLE 1.8

In the year 2030, the oldest "baby boomers" (born in 1946) will be 84 years old, and the oldest "Gen-Xers" (born in 1965) will be eligible to collect Canada Pension Plan (CPP) benefits. How will this affect the consumer trends in the next 25 years? Statistics Canada gives projections for age group 65–69 years, as shown in Table 1.6 below. Construct a line chart to illustrate the data. What is the effect of stretching and shrinking the vertical axis on the line chart?

TABLE 1.6 **Population Growth Projections**

Year	2006	2011	2016	2021	2026	2031
Population (thousands)	1227.3	1513.1	1942.1	2184.7	2466.6	2527.6

Solution The quantitative variable population is measured over six time intervals, creating a *time series* that you can graph with a line chart. The time intervals are marked on the horizontal axis and the population on the vertical axis. The data points are then connected by line segments to form the line charts in Figure 1.8. Notice the marked difference in the vertical scales of the two graphs. *Shrinking* the scale on the vertical axis causes large changes to appear small, and vice versa. To avoid misleading conclusions, you must look carefully at the scales of the vertical and horizontal axes. However, from both graphs you get a clear picture of the steadily increasing numbers in the early years of the twenty-first century.

NEED A TIP? Beware of stretching or shrinking axes when you look at a graph!

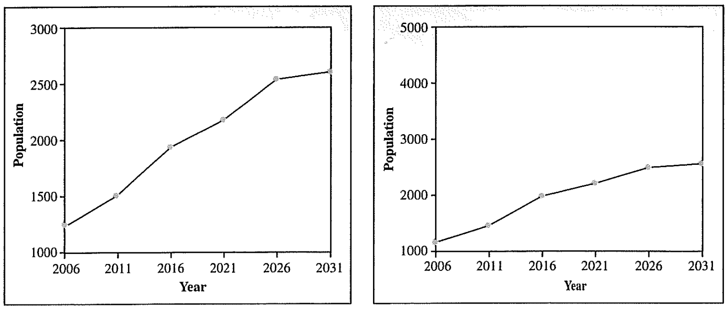


FIGURE 1.8 Line charts for Example 1.8

No exercises for Secton 5.2.

Reading for Section 5.3:

The following reading is excerpted from:

> DeVeaux, Velleman, and Bock. Intro Stats. 3rd edition, Pearson, 2009, pages 48-50, 53-56, 78-80, 83, 86, A-5, A-6.

Section 5.3: Histograms

CHAPTER 4

Displaying and Summarizing Quantitative Data

Where are we going?

If someone asked you to summarize a variable, what would you say? You might start by making a picture. For quantitative data, that first picture would probably be a histogram. We've all looked at histograms, but what should we look *for*? We'll describe the histogram, and we'll often do more—reporting numerical summaries of the center and spread as well. Spread measures are a bit less common, but in Statistics they are even more important. This chapter is where we'll first encounter the single most important calculated value in all of Statistics.

Tsunamis are potentially destructive waves that can occur when the sea floor is suddenly and abruptly deformed. They are most often caused by earthquakes beneath the sea that shift the earth's crust, displacing a large mass of water.

The tsunami of December 26, 2004, with epicenter off the west coast of Sumatra, was caused by an earthquake of magnitude 9.0 on the Richter scale. It killed an estimated 297,248 people, making it the most disastrous tsunami on record. But was the earthquake that caused it truly extraordinary, or did it just happen at an unlucky place and time? The U.S. National Geophysical Data Center[1] has information on more than 2400 tsunamis dating back to 2000 B.C.E., and we have estimates of the magnitude of the underlying earthquake for 1240 of them. What can we learn from these data?

Histograms

Let's start with a picture. For categorical variables, it is easy to draw the distribution because each category is a natural "pile." But for quantitative variables, there's no obvious way to choose piles. So, usually, we slice up all the possible values into equal-width bins. We then count the number of cases that fall into

[1] www.ngdc.noaa.gov

WHO	2410 earthquakes known to have caused tsunamis for which we have data or good estimates
WHAT	Magnitude (Richter scale [2]), depth (m), date, location, and other variables
WHEN	From 2000 B.C.E. to the present
WHERE	All over the earth

each bin. The bins, together with these counts, give the **distribution** of the quantitative variable and provide the building blocks for the histogram. By representing the counts as bars and plotting them against the bin values, the **histogram** displays the distribution at a glance.

For example, here are the *Magnitudes* (on the Richter scale) of the 2410 earthquakes in the NGDC data:

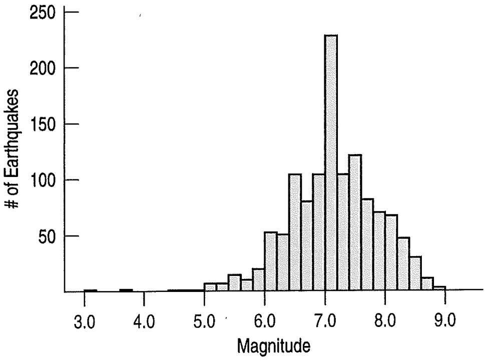


FIGURE 4.1

A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

Like a bar chart, a histogram plots the bin counts as the heights of bars. In this histogram of earthquake magnitudes, each bin has a width of 0.2, so, for example, the height of the tallest bar says that there were about 230 earthquakes with magnitudes between 7.0 and 7.2. In this way, the histogram displays the entire distribution of earthquake magnitudes.

HOW DO HISTOGRAMS WORK?

If you make a histogram by hand, you'll need to decide the endpoints of the bins. Usually, it helps to make them come out to "nice" numbers that are easy to think about. The standard rule for a value that falls exactly on a bin boundary is to put it into the next higher bin, so if a bin spans magnitudes 5.0 to 5.2, and the next goes from 5.2 to 5.4, you'd put an earthquake with magnitude 5.2 into the higher bin.

Different features of the distribution may appear more obvious at different bin width choices. When you use technology, it's usually easy to vary the bin width interactively so you can make sure that a feature you think you see isn't a consequence of a certain bin width choice.

One surprising feature of the earthquake magnitudes is the spike around magnitude 7.0. Only one other bin holds even half that many earthquakes. These values include historical data for which the magnitudes were estimated by experts and not measured by modern seismographs. Perhaps the experts thought 7 was a typical and reasonable value for a tsunami-causing earthquake when they lacked detailed information. That would explain the overabundance of magnitudes right at 7.0 rather than spread out near that value.

Does the distribution look as you expected? It is often a good idea to *imagine* what the distribution might look like before you make the display. That way you'll be less likely to be fooled by errors in the data or when you accidentally graph the wrong variable.

From the histogram, we can see that these earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, and some are as small as 3 and as big as 9. Now we can answer the question about the Sumatra tsunami. With a value of 9.0 it's clear that the earthquake that caused it was an extraordinarily powerful earthquake—one of the largest on record.[3]

The bar charts of categorical variables we saw in Chapter 3 had spaces between the bars to separate the counts of different categories. But in a histogram, the bins

[2] Technically, Richter scale values are in units of log dyne-cm. But the Richter scale is so common now that usually the units are assumed. The U.S. Geological Survey gives the background details of Richter scale measurements on its Web site www.usgs.gov/.

[3] Some experts now estimate the magnitude at between 9.1 and 9.3.

slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram,** replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.

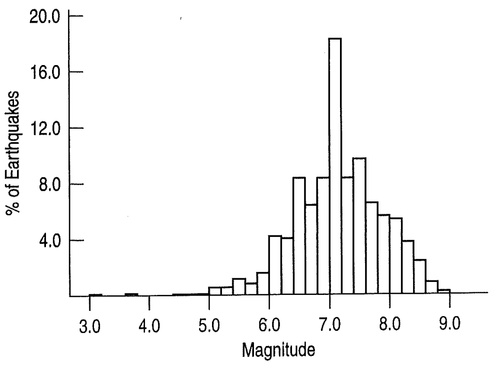


FIGURE 4.2

A relative frequency histogram looks.just like a frequency histogram except for the labels on the y-axis, which now show the percentage of earthquakes in each bin.

The Shape of a Distribution

The **mode** is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. What is the mode of the Kentucky Derby times? Well, seven races were timed at 122.2 seconds—more than any other race time. Should that be the mode? Probably not. For quantitative data, it makes more sense to use the term "mode" in the more general sense of the peak of the histogram rather than as a single summary value. In this sense, the important feature of the Kentucky Derby races is that there are two distinct modes, representing the two different versions of the race and warning us to consider those two versions separately.

1. *Does the histogram have a single, central hump or several separated humps?* These humps are called **modes.**[6] The earthquake magnitudes have a single mode at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal;** histograms with two peaks are **bimodal,** and those with three or more are called **multimodal.**[7] For example, here's a bimodal histogram.

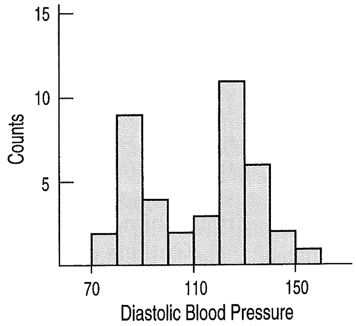


FIGURE 4.5
A bimodal histogram has two apparent peaks.

A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform.**

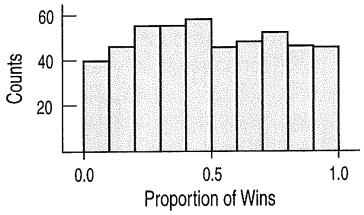
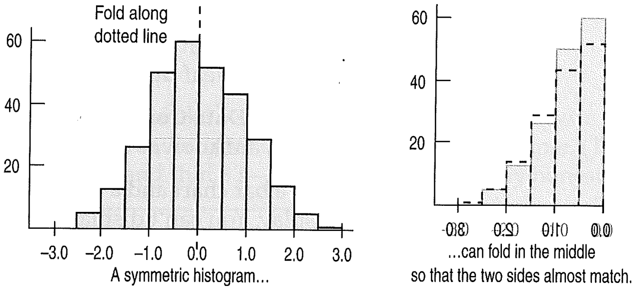


FIGURE 4.6
In a uniform histogram, the bars are all about the same height. The histogram doesn't appear to have a mode.

[6] Well, technically, it's the value on the horizontal axis of the histogram that is the mode, but anyone asked to point to the mode would point to the hump.
[7] Apparently, statisticians don't like to count past two.

You've heard of pie à la mode. Is there a connection between pie and the mode of a distribution? Actually, there is! The mode of a distribution is a *popular* value near which a lot of the data values gather. And "à la mode" means "in style"—*not* "with ice cream." That just happened to be a *popular* way to have pie in Paris around 1900.

2. *Is the histogram **symmetric**?* Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?



A symmetric histogram...

...can fold in the middle so that the two sides almost match.

FIGURE 4.7

The (usually) thinner ends of a distribution are called the **tails.** If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

A S *Activity:* **Attributes of Distribution Shape.** This activity and the others on this page show off aspects of distribution shape through animation and examples, then let you make and interpret histograms with your statistics package.

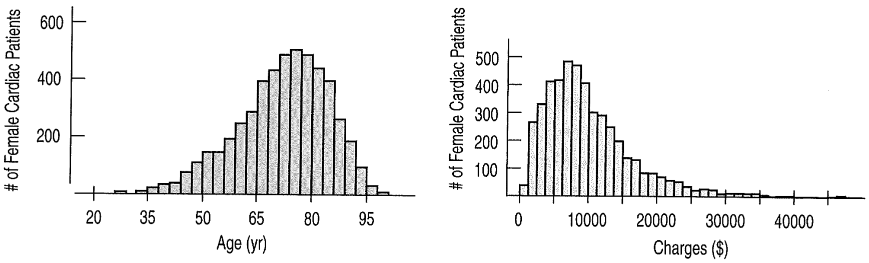
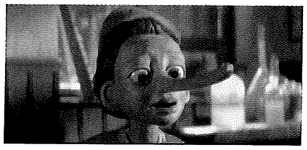


FIGURE 4.8

Two skewed histograms showing data on two variables for all female heart attack patients in New York state in one year. The blue one (age in years) is skewed to the left. The purple one (charges in $) is skewed to the right.



3. *Do any unusual features stick out?* Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or **outliers,** that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

   Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.

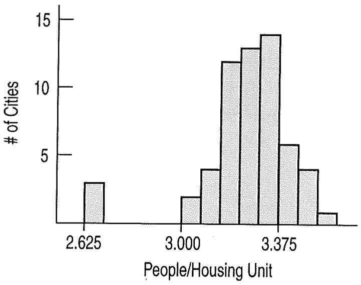


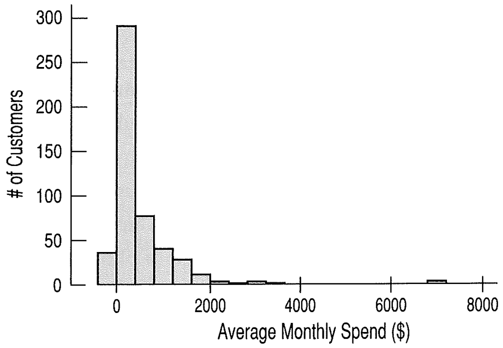
FIGURE 4.9

A histogram with outliers. There are three cities in the leftmost bar.

Are there any gaps in the distribution? The Kentucky Derby data that we saw in the dotplot on page 52 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes and encourage us to notice when the data may come from different sources or contain more than one group.

FOR EXAMPLE Describing Histograms

Question: A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data[8] on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.



Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about $7000, and some of the expenditures are negative.

Toto, I've a feeling we're not in math class anymore... When Dorothy and her dog Toto land in Oz, everything is more vivid and colorful, but also more dangerous and exciting. Dorothy has new choices to make. She can't always rely on the old definitions, and the yellow brick road has many branches. You may be coming to a similar realization about Statistics.

When we summarize data, our goal is usually more than just developing a detailed knowledge of the data we have at hand. Scientists generally don't care about the particular guinea pigs they've treated, but rather about what their reactions say about how animals (and, perhaps, humans) would respond.

When you look at data, you want to know what the data say about the world, so you'd like to know whether the patterns you see in histograms and summary statistics generalize to other individuals and situations. You'll want to calculate summary statistics accurately, but then you'll also want to think about what they may say beyond just describing the data. And your knowledge about the world matters when you think about the overall meaning of your analysis.

It may surprise you that many of the most important concepts in Statistics are not defined as precisely as most concepts in mathematics. That's done on purpose, to leave room for judgment.

Because we want to see broader patterns rather than focus on the details of the data set we're looking at, we deliberately leave some statistical concepts a bit vague. Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the rest of the data to be considered an outlier—these are all somewhat vague concepts. And they all require judgment. You

[8] These data are real, but cannot be further identified for obvious privacy reasons.

may be used to finding a single correct and precise answer, but in Statistics, there may be more than one interpretation. That may make you a little uncomfortable at first, but soon you'll see that this room for judgment brings you enormous power and responsibility. It means that using your own knowledge and judgment and supporting your findings with statistical evidence and justifications entitles you to your own opinions about what you see.

JUST CHECKING

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

Section 5.3: Exercises

EXERCISES

1. **Statistics in print.** Find a histogram that shows the distribution of a variable in a newspaper or magazine article.
   a) Does the article identify the W's?
   b) Discuss whether the display is appropriate for the data.
   c) Discuss what the display reveals about the variable and its distribution.
   d) Does the article accurately describe and interpret the data? Explain.

2. **Not a histogram.** Find a graph other than a histogram that shows the distribution of a quantitative variable in a newspaper or magazine article.
   a) Does the article identify the W's?
   b) Discuss whether the display is appropriate for the data.
   c) Discuss what the display reveals about the variable and its distribution.
   d) Does the article accurately describe and interpret the data? Explain.

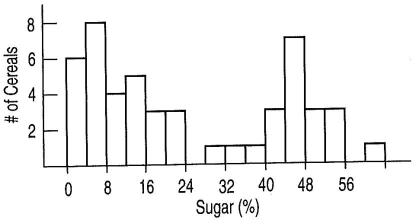
3. **In the news.** Find an article in a newspaper or a magazine that discusses an "average."
   a) Does the article discuss the W's for the data?
   b) What are the units of the variable?
   c) Is the average used the median or the mean? How can you tell?
   d) Is the choice of median or mean appropriate for the situation? Explain.

4. **In the news II.** Find an article in a newspaper or a magazine that discusses a measure of spread.
   a) Does the article discuss the W's for the data?
   b) What are the units of the variable?
   c) Does the article use the range, IQR, or standard deviation?
   d) Is the choice of measure of spread appropriate for the situation? Explain.

5. **Thinking about shape.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
   a) The number of speeding tickets each student in the senior class of a college has ever had.
   b) Players' scores (number of strokes) at the U.S. Open golf tournament in a given year.
   c) Weights of female babies born in a particular hospital over the course of a year.
   d) The length of the average hair on the heads of students in a large class.

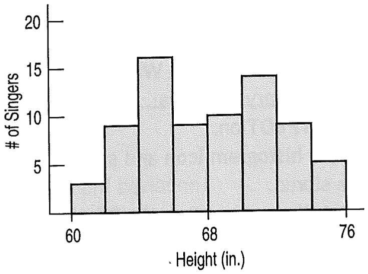
6. **More shapes.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
   a) Ages of people at a Little League game.
   b) Number of siblings of people in your class.
   c) Pulse rates of college-age males.
   d) Number of times each face of a die shows in 100 tosses.

7. **Sugar in cereals.** The histogram displays the sugar content (as a percent of weight) of 49 brands of breakfast cereals.



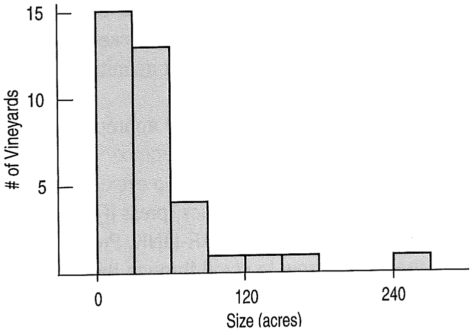
   a) Describe this distribution.
   b) What do you think might account for this shape?

8. **Singers.** The display shows the heights of some of the singers in a chorus, collected so that the singers could be positioned on stage with shorter ones in front and taller ones in back.



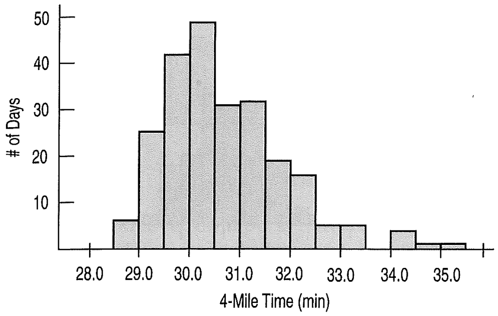
   a) Describe the distribution.
   b) Can you account for the features you see here?

9. **Vineyards.** The histogram shows the sizes (in acres) of 36 vineyards in the Finger Lakes region of New York.



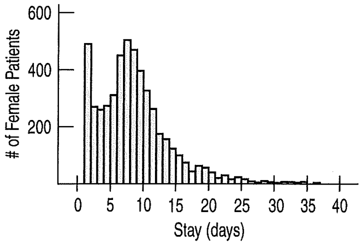
   a) Approximately what percentage of these vineyards are under 60 acres?
   b) Write a brief description of this distribution (shape, center, spread, unusual features).

10. **Run times.** One of the authors collected the times (in minutes) it took him to run 4 miles on various courses during a 10-year period. Here is a histogram of the times.



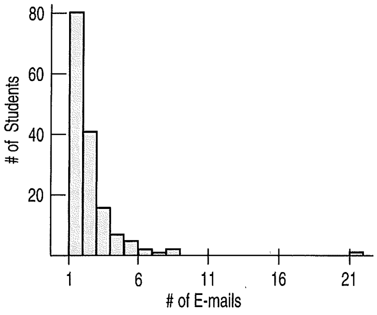
Describe the distribution and summarize the important features. What is it about running that might account for the shape you see?

11. **Heart attack stays.** The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospitals in New York during one year with a primary diagnosis of acute myocardial infarction (heart attack).



a) From the histogram, would you expect the mean or median to be larger? Explain.
b) Write a few sentences describing this distribution (shape, center, spread, unusual features).

12. **E-mails.** A university teacher saved every e-mail received from students in a large Introductory Statistics class during an entire term. He then counted, for each student who had sent him at least one e-mail, how many e-mails each student had sent.



a) From the histogram, would you expect the mean or the median to be larger? Explain.
b) Write a few sentences describing this distribution (shape, center, spread, unusual features).

13. **Super Bowl points.** How many points do football teams score in the Super Bowl? Here are the total numbers of points scored by both teams in each of the first 41 Super Bowl games:

45, 47, 23, 30, 29, 27, 21, 31, 22, 38, 46, 37, 66, 50, 37, 47, 44, 47, 54, 56, 59, 52, 36, 65, 39, 61, 69, 43, 75, 44, 56, 55, 53, 39, 41, 37, 69, 61, 45, 31, 46

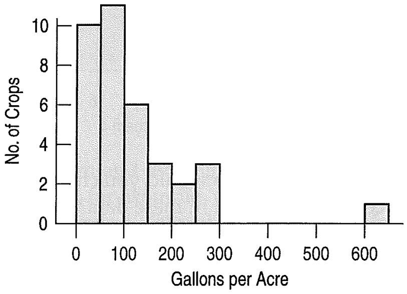
a) Find the median.
b) Find the quartiles.

14. **Super Bowl wins.** In the Super Bowl, by how many points does the winning team outscore the losers? Here are the winning margins for the first 41 Super Bowl games:

25, 19, 9, 16, 3, 21, 7, 17, 10, 4, 18, 17, 4, 12, 17, 5, 10, 29, 22, 36, 19, 32, 4, 45, 1, 13, 35, 17, 23, 10, 14, 7, 15, 7, 27, 3, 27, 3, 3, 11, 12

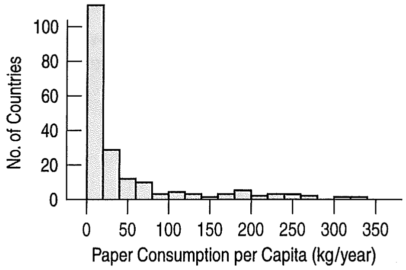
a) Find the median.
b) Find the quartiles.

15. **Oil production.** The histogram shows amount of oil produced (in gallons) from an acre of land in the United States from 36 different crops.



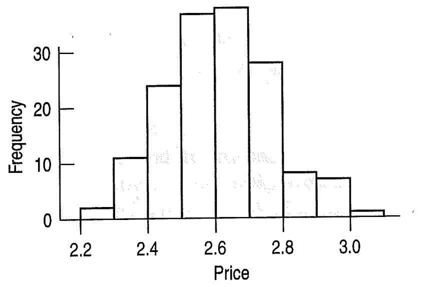
Which summary statistics would you choose to summarize the center and spread in these data? Why?

16. **Paper consumption.** The histogram shows the 2004 per capita consumption of paper for 195 countries around the world (in *kg per person per year*). (www.swivel.com)



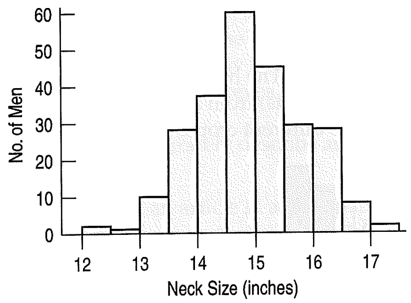
Which summary statistics would you choose to summarize the center and spread in these data? Why?

17. **Pizza prices.** The histogram shows the distribution of the prices of a small, plain pizza (in $) for 156 weeks in Dallas, Texas.



Which summary statistics would you choose to summarize the center and spread in these data? Why?

18. Neck size. The histogram shows the neck sizes (in inches) of 250 men recruited for a health study in Utah.



Which summary statistics would you choose to summarize the center and spread in these data? Why?

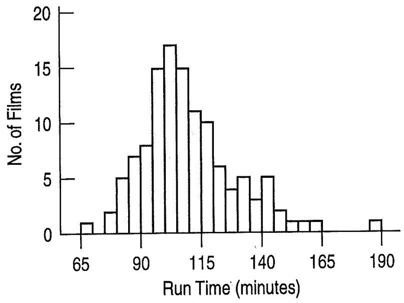
19. Pizza prices again. Look again at the histogram of the pizza prices in Exercise 17.

a) Is the mean closer to $2.40, $2.60, or $2.80? Why?
b) Is the standard deviation closer to $0.15, $0.50, or $1.00? Explain.

20. Neck sizes again. Look again at the histogram of men's neck sizes in Exercise 18.

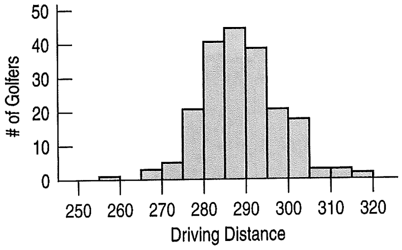
a) Is the mean closer to 14, 15, or 16 inches? Why?
b) Is the standard deviation closer to 1 inch, 3 inches, or 5 inches? Explain.

21. Movie lengths. The histogram shows the running times in minutes of 122 feature films released in 2005.



a) You plan to see a movie this weekend. Based on these movies, how long do you expect a typical movie to run?
b) Would you be surprised to find that your movie ran for 2 1/2 hours (150 minutes)?
c) Which would you expect to be higher, the mean or the median run time for all movies? Why?

22. Golf drives. The display shows the average drive distance (in yards) for 202 professional golfers on the men's PGA tour.



a) Describe this distribution.
b) Approximately what proportion of professional male golfers drive, on average, less than 280 yards?
c) Estimate the mean by examining the histogram.
d) Do you expect the mean to be smaller than, approximately equal to, or larger than the median? Why?

23. Summaries. Here are costs of 10 electric smoothtop ranges rated very good or excellent by *Consumer Reports* in August 2002:

$850 900 1400 1200 1050 1000 750 1250 1050 565

Find these statistics *by hand* (no calculator!):

a) mean
b) median and quartiles
c) range and IQR

24. More summaries. Here are the annual numbers of deaths from tornadoes in the United States from 1990 through 2000 (www.noaa.gov):

53 39 39 33 69 30 25 67 130 94 40

Find these statistics *by hand* (no calculator!):

a) mean
b) median and quartiles
c) range and IQR

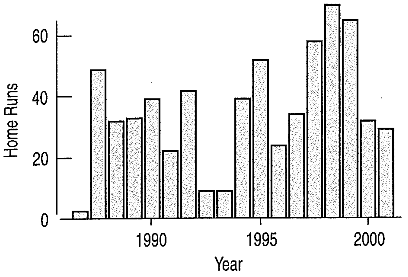
25. Mistake. A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as $2,000,000 instead of $200,000. Explain how this error will affect these summary statistics for the company payroll:

a) measures of center: median and mean.
b) measures of spread: range, IQR, and standard deviation.

26. Cold weather. A meteorologist preparing a talk about global warming compiled a list of weekly low temperatures (in degrees Fahrenheit) he observed at his southern Florida home last year. The coldest temperature for any week was 36°F but he inadvertently recorded the Celsius value of 2°. Assuming that he correctly listed all the other temperatures, explain how this error will affect these summary statistics:

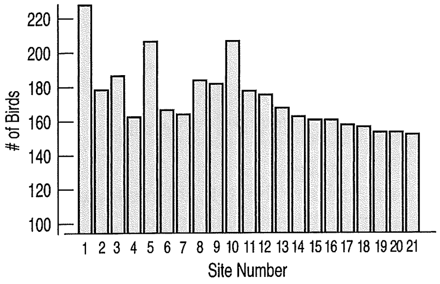
a) measures of center: mean and median.
b) measures of spread: range, IQR, and standard deviation.

43. **Home runs, again.** Students were asked to make a histogram of the number of home runs hit by Mark McGwire from 1986 to 2001 (see Exercise 39). One student submitted the following display:



a) Comment on this graph.
b) Create your own histogram of the data.

44. **Return of the birds.** Students were given the assignment to make a histogram of the data on bird counts reported in Exercise 40. One student submitted the following display:



a) Comment on this graph.
b) Create your own histogram of the data.

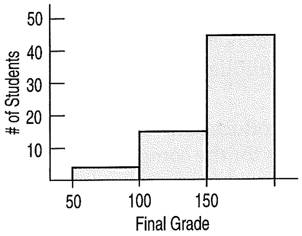
45. **Acid rain.** Two researchers measured the pH (a scale on which a value of 7 is neutral and values below 7 are acidic) of water collected from rain and snow over a 6-month period in Allegheny County, Pennsylvania. Describe their data with a graph and a few sentences:

4.57	5.62	4.12	5.29	4.64	4.31	4.30	4.39	4.45
5.67	4.39	4.52	4.26	4.26	4.40	5.78	4.73	4.56
5.08	4.41	4.12	5.51	4.82	4.63	4.29	4.60	

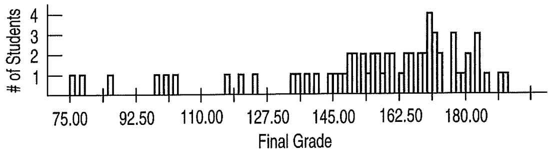
46. **Marijuana 2003.** In 2003 the Council of Europe published a report entitled *The European School Survey Project on Alcohol and Other Drugs* (www.espad.org). Among other issues, the survey investigated the percentages of 16-year-olds who had used marijuana. Shown here are the results for 20 European countries. Create an appropriate graph of these data, and describe the distribution.

Country	Percentage	Country	Percentage
Austria	21%	Italy	27%
Belgium	32%	Latvia	16%
Bulgaria	21%	Lithuania	13%
Croatia	22%	Malta	10%
Cyprus	4%	Netherlands	28%
Czech Republic	44%	Norway	9%
Denmark	23%	Poland	18%
Estonia	23%	Portugal	15%
Faroe Islands	9%	Romania	3%
Finland	11%	Russia	22%
France	22%	Slovak Republic	27%
Germany	27%	Slovenia	28%
Greece	6%	Sweden	7%
Greenland	27%	Switzerland	40%
Hungary	16%	Turkey	4%
Iceland	13%	Ukraine	21%
Ireland	39%	United Kingdom	38%
Isle of Man	39%		

47. **Final grades.** A professor (of something other than Statistics!) distributed the following histogram to show the distribution of grades on his 200-point final exam. Comment on the display.

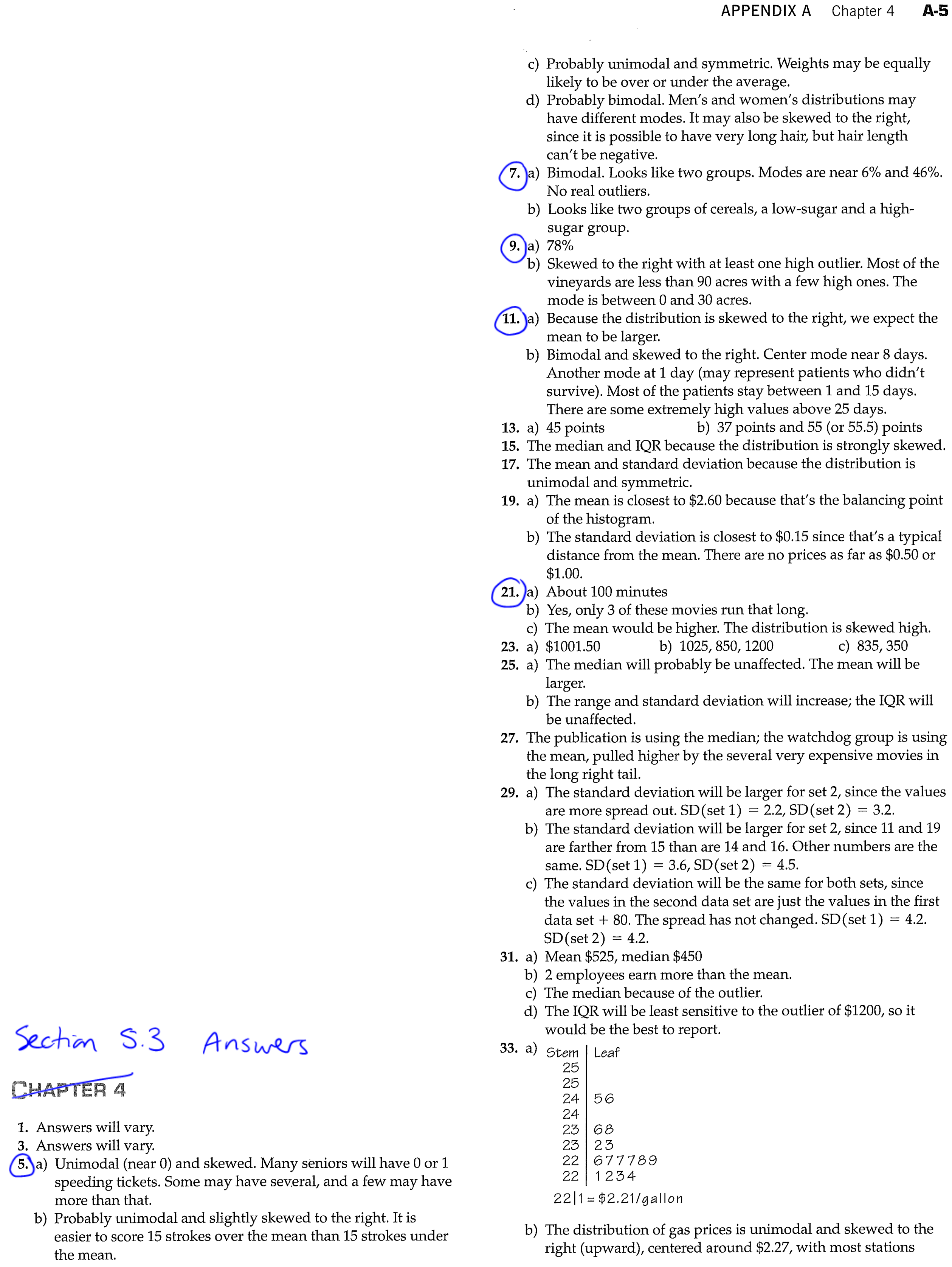


48. **Final grades revisited.** After receiving many complaints about his final-grade histogram from students currently taking a Statistics course, the professor from Exercise 47 distributed the following revised histogram:



a) Comment on this display.
b) Describe the distribution of grades.

49. **Zip codes.** Holes-R-Us, an Internet company that sells piercing jewelry, keeps transaction records on its sales. At a recent sales meeting, one of the staff presented a histogram of the zip codes of the last 500 customers, so that the staff might understand where sales are coming from. Comment on the usefulness and appropriateness of the display.



Section 5.3 Answers

~~CHAPTER~~ 4

1. Answers will vary.
3. Answers will vary.
5. a) Unimodal (near 0) and skewed. Many seniors will have 0 or 1 speeding tickets. Some may have several, and a few may have more than that.
b) Probably unimodal and slightly skewed to the right. It is easier to score 15 strokes over the mean than 15 strokes under the mean.
c) Probably unimodal and symmetric. Weights may be equally likely to be over or under the average.
d) Probably bimodal. Men's and women's distributions may have different modes. It may also be skewed to the right, since it is possible to have very long hair, but hair length can't be negative.

7. a) Bimodal. Looks like two groups. Modes are near 6% and 46%. No real outliers.
b) Looks like two groups of cereals, a low-sugar and a high-sugar group.

9. a) 78%
b) Skewed to the right with at least one high outlier. Most of the vineyards are less than 90 acres with a few high ones. The mode is between 0 and 30 acres.

11. a) Because the distribution is skewed to the right, we expect the mean to be larger.
b) Bimodal and skewed to the right. Center mode near 8 days. Another mode at 1 day (may represent patients who didn't survive). Most of the patients stay between 1 and 15 days. There are some extremely high values above 25 days.

13. a) 45 points b) 37 points and 55 (or 55.5) points

15. The median and IQR because the distribution is strongly skewed.

17. The mean and standard deviation because the distribution is unimodal and symmetric.

19. a) The mean is closest to \$2.60 because that's the balancing point of the histogram.
b) The standard deviation is closest to \$0.15 since that's a typical distance from the mean. There are no prices as far as \$0.50 or \$1.00.

21. a) About 100 minutes
b) Yes, only 3 of these movies run that long.
c) The mean would be higher. The distribution is skewed high.

23. a) \$1001.50 b) 1025, 850, 1200 c) 835, 350

25. a) The median will probably be unaffected. The mean will be larger.
b) The range and standard deviation will increase; the IQR will be unaffected.

27. The publication is using the median; the watchdog group is using the mean, pulled higher by the several very expensive movies in the long right tail.

29. a) The standard deviation will be larger for set 2, since the values are more spread out. SD(set 1) = 2.2, SD(set 2) = 3.2.
b) The standard deviation will be larger for set 2, since 11 and 19 are farther from 15 than are 14 and 16. Other numbers are the same. SD(set 1) = 3.6, SD(set 2) = 4.5.
c) The standard deviation will be the same for both sets, since the values in the second data set are just the values in the first data set + 80. The spread has not changed. SD(set 1) = 4.2. SD(set 2) = 4.2.

31. a) Mean \$525, median \$450
b) 2 employees earn more than the mean.
c) The median because of the outlier.
d) The IQR will be least sensitive to the outlier of \$1200, so it would be the best to report.

33. a)

Stem	Leaf
25	
25	
24	56
24	
23	68
23	23
22	677789
22	1234

22|1 = \$2.21/gallon

b) The distribution of gas prices is unimodal and skewed to the right (upward), centered around \$2.27, with most stations charging between $2.26 and $2.33 per gallon. The lowest and highest prices were $2.21 and $2.46.

c) There are two high prices separated from the other gas stations by a gap.

35. a) The histogram of height is most nearly symmetric and shows no outliers. That makes it the best candidate for summarizing with a mean.

b) The histogram of sip size shows a high outlier. The standard deviation is sensitive to outliers, so we'd prefer to use the IQR for this one.

37. a) Since these data are strongly skewed to the right, the median and IQR are the best statistics to report.

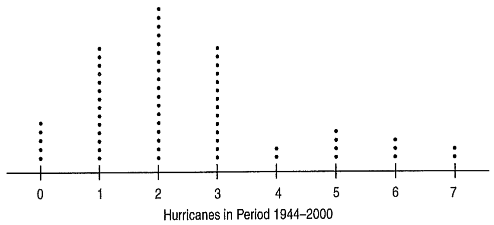
b) The mean will be larger than the median because the data are skewed to the right.

c) The median is 4 million. The IQR is 4.5 million (Q3 = 6 million, Q1 = 1.5 million).

d) The distribution of populations of the states and Washington, DC, is unimodal and skewed to the right. The median population is 4 million. One state is an outlier, with a population of 34 million.

39. Skewed to the right, mode in low 30s. Three low outliers, then a gap from 9 to 22.

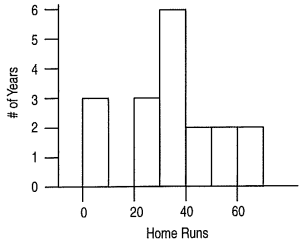
41. a)



b) Slightly skewed to the right. Unimodal, mode near 2. Possibly a second mode near 5. No outliers.

43. a) This is not a histogram. The horizontal axis should split the number of home runs hit in each year into bins. The vertical axis should show the number of years in each bin.

b)



45. Skewed to the right, possibly bimodal with one fairly symmetric group near 4.4, another at 5.6. Two outliers in middle seem not to belong to either group.

Stem	Leaf
57	8
56	27
55	1
54	
53	
52	9
51	
50	8
49	
48	2
47	3
46	034
45	267
44	015
43	0199
42	669
41	22

41|2 = 4.12 pH

47. Histogram bins are too wide to be useful.

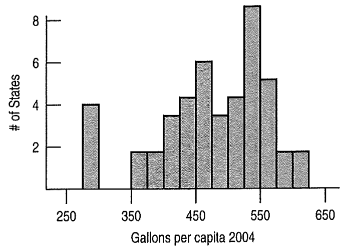
49. Neither appropriate nor useful. Zip codes are categorical data, not quantitative. But they do contain *some* information. The leading digit gives a rough East-to-West placement in the United States. So we see that they have almost no customers in the Northeast, but a bar chart by leading digit would be more appropriate.

51. a) Median 239, IQR 9, Mean 237.6, SD 5.7

b) Because it's skewed to the left, probably better to report Median and IQR.

c) Skewed to the left; may be bimodal. The center is around 239. The middle 50% of states scored between 233 and 242. Alabama, Mississippi, and New Mexico scores were much lower than other states' scores.

53. In the year 2004, per capita gasoline use by state in the United States averaged around 500 gallons per person (mean 488.7, median 500.5). States varied in per capita consumption, with a standard deviation of 68.7 gallons. The only outlier is New York. The IQR of 96.9 gallons shows that 50% of the states had per capita consumption of between 447.5 and 544.4 gallons. The data appear to be bimodal, so the median and IQR are better choices of summary statistics.



55. a) Although numeric codes have been assigned to the different titles, these data are categorical, not quantitative. The mean of 54.41 is meaningless.

b) The typical reasons are skewness and/or outliers.

c) No. Here the numbers are just codes. Most of the people probably had titles of Mr. or Mrs., making the "median" 1, but these summary statistics are meaningless.

Reading for Section 5.4:

The following reading is excerpted from:

> Sullivan. Statistics: Informed Decisions Using Data. 3rd edition, Prentice Hall, 2010, pages 110-118, ANS-14, ANS-15.

Section 5.4:

2.4 GRAPHICAL MISREPRESENTATIONS OF DATA

Objective 1 Describe what can make a graph misleading or deceptive

1 Describe What Can Make a Graph Misleading or Deceptive

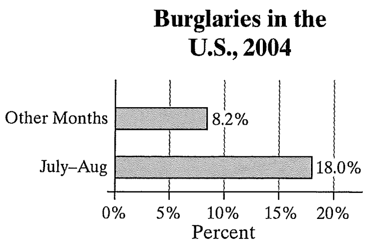
> Statistics: The only science that enables different experts using the same figures to draw different conclusions.—EVAN ESAR

Often, statistics gets a bad rap for having the ability to manipulate data to support any position desired. One method of distorting the truth is through graphics. We mentioned in Section 2.1 how visual displays send more powerful messages than raw data or even tables of data. Since graphics are so powerful, care must be taken both in constructing graphics and interpreting the messages they are trying to convey. Sometimes graphics *mislead*; other times they *deceive*. We will call graphs misleading if they unintentionally create an incorrect impression. We consider graphs deceptive if they purposely attempt to create an incorrect impression. Regardless of the intentions of the graph's creator, an incorrect impression on the reader's part can have serious consequences. Therefore, it is important to be able to recognize misleading and deceptive graphs.

The most common graphical misrepresentation of data is accomplished through manipulation of the scale of the graph, typically in the form of an inconsistent scale or a misplaced origin. Increments between tick marks should remain constant, and scales for comparative graphs should be the same. In addition, readers will usually assume that the baseline, or zero point, is at the bottom of the graph. Starting the graph at a higher or lower value can be misleading.

EXAMPLE 1 Misrepresentation of Data

Figure 21



Problem: A home security company puts out a summer ad campaign with the slogan "When you leave for vacation, burglars leave for work." According to the FBI, roughly 18% of home burglaries in 2004 occurred during the peak vacation months of July and August. The advertisement contains the graph shown in Figure 21. Explain what is wrong with the graphic.

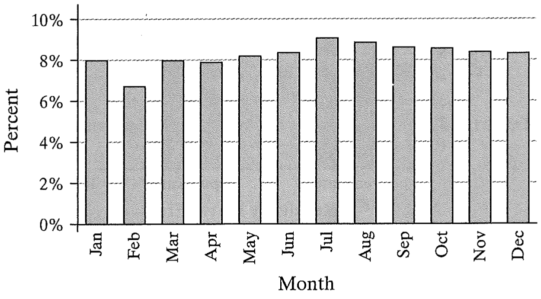
Approach: We need to look at the graph for any characteristics that may mislead a reader, such as inconsistent scales or poorly defined categories.

Solution: Let's consider how the categories of data are defined. The sum of the percentages (the relative frequencies) over all 12 months should be 1. Because $10(0.082) + 0.18 = 1$, it is clear that the bar for Other Months represents an average percent for *each* month, while the bar for July–August represents the average percent for the months July and August *combined*. By combining months, the unsuspecting reader is mislead into thinking that July and August each have a much higher burglary rate.

Figure 22 gives a better picture of the burglary distribution. While there is an increase during the months of July and August, the increase is not as dramatic as the bar graph in Figure 21 implies. In fact, Figure 21 would be considered deceitful because the security company is intentionally trying to convince consumers that July and August are much higher burglary months.

Figure 22

Burglaries in the U.S., 2004



Source: FBI, Crime in the United States, 2004

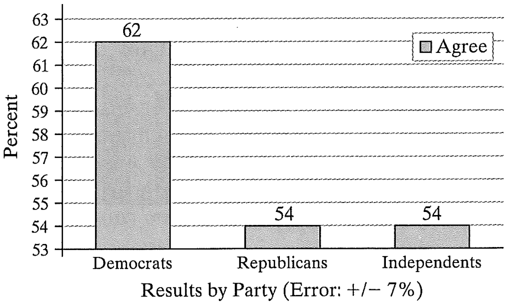
Now Work Problem 5

EXAMPLE 2 Misrepresentation of Data by Manipulating the Vertical Scale

Problem: In 2005, Terri Schiavo was the center of a right-to-die battle that drew international attention. At issue was whether her husband had the right to remove her feeding tube on which she had been dependent for the previous 15 years. A CNN/USA Today/Gallup poll conducted March 18–20, 2005, asked respondents, "*As you may know, on Friday the feeding tube keeping Terri Schiavo alive was removed. Based on what you have heard or read about the case, do you think that the feeding tube should or should not have been removed?*" The results were presented in a graph similar to Figure 23. Explain what is wrong with the graphic.

Figure 23

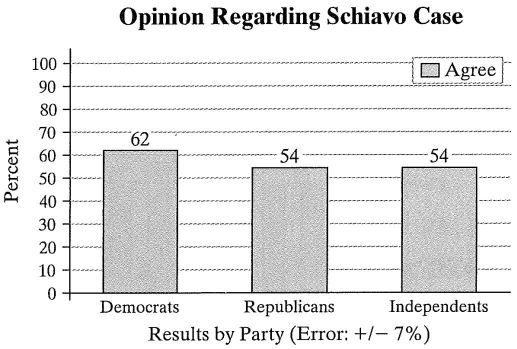
Opinion Regarding Schiavo Case



Approach: We need to look at the graph for any characteristics that may mislead a reader, such as manipulation of the vertical scale.

Solution: The graphic seems to indicate that Democrats overwhelmingly supported the removal of the feeding tube, much more so than either Republicans or Independents. Because the vertical scale does not begin at 0, it may appear that Democrats are 9 times more likely to support the decision (because the bar is 9 times as high as the others) when there is really only an 8 percentage point difference. The dramatic difference in bar heights overshadows the data being presented. Note that the majority of each party sampled supported the decision to remove the feeding tube. In addition, given a ±7 percentage point margin of error for each sample, the actual difference of 8 percentage points would not be *statistically significant* (we will learn more about this later in the text). Ultimately, CNN posted a corrected graphic similar to the one in Figure 24. Note how starting the vertical scale at 0 allows for a more accurate comparison.

Figure 24

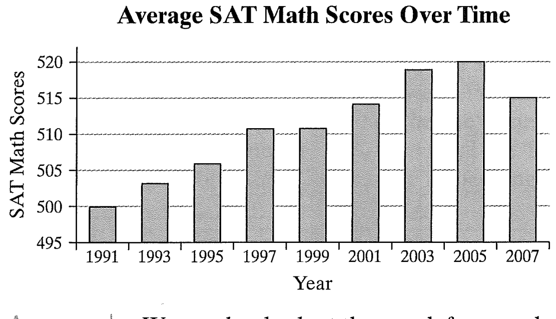


Recall from Section 1.5 that the order of words in a question can affect responses and lead to potential *response bias*. In the question presented in Example 2, the order of the choices ("should," and "should not") were not rotated. The first choice given was "should," and the majority of respondents stated that the feeding tube should be removed. It is possible that the position of the choice in the question could have affected the responses. A better way to present the question would be, "*As you may know, on Friday the feeding tube keeping Terri Schiavo alive was removed. Based on what you have heard or read about the case, do you [Rotated – agree (or) disagree] that the feeding tube should have been removed?*"

EXAMPLE 3 Misrepresentation of Data by Manipulating the Vertical Scale

Problem: The time-series graph shown in Figure 25 depicts the average SAT math scores of college-bound seniors for the years 1991–2007. Determine why this graph might be considered misrepresentative. (*Source*: *College Board*, *College-Bound Seniors*, 2007)

Figure 25



Approach: We need to look at the graph for any characteristics that may mislead a reader, such as manipulation of the vertical scale.

Solution: The graph in the figure may lead a reader to believe that SAT math scores have increased substantially since 1991. While SAT math scores have been increasing, they have not doubled or quadrupled (since the point for 2007 is 4 times as high as the point for 1991). We notice in the figure that the vertical axis begins at 495 instead of the baseline of 200 (the minimum score for the math portion of the SAT). This type of scaling is common when the smallest observed data value is a rather large number. It is not necessarily done purposely to confuse or mislead the reader. Often, the main purpose in graphs (particularly time-series graphs) is to discover a trend, rather than the actual differences in the data. The trend is clearer in Figure 25 than in Figure 26, where the vertical axis begins at the baseline. Remember that the goal of a good graph is to make the data stand out. When displaying time-series data, as in this example, it is better to use a time-series plot to discover any trends. In addition, instead of beginning the axis of a graph at 0 as in Figure 26, scales are frequently truncated so they begin at a value slightly less than the smallest value in the data set. There is nothing inherently wrong with doing this, but special care must be taken to make the reader aware of the scaling that is used. Figure 27 shows the proper construction of the graph of the SAT math scores, with the graph beginning at 495. The symbol ϟ is used to indicate that the scale has been truncated and the graph has a gap in it. Notice that the lack of bars allows us to focus on the trend in the data, rather than the relative size (or area) of the bars.

Figure 26

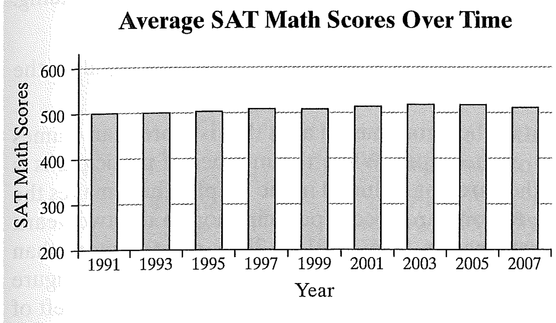
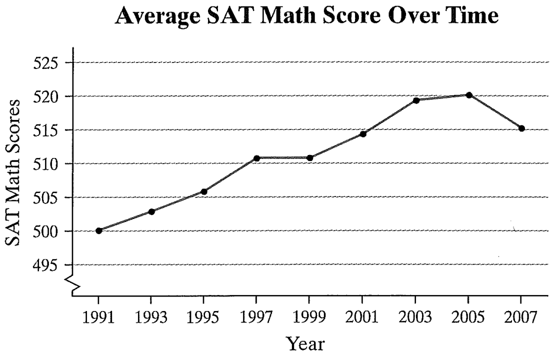


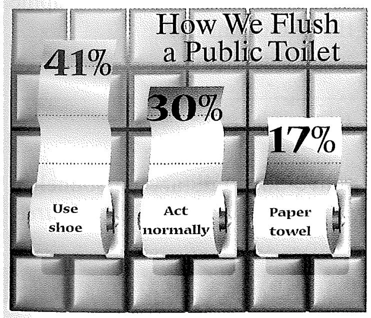
Figure 27



Now Work Problem 3

EXAMPLE 4 Misrepresentation of Data

Figure 28



Problem: The bar graph illustrated in Figure 28 is a *USA Today*-type graph. A survey was conducted by Impulse Research for Quilted Northern Confidential in which individuals were asked how they would flush a toilet when the facilities are not sanitary. What's wrong with the graphic?

Approach: We need to compare the vertical scales of each bar to see if they accurately depict the percentages given.

Solution: First, it is unclear whether the bars include the roll of toilet paper or not. In either case, the roll corresponding to "use shoe" should be 2.4 (= 41/17) times longer than the roll corresponding to "paper towel." If we include the roll of toilet paper, then the bar corresponding to "use shoe" is less than double the length of "paper towel." If we do not include the roll of toilet paper, then the bar corresponding to "use shoe" is almost exactly double the length of the bar corresponding to "paper towel." The vertical scaling is incorrect.

Now Work Problem 11

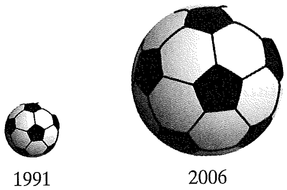
Newspapers, magazines, and Websites often go for a "wow" factor when displaying graphs. In many cases, the graph designer is more interested in catching the reader's eye than making the data stand out. The two most commonly used tactics are 3-D graphs and pictograms (graphs that use pictures to represent the data). The use of 3-D effects is strongly discouraged, because such graphs are often difficult to read, add little value to the graph, and distract the reader from the data.

When comparing bars that represent different quantities, our eyes are really comparing the *areas* of the bars. In our discussion of bar graphs and histograms, we emphasized that the bars or classes should be of the same width. The advantage of having uniform width is that the area of the bar is then proportional to its height, so we can simply compare the heights of the bars for the different quantities. However, when we use two-dimensional pictures in place of the bars, it is not possible to obtain a uniform width. To avoid distorting the picture when values increase or decrease, both the height and width of the picture must be adjusted. This often leads to misleading graphs.

EXAMPLE 5 Misleading Graphs

Figure 29

Soccer Participation



Problem: Soccer continues to grow in popularity as a sport in the United States. High-profile players such as Mia Hamm and Landon Donovan have helped to generate renewed interest in the sport at various age levels. In 1991 there were approximately 10 million participants in the United States aged 7 years or older. By 2006 this number had climbed to 14 million. To illustrate this increase, we could create a graphic like the one shown in Figure 29. Describe how the graph may be misleading. (*Source*: U.S. Census Bureau; National Sporting Goods Association.)

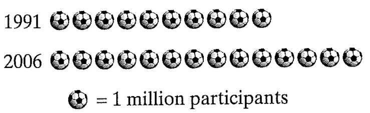
Approach: We look for characteristics of the graph that seem to manipulate the facts, such as an incorrect depiction of the size of the graphics.

Solution: The graph on the right of the figure has an area that is more than 4 times the area of the graph on the left of the figure. While the number of participants is given in the problem statement, they are not included in the graph, which makes the reader rely on the graphic alone to compare soccer participation in the two years. There was a 40% increase in participation from 1991 to 2006, not the more than 300% indicated by the graphic. To be correct, the graph on the right of the figure should have an area that is only 40% more than the area of the graph on the left of the figure. Adding the data values to the graphic would help reduce the chance of misinterpretation due to the oversized graph.

Now Work Problem 17

Figure 30

Soccer Participation



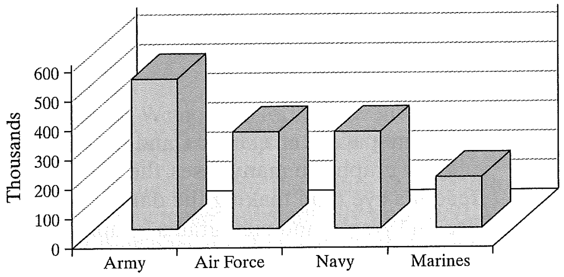
A variation on pictograms is to use a smaller picture repeatedly, with each picture representing a certain quantity. For example, we could present the data from Figure 29 by using a smaller soccer ball to represent 1 million participants. The resulting graphic is displayed in Figure 30. Note how the uniform size of the graphic allows us to make a more accurate comparison of the two quantities.

EXAMPLE 6 Misleading Graphs

Problem: Figure 31 represents the number of active-duty military personnel in the United States as of August 2007. Describe how this graph is misleading. (*Source*: infoplease.com.)

Figure 31

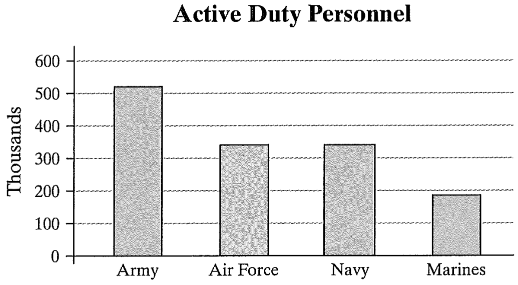
Active Duty Personnel, 2007 (Aug)



Approach: Again, we look for characteristics of the graph that seem to distort the facts or distract the reader.

Solution: The three-dimensional bar graph in the figure may draw the reader's attention, but the bars seem to stand out more than the data they represent. The perspective angle of the graph makes it difficult to estimate the data values being presented, actually resulting in estimates that are typically lower than the true values. This in turn makes comparison of the data difficult. The only dimension that matters is bar height, so this is what should be emphasized. Figure 32 displays the same data in a two-dimensional bar graph. Which graphic is easier to read?

Figure 32



The material presented in this section is by no means all-inclusive. There are many ways to create graphs that mislead. Two popular texts written about ways that graphs mislead or deceive are *How to Lie with Statistics* (W. W. Norton & Company, Inc., 1982) by Darrell Huff and *The Visual Display of Quantitative Information* (Graphics Press, 2001) by Edward Tufte.

We conclude this section with some guidelines for constructing good graphics.

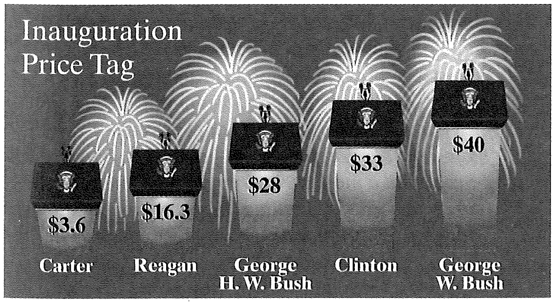
- Title and label the graphic axes clearly, providing explanations if needed. Include units of measurement and a data source when appropriate.
- Avoid distortion. Never lie about the data.
- Minimize the amount of white space in the graph. Use the available space to let the data stand out. If scales are truncated, be sure to clearly indicate this to the reader.
- Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader.
- Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion of the graph to draw attention to that area. Don't try to force the reader to any specific part of the graph. Let the data speak for themselves.
- Avoid relative graphs that are devoid of data or scales.

Section 5.4 Exercises

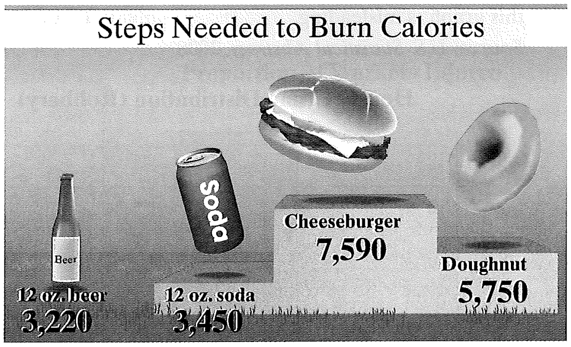
2.4 ASSESS YOUR UNDERSTANDING

Applying the Concepts

1. **Inauguration Cost** The following is a *USA Today*-type graph. Explain how it is misleading.



2. **Burning Calories** The following is a *USA Today*-type graph.

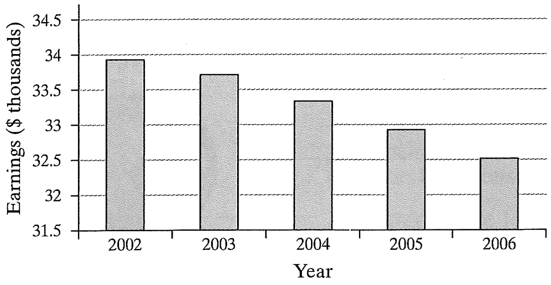


(a) Explain how it is misleading.
(b) What could be done to improve the graphic?

3. **Median Earnings** The following graph shows the median NW earnings for females from 2002 to 2006 in constant 2006 dollars.

Source: U.S. Census Bureau, Income, Poverty, and Health Insurance Coverage in the United States, 2006

Median Earnings for Females



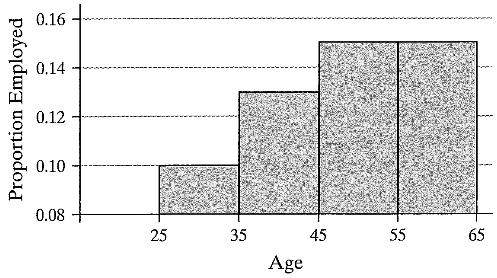
(a) How is the bar graph misleading? What does the graph seem to convey?

(b) Redraw the graph so that it is not misleading. What does the new graph seem to convey?

4. **Union Membership** The following relative frequency histogram represents the proportion of employed people aged 25 to 64 years old who were members of a union.

Source: U.S. Bureau of Labor Statistics

Union Membership



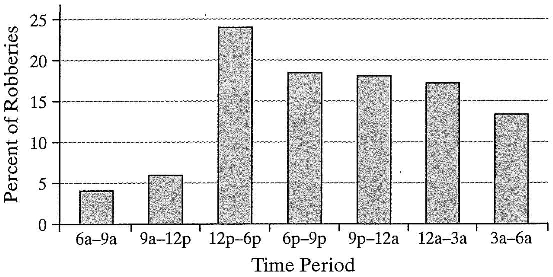
(a) Describe how this graph is misleading. What might a reader conclude from the graph?

(b) Redraw the histogram so that it is not misleading.

5. **Robberies** A newspaper article claimed that the afternoon NW hours were the worst in terms of robberies and provided the following graph in support of this claim. Explain how this graph is misleading.

Source: *U.S. Statistical Abstract*, 2008

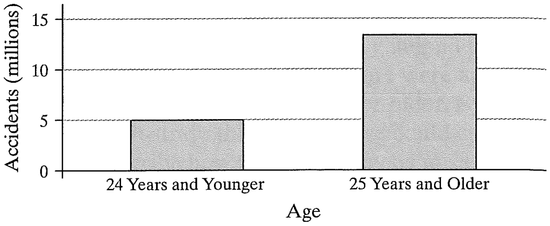
Hourly Crime Distribution (Robbery)



6. **Car Accidents** An article in a student newspaper claims that younger drivers are safer than older drivers and provides the following graph to support the claim. Explain how this graph is misleading.

Source: *U.S. Statistical Abstract*, 2008

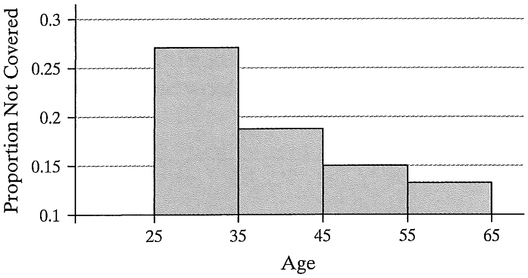
Number of Motor Vehicle Accidents, 2005



7. **Health Insurance** The following relative frequency histogram represents the proportion of people aged 25 to 64 years old not covered by any health insurance in 2006.

Source: U.S. Census Bureau

Proportion Not Covered by Health Insurance



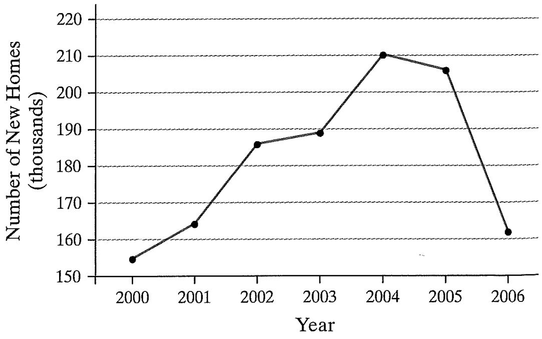
(a) Describe how this graph is misleading. What might a reader conclude from the graph?

(b) Redraw the histogram so that it is not misleading.

8. **New Homes** The following time-series plot shows the number of new homes built in the Midwest from 2000 to 2006.

Source: *U.S. Statistical Abstract*, 2008

New Homes in Midwest



(a) Describe how this graph is misleading.

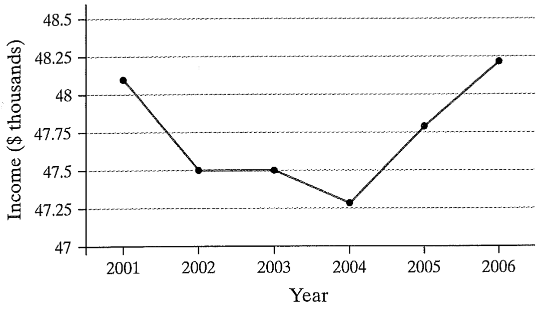
(b) What is the graph trying to convey?

(c) In January 2006, the National Association of Realtors reported, "*A lot of demand has been met over the last five years, and a modest rise in mortgage interest rates is causing some market cooling. Along with regulatory tightening on nontraditional mortgages, there will be fewer investors in the market this year.*" Does the graph support this view? Explain why or why not.

9. **Median Income** The following time-series plot shows the median household income for the years 2001 to 2006 in constant 2006 dollars.

Source: U.S. Census Bureau

U.S. Median Household Income

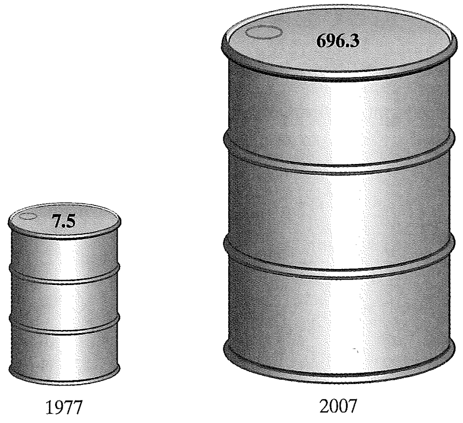


(a) Describe how the graph is misleading.
(b) What is the graph trying to convey?
(c) Redraw the graph so that Median Household Income appears to be relatively stable for the years shown.

10. **You Explain It! Oil Reserves** The U.S. Strategic Oil Reserve is a government-owned stockpile of crude oil. It was established after the oil embargo in the mid-1970s and is meant to serve as a national defense fuel reserve, as well as to offset reductions in commercial oil supplies that would threaten the U.S. economy.

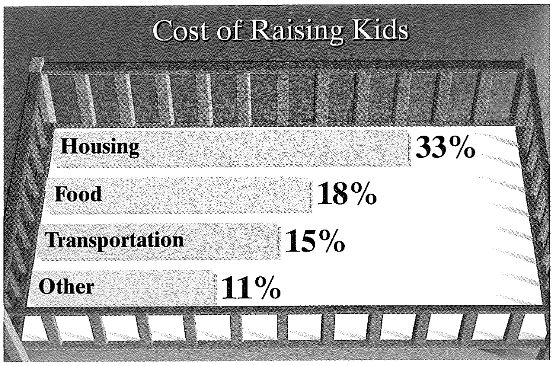
Source: U.S. Energy Information Administration

U.S. Strategic Oil Reserves
(millions of barrels)



(a) How many times larger should the graphic for 2007 be than the 1977 graphic (to the nearest whole number)?
(b) The United States imported approximately 10.1 million barrels of oil per day in 2007. At that rate, assuming no change in U.S. oil production, how long would the U.S. strategic oil reserve last if no oil were imported?

11. **Cost of Kids** The following is a *USA Today*-type graph based on data from the Department of Agriculture. It represents the percentage of income a middle-income family will spend on their children.



(a) How is the graphic misleading?
(b) What could be done to improve the graphic?

12. **Electricity** The following table gives the average per kilowatt-hour prices of electricity in the United States for the years 2001 to 2007.

Source: U.S. Energy Information Administration

Year	2001	2002	2003	2004	2005	2006	2007
Price per kWh (cents)	8.58	8.44	8.72	8.95	9.45	10.40	10.65

(a) Construct a misleading graph indicating that the price per kilowatt-hour has more than tripled since 2001.
(b) Construct a graph that is not misleading.

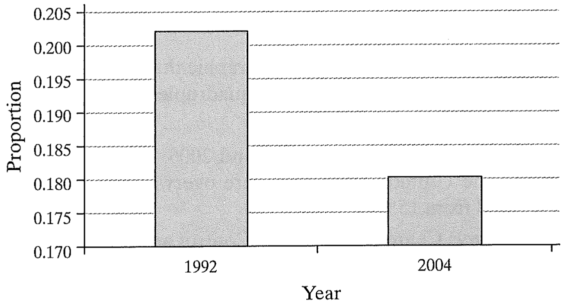
13. **ACT Composite** The following table gives the average ACT composite scores for the years 2003–2007.

Year	2003	2004	2005	2006	2007
Average ACT composite	20.8	20.9	20.9	21.1	21.2

(a) Construct a misleading time-series plot that indicates the average ACT composite score has risen sharply over the given time period.
(b) Construct a time-series plot that is not misleading.
(c) Which of the two graphs would you prefer if you were merely looking for trends in the data? Explain.

14. **Worker Injury** The safety manager at Klutz Enterprises provides the following graph to the plant manager and claims that the rate of worker injuries has been reduced by 67% over a 12-year period. Does the graph support his claim? Explain.

Proportion of Workers Injured



15. Health Care Expenditures The following data represent health care expenditures as a percentage of the U.S. gross domestic product (GDP) from 2001 to 2007. Gross domestic product is the total value of all goods and services created during the course of the year.

Source: Center for Medicare and Medicad Services, Office of the Actuary

Year	Health Care as a Percent of GDP
2001	14.5
2002	15.3
2003	15.8
2004	15.9
2005	16.0
2006	16.0
2007	16.2

(a) Construct a time-series plot that a politician would create to support the position that health care expenditures, as a percentage of GDP, are increasing and must be slowed.

(b) Construct a time-series plot that the health care industry would create to refute the opinion of the politician.

(c) Construct a time-series plot that is not misleading.

16. Motor Vehicle Death Rates The following data represent the number of motor vehicle deaths (within 30 days of accident) and the traffic death rates (number of deaths per 100,000 licensed drivers) from 2001 to 2005.

Year	Motor Vehicle Deaths (in thousands)	Traffic Death Rate (per 100,000 licensed drivers)
2001	42.2	22.1
2002	43.0	22.0
2003	42.9	21.9
2004	42.8	21.5
2005	43.4	21.7

Source: *U.S. Statistical Abstract*, 2008

(a) Construct a time-series graph to support the belief that the roads are becoming less safe.

(b) Construct a time-series graph to support the belief that the roads are becoming safer.

(c) Which graph do you feel better represents the situation?

17. Gas Hike NW The average per gallon price for regular unleaded gasoline in the United States rose from $1.46 in 2001 to $4.01 in 2008.

Source: U.S. Energy Information Administration

(a) Construct a graphic that is not misleading to depict this situation.

(b) Construct a misleading graphic that makes it appear the average price roughly quadrupled between 2001 and 2007.

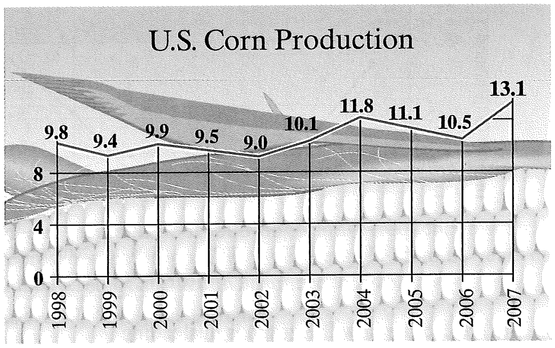
18. Overweight Between 1980 and 2006, the number of adults in the United States who were overweight more than doubled from 15% to 34%.

Source: Centers for Disease Control and Prevention

(a) Construct a graphic that is not misleading to depict this situation.

(b) Construct a misleading graphic that makes it appear that the percent of overweight adults has more than quadrupled between 1980 and 2006.

19. **Corn Production** The following *USA Today*-type graphic illustrates U.S. corn production in billions of bushels for the years 1998 to 2007.

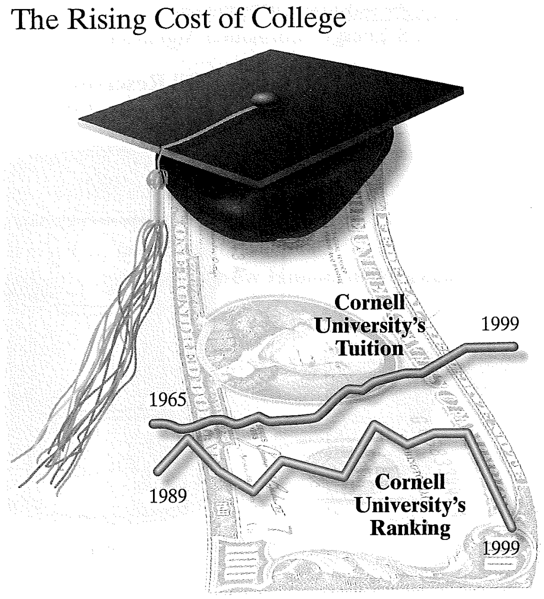


(a) What type of graph is being displayed?

(b) Describe some of the problems with this graphic.

(c) Construct a new graphic that is not misleading and makes the data stand out.

20. Putting It Together: College Costs The cover of the *Ithaca Times* from December 7, 2000 is shown.



(a) Identify the two variables being graphed and describe them in terms of type and measurement level.

(b) What type of data collection method was likely used to create this graph?

(c) What type of graph is displayed?

(d) What message does the graph convey to you? How might this graph be misleading?

(e) Describe at least three things that are wrong with the graph.

Section 5.4:

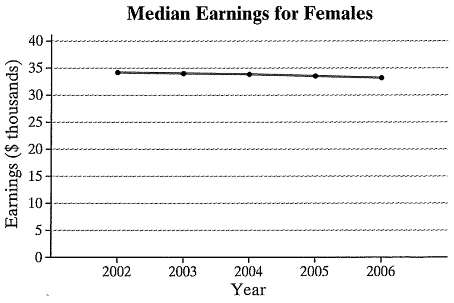
Answers

2.4 Assess Your Understanding (page 115)

1. The lengths of the bars are not proportional. For example, the bar representing the cost of Clinton's inauguration should be slightly more than 9 times the one for Carter's cost and twice as long as the bar representing Reagan's cost.

3. **(a)** The vertical axis starts at 31.5 instead of 0. This tends to indicate that the median earnings for females decreased at a faster rate than they actually did.

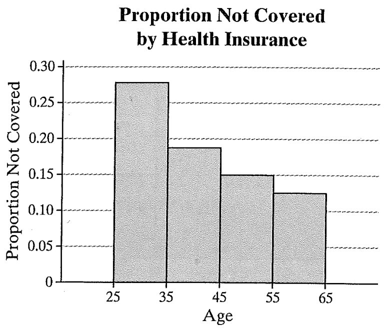
(b) This graph indicates that the median earnings for females have decreased slightly over the given time period.



5. The bar for 12p–6p covers twice as many hours as the other bars. By combining two 3-hour periods, this bar looks larger compared to the others, making afternoon hours look more dangerous. When the bar is split into two periods, the graph may give a different impression.

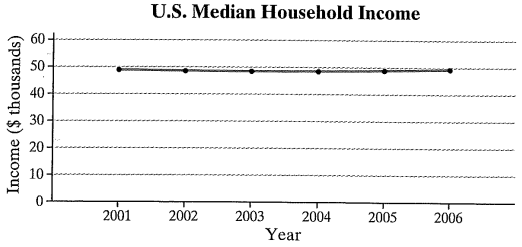
7. **(a)** The vertical axis starts at 0.1 instead of 0. This might cause the reader to conclude, for example, that the proportion of people aged 25 to 34 who are not covered by health insurance is more than twice the proportion for those aged 45 to 54 years.

(b)



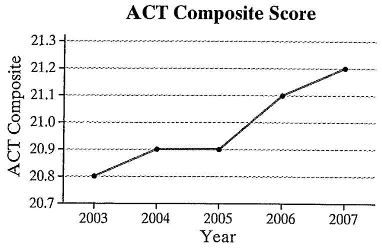
9. **(a)** The vertical axis starts at 47 without indicating a gap.
(b) It may convey that the median household income is increasing after a period of decline.

(c)

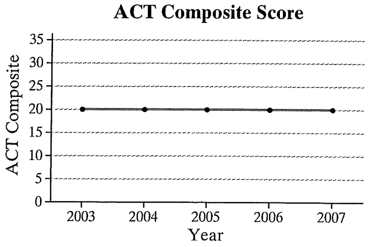


11. **(a)** The bar for housing should be a little more than twice the length of the bar for transportation, but it is not.
(b) Adjust the graph so that the lengths of the bars are proportional.

13. **(a)**

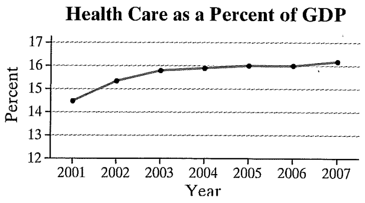


(b)

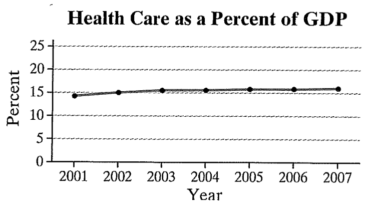


(c) Graph (a) is preferred because the trend can be seen.

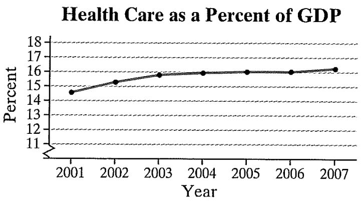
15. **(a)** The politician's view:



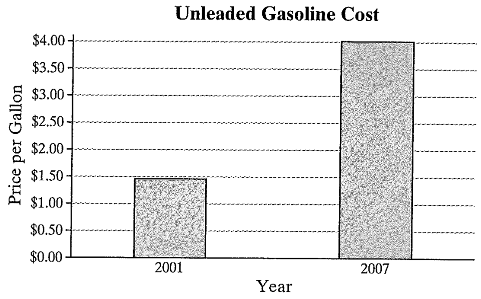
(b) The health care industry's view:



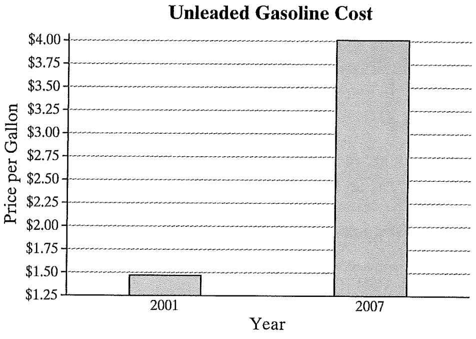
(c) A view that is not misleading:



17. **(a)** Graphic that is not misleading:

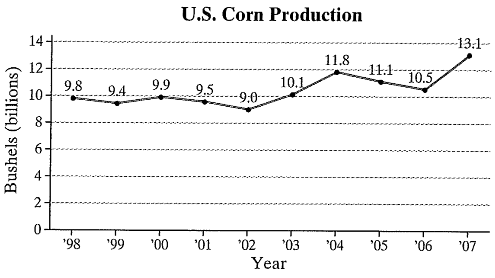


(b) Graphic that is misleading (graphics may vary):



19. **(a)** Time series
(b) The graph is too cluttered; the axes are not labeled; the grid stands out more than the data.

(c) Graph that is not misleading:



Chapter 6

Describing Data with Numbers

Reading for Section 6.1:

The following reading is excerpted from:

> Sullivan. Statistics: Informed Decisions Using Data. 3rd edition, Prentice Hall, 2010, pages 129-135, 137-138, 141-142, ANS-21.

Section ~~2.1~~ 6.1

3.1 MEASURES OF CENTRAL TENDENCY

Preparing for This Section Before getting started, review the following:

- Population versus sample (Section 1.1, p. 5)
- Parameter versus statistic (Section 1.1, p. 5)
- Quantitative data (Section 1.1, p. 9)
- Qualitative data (Section 1.1, p. 9)
- Simple random sampling (Section 1.3, pp. 22–27)

Objectives

1 Determine the arithmetic mean of a variable from raw data

2 Determine the median of a variable from raw data

3 Explain what it means for a statistic to be resistant

4 Determine the mode of a variable from raw data

A measure of central tendency numerically describes the average or typical data value. We hear the word *average* in the news all the time:

- The average miles per gallon of gasoline of the 2008 Chevrolet Corvette in city driving is 15 miles.
- According to the U.S. Census Bureau, the national average commute time to work in 2006 was 24.0 minutes.
- According to the U.S. Census Bureau, the average household income in 2006 was $48,201.
- The average American woman is 5′4″ tall and weighs 142 pounds.

In this chapter, we discuss three measures of central tendency: the *mean*, the *median*, and the *mode*. While other measures of central tendency exist, these three are the most widely used. When the word *average* is used in the media (newspapers, reporters, and so on) it usually refers to the mean. But beware! Some reporters use the term *average* to refer to the median or mode. As we shall see, these three measures of central tendency can give very different results!

CAUTION Whenever you hear the word *average*, be aware that the word may not always be referring to the mean. One average could be used to support one position, while another average could be used to support a different position.

1 Determine the Arithmetic Mean of a Variable from Raw Data

When used in everyday language, the word *average* often represents the arithmetic mean. To compute the arithmetic mean of a set of data, the data must be quantitative.

Definitions

The **arithmetic mean** of a variable is computed by determining the sum of all the values of the variable in the data set and dividing by the number of observations. The **population arithmetic mean**, μ (pronounced "mew"), is computed using all the individuals in a population. The population mean is a parameter.

The **sample arithmetic mean**, $\bar{x}$ (pronounced "x-bar"), is computed using sample data. The sample mean is a statistic.

While other types of means exist (see Problems 45 and 46), the arithmetic mean is generally referred to as the **mean**. We will follow this practice for the remainder of the text.

Typically, Greek letters are used to represent parameters, and Roman letters are used to represent statistics. Statisticians use mathematical expressions to describe the method for computing means.

In Other Words
To find the mean of a set of data, add up all the observations and divide by the number of observations.

If $x_1, x_2, \ldots, x_N$ are the N observations of a variable from a population, then the population mean, μ, is

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N} \qquad \textbf{(1)}$$

If $x_1, x_2, \ldots, x_n$ are n observations of a variable from a sample, then the sample mean, $\overline{x}$, is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n} \qquad \textbf{(2)}$$

Note that N represents the size of the population, while n represents the size of the sample. The symbol Σ (the Greek letter capital sigma) tells us the terms are to be added. The subscript i is used to make the various values distinct and does not serve as a mathematical operation. For example, x_1 is the first data value, x_2 is the second, and so on.

Let's look at an example to help distinguish the population mean and sample mean.

EXAMPLE 1 Computing a Population Mean and a Sample Mean

Table 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Problem: The data in Table 1 represent the first exam score of 10 students enrolled in a section of Introductory Statistics.

(a) Compute the population mean.

(b) Find a simple random sample of size $n = 4$ students.

(c) Compute the sample mean of the sample obtained in part (b).

Approach

(a) To compute the population mean, we add up all the data values (test scores) and then divide by the number of individuals in the population.

(b) Recall from Section 1.3 that we can use either Table I in Appendix A, a calculator with a random-number generator, or computer software to obtain simple random samples. We will use a TI-84 Plus graphing calculator.

(c) The sample mean is found by adding the data values that correspond to the individuals selected in the sample and then dividing by $n = 4$, the sample size.

Solution

(a) We compute the population mean by adding the scores of all 10 students:

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + \cdots + x_{10} \\ &= 82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88 \\ &= 790\end{aligned}$$

Divide this result by 10, the number of students in the class.

$$\mu = \frac{\sum x_i}{N} = \frac{790}{10} = 79$$

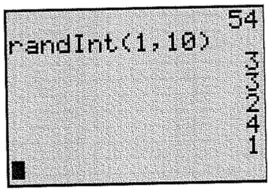
Although it was not necessary in this problem, we will agree to round the mean to one more decimal place than that in the raw data.

(b) To find a simple random sample of size $n = 4$ from a population whose size is $N = 10$, we will use the TI-84 Plus random-number generator with a seed of 54. (Recall that this gives the starting point that the calculator uses to generate the list of random numbers.) Figure 1 shows the students in the sample. Bilal (90), Ryanne (77), Pam (71), and Michelle (82) are in the sample.

(c) We compute the sample mean by first adding the scores of the individuals in the sample.

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + x_4 \\ &= 90 + 77 + 71 + 82 \\ &= 320\end{aligned}$$

Figure 1



```
             54
randInt(1,10)
              3
              3
              2
              4
              1
```

Divide this result by 4, the number of individuals in the sample.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{320}{4} = 80$$

Now Work Problem 23

IN CLASS ACTIVITY

Population Mean versus Sample Mean

Treat the students in the class as a population. All the students in the class should determine their pulse rates.

(a) Compute the population mean pulse rate.

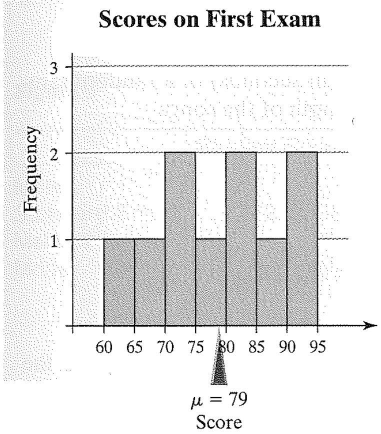
(b) Obtain a simple random sample of $n = 4$ students and compute the sample mean. Does the sample mean equal the population mean?

(c) Obtain a second simple random sample of $n = 4$ students and compute the sample mean. Does the sample mean equal the population mean?

(d) Are the sample means the same? Why?

It is helpful to think of the mean of a data set as the center of gravity. In other words, the mean is the value such that a histogram of the data is perfectly balanced, with equal weight on each side of the mean. Figure 2 shows a histogram of the data in Table 1 with the mean labeled. The histogram balances at $\mu = 79$.

Figure 2

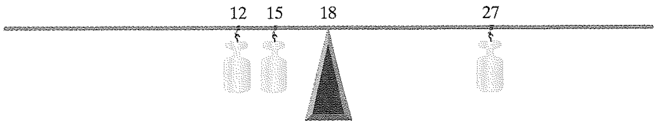


IN CLASS ACTIVITY

The Mean as the Center of Gravity

Find a yardstick, a fulcrum, and three objects of equal weight (maybe 1-kilogram weights from the physics department). Place the fulcrum at 18 inches so that the yardstick balances like a teeter-totter. Now place one weight on the yardstick at 12 inches, another at 15 inches, and the third at 27 inches. See Figure 3.

Figure 3



Does the yardstick balance? Now compute the mean of the location of the three weights. Compare this result with the location of the fulcrum. Conclude that the mean is the center of gravity of the data set.

2 Determine the Median of a Variable from Raw Data

A second measure of central tendency is the median. To compute the median of a set of data, the data must be quantitative.

Definition

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use *M* to represent the median.

In Other Words

To help remember the idea behind the median, think of the median of a highway; it divides the highway in half. So the median divides the data in half, with at most half the data below the median and at most half above it.

To determine the median of a set of data, we use the following steps:

Steps in Finding the Median of a Data Set

Step 1: Arrange the data in ascending order.

Step 2: Determine the number of observations, n.

Step 3: Determine the observation in the middle of the data set.

- If the number of observations is odd, then the median is the data value that is exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{n+1}{2}$ position.
- If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the $\frac{n}{2}$ position and the $\frac{n}{2}+1$ position.

EXAMPLE 2 Determining the Median of a Data Set with an Odd Number of Observations

Problem: The data in Table 2 represent the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

Approach: We will follow the steps listed above.

Solution

Table 2

Song Name	Length
"Sister Golden Hair"	201
"Black Water"	257
"Free Bird"	284
"The Hustle"	208
"Southern Nights"	179
"Stayin' Alive"	222
"We Are Family"	217
"Heart of Glass"	206
"My Sharona"	240

Step 1: Arrange the data in ascending order:

$$179, 201, 206, 208, 217, 222, 240, 257, 284$$

Step 2: There are $n = 9$ observations.

Step 3: Since there are an odd number of observations, the median will be the observation exactly in the middle of the data set. The median, M, is 217 seconds (the $\frac{n+1}{2} = \frac{9+1}{2} = 5$th data value). We list the data in ascending order, with the median in blue.

$$179, 201, 206, 208, 217, 222, 240, 257, 284$$

Notice there are four observations to the left and four observations to the right of the median.

EXAMPLE 3 Determining the Median of a Data Set with an Even Number of Observations

Problem: Find the median score of the data in Table 1 on page 130.

Approach: We will follow the steps listed above.

Solution

Step 1: Arrange the data in ascending order:

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$

Step 2: There are $n = 10$ observations.

Step 3: Because there are $n = 10$ observations, the median will be the mean of the two middle observations. The median is the mean of the fifth $\left(\frac{n}{2} = \frac{10}{2} = 5\right)$ and sixth $\left(\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6\right)$ observations with the data written in ascending order. So the median is the mean of 77 and 82:

$$M = \frac{77 + 82}{2} = 79.5$$

Notice that there are five observations to the left and five observations to the right of the median, as follows:

$$62, 68, 71, 74, 77, \underset{\boxed{M = 79.5}}{\uparrow} 82, 84, 88, 90, 94$$

We conclude that 50% (or half) of the students scored less than 79.5 and 50% (or half) of the students scored above 79.5.

Now compute the median of the data in Problem 15 by hand

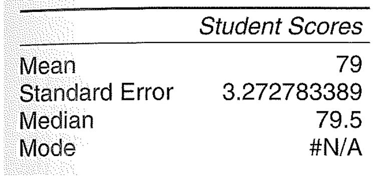
EXAMPLE 4 Finding the Mean and Median Using Technology

Problem: Use statistical software or a calculator to determine the population mean and median of the student test score data in Table 1 on page 130.

Approach: We will use Excel to obtain the mean and median. The steps for calculating measures of central tendency using the TI-83/84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step-by-Step on page 142.

Solution: Figure 4 shows the output obtained from Excel.

Figure 4



	Student Scores
Mean	79
Standard Error	3.272783389
Median	79.5
Mode	#N/A

3 Explain What It Means for a Statistic to be Resistant

Thus far, we have discussed two measures of central tendency, the mean and the median. Perhaps you are asking yourself which measure is better. It depends.

EXAMPLE 5 Comparing the Mean and the Median

Problem: Yolanda wants to know how much time she typically spends on her cell phone. She goes to her phone's website and records the phone call length for a random sample of 12 calls and obtains the data in Table 3. Find the mean and median length of a cell phone call. Which measure of central tendency better describes the length of a typical phone call?

Table 3

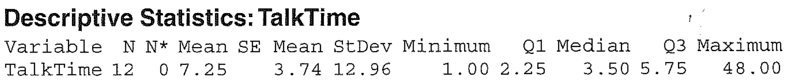
1	7	4	1
2	4	3	48
3	5	3	6

Source: Yolanda Sullivan's cell phone records

Approach: We will find the mean and median using MINITAB. To help judge which is the better measure of central tendency, we will also draw a dot plot of the data using MINITAB.

Solution: Figure 5 indicates that the mean talk time is $\bar{x} = 7.3$ minutes and the median talk time is 3.5 minutes. Figure 6 shows a dot plot of the data using MINITAB.

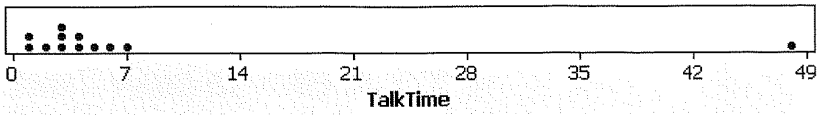
Figure 5



Descriptive Statistics: TalkTime

```
Variable  N N* Mean SE Mean StDev Minimum   Q1 Median   Q3 Maximum
TalkTime 12  0 7.25    3.74 12.96    1.00 2.25   3.50 5.75   48.00
```

Figure 6



Which measure of central tendency do you think better describes the typical amount of time Yolanda spends on a phone call? Given the fact that only one phone call has a talk time greater than the mean, we conclude that the mean is not representative of the typical talk time. So the median is the better measure of central tendency.

Look back at the data in Table 3. Suppose Yolanda's 48-minute phone call was actually a 5-minute phone call. Then the mean talk time would be 3.7 minutes and the median talk time would still be 3.5 minutes. So the one extreme observation (48 minutes) can cause the mean to increase substantially, but have no effect on the median. In other words, the mean is sensitive to extreme values while the median is not. In fact, if Yolanda's 48-minute phone call had actually been a 148-minute phone call, the median would still be 3.5 minutes, but the mean would increase to 15.6 minutes. The median is unchanged because it is based on the value of the middle observation, so the value of the largest observation does not play a role in its computation. Because extreme values do not affect the value of the median, we say that the median is *resistant.*

Definition A numerical summary of data is said to be **resistant** if extreme values (very large or small) relative to the data do not affect its value substantially.

So the median is resistant, while the mean is not resistant.

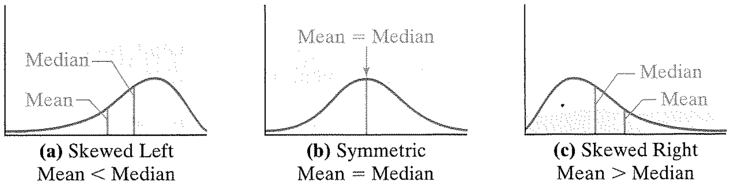
When data are either skewed left or skewed right, there are extreme values in the tail, which tend to pull the mean in the direction of the tail. For example, in skewed-right distributions, there are large observations in the right tail. These observations tend to increase the value of the mean, while having little effect on the median. Similarly, in distributions that are skewed left, the mean will tend to be smaller than the median. In distributions that are symmetric, the mean and the median are close in value. We summarize these ideas in Table 4 and Figure 7.

Table 4

Relation Between the Mean, Median, and Distribution Shape

Distribution Shape	**Mean versus Median**
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

Figure 7
Mean or median versus skewness



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

A word of caution is in order. The relation between the mean, median, and skewness are guidelines. The guidelines tend to hold up well for continuous data, but, when the data are discrete, the rules can be easily violated. See Problem 47.*

A question you may be asking yourself is, "Why would I ever compute the mean?" After all, the mean and median are close in value for symmetric data, and

*This idea is discussed in "Mean, Median, and Skew: Correcting a Textbook Rule" by Paul T. von Hippel. *Journal of Statistics Education,* Volume 13, Number 2 (2005).

the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of the statistical inference that we perform is based on the mean. We will have more to say about this in Chapter 8.

EXAMPLE 6 Describing the Shape of a Distribution

Table 5

5.8	7.4	9.2	7.0	8.5	7.6
7.9	7.8	7.9	7.7	9.0	7.1
8.7	7.2	6.1	7.2	7.1	7.2
7.9	5.9	7.0	7.8	7.2	7.5
7.3	6.4	7.4	8.2	9.1	7.3
9.4	6.8	7.0	8.1	8.0	7.5
7.3	6.9	6.9	6.4	7.8	8.7
7.1	7.0	7.0	7.4	8.2	7.2
7.6	6.7				

Problem: The data in Table 5 represent the birth weights (in pounds) of 50 randomly sampled babies.

(a) Find the mean and the median.

(b) Describe the shape of the distribution.

(c) Which measure of central tendency better describes the average birth weight?

Approach

(a) This can be done either by hand or technology. We will use a TI-84 Plus to compute the mean and the median.

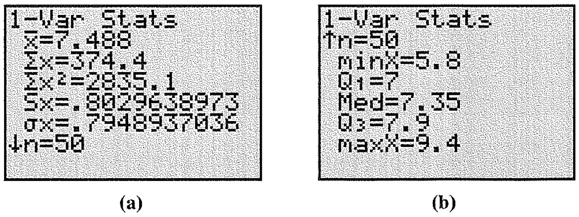
(b) We will draw a histogram to identify the shape of the distribution.

(c) If the data are roughly symmetric, the mean is a good measure of central tendency. If the data are skewed, the median is a good measure of central tendency.

Solution

(a) Using a TI-84 Plus, we find $\bar{x} = 7.49$ and $M = 7.35$. See Figure 8.

Figure 8



```
1-Var Stats
 x̄=7.488
 Σx=374.4
 Σx²=2835.1
 Sx=.8029638973
 σx=.7948937036
↓n=50
```

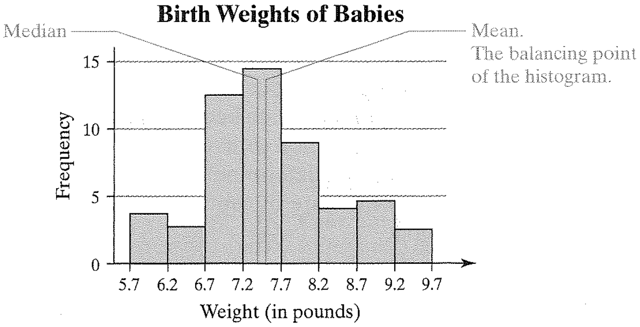
(a)

```
1-Var Stats
↑n=50
 minX=5.8
 Q1=7
 Med=7.35
 Q3=7.9
 maxX=9.4
```

(b)

(b) See Figure 9 for the frequency histogram with the mean and median labeled. The distribution is bell shaped. We have further evidence of the shape because the mean and median are close to each other.

Figure 9 Birth weights of 50 randomly selected babies



(c) Because the mean and median are close in value, we use the mean as the measure of central tendency.

Now Work Problem 27

Summary

We conclude this section with the following chart, which addresses the circumstances under which each measure of central tendency should be used.

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\Sigma x_i}{N}$ Sample mean: $\bar{x} = \frac{\Sigma x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and divide the data set in half	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or skewed right

Section ~~3.1~~ 6.1 Exercises

3.1 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. What does it mean if a statistic is resistant? Why is the median resistant, but the mean is not?

2. In the 2000 census conducted by the U.S. Census Bureau, two average household incomes were reported: $41,349 and $55,263. One of these averages is the mean and the other is the median. Which is the mean? Support your answer.

3. The U.S. Department of Housing and Urban Development (HUD) uses the median to report the average price of a home in the United States. Why do you think HUD uses the median?

4. A histogram of a set of data indicates that the distribution of the data is skewed right. Which measure of central tendency will likely be larger, the mean or the median? Why?

5. If a data set contains 10,000 values arranged in increasing order, where is the median located?

6. *True or False*: A data set will always have exactly one mode.

Skill Building

In Problems 7–10, find the population mean or sample mean as indicated.

7. Sample: 20, 13, 4, 8, 10

8. Sample: 83, 65, 91, 87, 84

9. Population: 3, 6, 10, 12, 14

10. Population: 1, 19, 25, 15, 12, 16, 28, 13, 6

11. For Super Bowl XL, CBS television sold 65 ad slots for a total revenue of roughly $162.5 million. What was the mean price per ad slot?

12. The median for the given set of six ordered data values is 26.5. What is the missing value? 7 12 21 ______ 41 50

13. **Crash Test Results** The Insurance Institute for Highway Safety crashed the 2007 Audi A4 four times at 5 miles per hour. The costs of repair for each of the four crashes were

$976, $2038, $918, $1899

Compute the mean, median, and mode cost of repair.

14. **Cell Phone Use** The following data represent the monthly cell phone bill for my wife's phone for six randomly selected months.

$35.34, $42.09, $39.43, $38.93, $43.39, $49.26

Compute the mean, median, and mode phone bill.

15. **Concrete Mix** A certain type of concrete mix is designed to withstand 3,000 pounds per square inch (psi) of pressure. The strength of concrete is measured by pouring the mix into casting cylinders 6 inches in diameter and 12 inches tall. The concrete is allowed to set for 28 days. The concrete's strength is then measured. The following data represent the strength of nine randomly selected casts (in psi).

3960, 4090, 3200, 3100, 2940, 3830, 4090, 4040, 3780

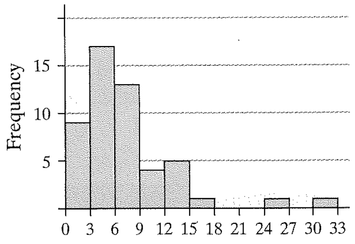
Compute the mean, median, and ~~mode~~ strength of the concrete (in psi).

16. **Flight Time** The following data represent the flight time (in minutes) of a random sample of seven flights from Las Vegas, Nevada, to Newark, New Jersey, on Continental Airlines.

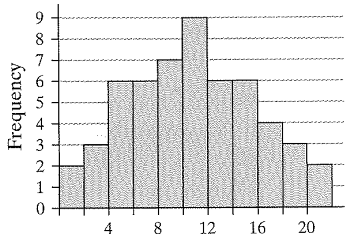
282, 270, 260, 266, 257, 260, 267

Compute the mean, median, and mode flight time.

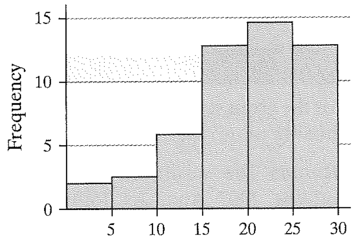
17. For each of the three histograms shown, determine whether the mean is greater than, less than, or approximately equal to the median. Justify your answer.



(a)



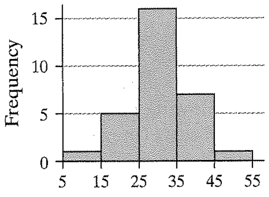
(b)



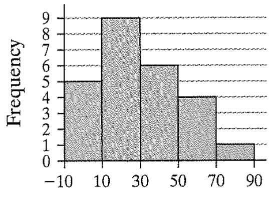
(c)

18. Match the histograms shown to the summary statistics:

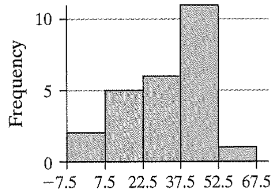
	Mean	Median
I	42	42
II	31	36
III	31	26
IV	31	32



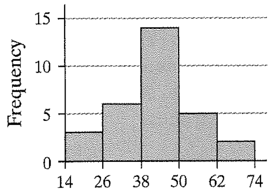
(a)



(b)



(c)



(d)

19. **Mean versus Median Applet** Load the mean versus median applet that is located on the CD that accompanies the text. Change the lower limit to 0 and the upper limit to 10 and click Update.

(a) Create a data set of at least eight observations such that the mean and median are roughly 2.

(b) Add a single observation near 9. How does this new value affect the mean? How does this new value affect the median?

(c) Change the upper limit to 25 and click Update. Remove the single value added from part (b). Add a single observation near 24. How does this new value affect the mean? the median?

(d) Refresh the page. Change the lower limit to 0 and the upper limit to 50 and click Update. Create a data set of at least eight observations such that the mean and median are roughly 40.

(e) Add a single observation near 35. How does this new value affect the mean? the median? Now "grab" this point with your mouse cursor and drag it toward 0. What happens to the value of the mean? What happens to the value of the median? Why?

20. **Mean versus Median Applet** Load the mean versus median applet that is located on the CD that accompanies the text. Change the lower limit to 0 and the upper limit to 10 and click Update.

(a) Create a data set of at least 10 observations such that the mean equals the median.

(b) Create a data set of at least 10 observations such that the mean is greater than the median.

(c) Create a data set of at least 10 observations such that the mean is less than the median.

(d) Comment on the shape of each distribution from parts (a)–(c).

(e) Can you create a distribution that is skewed left, but has a mean that is greater than the median?

Applying the Concepts

21. **pH in Water** The acidity or alkalinity of a solution is measured using pH. A pH less than 7 is acidic; a pH greater than 7 is alkaline. The following data represent the pH in samples of bottled water and tap water.

Tap	7.64	7.45	7.47	7.50	7.68	7.69
	7.45	7.10	7.56	7.47	7.52	7.47
Bottled	5.15	5.09	5.26	5.20	5.02	5.23
	5.28	5.26	5.13	5.26	5.21	5.24

Source: Emily McCarney, student at Joliet Junior College

~~(a)~~ Compute the mean, median, and mode pH for each type of water. Comment on the differences between the two water types.

(b) Suppose the pH of 7.10 in tap water was incorrectly recorded as 1.70. How does this affect the mean? the median? What property of the median does this illustrate?

22. **Reaction Time** In an experiment conducted online at the University of Mississippi, study participants are asked to react to a stimulus. In one experiment, the participant must

781	1,038	453	1,446	3,082
501	451	1,826	1,348	3,001
1,342	1,889	580	0	2,909
2,883	480	1,664	1,064	2,978
149	1,291	507	261	540
543	87	798	673	2,862
1,692	1,783	2,186	398	526
730	2,324	2,823	1,676	4,148

Source: Ashley Hudson, student at Joliet Junior College

Determine the shape of the distribution of new-car profit by drawing a frequency histogram. Compute the mean and median. Which measure of central tendency better describes the profit?

33. Political Views A sample of 30 registered voters was surveyed in which the respondents were asked, "Do you consider your political views to be conservative, moderate, or liberal?" The results of the survey are shown in the table.

Liberal	Conservative	Moderate
Moderate	Liberal	Moderate
Liberal	Moderate	Conservative
Moderate	Conservative	Moderate
Moderate	Moderate	Liberal
Liberal	Moderate	Liberal
Conservative	Moderate	Moderate
Liberal	Conservative	Liberal
Liberal	Conservative	Liberal
Conservative	Moderate	Conservative

Source: Based on data from the General Social Survey

(a) Determine the mode political view.

(b) Do you think it would be a good idea to rotate the choices conservative, moderate, or liberal in the question? Why?

34. Hospital Admissions The following data represent the diagnosis of a random sample of 20 patients admitted to a hospital.

Cancer	Motor vehicle accident	Congestive heart failure
Gunshot wound	Fall	Gunshot wound
Gunshot wound	Motor vehicle accident	Gunshot wound
Assault	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Gunshot wound	Motor vehicle accident
Fall	Gunshot wound	

Source: Tamela Ohm, student at Joliet Junior College

Determine the mode diagnosis.

35. Resistance and Sample Size Each of the following three data sets represents the IQ scores of a random sample of adults. IQ scores are known to have a mean and median of 100. For each data set, compute the mean and median. For each data set recalculate the mean and median, assuming that the individual whose IQ is 106 is accidentally recorded as 160. For each sample size, state what happens to the mean and the median? Comment on the role the number of observations plays in resistance.

Sample of Size 5				
106	92	98	103	100

Sample of Size 12					
106	92	98	103	100	102
98	124	83	70	108	121

Sample of Size 30					
106	92	98	103	100	102
98	124	83	70	108	121
102	87	121	107	97	114
140	93	130	72	81	90
103	97	89	98	88	103

36. Mr. Zuro finds the mean height of all 14 students in his statistics class to be 68.0 inches. Just as Mr. Zuro finishes explaining how to get the mean, Danielle walks in late. Danielle is 65 inches tall. What is the mean height of the 15 students in the class?

37. A researcher with the Department of Energy wants to determine the mean natural gas bill of households throughout the United States. He knows the mean natural gas bill of households for each state, so he adds together these 50 values and divides by 50 to arrive at his estimate. Is this a valid approach? Why or why not?

38. Net Worth According to the *Statistical Abstract of the United States*, the mean net worth of all households in the United States in 2004 was $448,200, while the median net worth was $93,100.

(a) Which measure do you believe better describes the typical U.S. household's net worth? Support your opinion.

(b) What shape would you expect the distribution of net worth to have? Why?

(c) What do you think causes the disparity in the two measures of central tendency?

39. You are negotiating a contract for the Players Association of the NBA. Which measure of central tendency will you use to support your claim that the average player's salary needs to be increased? Why? As the chief negotiator for the owners, which measure would you use to refute the claim made by the Players Association?

40. In January 2008, the mean amount of money lost per visitor to a local riverboat casino was $135. Do you think the median was more than, less than, or equal to this amount? Why?

41. Missing Exam Grade A professor has recorded exam grades for 20 students in his class, but one of the grades is no longer readable. If the mean score on the exam was 82 and the mean of the 19 readable scores is 84, what is the value of the unreadable score?

42. For each of the following situations, determine which measure of central tendency is most appropriate and justify your reasoning.

(a) Average price of a home sold in Pittsburgh, Pennsylvania in 2009

(b) Most popular major for students enrolled in a statistics course

(c) Average test score when the scores are distributed symmetrically

(d) Average test score when the scores are skewed right

(e) Average income of a player in the National Football League

(f) Most requested song at a radio station

43. Linear Transformations Benjamin owns a small Internet business. Besides himself, he employs nine other people. The salaries earned by the employees are given next in thousands of dollars (Benjamin's salary is the largest, of course):

30, 30, 45, 50, 50, 50, 55, 55, 60, 75

(a) Determine the mean, median, and ~~mode~~ for salary.

(b) Business has been good! As a result, Benjamin has a total of $25,000 in bonus pay to distribute to his employees. One option for distributing bonuses is to give each employee (including himself) $2,500. Add the bonuses under this plan to the original salaries to create a new data set. Recalculate the mean, median, and ~~mode~~. How do they compare to the originals?

(c) As a second option, Benjamin can give each employee a bonus of 5% of his or her original salary. Add the bonuses under this second plan to the original salaries to create a new data set. Recalculate the mean, median, and ~~mode~~. How do they compare to the originals?

(d) As a third option, Benjamin decides not to give his employees a bonus at all. Instead, he keeps the $25,000 for himself. Use this plan to create a new data set. Recalculate the mean, median, and ~~mode~~. How do they compare to the originals?

44. Linear Transformations Use the five test scores of 65, 70, 71, 75, and 95 to answer the following questions:

(a) Find the sample mean.

(b) Find the median.

(c) Which measure of central tendency best describes the typical test score?

(d) Suppose the professor decides to curve the exam by adding 4 points to each test score. Compute the sample mean based on the adjusted scores.

(e) Compare the unadjusted test score mean with the curved test score mean. What effect did adding 4 to each score have on the mean?

45. Trimmed Mean Another measure of central tendency is the trimmed mean. It is computed by determining the mean of a data set after deleting the smallest and largest observed values. Compute the trimmed mean for the data in Problem 29. Is the trimmed mean resistant? Explain.

46. Midrange The midrange is also a measure of central tendency. It is computed by adding the smallest and largest observed values of a data set and dividing the result by 2; that is,

$$\text{Midrange} = \frac{\text{largest data value} + \text{smallest data value}}{2}$$

Compute the midrange for the data in Problem 29. Is the midrange resistant? Explain.

47. Putting It Together: Shape, Mean and Median As part of a semester project in a statistics course, Carlos surveyed a sample of 40 high school students and asked, "How many days in the past week have you consumed an alcoholic beverage?" The results of the survey are shown next.

0	0	1	4	1	1	1	5	1	3
0	1	0	1	0	4	0	1	0	1
0	0	0	0	2	0	0	0	0	0
1	0	2	0	0	0	1	2	1	1
2	0	1	0	1	3	1	1	0	3

(a) Is this data discrete or continuous?

(b) Draw a histogram of the data and describe its shape.

(c) Based on the shape of the histogram, do you expect the mean to be more than, equal to, or less than the median?

(d) Compute the mean and the median. What does this tell you?

(e) Determine the mode.

(f) Do you believe that Carlos' survey suffers from sampling bias? Why?

Section ~~3.1~~ 6.1 Answers

CHAPTER 3 Numerically Summarizing Data

3.1 Assess Your Understanding (page 137)

1. A statistic is resistant if it is not sensitive to extreme values. The median is resistant because it is a positional measure of central tendency, and increasing the largest value or decreasing the smallest value does not affect the position of the center. The mean is not resistant because it is a function of the sum of the values of data. Changing the magnitude of one value changes the sum of the values.

3. HUD uses the median because the data are skewed. Explanations will vary.

5. The median is between the 5000th and the 5001st ordered values.

7. $\bar{x} = 11$

9. $\mu = 9$

11. Mean price per ad slot was \$2.5 million.

13. Mean cost is \$1,457.75; median cost is \$1,437.50; there is no mode cost.

15. The mean, median, and mode strengths are 3,670, 3,830, and 4,090 pounds per square inch, respectively.

17. **(a)** mean > median
(b) mean = median
(c) mean < median
Justification will vary.

21. **(a)** Tap: $\bar{x} = 7.50$; $M = 7.485$; mode = 7.47. Bottled: $\bar{x} = 5.194$; $M = 5.22$; mode = 5.26
(b) $\bar{x} = 7.05$; $M = 7.485$; the median is ~~resistant~~. does not depend on outliers

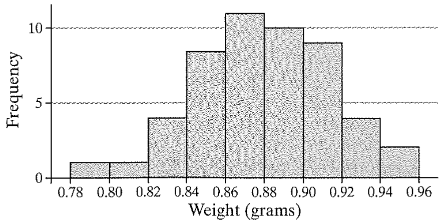
23. **(a)** The mean pulse rate is 72.2 beats per minute.
(b) Samples and sample means will vary.
(c) Answers will vary.

25. **(a)** 1,813,654.5 thousand metric tons
(b) Per capita is better because it adjusts for CO_2 emissions for population. After all, countries with more people will, in general, have higher CO_2 emissions.
(c) Mean = 2.814 thousand metric tons; median = 2.67 thousand metric tons; mean

27. The distribution is symmetric. The mean is the better measure of central tendency.

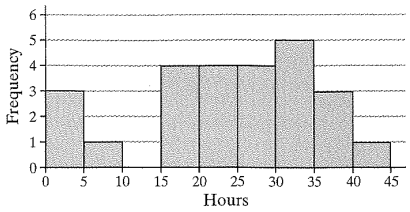
29. $\bar{x} = 0.875$ gram; $M = 0.875$ gram. The distribution is symmetric, so the mean is the better measure of central tendency.

Weight of Plain M&Ms



31. The distribution is skewed left; $\bar{x} = 22$ hours; $M = 25$ hours. The median is the better measure of central tendency.

Hours Worked per Week



33. **(a)** Moderate **(b)** Yes, to avoid response bias

35. Sample of size 5: All data recorded correctly: $\bar{x} = 99.8$; $M = 100$; 106 recorded as 160: $\bar{x} = 110.6$; $M = 100$

Sample of size 12: All data recorded correctly: $\bar{x} = 100.4$; $M = 101$; 106 recorded as 160: $\bar{x} = 104.9$; $M = 101$

Sample of size 30: All data recorded correctly: $\bar{x} = 100.6$; $M = 99$; 106 recorded as 160: $\bar{x} = 102.4$; $M = 99$

For each sample size, the mean becomes larger, but the median remains constant. As the sample size increases, the affect of the misrecorded data on the mean decreases.

37. No. Each state has a different population size. This must be taken into account.

39. The salary distribution is skewed right, so the players' negotiator would want to use the median salary; the owners' negotiator would use the mean salary to refute the players' claim.

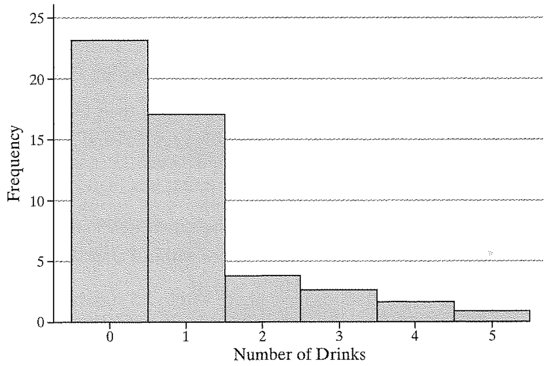
41. The unreadable score is 44.

43. **(a)** Mean = \$50,000; median = \$50,000; mode = \$50,000
(b) New data set: 32.5, 32.5, 47.5, 52.5, 52.5, 52.5, 57.5, 57.5, 62.5, 77.5; mean = \$52,500; median = \$52,500; mode = \$52,500. All three measures increased by \$2,500.
(c) New data set: 31.5, 31.5, 47.25, 52.5, 52.5, 52.5, 57.75, 57.75, 63, 78.75; mean = \$52,500; median = \$52,500; mode = \$52,500. All three measures increased by 5%.
(d) New data set: 30, 30, 45, 50, 50, 50, 55, 55, 60, 100; mean = \$52,500; median = \$50,000; mode = \$50,000. The mean increased by \$2,500, but the median and the mode remained at \$50,000.

45. The trimmed mean is 0.875. Explanations will vary.

47. **(a)** Discrete
(b)

Number of Days High School Students Consume Alcohol Each Week



(c) Since the data are skewed right, we would expect the mean to be greater than the median.
(d) Mean: 0.94; median: 1; the mean can be less than the median in skewed-right data. Therefore, using the rule *mean greater than median implies the data are skewed right* does not always work.
(e) 0
(f) Yes. It is difficult to get truthful responses to this type of question. Carlos would need to ensure that the identity of the respondents is anonymous.

Reading for Section 6.2: The following reading is excerpted from:

Wrean. Online Textbook for MATH 163. Camosun College, 2015.

Section 6.2:

Measures of variability

Measures of variability (or spread) are statistical quantities that describe the width of the distribution. The two we will be looking at are the range and the standard deviation.

Range

The range of a set of data points is the difference between the highest and lowest value.

Example

Calculate the range of the following data set: 2, 6, 7, 3, 3.

Answer: range = highest – lowest = 7 – 2 = 5.

The range is very easy to calculate and visualize, but can be problematic: if you have an outlying point, it can distort the range quantity and make it not as meaningful as you might want. That's why we use another quantity which has a more slightly complicated calculation, the standard deviation.

Standard Deviation

Let's consider our data set 2, 6, 7, 3, 3, which has a mean of 4.2. One way we could determine how "wide" the distribution is would be to calculate how far each data point is from the centre-line (the mean). So we could calculate $\left(x_i - \bar{x}\right)$ as a measure of how far each data point is from the centre. For example, when x_i=2, then $\left(x_i - \bar{x}\right) = 2 - 4.2 = -2.2$. If we were to add up all of these quantities, we should get zero, because all of the points on the left side of the distribution should cancel out the points on the right side of the distribution. But if we were to square this quantity to get all positive values, then the sum would be a direct measure of the distance of each point from the centre line (since distances are always positive). The sum of these squares divided by the number of data points gives a value called the **variance**. However, since the units of the variance are the squares of the units of the original measurements, we prefer to take the square root of this quantity, which is called the standard deviation.

For a sample of measurements, the standard deviation *s* is equal to

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

where $\bar{x}$ is the mean of the sample and n is the number of measurements.

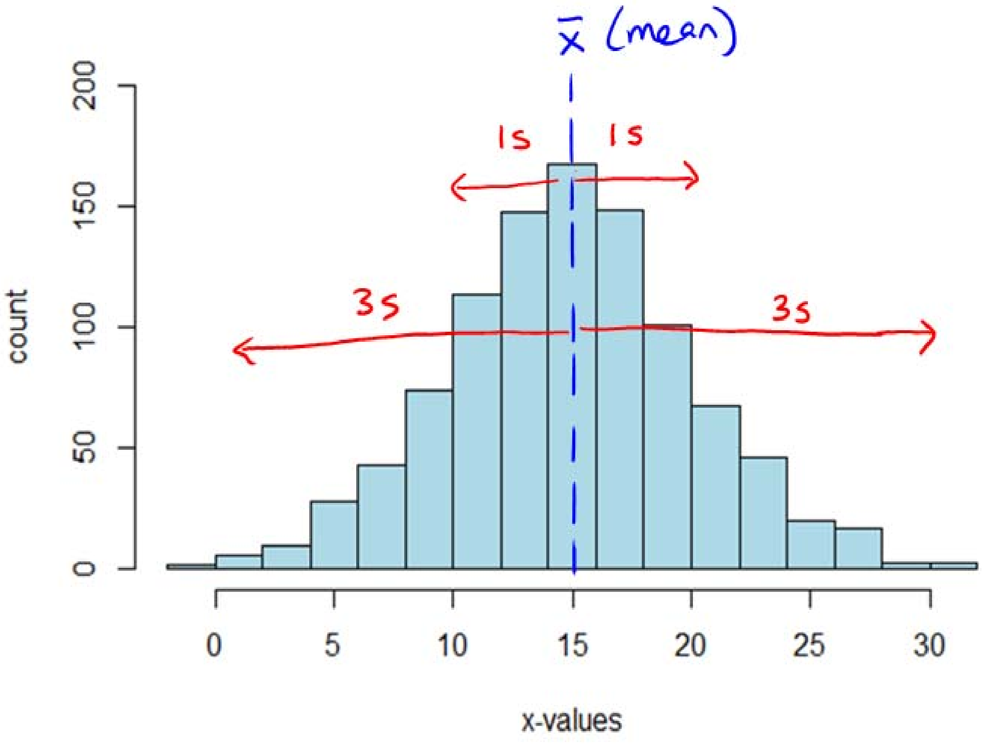
For measurements on an entire population, the standard deviation is given the symbol σ and is given by

$$\sigma = \sqrt{\frac{\sum\left(x_i - \mu\right)^2}{n}}$$

where μ is the mean of the population and n is the number of measurements.

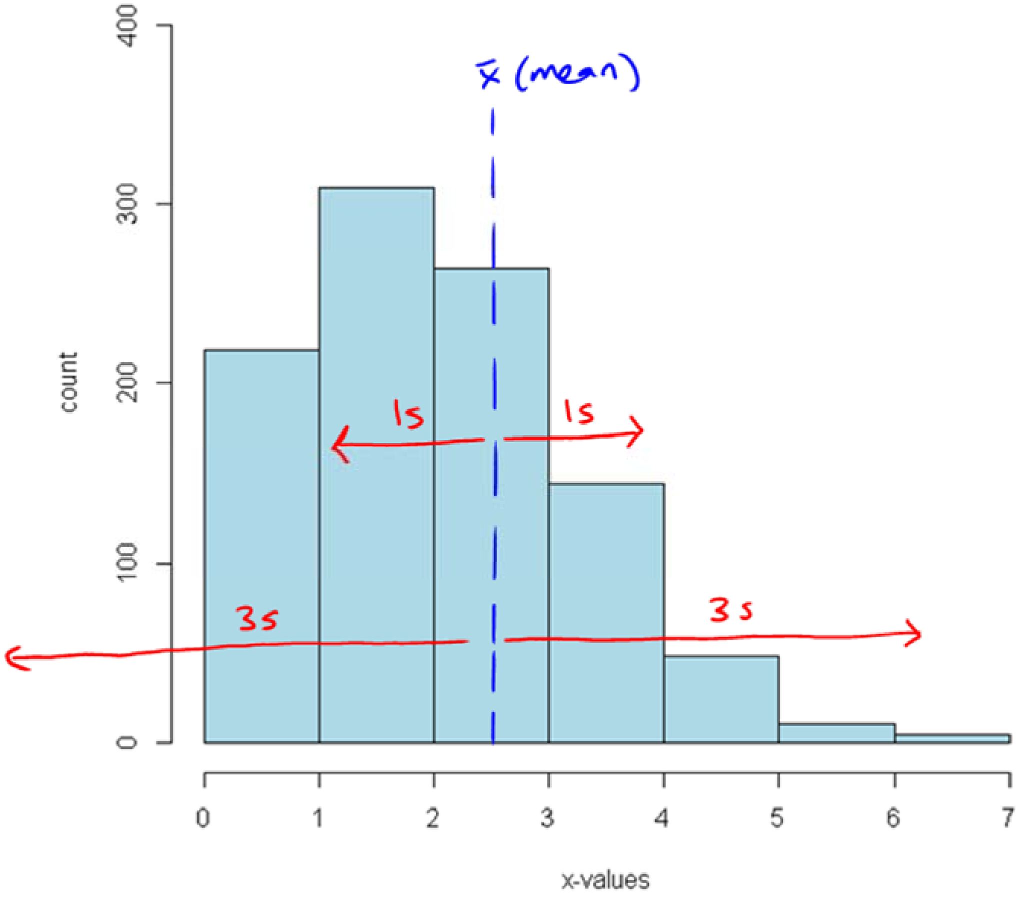
Let's see what this looks like with some data. Suppose we have a population whose data is symmetrically distributed with only one peak. (Distributions with only one peak are said to be unimodal.) This data set has a mean $\bar{x}$ of 15, a median of 15, and a standard deviation s of 5. You can see that the bulk of the measurements lie within one standard deviation of the mean, and almost all of them lie within three standard deviations from the mean.

Figure 1: A Symmetrical, Unimodal Distribution



Compare the previous histogram with the next one, which has an asymmetrical distribution with mean $\bar{x}$ of 2.5 and standard deviation s of 1.4. Even though this one has considerable asymmetry (the median is 2), the bulk of the data still lies within one standard deviation of the mean, and virtually all of the data lies within three standard deviations of the mean.

Figure 2: An Asymmetrical Distribution



Section 6.2 Exercises

Section 6.4: Measures of Centre and of Variability

Exercises

1. The top ten movies (Skyfall, The Hobbit, etc.) and their profits (in millions of dollars) from last weekend are reported in Monday Magazine. Calculate the mean and median for this data.

   profits: 1.4, 4.1, 1.2, 1.3, 5.8, 5.0, 2.6, 1.8, 2.9, 5.9, 2.5, 5.3

2. Calculate the mean and median for the data set: 35, 47, 29, 42, 38, 39, 42.

3. Pat finds the mean height of all twelve students in her physics class to be 68.0 inches. Just as she's finished that calculation, one more student walks in late. If that student is 63.0 inches tall, what is the mean height of all thirteen students?

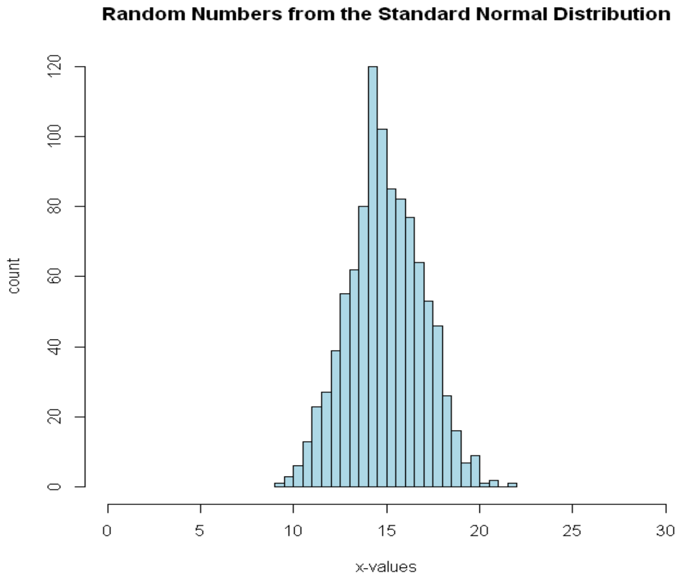
4. The Victoria Real Estate Board claims that in October of 2012, the average cost of a single-family home in Greater Victoria was $592,000, while the median was $527,000. Why is the mean greater than the median for housing prices? Explain.

5. Tom is running a small business with five employees, including himself. The salaries of the five people (in thousands of dollars) are 30, 45, 50, 55, and 75, with Tom making the highest salary.

   a) calculate the mean and median of these salaries
   b) if Tom gives everyone a $2000 bonus, what happens to the mean and median?
   c) if Tom gives everyone a 5% raise, what happens to the mean and median?
   d) if Tom decides to keep everyone else's salary the same, but raise his own salary by $10,000, what happens to the mean and the median?

~~6.~~ Consider the following histogram. Is the standard deviation equal to

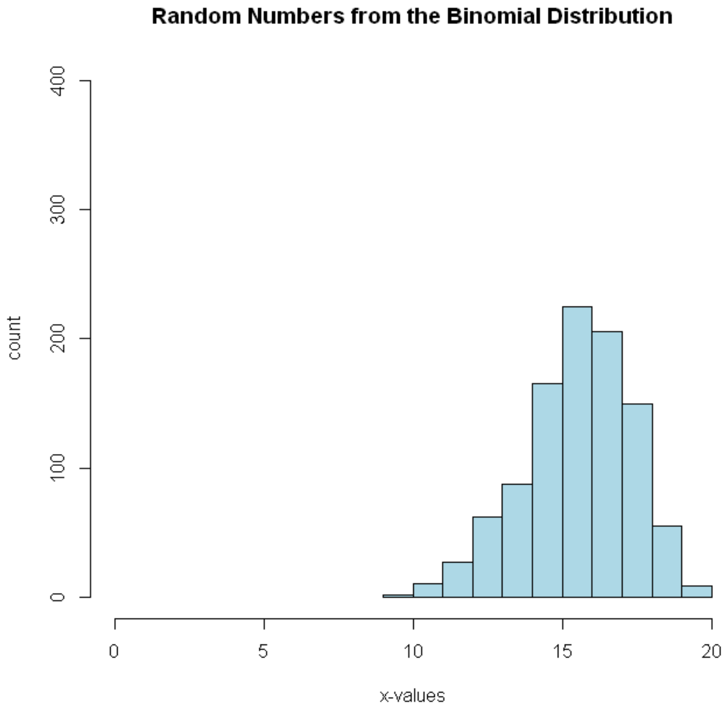
a) 0.5
b) 2
c) 15
d) 20



7. Consider the following data set: 7, 7, 7, 7, 7, and 7. What is the mean and the median? What is the range? Without calculating it, what would be the standard deviation?

~~8.~~ Consider the following histogram. Is the standard deviation equal to

a) 1
b) 2
c) 5
d) 10
e) 15
f) 20



9. Pat, when entering quiz scores into her spreadsheet, accidentally put an extra zero on the end of one student's score (making it 380/40 instead of 38/40), and then calculated the mean, median, range, and standard deviation for the section. She then noticed her mistake and recalculated all of the quantities. For the following quantities, state whether the corrected value will be higher, lower, or the same as the value calculated with the incorrect quiz score:

a) mean
b) median
c) range
d) standard deviation

10. Consider the following sets of data. Without calculating any values, state which set will have the higher standard deviation (or will they be the same?).

a) Set 1: 2, 3, 9, 16, 17 Set 2: 2, 8, 9, 10, 17
b) Set 1: 2, 3, 9, 16, 17 Set 2: 3, 4, 10, 17, 18

Sections 6.2 Answers:

Section 6~~.4~~: Statistical Quantities

Solutions

1. The mean is 3.31667, or just 3.3. There are twelve points, so the median is the 12/2+1/2=6.5th point, which means the average of the 6th and 7th points. Therefore, the median is (2.6+2.9)/2 = 2.75.

2. The mean is 38.8571. (You can round to 38.9 if you like.)

3. To find the mean, we want the sum of all of the heights divided by the total number of students. Since the average of the twelve students is 68.0 inches, the total of all of those heights is just 68.0 times 12, which is 816.0 inches. Adding the height of the thirteenth student brings the total to 879.0 inches, then dividing by 13 gives a mean of 67.6 inches.

4. The histogram of Victoria housing prices will not be symmetrical: there is a lower limit for the price of single-family homes, while there can be house prices in the millions of dollars. Just a few very expensive homes will bring up the mean but not affect the median in any way, which is why the mean is greater than the median.

5. The means and medians are:

    a) mean = $51,000 and median = $50,000
    b) the mean and median will each increase by $2000: mean is now $53,000 and the median $52,000
    c) the mean and median will both increase by a factor of 1.05 (they are multiplied by 1.05): mean is now $53,550 and median is $52,500
    d) the mean will become $53,000 but the median will stay the same

~~6.~~ Looking at the histogram, you can estimate the standard deviation by picking a "width" about the mean/average that most of the data points fall within. From this histogram, the standard deviation is about half of 5, since most of the data falls between approximately 12.5 and 17.5 (ish). And the closest value given that matches that is (b) 2.

7. The mean and median are both 7. The range is 0. The standard deviation is also 0, since all points lie exactly on the mean and $(x-\overline{x})$ is zero for each point.

~~8.~~ Using the same reasoning as for question 6, most of the data seems to fall between 12.5 and 17.5, so the standard deviation is around 2.5 (ish). So the closest option given is (b) again.

9. New values:

    a) The corrected mean will be lower, since one value was lowered.

b) The median will remain unchanged (assuming that the 38/40 was in the upper half of the scores to begin with, so changing it to 380 and back won't affect that)

c) The corrected range will be lower, since the highest point has changed.

d) The standard deviation will be lower, since the corrected point's distance from the mean is lower than the uncorrected value.

10. a) Set 1's values are farther from the mean on average than Set 2's data points. So Set 1 will have a higher standard deviation.

b) Set 2's data points are just Set 1's points moved up by 1 unit. So each point's distance from the mean will be the same as Set 1, and the standard deviations will be the same also.

Reading for Sections 6.3 and 6.4:

The following readings are excerpted from:

> Bluman and Mayer. Elementary Statistics: A step by step approach. 2nd Canadian edition, Mc-Graw Hill Ryerson, 2011, pages 110-113, 116-119, 128-130, 725.

Section 6.3:

For many data sets, almost all data values will fall within 2 standard deviations of the mean. Better approximations can be obtained by using Chebyshev's theorem and the empirical rule. These are explained next.

Chebyshev's Theorem

As stated previously, the variance and standard deviation of a variable can be used to determine the spread, or dispersion, of a variable. That is, the larger the variance or standard deviation, the more the data values are dispersed. For example, if two variables measured in the same units have the same mean, say, 70, and variable 1 has a standard deviation of 1.5 while variable 2 has a standard deviation of 10, then the data for variable 2 will be more spread out than the data for variable 1. *Chebyshev's theorem,* developed by the Russian mathematician Pafnuty L. Chebyshev (1821–1894), specifies the proportions of the spread in terms of the standard deviation.

note: this text uses $k > 1$, and that what we'll use for this class. other sources may use $k \geq 1$. See your instructor if you want to know it means when $k = 1$.

Chebyshev's theorem The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $1 - 1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

This theorem states that at least three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean of the data set. This result is found by substituting $k = 2$ in the expression.

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

For the example in which variable 1 has a mean of 70 and a standard deviation of 1.5, at least three-fourths, or 75%, of the data values fall between 67 and 73. These values are found by adding 2 standard deviations to the mean and subtracting 2 standard deviations from the mean, as shown:

$$70 + 2(1.5) = 70 + 3 = 73$$

and

$$70 - 2(1.5) = 70 - 3 = 67$$

For variable 2, at least three-fourths, or 75%, of the data values fall between 50 and 90. Again, these values are found by adding and subtracting, respectively, 2 standard deviations to and from the mean.

$$70 + 2(10) = 70 + 20 = 90$$

and

$$70 - 2(10) = 70 - 20 = 50$$

Unusual Stat

At birth men outnumber women by 2%. By age 25, the number of men living is about equal to the number of women living. By age 65, there are 14% more women living than men.

Furthermore, the theorem states that at least eight-ninths, or 88.89%, of the data values will fall within 3 standard deviations of the mean. This result is found by letting $k = 3$ and substituting in the expression.

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 88.89\% \text{ (rounded)}$$

For variable 1, at least eight-ninths, or approximately 88.89%, of the data values fall between 65.5 and 74.5, since

$$70 + 3(1.5) = 70 + 4.5 = 74.5$$

and

$$70 - 3(1.5) = 70 - 4.5 = 65.5$$

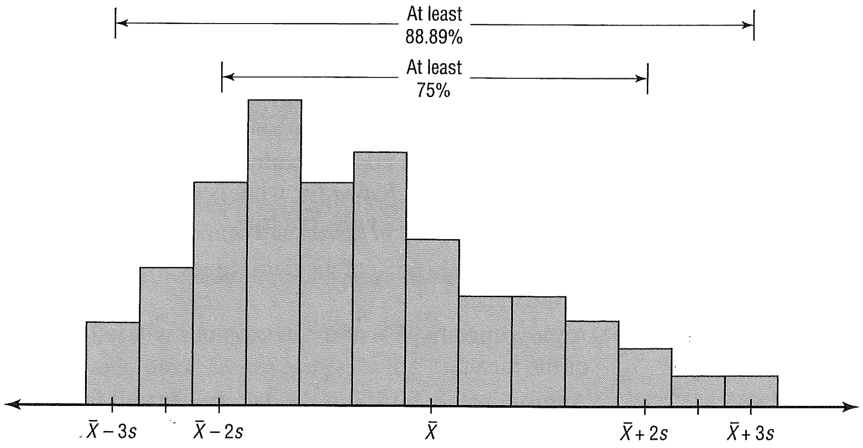
For variable 2, at least eight-ninths, or approximately 88.89%, of the data values fall between 40 and 100.

This theorem can be applied to any distribution regardless of its shape (see Figure 3–3).

Examples 3–24 and 3–25 illustrate the application of Chebyshev's theorem.

Figure 3–3

Chebyshev's Theorem



Example 3–24

House Prices

The mean price of houses in a certain neighbourhood is \$360,000, and the standard deviation is \$45,000. Find the price range for which at least 75% of the houses will sell.

Solution

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$360{,}000 + 2(\$45{,}000) = \$360{,}000 + \$90{,}000 = \$450{,}000$$

and

$$\$360{,}000 - 2(\$45{,}000) = \$360{,}000 - \$90{,}000 = \$270{,}000$$

Therefore, at least 75% of all homes sold in the area will have a price range from \$270,000 and \$450,000.

Chebyshev's theorem can be used to find the minimum percentage of data values that will fall between any two given values. The procedure is shown in Example 3–25.

Example 3–25

Travel Allowances

A survey of local companies found that the mean amount of travel allowance for executives was \$0.25 per kilometre. The standard deviation was \$0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

Solution

Step 1 Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

Step 2 Divide the difference by the standard deviation to get *k*.

$$k = \frac{0.05}{0.02} = 2.5$$

Step 3 Use Chebyshev's theorem to find the percentage.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \qquad \text{or} \qquad 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

The Empirical (Normal) Rule

Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule,** are true.

Approximately 68% of the data values will fall within 1 standard deviation of the mean.

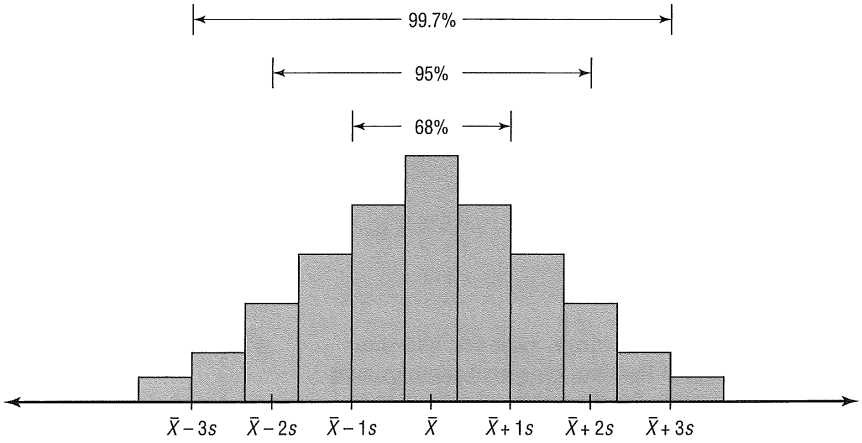
Approximately 95% of the data values will fall within 2 standard deviations of the mean.

Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 (480 + 90 = 570 and 480 − 90 = 390). Approximately 95% of the scores will fall between 300 and 660 (480 + 2 · 90 = 660 and 480 − 2 · 90 = 300). Approximately 99.7% will fall between 210 and 750 (480 + 3 · 90 = 750 and 480 − 3 · 90 = 210). See Figure 3–4. (The empirical rule is explained in greater detail in Chapter 7.)

Figure 3–4

The Empirical Rule



3–2 Applying the Concepts

Blood Pressure

The table lists means and standard deviations. The mean is the number before the plus/minus, and the standard deviation is the number after the plus/minus. The results are from a study attempting to find the average blood pressure of older adults. Use the results to answer the questions.

	Normotensive		**Hypertensive**	
	Men **(n = 1200)**	**Women** **(n = 1400)**	**Men** **(n = 1100)**	**Women** **(n = 1300)**
Age	55 ± 10	55 ± 10	60 ± 10	64 ± 10
Blood pressure (mm Hg)				
Systolic	123 ± 9	121 ± 11	153 ± 17	156 ± 20
Diastolic	78 ± 7	76 ± 7	91 ± 10	88 ± 10

1. Apply Chebyshev's theorem to the systolic blood pressure of normotensive men. At least how many of the men in the study fall within 1 standard deviation of the mean?
2. At least how many of those men in the study fall within 2 standard deviations of the mean?

Assume that blood pressure is normally distributed among older adults. Answer the following questions, using the empirical rule instead of Chebyshev's theorem.

3. Give ranges for the diastolic blood pressure (normotensive and hypertensive) of older women.
4. Do the normotensive, male, systolic blood pressure ranges overlap with the hypertensive, male, systolic, blood pressure ranges?

See page 145 for the answers.

Section 6.3: Exercises

26. Dog Reaction Times Find the variance and standard deviation for the two distributions in Exercise 8 in Section 2–2 and Exercise 18 in Section 2–2. Compare the variation of the data sets. Decide if one data set is more variable than the other.

27. Photocopier Service Calls This frequency distribution represents the data obtained from a sample of photocopier service technicians. The values are the days between service calls on 80 photocopy machines.

Days between calls	Frequency
25.5–28.5	5
28.5–31.5	9
31.5–34.5	32
34.5–37.5	20
37.5–40.5	12
40.5–43.5	2

28. Exam Scores The average score of the students in one calculus class is 110, with a standard deviation of 5; the average score of students in a statistics class is 106, with a standard deviation of 4. Which class is more variable in terms of scores?

29. Suspension Bridges The data show the lengths (in metres) of suspension bridges in the eastern part of North America and western part of North America. Compare the variability of the two samples.

East:	1298, 1067, 375, 655, 610, 533
West:	1250, 853, 704, 472, 457, 368

Source: World Almanac and Book of Facts.

30. Exam Scores The average score on an English final examination was 85, with a standard deviation of 5; the average score on a history final exam was 110, with a standard deviation of 8. Which class was more variable?

31. Accountants' Ages The average age of the accountants at Three Rivers Corp. is 26 years, with a standard deviation of 6 years; the average salary of the accountants is $31,000, with a standard deviation of $4000. Compare the variations of age and income.

32. Using Chebyshev's theorem, solve these problems for a distribution with a mean of 80 and a standard deviation of 10.

a. At least what percentage of values will fall between 60 and 100?

b. At least what percentage of values will fall between 65 and 95?

33. The mean of a distribution is 20 and the standard deviation is 2. Use Chebyshev's theorem.

a. At least what percentage of the values will fall between 10 and 30?

b. At least what percentage of the values will fall between 12 and 28?

34. In a distribution of 200 values, the mean is 50 and the standard deviation is 5. Use Chebyshev's theorem.

a. At least how many values will fall between 30 and 70?

b. At most how many values will be less than 40 or more than 60?

35. Fast-Food Industry Wages A sample of hourly wages of employees in the fast-food industry has a mean of $8.26 and a standard deviation of $0.33. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall.

36. Time Spent Online Adult Canadians spend an average of 2.7 hours per day online. Assuming a standard deviation of 30 minutes, find the range in which at least 88.89% of adult Canadian users spend online. Use Chebyshev's theorem.

Source: Ipsos, *News and Polls,* "Canadian Teenagers Are Leading the Online Revolution? Maybe Not...," February 27, 2008.

37. Cereal Potassium per Serving A survey of a number of the leading brands of cereal shows that the mean content of potassium per serving is 95 milligrams, and the standard deviation is 2 milligrams. Find the range in which at least 88.89% of the data will fall. Use Chebyshev's theorem.

38. Solid Waste Production The average college student produces 290 kilograms of solid waste each year, including 500 disposable cups and 140 kilograms of paper. If the standard deviation is approximately 36 kilograms, within what weight limits will at least 75% of all students' garbage lie?

Source: Environmental Sustainability Committee, www.esc.mtu.edu.

39. Trials to Learn a Maze The average of the number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of 4 to 20 trials.

40. Farm Size The average farm in Canada in 2005 contained 295 hectares. Assume a standard deviation of 16 hectares. Use Chebyshev's theorem to find the minimum percentage of farms that fell in the range of 255 to 335 hectares.

Source: Agriculture and Agri-Food Canada, *Special Features: Census of Agriculture Summary.*

41. Fresh Food Consumption The average yearly per capita consumption of fresh fruit in Canada is 68.8 kilograms. Suppose that the distribution of fruit amounts consumed is bell-shaped with a standard deviation equal to 3.1 kilograms. What percentage of Canadians would you expect to consume more than 75 kilograms of fresh fruit per year?

Source: Statistics Canada.

42. Faculty Work Hours The average full-time faculty member in a post-secondary degree-granting institution works an average of 53 hours per week.

a. If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week?

b. If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week?

Source: National Center for Education Statistics.

Extending the Concepts »

43. Serum Cholesterol Levels For this data set, find the mean and standard deviation of the variable. The data represent the serum cholesterol levels of 30 individuals. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

211	240	255	219	204
200	212	193	187	205
256	203	210	221	249
231	212	236	204	187
201	247	206	187	200
237	227	221	192	196

44. Ages of Consumers For this data set, find the mean and standard deviation of the variable. The data represent the ages of 30 customers who ordered a product advertised on television. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

42	44	62	35	20
30	56	20	23	41
55	22	31	27	66
21	18	24	42	25
32	50	31	26	36
39	40	18	36	22

45. Using Chebyshev's theorem, complete the table to find the minimum percentage of data values that fall within k standard deviations of the mean.

k	1.5	2	2.5	3	3.5
Percentage					

46. Use this data set: 10, 20, 30, 40, 50.

a. Find the standard deviation.

b. Add 5 to each value, and then find the standard deviation.

c. Subtract 5 from each value and find the standard deviation.

d. Multiply each value by 5 and find the standard deviation.

e. Divide each value by 5 and find the standard deviation.

f. Generalize the results of parts *b* through *e*.

g. Compare these results with those in Exercise 38.

47. The mean deviation is found by using this formula:

$$\text{Mean deviation} = \frac{\Sigma|X - \overline{X}|}{n}$$

where

X = value
$\overline{X}$ = mean
n = number of values
$||$ = absolute value

Find the mean deviation for these data.

5, 9, 10, 11, 11, 12, 15, 18, 20, 22

48. A measure to determine the skewness of a distribution is called the *Pearson coefficient of skewness.* The formula is

$$\text{Skewness} = \frac{3(\overline{X} - \text{MD})}{s}$$

The values of the coefficient usually range from -3 to $+3$. When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed, it is positive; and when the distribution is negatively skewed, it is negative.

Using the formula, find the coefficient of skewness for each distribution, and describe the shape of the distribution.

a. Mean = 10, median = 8, standard deviation = 3.

b. Mean = 42, median = 45, standard deviation = 4.

c. Mean = 18.6, median = 18.6, standard deviation = 1.5.

d. Mean = 98, median = 97.6, standard deviation = 4.

49. All values of a data set must be within $s\sqrt{n - 1}$ of the mean. If a person collected 25 data values that had a mean of 50 and a standard deviation of 3 and you saw that one data value was 67, what would you conclude?

Section ~~2.3~~ 6.3: Answers

17. R = 11,263; 7,436,475.0; 2727.0

19. 133.6; 11.6 **21.** 45.93; 6.78

23. 211.2; 14.5 **25.** 211.2; 14.5;

No, the variability of the lifetimes of the batteries is quite large.

27. 11.7; 3.4

29. For West, CVar = 46.4%. For East, CVar = 48.3%. The data for East are more variable.

31. 23.1%; 12.9%.
The age is more variable.

33. *a.* 96% *b.* 93.75%

35. \$7.60–\$8.92 **37.** 89–101

39. 86% **41.** 2.5%

43. $n = 30$ $\overline{X} = 214.97$ $s = 20.76$. At least 75% of the data values will fall between $\overline{X} \pm 2s$.
$\overline{X} - 2(20.76) = 214.97 - 41.52 = 173.45$ and
$\overline{X} + 2(20.76) = 214.97 + 41.52 = 256.49$

In this case all 30 values fall within this range; hence Chebyshev's theorem is correct for this example.

45. 56%; 75%; 84%; 88.89%; 92%

47. 4.36

49. It must be an incorrect data value, since it is beyond the range using the formula $s\sqrt{n-1}$.

Section 6.4:

3–3 Measures of Position

LO3

Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it. This section discusses these measures of position.

Standard Scores

There is an old saying, "You can't compare apples and oranges." But with the use of statistics, it can be done to some extent. Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of both scores' positions relative to their respective evaluations' average scores and dispersion can be made. This comparison uses the mean and standard deviation and is called a *standard score* or *z score*. (We also use *z* scores in later chapters.)

A ***z* score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{Value} - \text{Mean}}{\text{Standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \overline{X}}{s}$$

For populations, the formula is

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

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

For the purpose of this book, it will be assumed that when we find *z* scores, the data were obtained from samples.

Example 3–26

Test Scores

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

Solution

First, find the *z* scores. For calculus the *z* score is

$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the *z* score is

$$z = \frac{30 - 25}{5} = 1.0$$

Since the *z* score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

Note that if the *z* score is positive, the score is above the mean. If the *z* score is 0, the score is the same as the mean. And if the *z* score is negative, the score is below the mean.

Example 3–27

Test Scores

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\overline{X} = 40$	$s = 5$
Test B	$X = 94$	$\overline{X} = 100$	$s = 10$

Solution

For test A,

$$z = \frac{X - \overline{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.

When all data for a variable are transformed into z scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A z score, then, is actually the number of standard deviations each value is from the mean for a specific distribution. In Example 3–26, the calculus score of 65 was actually 1.5 standard deviations above the mean of 50. This will be explained in greater detail in Chapter 7.

Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

Percentiles divide a data set into 100 equal parts in which the pth percentile is a value that at most p% of the observations in the data set are less than this value and the remainder are greater.

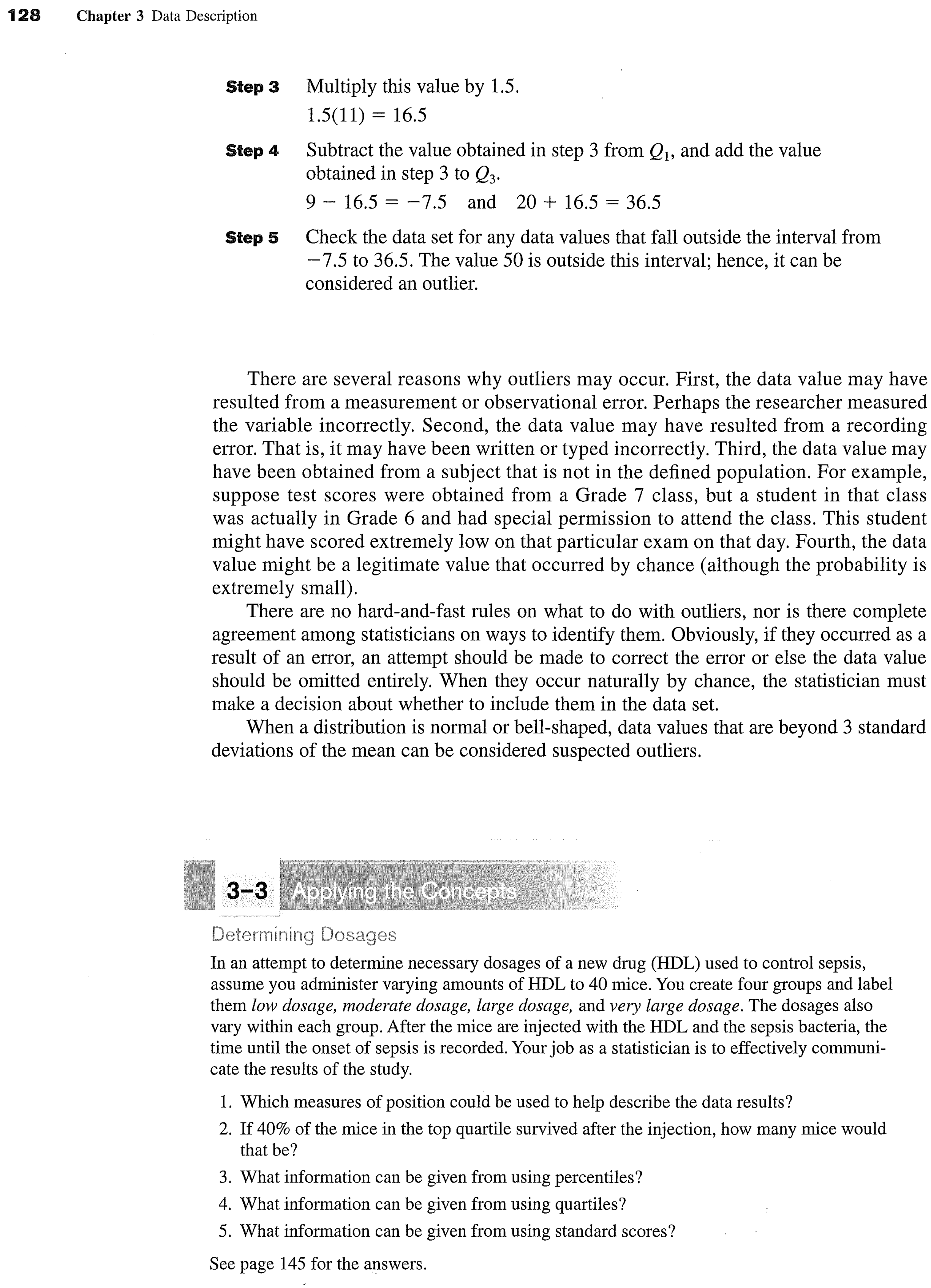
In many situations, the graphs and tables showing the percentiles for various measures such as test scores, heights, weights, and body mass index have already been completed. Table 3–3 shows the percentile ranks for scaled scores on the *Test of English as a Foreign Language* (TOEFL). For example, if a student had a scaled score of 60 for Section 1 (listening comprehension), that student would have a percentile rank of 79. Hence, that student did better than 79% of the students who took section 1 of the exam.

Figure 3–5 shows percentiles in graphical form of the body mass index (BMI) of girls from ages 5 to 19. The BMI is the ratio of a person's weight in kilograms (kg) to the square of their height in metres (m), or kg/m^2. To find the percentile rank of an 18-year-old girl whose BMI score is 25, start at the 25 BMI score on the left axis and move horizontally to the right. Find 18 on the horizontal axis and move up vertically. The two lines meet at the 85th percentile curved line; hence, an 18-year-old girl with a 25 BMI score is in the 85th percentile for her age group. If the lines do not meet exactly on one of the curved percentile lines, then the percentile rank must be approximated.

Percentiles are also used to compare an individual's test score with the national norm. For example, the *Canadian Achievement Test* (CAT), offered by the Canadian Test Centre (CTC) Educational Assessment Services, measures outcomes of skill sets in reading, language, spelling, and mathematics. Parents receive a report indicating a student's nationalized percentile ranking compared to students at the same grade level.

Interesting Facts

The highest recorded temperature on Earth was 57.8°C in Libya in 1922. The lowest recorded temperature on Earth was −89.2°C in Antarctica in 1983.



Step 3 Multiply this value by 1.5.

$1.5(11) = 16.5$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$9 - 16.5 = -7.5$ and $20 + 16.5 = 36.5$

Step 5 Check the data set for any data values that fall outside the interval from -7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

There are several reasons why outliers may occur. First, the data value may have resulted from a measurement or observational error. Perhaps the researcher measured the variable incorrectly. Second, the data value may have resulted from a recording error. That is, it may have been written or typed incorrectly. Third, the data value may have been obtained from a subject that is not in the defined population. For example, suppose test scores were obtained from a Grade 7 class, but a student in that class was actually in Grade 6 and had special permission to attend the class. This student might have scored extremely low on that particular exam on that day. Fourth, the data value might be a legitimate value that occurred by chance (although the probability is extremely small).

There are no hard-and-fast rules on what to do with outliers, nor is there complete agreement among statisticians on ways to identify them. Obviously, if they occurred as a result of an error, an attempt should be made to correct the error or else the data value should be omitted entirely. When they occur naturally by chance, the statistician must make a decision about whether to include them in the data set.

When a distribution is normal or bell-shaped, data values that are beyond 3 standard deviations of the mean can be considered suspected outliers.

3–3 Applying the Concepts

Determining Dosages

In an attempt to determine necessary dosages of a new drug (HDL) used to control sepsis, assume you administer varying amounts of HDL to 40 mice. You create four groups and label them *low dosage, moderate dosage, large dosage,* and *very large dosage*. The dosages also vary within each group. After the mice are injected with the HDL and the sepsis bacteria, the time until the onset of sepsis is recorded. Your job as a statistician is to effectively communicate the results of the study.

1. Which measures of position could be used to help describe the data results?
2. If 40% of the mice in the top quartile survived after the injection, how many mice would that be?
3. What information can be given from using percentiles?
4. What information can be given from using quartiles?
5. What information can be given from using standard scores?

See page 145 for the answers.

Section 6.4 Exercises

Exercises 3–3

1. What is a z score?

2. Define *percentile rank.*

3. What is the difference between a percentage and a percentile?

4. Define *quartile.*

5. What is the relationship between quartiles and percentiles?

6. What is a decile?

7. How are deciles related to percentiles?

8. To which percentile, quartile, and decile does the median correspond?

9. **Vacation Days** If the average number of vacation days for a selection of various countries has a mean of 29.4 days and a standard deviation of 8.6, find the z scores for the average number of vacation days in each of these countries.

Canada	26 days
Italy	42 days
United States	13 days

Source: Infoplease: www.infoplease.com.

10. **Reaction Time of Sprinters** The mean reaction time to the starting pistol for world-class sprinters is 153 milliseconds (ms) with a standard deviation of 28 ms. Find the corresponding z score for each sprinter's reaction time (in ms).

a. 195 *d.* 241.2
b. 90 *e.* 88.6
c. 139

Source: Kevin Duffy's Home Page, "Reaction Times and Sprint False Starts."

11. **Exam Scores** A final examination for a psychology course has a mean of 84 and a standard deviation of 4. Find the corresponding z score for each raw score.

a. 87 *d.* 76
b. 79 *e.* 82
c. 93

12. **Temperature of the Human Body** The healthy human body has a mean temperature of 36.8°C, with a standard deviation of 0.7°C. Find the corresponding z score for the following body temperatures (°C).

a. 37.5
b. 36.1
c. 34.77
d. 37.85
e. 38.41

Source: NationMaster.com, *Normal Human Body Temperature.*

13. **Exam Scores** A student scored 61 on the chemistry final exam, which had a mean of 54 and a standard deviation of 3.5, and she scored 85 on the biology final with a mean of 79 and a standard deviation of 2.5. On which exam did she perform relatively better?

14. **Marathon Run** An amateur male ran a marathon in 3 hours and 51 minutes; the mean time of the male runners was 4 hours and 30 minutes with a standard deviation of 39 minutes. An amateur female ran the same marathon in 4 hours and 40 minutes; the mean time of the female runners was 5 hours and 10 minutes with a standard deviation of 50 minutes. Which marathon runner did relatively better with respect to their gender? *Note*: A lower relative position is better in a marathon run.

Source: Marathon Training Expert.com, *What Is the Average Marathon Time?*

15. **Graduate Record Exam** A Canadian student applying to an American college wrote the three-part Graduate Record Exam (GRE). The student scores were as follows. In which part of the GRE did the student do relatively better?

Verbal reasoning	Quantitative reasoning	Analytical writing
$X = 593$	$X = 811$	$X = 5.2$
$\overline{X} = 462$	$\overline{X} = 584$	$\overline{X} = 4.0$
$s = 119$	$s = 151$	$s = 0.9$

Source: Educational Testing Services, *GRE* and *Interpreting Your GRE Scores, 2008–2009.*

16. **College Room and Board Costs** Room and board costs for selected schools are summarized in this distribution. Find the approximate cost of room and board corresponding to each of the following percentiles.

Costs (in dollars)	Frequency
3000.5–4000.5	5
4000.5–5000.5	6
5000.5–6000.5	18
6000.5–7000.5	24
7000.5–8000.5	19
8000.5–9000.5	8
9000.5–10,000.5	5

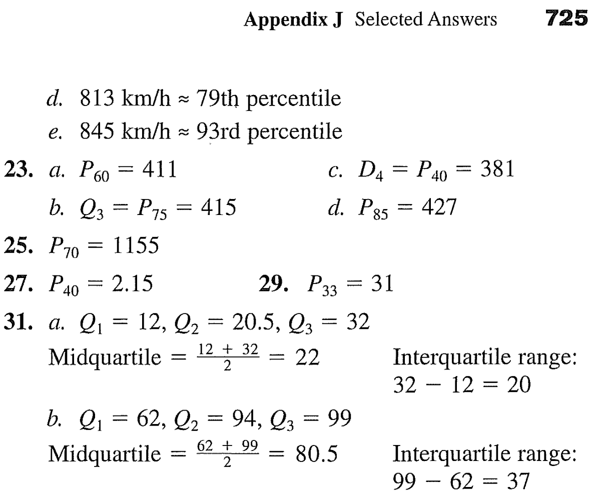
a. 30th percentile
b. 50th percentile
c. 75th percentile
d. 90th percentile

Source: World Almanac.

17. Using the data in Exercise 16, find the approximate percentile rank of each of the following costs.

a. 5500 *c.* 6500
b. 7200 *d.* 8300

Section 6.4 Answers



Appendix J Selected Answers **725**

d. 813 km/h ≈ 79th percentile

e. 845 km/h ≈ 93rd percentile

23. *a.* $P_{60} = 411$ *c.* $D_4 = P_{40} = 381$

b. $Q_3 = P_{75} = 415$ *d.* $P_{85} = 427$

25. $P_{70} = 1155$

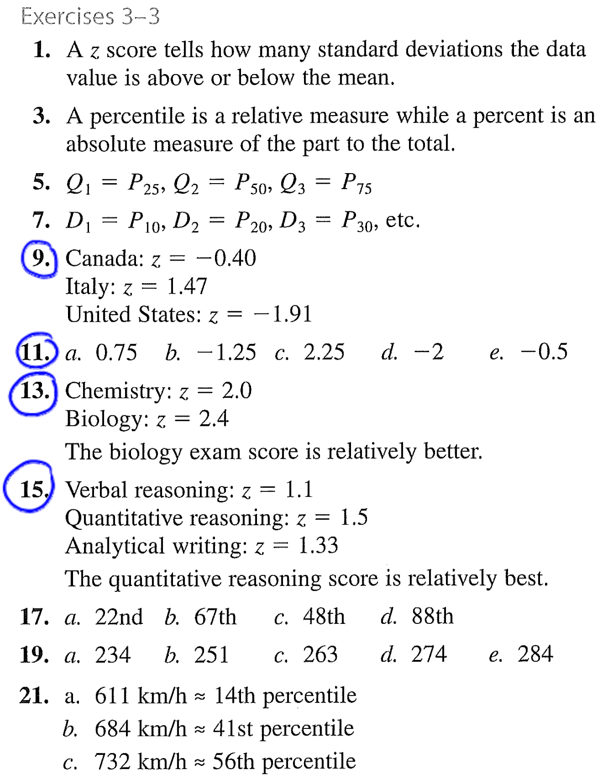
27. $P_{40} = 2.15$ **29.** $P_{33} = 31$

31. *a.* $Q_1 = 12$, $Q_2 = 20.5$, $Q_3 = 32$

Midquartile $= \frac{12 + 32}{2} = 22$ Interquartile range: $32 - 12 = 20$

b. $Q_1 = 62$, $Q_2 = 94$, $Q_3 = 99$

Midquartile $= \frac{62 + 99}{2} = 80.5$ Interquartile range: $99 - 62 = 37$



Exercises 3–3

1. A z score tells how many standard deviations the data value is above or below the mean.

3. A percentile is a relative measure while a percent is an absolute measure of the part to the total.

5. $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$

7. $D_1 = P_{10}$, $D_2 = P_{20}$, $D_3 = P_{30}$, etc.

9. Canada: $z = -0.40$
Italy: $z = 1.47$
United States: $z = -1.91$

11. *a.* 0.75 *b.* −1.25 *c.* 2.25 *d.* −2 *e.* −0.5

13. Chemistry: $z = 2.0$
Biology: $z = 2.4$
The biology exam score is relatively better.

15. Verbal reasoning: $z = 1.1$
Quantitative reasoning: $z = 1.5$
Analytical writing: $z = 1.33$
The quantitative reasoning score is relatively best.

17. *a.* 22nd *b.* 67th *c.* 48th *d.* 88th

19. *a.* 234 *b.* 251 *c.* 263 *d.* 274 *e.* 284

21. a. 611 km/h ≈ 14th percentile

b. 684 km/h ≈ 41st percentile

c. 732 km/h ≈ 56th percentile

Chapter 7

Producing Data

Reading for Chapter 7: The following reading is excerpted from:

> Bluman and Mayer. Elementary Statistics: A step by step approach. Canadian edition, Mc-Graw Hill Ryerson, 2008, pages 9-18, 22-26, 727-728.

Section 7.1

1–4 Data Collection and Sampling Techniques

Objective 5

Identify the four basic sampling techniques.

In research, statisticians use data in many different ways. As stated previously, data can be used to describe situations or events. For example, a manufacturer might want to know something about the consumers who will be purchasing his product so he can plan an effective marketing strategy. In another situation, the management of a company might survey its employees to assess their needs in order to negotiate a new contract with the employees' union. Data can be used to determine whether the educational goals of a school district are being met. Finally, trends in various areas, such as the stock market, can be analyzed, enabling prospective buyers to make more intelligent decisions concerning what stocks to purchase. These examples illustrate a few situations where collecting data will help people make better decisions on courses of action.

Data can be collected in a variety of ways. One of the most common methods is through the use of surveys. Surveys can be done by using a variety of methods. Three of the most common methods are the telephone survey, the mailed questionnaire, and the personal interview.

Telephone surveys have an advantage over personal interview surveys in that they are less costly. Also, people may be more candid in their opinions since there is no face-to-face contact. A major drawback to the telephone survey is that some people in the population will not have phones or will not answer when the calls are made; hence, not all people have a chance of being surveyed. Also, many people now have unlisted numbers and cellphones, so they cannot be surveyed. Finally, even the tone of the voice of the interviewer might influence the response of the person who is being interviewed.

Mailed questionnaire surveys can be used to cover a wider geographic area than telephone surveys or personal interviews since mailed questionnaire surveys are less expensive to conduct. Also, respondents can remain anonymous if they desire. Disadvantages of mailed questionnaire surveys include a low number of responses and inappropriate answers to questions. Another drawback is that some people may have difficulty reading or understanding the questions.

Historical Note

A pioneer in census taking was Pierre-Simon de Laplace. In 1780, he developed the Laplace method of estimating the population of a country. The principle behind his method was to take a census of a few selected communities and to determine the ratio of the population to the number of births in these communities. (Good birth records were kept.) This ratio would be used to multiply the number of births in the entire country to estimate the number of citizens in the country.

Personal interview surveys have the advantage of obtaining in-depth responses to questions from the person being interviewed. One disadvantage is that interviewers must be trained in asking questions and recording responses, which makes the personal interview survey more costly than the other two survey methods. Another disadvantage is that the interviewer may be biased in his or her selection of respondents.

Data can also be collected in other ways, such as *surveying records* or *direct observation* of situations.

As stated in Section 1–2, researchers use samples to collect data and information about a particular variable from a large population. Using samples saves time and money and, in some cases, enables the researcher to get more detailed information about a particular subject. Samples cannot be selected in haphazard ways because the information obtained might be biased. For example, interviewing people on a street corner during the day would not include responses from people working in offices at that time or from people attending school; hence, not all subjects in a particular population would have a chance of being selected.

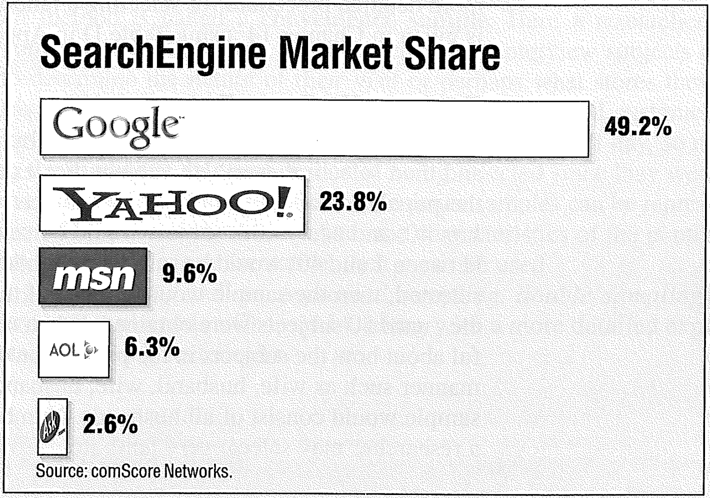
To obtain samples that are unbiased—i.e., give each subject in the population an equally likely chance of being selected—statisticians use four basic methods of sampling: random, systematic, stratified, and cluster sampling.

Random Sampling

Random samples are selected by using chance methods or random numbers. One such method is to number each subject in the population. Then place numbered cards in a bowl, mix them thoroughly, and select as many cards as needed. The subjects whose numbers are selected constitute the sample. Since it is difficult to mix the cards thoroughly, there is a chance of obtaining a biased sample. For this reason, statisticians use another method of obtaining numbers. They generate random numbers with a computer or calculator. Before the invention of computers, random numbers were obtained from tables.

Speaking of Statistics

This study of Internet search engine usage in North America was conducted by comScore Networks. Refer to the sampling methods and identify the probable method used to collect the usage data.



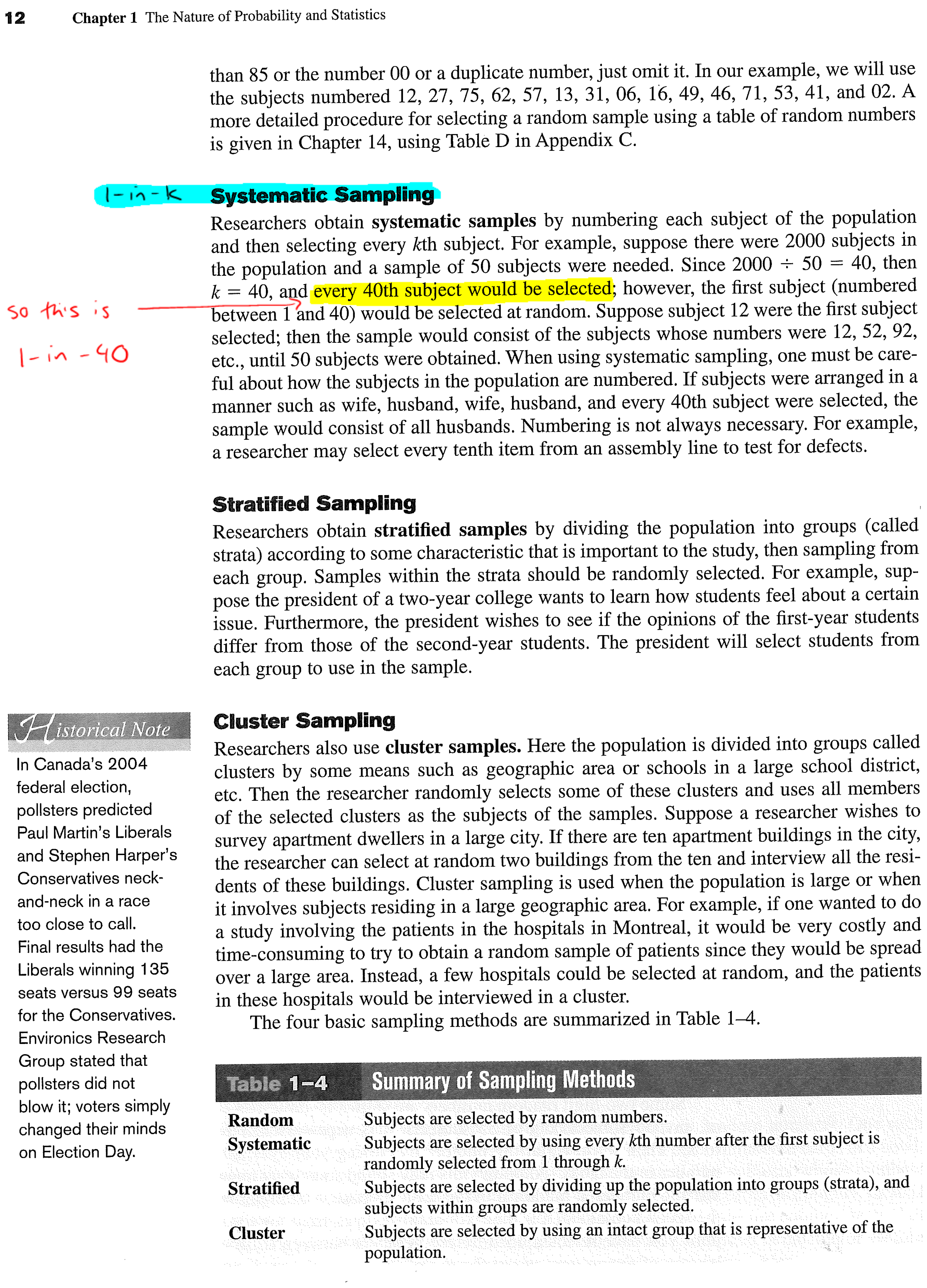
SearchEngine Market Share

Source: comScore Networks.

Table 1–3 Random Numbers

79	41	71	93	60	35	04	67	96	04	79	10	86
26	52	53	13	43	50	92	09	87	21	83	75	17
18	13	41	30	56	20	37	74	49	56	45	46	83
19	82	02	69	34	27	77	34	24	93	16	77	00
14	57	44	30	93	76	32	13	55	29	49	30	77
29	12	18	50	06	33	15	79	50	28	50	45	45
01	27	92	67	93	31	97	55	29	21	64	27	29
55	75	65	68	65	73	07	95	66	43	43	92	16
84	95	95	96	62	30	91	64	74	83	47	89	71
62	62	21	37	82	62	19	44	08	64	34	50	11
66	57	28	69	13	99	74	31	58	19	47	66	89
48	13	69	97	29	01	75	58	05	40	40	18	29
94	31	73	19	75	76	33	18	05	53	04	51	41
00	06	53	98	01	55	08	38	49	42	10	44	38
46	16	44	27	80	15	28	01	64	27	89	03	27
77	49	85	95	62	93	25	39	63	74	54	82	85
81	96	43	27	39	53	85	61	12	90	67	96	02
40	46	15	73	23	75	96	68	13	99	49	64	11

Some two-digit random numbers are shown in Table 1–3. To select a random sample of, say, 15 subjects out of 85 subjects, it is necessary to number each subject from 1 to 85. Then select a starting number by closing your eyes and placing your finger on a number in the table. (Although this may sound somewhat unusual, it enables us to find a starting number at random.) In this case suppose your finger landed on the number 12 in the second column. (It is the sixth number down from the top.) Then proceed downward until you have selected 15 different numbers between 01 and 85. When you reach the bottom of the column, go to the top of the next column. If you select a number greater



than 85 or the number 00 or a duplicate number, just omit it. In our example, we will use the subjects numbered 12, 27, 75, 62, 57, 13, 31, 06, 16, 49, 46, 71, 53, 41, and 02. A more detailed procedure for selecting a random sample using a table of random numbers is given in Chapter 14, using Table D in Appendix C.

1-in-k

Systematic Sampling

Researchers obtain **systematic samples** by numbering each subject of the population and then selecting every *k*th subject. For example, suppose there were 2000 subjects in the population and a sample of 50 subjects were needed. Since 2000 ÷ 50 = 40, then $k = 40$, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at random. Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until 50 subjects were obtained. When using systematic sampling, one must be careful about how the subjects in the population are numbered. If subjects were arranged in a manner such as wife, husband, wife, husband, and every 40th subject were selected, the sample would consist of all husbands. Numbering is not always necessary. For example, a researcher may select every tenth item from an assembly line to test for defects.

so this is 1-in-40

Stratified Sampling

Researchers obtain **stratified samples** by dividing the population into groups (called strata) according to some characteristic that is important to the study, then sampling from each group. Samples within the strata should be randomly selected. For example, suppose the president of a two-year college wants to learn how students feel about a certain issue. Furthermore, the president wishes to see if the opinions of the first-year students differ from those of the second-year students. The president will select students from each group to use in the sample.

Cluster Sampling

Researchers also use **cluster samples.** Here the population is divided into groups called clusters by some means such as geographic area or schools in a large school district, etc. Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples. Suppose a researcher wishes to survey apartment dwellers in a large city. If there are ten apartment buildings in the city, the researcher can select at random two buildings from the ten and interview all the residents of these buildings. Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area. For example, if one wanted to do a study involving the patients in the hospitals in Montreal, it would be very costly and time-consuming to try to obtain a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected at random, and the patients in these hospitals would be interviewed in a cluster.

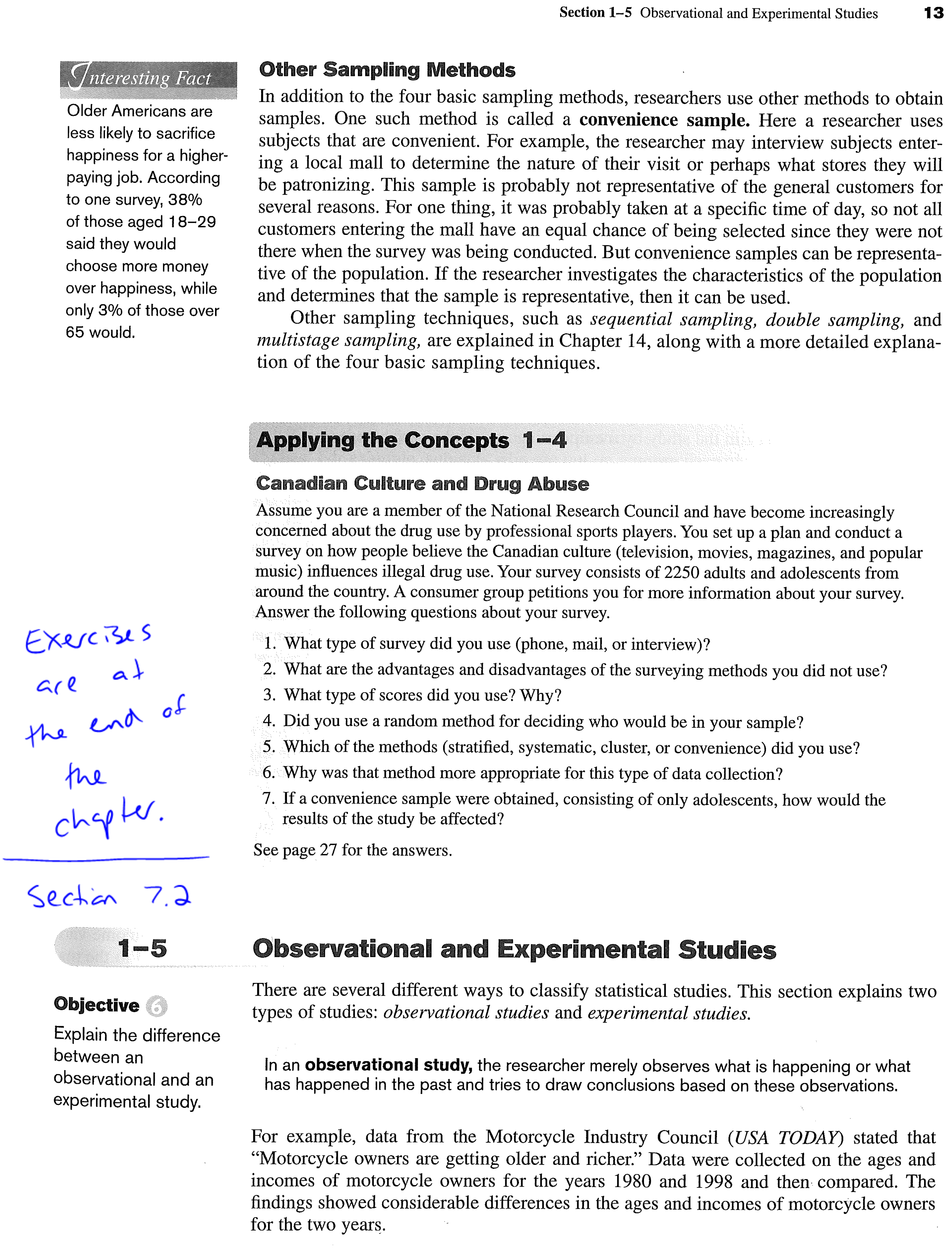
The four basic sampling methods are summarized in Table 1–4.

Historical Note

In Canada's 2004 federal election, pollsters predicted Paul Martin's Liberals and Stephen Harper's Conservatives neck-and-neck in a race too close to call. Final results had the Liberals winning 135 seats versus 99 seats for the Conservatives. Environics Research Group stated that pollsters did not blow it; voters simply changed their minds on Election Day.

Table 1–4 Summary of Sampling Methods

Random	Subjects are selected by random numbers.
Systematic	Subjects are selected by using every *k*th number after the first subject is randomly selected from 1 through *k*.
Stratified	Subjects are selected by dividing up the population into groups (strata), and subjects within groups are randomly selected.
Cluster	Subjects are selected by using an intact group that is representative of the population.



Interesting Fact

Older Americans are less likely to sacrifice happiness for a higher-paying job. According to one survey, 38% of those aged 18–29 said they would choose more money over happiness, while only 3% of those over 65 would.

Other Sampling Methods

In addition to the four basic sampling methods, researchers use other methods to obtain samples. One such method is called a **convenience sample.** Here a researcher uses subjects that are convenient. For example, the researcher may interview subjects entering a local mall to determine the nature of their visit or perhaps what stores they will be patronizing. This sample is probably not representative of the general customers for several reasons. For one thing, it was probably taken at a specific time of day, so not all customers entering the mall have an equal chance of being selected since they were not there when the survey was being conducted. But convenience samples can be representative of the population. If the researcher investigates the characteristics of the population and determines that the sample is representative, then it can be used.

Other sampling techniques, such as *sequential sampling, double sampling,* and *multistage sampling,* are explained in Chapter 14, along with a more detailed explanation of the four basic sampling techniques.

Applying the Concepts 1–4

Canadian Culture and Drug Abuse

Assume you are a member of the National Research Council and have become increasingly concerned about the drug use by professional sports players. You set up a plan and conduct a survey on how people believe the Canadian culture (television, movies, magazines, and popular music) influences illegal drug use. Your survey consists of 2250 adults and adolescents from around the country. A consumer group petitions you for more information about your survey. Answer the following questions about your survey.

1. What type of survey did you use (phone, mail, or interview)?
2. What are the advantages and disadvantages of the surveying methods you did not use?
3. What type of scores did you use? Why?
4. Did you use a random method for deciding who would be in your sample?
5. Which of the methods (stratified, systematic, cluster, or convenience) did you use?
6. Why was that method more appropriate for this type of data collection?
7. If a convenience sample were obtained, consisting of only adolescents, how would the results of the study be affected?

See page 27 for the answers.

Exercises are at the end of the chapter.

Section 7.2

1–5 Observational and Experimental Studies

Objective 6

Explain the difference between an observational and an experimental study.

There are several different ways to classify statistical studies. This section explains two types of studies: *observational studies* and *experimental studies.*

In an **observational study,** the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

For example, data from the Motorcycle Industry Council (*USA TODAY*) stated that "Motorcycle owners are getting older and richer." Data were collected on the ages and incomes of motorcycle owners for the years 1980 and 1998 and then compared. The findings showed considerable differences in the ages and incomes of motorcycle owners for the two years.

In this study, the researcher merely observed what had happened to the motorcycle owners over a period of time. There was no type of research intervention.

In an **experimental study,** the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Interesting Fact

The average age of a Harley-Davidson motorcycle owner is 52 years.

For example, a study conducted at Virginia Polytechnic Institute and presented in *Psychology Today* divided 56 female undergraduate students into two groups and had the students perform as many sit-ups as possible in 90 sec. The first group was told only to "Do your best," while the second group was told to try to increase the actual number of sit-ups done each day by 10%. After four days, the subjects in the group who were given the vague instructions to "Do your best" averaged 43 sit-ups, while the group that was given the more specific instructions to increase the number of sit-ups by 10% averaged 56 sit-ups by the last day's session. The conclusion then was that athletes who were given specific goals performed better than those who were not given specific goals.

This study is an example of a statistical experiment since the researchers intervened in the study by manipulating one of the variables, namely, the type of instructions given to each group.

In a true experimental study, the subjects should be assigned to groups randomly. Also, the treatments should be assigned to the groups at random. In the sit-up study, the article did not mention whether the subjects were randomly assigned to the groups.

Sometimes when random assignment is not possible, researchers use intact groups. These types of studies are done quite often in education where already intact groups are available in the form of existing classrooms. When these groups are used, the study is said to be a **quasi-experimental study.** The treatments, though, should be assigned at random. Most articles do not state whether random assignment of subjects was used.

Statistical studies usually include one or more *independent variables* and one *dependent variable.*

The **independent variable** in an experimental study is the one that is being manipulated by the researcher. The independent variable is also called the **explanatory variable.** The response variable is called the **dependent variable** or the **outcome variable.**

The outcome variable is the variable that is studied to see if it has changed significantly due to the manipulation of the independent variable. For example, in the sit-up study, the researchers gave the groups two different types of instructions, general and specific. Hence, the independent variable is the type of instruction. The dependent variable, then, is the response variable, that is, the number of sit-ups each group was able to perform after four days of exercise. If the differences in the dependent or outcome variable are large and other factors are equal, these differences can be attributed to the manipulation of the independent variable. In this case, specific instructions were shown to increase athletic performance.

In the sit-up study, there were two groups. The group that received the special instruction is called the **treatment group** while the other is called the **control group.** The treatment group receives a specific treatment (in this case, instructions for improvement) while the control group does not.

Both types of statistical studies have advantages and disadvantages. Experimental studies have the advantage that the researcher can decide how to select subjects and how to assign them to specific groups. The researcher can also control or manipulate the independent variable. For example, in studies that require the subjects to consume a certain amount of medicine each day, the researcher can determine the precise dosages and, if necessary, vary the dosage for the groups.

Interesting Fact

Ipsos-Reid, on behalf of Microsoft Canada, conducted a poll of teens. Based on responses, one in five teens in the study has their own website.

There are several disadvantages to experimental studies. First, they may occur in unnatural settings, such as laboratories and special classrooms. This can lead to several problems. One such problem is that the results might not apply to the natural setting. The age-old question then is, "This mouthwash may kill 10,000 germs in a test tube, but how many germs will it kill in my mouth?"

Another disadvantage with an experimental study is the **Hawthorne effect.** This effect was discovered in 1924 in a study of workers at the Hawthorne plant of the Western Electric Company. In this study, researchers found that the subjects who knew they were participating in an experiment actually changed their behaviour in ways that affected the results of the study.

Another problem is called *confounding of variables.*

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other.

Researchers try to control most variables in a study, but this is not possible in some studies. For example, subjects who are put on an exercise program might also improve their diet unbeknownst to the researcher and perhaps improve their health in other ways not due to exercise alone. Then diet becomes a confounding variable.

Observational studies also have advantages and disadvantages. One advantage of an observational study is that it usually occurs in a natural setting. For example, researchers can observe people's driving patterns on streets and highways in large cities. Another advantage of an observational study is that it can be done in situations where it would be unethical or downright dangerous to conduct an experiment. Using observational studies, researchers can study suicides, rapes, murders, etc. In addition, observational studies can be done using variables that cannot be manipulated by the researcher, such as drug users versus nondrug users and right-handedness versus left-handedness.

Observational studies have disadvantages, too. As mentioned previously, since the variables are not controlled by the researcher, a definite cause-and-effect situation cannot be shown since other factors may have had an effect on the results. Observational studies can be expensive and time-consuming. For example, if one wanted to study the habitat of lions in Africa, one would need a lot of time and money, and there would be a certain amount of danger involved. Finally, since the researcher may not be using his or her own measurements, the results could be subject to the inaccuracies of those who collected the data. For example, if the researchers were doing a study of events that occurred in the 1800s, they would have to rely on information and records obtained by others from a previous era. There is no way to ensure the accuracy of these records.

When you read the results of statistical studies, decide if the study was observational or experimental. Then see if the conclusion follows logically, based on the nature of these studies.

No matter what type of study is conducted, two studies on the same subject sometimes have conflicting conclusions. Why might this occur? An article entitled "Bottom Line: Is It Good for You?" (*USA TODAY Weekend*) states that in the 1960s studies suggested that margarine was better for the heart than butter since margarine contains less saturated fat and users had lower cholesterol levels. In a 1980 study, researchers found that butter was better than margarine since margarine contained trans-fatty acids, which are worse for the heart than butter's saturated fat. Then in a 1998 study, researchers found that margarine was better for a person's health. Now, what is to be believed? Should one use butter or margarine?

Unusual Stat

Ninety-seven percent of tea in Canada is consumed hot; in the United States, 85% of tea is served iced.

The answer here is to take a closer look at these studies. Actually, it is not a choice between butter or margarine that counts, but the type of margarine used. In the 1980s, studies showed that solid margarine contains trans-fatty acids, and scientists believe that they are worse for the heart than butter's saturated fat. In the 1998 study, liquid margarine was used. It is very low in trans-fatty acids, and hence it is more healthful than butter because trans-fatty acids have been shown to raise cholesterol. Hence, the conclusion is to use liquid margarine instead of solid margarine or butter.

Before decisions based on research studies are made, it is important to get all the facts and examine them in light of the particular situation.

Applying the Concepts 1–5

Just a Pinch Between Your Cheek and Gum

As the evidence on the adverse effects of cigarette smoke grew, people tried many different ways to quit smoking. Some people tried chewing tobacco or, as it was called, smokeless tobacco. A small amount of tobacco was placed between the cheek and gum. Certain chemicals from the tobacco were absorbed into the bloodstream and gave the sensation of smoking cigarettes. This prompted studies on the adverse effects of smokeless tobacco. One study in particular used 40 university students as subjects. Twenty were given smokeless tobacco to chew, and 20 given a substance that looked and tasted like smokeless tobacco, but did not contain any of the harmful substances. The students were randomly assigned to one of the groups. The students' blood pressure and heart rate were measured before they started chewing and 20 minutes after they had been chewing. A significant increase in heart rate occurred in the group that chewed the smokeless tobacco. Answer the following questions.

Exercises are at the end of the chapter

1. What type of study was this (observational, quasi-experimental, or experimental)?
2. What are the independent and dependent variables?
3. Which was the treatment group?
4. Could the students' blood pressures be affected by knowing that they are part of a study?
5. List some possible confounding variables.
6. Do you think this is a good way to study the effect of smokeless tobacco?

See page 28 for the answers.

Section 7.3

1–6 Uses and Misuses of Statistics

Objective 7

Explain how statistics can be used and misused.

As explained previously, statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypotheses, and make estimates about population characteristics. However, there is another aspect of statistics, and that is the misuse of statistical techniques to sell products that don't work properly, to attempt to prove something true that is really not true, or to get our attention by using statistics to evoke fear, shock, and outrage.

There are two sayings that have been around for a long time that illustrate this point:

"There are three types of lies—lies, damn lies, and statistics."

"Figures don't lie, but liars figure."

Just because we read or hear the results of a research study or an opinion poll in the media, this does not mean that these results are reliable or that they can be applied to any and all situations. For example, reporters sometimes leave out critical details such as the size of the sample used or how the research subjects were selected. Without this information, one cannot properly evaluate the research and properly interpret the conclusions of the study or survey.

It is the purpose of this section to show some ways that statistics can be misused. One should not infer that all research studies and surveys are suspect, but that there are many factors to consider when making decisions based on the results of research studies and surveys. Here are some ways that statistics can be misrepresented.

Suspect Samples

The first thing to consider is the sample that was used in the research study. Sometimes researchers use very small samples to obtain information. Several years ago, advertisements contained such statements as "Three out of four doctors surveyed recommend brand such and such." If only four doctors were surveyed, the results could have been obtained by chance alone; however, if 100 doctors were surveyed, the results might be quite different.

Not only is it important to have a sample size that is large enough, but also it is necessary to see how the subjects in the sample were selected. Studies using volunteers sometimes have a built-in bias. Volunteers generally do not represent the population at large. Sometimes they are recruited from a particular socioeconomic background, and sometimes unemployed people volunteer for research studies to get a stipend. Studies that require the subjects to spend several days or weeks in an environment other than their home or workplace automatically exclude people who are employed and cannot take time away from work. Sometimes only college students or retirees are used in studies. In the past, many studies have used only men, but have attempted to generalize the results to both men and women. Opinion polls that require a person to phone or mail in a response most often are not representative of the population in general, since only those with strong feelings for or against the issue usually call or respond by mail.

Another type of sample that may not be representative is the convenience sample. Educational studies sometimes use students in intact classrooms since it is convenient. Quite often, the students in these classrooms do not represent the student population of the entire school district.

When results are interpreted from studies using small samples, convenience samples, or volunteer samples, care should be used in generalizing the results to the entire population.

Ambiguous Averages

In Chapter 3, you will learn that there are four commonly used measures that are loosely called *averages.* They are the *mean, median, mode,* and *midrange.* For the same data set, these averages can differ markedly. People who know this can, without lying, select the one measure of average which lends the most evidence to support their position. This fact is explained on page 88 of Chapter 3.

Changing the Subject

Another type of statistical distortion can occur when different values are used to represent the same data. For example, one political candidate who is running for reelection might say, "During my administration, expenditures increased a mere 3%." His opponent, who is trying to unseat him, might say, "During my opponent's administration, expenditures have increased a whopping $6,000,000." Here both figures are correct; however, expressing a 3% increase as $6,000,000 makes it sound like a very large increase. Here again, ask yourself, Which measure best represents the data?

Detached Statistics

A claim that uses a detached statistic is one in which no comparison is made. For example, you may hear a claim such as "Our brand of crackers has one-third fewer calories." Here, no comparison is made. One-third fewer calories than what? Another example is a claim that uses a detached statistic such as "Brand A aspirin works four times faster." Four times faster than what? When you see statements such as this, always ask yourself, Compared to what?

Implied Connections

Many claims attempt to imply connections between variables that may not actually exist. For example, consider the following statement: "Eating fish may help to reduce your cholesterol." Notice the words *may help*. There is no guarantee that eating fish will definitely help you reduce your cholesterol.

"Studies suggest that using our exercise machine will reduce your weight." Here the word *suggest* is used; and again, there is no guarantee that you will lose weight by using the exercise machine advertised.

Another claim might say, "Taking calcium will lower blood pressure in some people." Note the word *some* is used. You may not be included in the group of "some" people. Be careful when you draw conclusions from claims that use words such as *may, in some people,* and *might help.*

Misleading Graphs

Statistical graphs give a visual representation of data that enables viewers to analyze and interpret data more easily than by simply looking at numbers. In Chapter 2, you will see how some graphs are used to represent data. However, if graphs are drawn inappropriately, they can misrepresent the data and lead the reader to false conclusions. The misuse of graphs is also explained in Chapter 2, on pages 58–61.

Faulty Survey Questions

When analyzing the results of a survey using questionnaires, you should be sure that the questions are properly written since the way questions are phrased can often influence the way people answer them. For example, the responses to a question such as "Do you feel that the North Huntingdon School District should build a new football stadium?" might be answered differently than a question such as "Do you favour increasing school taxes so that the North Huntingdon School District can build a new football stadium?" Each question asks something a little different, and the responses could be radically different. When you read and interpret the results obtained from questionnaire surveys, watch out for some of these common mistakes made in the writing of the survey questions.

In Chapter 14, you will find some common ways that survey questions could be misinterpreted by those responding and could therefore result in incorrect conclusions.

To restate the premise of this section, statistics, when used properly, can be beneficial in obtaining much information, but when used improperly, can lead to much misinformation. It is like your automobile. If you use your automobile to get to school or work or to go on a vacation, that's good. But if you use it to run over your neighbour's dog because it barks all night long and tears up your flower garden, that's not so good!

Review Exercises

Note: **All odd-numbered problems and even-numbered problems marked with "ans" are included in the answer section at the end of this book.**

1. Name and define the two areas of statistics.
2. What is probability? Name two areas where probability is used.
3. Suggest some ways statistics can be used in everyday life.
4. Explain the differences between a sample and a population.
5. Why are samples used in statistics?
6. In each of these statements, tell whether descriptive or inferential statistics have been used.

   a. By the year 2010, 87 million homes worldwide will be watching high-definition television broadcasts (HDTV) (Source: www.electronics.ca).
   b. 83% of Canadian drowning victims are male. (Source: www.lifesaving.org)
   c. The value of Canada's building permits in December 2005 set a new record at $6.5 billion (Source: Statistics Canada).
   d. The median length of prison sentences in adult criminal court for property crimes in Canada in 2003 was 42 days (Source: Statistics Canada).
   e. Flu shots may also protect against heart disease and stroke (Source: Heart and Stroke Foundation).
   f. By 2031, 25% of Canada's population will be 65 years of age or over (Source: Statistics Canada).
   g. Average health-care spending per capita by provincial and territorial governments in 2005–2006 was $2845 (Source: Canadian Institute for Healthcare Information).
   h. The projected budget surplus in 2006–2007 for Alberta was $4.1 billion (Source: Government of Alberta).

7. Classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

   a. Pages in the city of Winnipeg telephone book.
   b. Rankings of tennis players.
   c. Weights of air conditioners.
   d. Temperatures inside ten refrigerators.
   e. Salaries of the top five CEOs in the United States.
   f. Ratings of eight local plays (poor, fair, good, excellent).
   g. Times required for mechanics to do a tune-up.
   h. Ages of students in a classroom.
   i. Marital status of patients in a doctor's office.
   j. Horsepower of tractor engines.

8. **(ans)** Classify each variable as qualitative or quantitative.

   a. Number of bicycles sold in one year by a large sporting goods store.
   b. Colours of baseball caps in a store.
   c. Time it takes to cut a lawn.
   d. Capacity in cubic metres of six truck beds.
   e. Classification of children in a day-care centre (infant, toddler, preschool).
   f. Weights of fish caught in Lake George.
   g. Marital status of faculty members in a large university.

9. Classify each variable as discrete or continuous.

   a. Number of doughnuts sold each day by Doughnut Heaven.
   b. Water temperatures of six swimming pools in Saskatoon on a given day.
   c. Weights of cats in a pet shelter.
   d. Lifetime (in hours) of 12 flashlight batteries.
   e. Number of cheeseburgers sold each day by a hamburger stand on a college campus.
   f. Number of DVDs rented each day by a video store.
   g. Daily volume of raw sewage and storm water (in litres) entering St. John's Harbour.

10. Give the boundaries of each value.

a. 42.8 kilometres
b. 1.6 millilitres
c. 5.36 grams
d. 18 kilograms
e. 13.8 °C
f. 40 centimetres

11. Name and define the four basic sampling methods.

12. **(ans)** Classify each sample as random, systematic, stratified, or cluster.

a. In a large school district, all teachers from two buildings are interviewed to determine whether they believe the students have less homework to do now than in previous years.

b. Every seventh customer entering a shopping mall is asked to select her or his favourite store.

c. Nursing supervisors are selected using random numbers in order to determine annual salaries.

d. Every 100th hamburger manufactured is checked to determine its fat content.

e. Mail carriers of a large city are divided into four groups according to gender (male or female) and according to whether they walk or ride on their routes. Then ten are selected from each group and interviewed to determine whether they have been bitten by a dog in the last year.

13. Give three examples each of nominal, ordinal, interval, and ratio data.

14. For each of these statements, define a population and state how a sample might be obtained.

a. The average cost of an airline meal is $4.55 (Source: *Everything Has Its Price,* Richard E. Donley, Simon and Schuster).

b. Some 25% of Canadian children are obese today (Source: *Reader's Digest, Canada*).

c. Every ten minutes, two people die in car crashes and 170 are injured (Source: National Safety Council estimates).

d. When older people with mild to moderate hypertension were given mineral salt for six months, the average blood pressure reading dropped by 8 points systolic and 3 points diastolic (Source: *Prevention*).

e. The average amount spent per gift for Mom on Mother's Day is $25.95 (Source: The Gallup Organization).

15. Select a newspaper or magazine article that involves a statistical study, and write a paper answering these questions.

a. Is this study descriptive or inferential? Explain your answer.

b. What are the variables used in the study? In your opinion, what level of measurement was used to obtain the data from the variables?

c. Does the article define the population? If so, how is it defined? If not, how could it be defined?

d. Does the article state the sample size and how the sample was obtained? If so, determine the size of the sample and explain how it was selected. If not, suggest a way it could have been obtained.

e. Explain *in your own words* what procedure (survey, comparison of groups, etc.) might have been used to determine the study's conclusions.

f. Do you agree or disagree with the conclusions? State your reasons.

16. Information from research studies is sometimes taken out of context. Explain why the claims of these studies might be suspect.

a. The average salary of the graduates of the class of 1980 is $32,500.

b. It is estimated that in Podunk there are 27,256 cats.

c. Only 3% of the men surveyed read *Cosmopolitan* magazine.

d. Based on a recent mail survey, 85% of the respondents favoured gun control.

e. A recent study showed that high school dropouts drink more coffee than students who graduated; therefore, coffee dulls the brain.

f. Since most automobile accidents occur within 24 kms of a person's residence, it is safer to make long trips.

17. Identify each study as being either observational or experimental.

a. Subjects were randomly assigned to two groups, and one group was given an herb and the other group a placebo. After six months, the numbers of respiratory tract infections each group had were compared.

b. A researcher stood at a busy intersection to see if the colour of the automobile that a person drives is related to running red lights.

c. A researcher finds that people who are more hostile have higher total cholesterol levels than those who are less hostile.

d. Subjects are randomly assigned to four groups. Each group is placed on one of four special diets—a low-fat diet, a high-fish diet, a combination of low-fat diet and high-fish diet, and a regular diet. After six months, the blood pressures of the groups are compared to see if diet has any effect on blood pressure.

18. Identify the independent variable(s) and the dependent variable for each of the studies in Exercise 17.

19. For each of the studies in Exercise 17, suggest possible confounding variables.

20. According to a pilot study of 20 people conducted at the University of Minnesota, daily doses of a compound called arabinogalactan over a period of six months resulted in a significant increase in the beneficial lactobacillus species of bacteria. Why can't it be concluded that the compound is beneficial for the majority of people?

21. Comment on the following statement, taken from a magazine advertisement: "In a recent clinical study, Brand ABC [actual brand will not be named] was proved to be 1950% better than creatine!"

22. In an ad for women, the following statement was made: "For every 100 women, 91 have taken the road less travelled." Comment on this statement.

23. In many ads for weight loss products, under the product claims and in small print, the following statement is made: "These results are not typical." What does this say about the product being advertised?

24. In an ad for moisturizing lotion, the following claim is made: " . . . it's the #1 dermatologist-recommended brand." What is misleading about this claim?

25. An ad for an exercise product stated: "Using this product will burn 74% more calories." What is misleading about this statement?

26. "Vitamin E is a proven antioxidant and may help in fighting cancer and heart disease." Is there anything ambiguous about this claim? Explain.

27. "Just 1 capsule of Brand X can provide 24 hours of acid control." (Actual brand will not be named.) What needs to be more clearly defined in this statement?

28. ". . . Male children born to women who smoke during pregnancy run a risk of violent and criminal behaviour that lasts well into adulthood." Can we infer that smoking during pregnancy is responsible for criminal behaviour in people?

29. In the 1980s, a study linked coffee to a higher risk of heart disease and pancreatic cancer. In the early 1990s, studies showed that drinking coffee posed minimal health threats. However, in 1994, a study showed that pregnant women who drank 3 or more cups of tea daily may be at risk for spontaneous abortion. In 1998, a study claimed that women who drank more than a half-cup of caffeinated tea every day may actually increase their fertility. In 1998, a study showed that over a lifetime, a few extra cups of coffee a day can raise blood pressure, heart rate, and stress (Source: "Bottom Line: Is It Good for You? Or Bad?" by Monika Guttman, *USA TODAY Weekend*). Suggest some reasons why these studies appear to be conflicting.

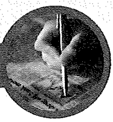
Extending the Concepts

30. Find an article that describes a statistical study, and identify the study as observational or experimental.

31. For the article that you used in Exercise 30, identify the independent variable(s) and dependent variable for the study.

32. For the article that you selected in Exercise 30, suggest some confounding variables that may have an effect on the results of the study.

Statistics Today



Are We Improving Our Diet?—Revisited

Researchers selected a *sample* of 23,699 adults in the United States, using phone numbers selected at *random,* and conducted a *telephone survey.* All respondents were asked six questions:

1. How often do you drink juices such as orange, grapefruit, or tomato?
2. Not counting juice, how often do you eat fruit?
3. How often do you eat green salad?
4. How often do you eat potatoes (not including french fries, fried potatoes, or potato chips)?
5. How often do you eat carrots?
6. Not counting carrots, potatoes, or salad, how many servings of vegetables do you usually eat?

Researchers found that men consumed fewer servings of fruits and vegetables per day (3.3) than women (3.7). Only 20% of the population consumed the recommended five or more daily servings. In addition, they found that youths and less-educated people consumed an even lower amount than the average.

Based on this study, they recommend that greater educational efforts are needed to improve fruit and vegetable consumption by North Americans and to provide environmental and institutional support to encourage increased consumption.

Source: Mary K. Serdula, M.D., et al., "Fruit and Vegetable Intake Among Adults in 16 States: Results of a Brief Telephone Survey," *American Journal of Public Health* 85, no. 2. Copyright by the American Public Health Association; Government of Canada.

More Exercises

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. Probability is used as a basis for inferential statistics.
2. The height of Sir John A. Macdonald is an example of a variable.
3. The highest level of measurement is the interval level.
4. When the population of college professors is divided into groups according to their rank (instructor, assistant professor, etc.) and then several are selected from each group to make up a sample, the sample is called a cluster sample.
5. The variable *age* is an example of a qualitative variable.
6. The weight of pumpkins is considered to be a continuous variable.
7. The boundary of a value such as 6 centimetres would be 5.9–6.1 centimetres.

Select the best answer.

8. The number of absences per year that a worker has is an example of what type of data?

   a. Nominal
   b. Qualitative
   c. Discrete
   d. Continuous

9. What are the boundaries of 25.6 grams?

   a. 25–26 grams
   b. 25.55–25.65 grams
   c. 25.5–25.7 grams
   d. 20–39 grams

10. A researcher divided subjects into two groups according to gender and then selected members from each group for her sample. What sampling method was the researcher using?

    a. Cluster
    b. Random
    c. Systematic
    d. Stratified

11. Data that can be classified according to colour are measured on what scale?

    a. Nominal
    b. Ratio
    c. Ordinal
    d. Interval

12. A study that involves no researcher intervention is called

    a. An experimental study.
    b. A noninvolvement study.
    c. An observational study.
    d. A quasi-experimental study.

13. A variable that interferes with other variables in the study is called

    a. A confounding variable.
    b. An explanatory variable.
    c. An outcome variable.
    d. An interfering variable.

Use the best answer to complete these statements.

14. Two major branches of statistics are ______ and ______.
15. Two uses of probability are ______ and ______.
16. The group of all subjects under study is called a(n) ______.
17. A group of subjects selected from the group of all subjects under study is called a(n) ______.
18. Three reasons why samples are used in statistics are

    a. ________ *b.* ________ *c.* ________.

19. The four basic sampling methods are

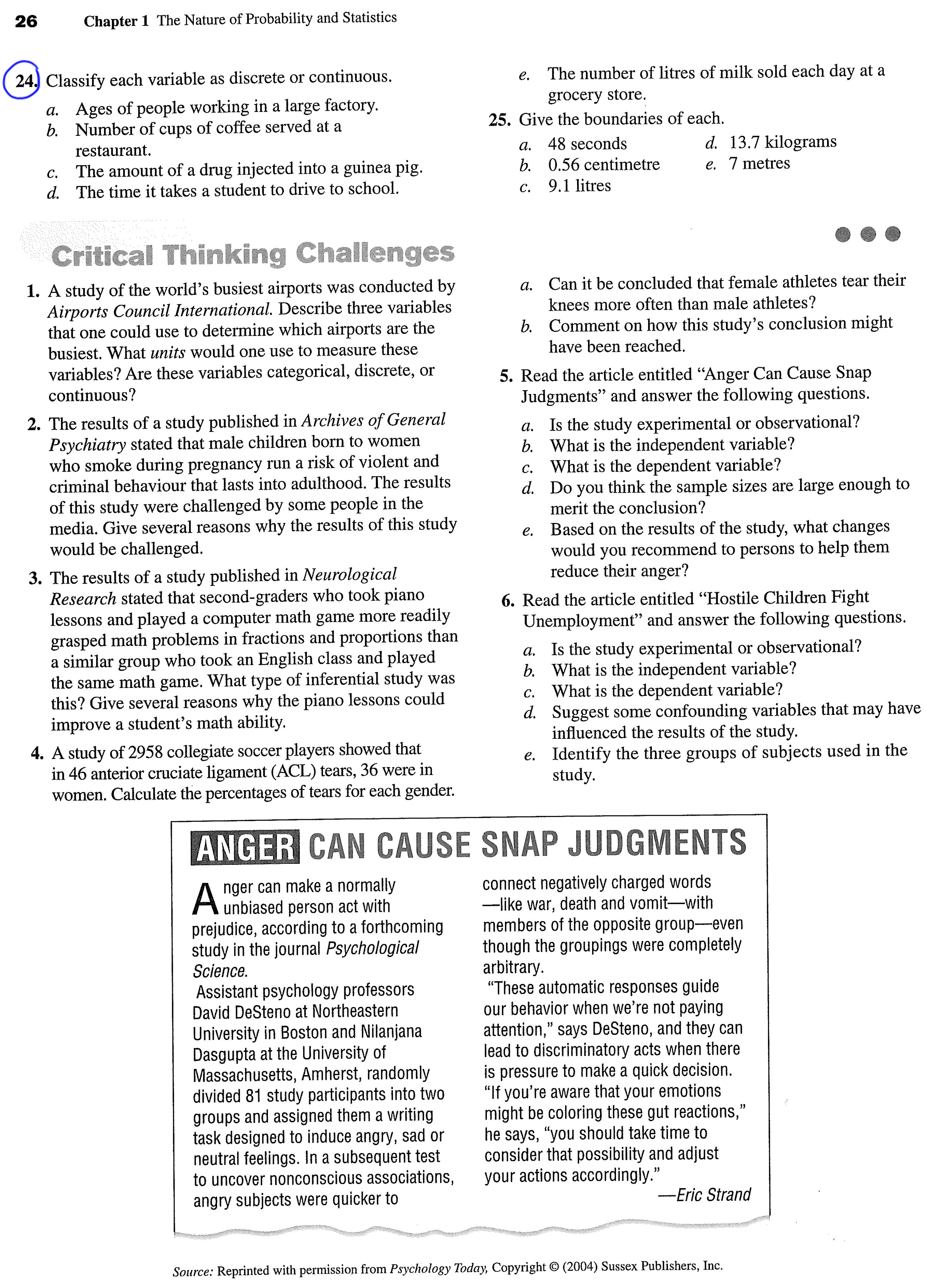
    a. ________ *b.* ________ *c.* ________ *d.* ________.

20. A study that uses intact groups when it is not possible to randomly assign participants to the groups is called a(n) ______ study.
21. In a research study, participants should be assigned to groups using ______ methods, if possible.
22. For each statement, decide whether descriptive or inferential statistics is used.

    a. The average life expectancy in New Zealand is 78.49 years. Source: *World Factbook 2004.*
    b. A diet high in fruits and vegetables will lower blood pressure. Source: Institute of Medicine.
    c. The total amount of estimated losses from hurricane Hugo was $4.2 billion. Source: Insurance Service Office.
    d. Researchers stated that the shape of a person's ears is related to the person's aggression. Source: *American Journal of Human Biology.*
    e. In 2013, the number of high school graduates will be 3.2 million students. Source: National Center for Education.

23. Classify each as nominal-level, ordinal-level, interval-level, or ratio-level measurement.

    a. Rating of movies as G, PG, and R.
    b. Number of candy bars sold on a fund drive.
    c. Classification of automobiles as subcompact, compact, standard, and luxury.
    d. Temperatures of hair dryers.
    e. Weights of suitcases on a commercial airline.



24. Classify each variable as discrete or continuous.
   - *a.* Ages of people working in a large factory.
   - *b.* Number of cups of coffee served at a restaurant.
   - *c.* The amount of a drug injected into a guinea pig.
   - *d.* The time it takes a student to drive to school.
   - *e.* The number of litres of milk sold each day at a grocery store.

25. Give the boundaries of each.
   - *a.* 48 seconds
   - *b.* 0.56 centimetre
   - *c.* 9.1 litres
   - *d.* 13.7 kilograms
   - *e.* 7 metres

Critical Thinking Challenges

1. A study of the world's busiest airports was conducted by *Airports Council International.* Describe three variables that one could use to determine which airports are the busiest. What *units* would one use to measure these variables? Are these variables categorical, discrete, or continuous?

2. The results of a study published in *Archives of General Psychiatry* stated that male children born to women who smoke during pregnancy run a risk of violent and criminal behaviour that lasts into adulthood. The results of this study were challenged by some people in the media. Give several reasons why the results of this study would be challenged.

3. The results of a study published in *Neurological Research* stated that second-graders who took piano lessons and played a computer math game more readily grasped math problems in fractions and proportions than a similar group who took an English class and played the same math game. What type of inferential study was this? Give several reasons why the piano lessons could improve a student's math ability.

4. A study of 2958 collegiate soccer players showed that in 46 anterior cruciate ligament (ACL) tears, 36 were in women. Calculate the percentages of tears for each gender.
   - *a.* Can it be concluded that female athletes tear their knees more often than male athletes?
   - *b.* Comment on how this study's conclusion might have been reached.

5. Read the article entitled "Anger Can Cause Snap Judgments" and answer the following questions.
   - *a.* Is the study experimental or observational?
   - *b.* What is the independent variable?
   - *c.* What is the dependent variable?
   - *d.* Do you think the sample sizes are large enough to merit the conclusion?
   - *e.* Based on the results of the study, what changes would you recommend to persons to help them reduce their anger?

6. Read the article entitled "Hostile Children Fight Unemployment" and answer the following questions.
   - *a.* Is the study experimental or observational?
   - *b.* What is the independent variable?
   - *c.* What is the dependent variable?
   - *d.* Suggest some confounding variables that may have influenced the results of the study.
   - *e.* Identify the three groups of subjects used in the study.

ANGER CAN CAUSE SNAP JUDGMENTS

Anger can make a normally unbiased person act with prejudice, according to a forthcoming study in the journal *Psychological Science.*

Assistant psychology professors David DeSteno at Northeastern University in Boston and Nilanjana Dasgupta at the University of Massachusetts, Amherst, randomly divided 81 study participants into two groups and assigned them a writing task designed to induce angry, sad or neutral feelings. In a subsequent test to uncover nonconscious associations, angry subjects were quicker to connect negatively charged words —like war, death and vomit—with members of the opposite group—even though the groupings were completely arbitrary.

"These automatic responses guide our behavior when we're not paying attention," says DeSteno, and they can lead to discriminatory acts when there is pressure to make a quick decision. "If you're aware that your emotions might be coloring these gut reactions," he says, "you should take time to consider that possibility and adjust your actions accordingly."

—*Eric Strand*

Source: Reprinted with permission from *Psychology Today,* Copyright © (2004) Sussex Publishers, Inc.

Appendix J

Selected Answers*

Chapter 1

Review Exercises

1. Descriptive statistics describes a set of data. Inferential statistics uses a set of data to make predictions about a population.

3. Answers will vary.

5. When the population is large, the researcher saves time and money using samples. Samples are used when the units must be destroyed.

7. *a.* ratio *b.* ordinal *c.* ratio *d.* interval *e.* ratio *f.* ordinal *g.* ratio *h.* ratio *i.* nominal *j.* ratio

8. *a.* quantitative *b.* qualitative *c.* quantitative *d.* quantitative *e.* qualitative *f.* quantitative *g.* qualitative

9. *a.* discrete *b.* continuous *c.* continuous *d.* continuous *e.* discrete *f.* continuous

11. Random samples are selected by using chance methods or random numbers. Systematic samples are selected by numbering each subject and selecting every *k*th number. Stratified samples are selected by dividing the population into groups and selecting from each group. Cluster samples are selected by using intact groups called clusters.

12. *a.* cluster *c.* random *e.* stratified
1-in-7 *b.* systematic *d.* systematic 1-in-100

13. Answers will vary.

15. Answers will vary.

17. *a.* experimental *b.* observational *c.* observational *d.* experimental

19. Answers will vary. Possible answers include:

a. overall health of participants, amount of exposure to infected individuals through the workplace or home

b. gender and/or age of driver, time of day

c. diet, general health, heredity factors

d. amount of exercise, heredity factors

21. Claims can be proven only if the entire population is used.

23. Since the results are not typical, the advertisers selected only a few people for whom the product worked extremely well.

25. "74% more calories" than what? No comparison group is stated.

27. What is meant by "24 hours of acid control"?

29. Possible reasons for conflicting results: The amount of caffeine in the coffee or tea or the brewing method.

31. Answers will vary.

Chapter Quiz

1. True
2. False
3. False
4. False
5. False
6. True
7. False
8. *c*
9. *b*
10. *d*
11. *a*
12. *c*
13. *a*
14. descriptive, inferential
15. gambling, insurance
16. population
17. sample

18. *a.* saves time *b.* saves money *c.* use when population is infinite

19. *a.* random *b.* systematic *c.* cluster *d.* stratified

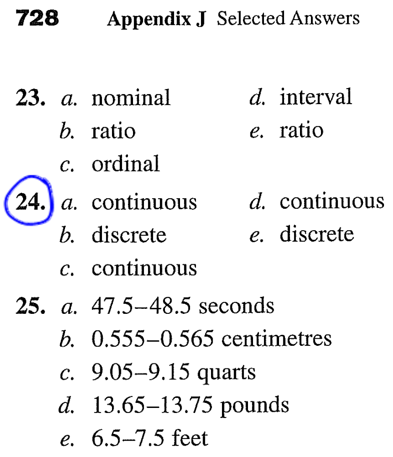
20. quasi-experimental

21. random

22. *a.* descriptive *b.* inferential *c.* descriptive *d.* inferential *e.* inferential

*Answers may vary due to rounding or use of technology.

Note: These answers to odd-numbered and selected even-numbered exercises include all quiz answers.



728 **Appendix J** Selected Answers

23. *a.* nominal
b. ratio
c. ordinal
d. interval
e. ratio

24. *a.* continuous
b. discrete
c. continuous
d. continuous
e. discrete

25. *a.* 47.5–48.5 seconds
b. 0.555–0.565 centimetres
c. 9.05–9.15 quarts
d. 13.65–13.75 pounds
e. 6.5–7.5 feet

Chapter 8

Introduction to Probability

Reading for Chapter 8: The following reading is exerpted from:

Patricia Wrean. Online Textbook for MATH 163. Camosun College, 2015.

Chapter 4

Probability

8.1 ~~4.1~~ Counting Techniques

Although the idea of counting the number of objects seems straightforward, there are a few tricky bits to it when you are looking at large quantities. Let's start by looking at an example.

Example: How many three-digit natural numbers are there?

Answer: Let's look at the list: $100, 101, 102, \ldots 999$. There are three methods we can use to determine the total number here.

Method #1: We could, for example, write the list as a sequence. We see that it's an arithmetic sequence with $d = 1$. Then

$$\begin{aligned}
&a_n = a_1 + (n-1)\,d \\
&999 = 100 + (n-1)1 \\
&899 = n - 1 \\
&n = 900
\end{aligned}$$

so there are 900 three-digit numbers in total.

Method #2: Since the numbers go up one-by-one, can use the nice summation notation trick of $(\text{last} - \text{first} + 1) = (999 - 100 + 1) = 900$ as well.

Method #3: Consider the digits __ __ __ . The first digit can be 1-9 for 9 choices, while the second and third can be 0-9 for

10 choices. Then you get $9 \times 10 \times 10 = 900$ numbers.

Example: How many three-digit natural numbers are divisible by 5?

Answer: The numbers are $100, 105, 110, \dots 995$.

Method #1 works very nicely:

$$\begin{aligned}
&a_n = a_1 + (n-1)\,d\\
&995 = 100 + (n-1)5\\
&895 = (n-1)\,5\\
&179 = n-1\\
&n = 180
\end{aligned}$$

Method #2: we can rewrite the sequence as $20 \times 5, 21 \times 5, 22 \times 5, \dots 199 \times 5$. Then we use $(\text{last} - \text{first} + 1) = (199 - 20 + 1) = 180$. (We can only use this formula when we are counting by steps of one. So we have to force our sequence into a counting-by-one step to use it.)

Method #3: We note that if the number is divisible by 5, then the last digit is either 0 or 5, for two choices. We then get $9 \times 10 \times 2 = 180$. However, be careful with this method! It will work well for numbers divisible by 1, 2, 5, and 10, because this eliminates digits in the last column. You can't use this method at all with most other divisors like 3, 4, 6, etc.

The reason Method #3 works is called the Multiplication Principle of Counting. If a question consists of a series of choices in which there are p possibilities for the first choice, q possibilities for the second choice, r for the third, etc., then the number of ways in which the question can be done is just $p \times q \times r \times \dots$ In other words, you just multiply together the number of ways each step can be done.

Example: How many BC licence plates for cars are there (barring reserved words, etc.)?

Answer: The patterns allowable are

letter-letter-letter number-number-number
number-number-number letter-letter-letter
letter-letter-number-number-number-letter

(we'll ignore personalized plates or reserved words, etc.) Looking at the first pattern, there are 26 choices for each letter and 10 for each number. So we've got $26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17,576,000$ plates.

The other two patterns will have the same number, for a total of $52,728,000$ plates using the two patterns.

Example: How many (250) area code phone numbers are there?

Answer: Phone numbers are of the form (250) ### - ####. To look at all possibilities, there are 10 digits for each #, so 10^7 or ten million possibilities.

Example: How many (250) area code phone numbers are there that don't start with zero?

Answer: Now we have only 9 choices for the first #, so 9×10^6 or nine million.

Example: How many (250) area code phone numbers are there that don't start with 911?

Answer: This one's more tricky. What we'll do is take the total number of phone numbers and subtract the number that **do** begin with 911. Numbers beginning with 911 will look like: $(250)911 - \#\#\#\#$, so we'll have 104 choices. Since there are $10,000,000$ numbers in total, the number that don't start with 911 is $(10,000,000 - 1000) = 9,999,000$.

So our rule is:

$$\text{number allowed} = \text{total number} - \text{number not allowed}$$

There's another rule that we should know regarding counting if we're using the word "or".

Example: How many numbers from 1 to 30 are a) divisible by 3? b) divisible by 5? c) divisible by 3 or 5?

Answer:

Well, let's try the brute force method.

Divisible by 3: $3, 6, 9, 12, 15, 18, 21, 24, 27, 30$ total: 10

Divisible by 5: $5, 10, 15, 20, 25, 30$ total: 6

Divisible by 3 or 5: number in first row (10) + number in second row (6) but we have to subtract 2, because otherwise we are counting 15 and 30 twice! So we get $10 + 6 - 2 = 14$.

This brings up the addition rule:

$$n(A \text{ or } B) = n(A) + n(B) - n(AB)$$

Now, if the two situations don't have overlap, then $n(AB)$ can be zero. We say then that the two situations are **mutually exclusive**.

Example: How many case-sensitive alpha-numeric passwords are there that have 6 or 7 characters?

Answer: First, let's look at what "case-sensitive, alpha-numeric" means. It means that capital letters are considered different from lowercase, so 52 letters instead of 26. Also, numbers are allowed, so 62 choices in total.

Number of 6-char passwords: $62 \times 62 \times 62 \times 62 \times 62 \times 62 = 62^6 = 5.68 \times 10^{10}$

Number of 7-char passwords: $627 = 3.52 \times 10^{12}$

Now, a password can either have 6 characters or 7 but not both, so to get the total, we just add $5.68 \times 10^{10} + 3.52 \times 10^{12} = 3.58 \times 10^{12}$.

Example: How many case-sensitive alpha-numeric passwords are there that have 6 characters and at least one number and letter?

Answer: This question is quite difficult as written. However, if we calculate instead the total number of passwords and subtract the number that don't have any numbers and the number that don't have any letters, we'll get the same result.

Number of 6-char passwords without any numbers: 52^6

Number of 6-char passwords without any letters: 10^6

So we get total $= 62^6 - 52^6 - 10^6 = 3.70 \times 10^{10}$

~~4.1.1~~ 8.1 Exercises

1. How many 2-digit numbers are

   [✓ (a)] even?

   (b) divisible by 7?

   (c) not divisible by 7?

2. How many 4-digit numbers are

   (a) divisible by 3?

   [✓ (b)] divisible by 5?

   (c) divisible by 3 and 5?

   (d) divisible by 3 or 5?

   (e) divisible by neither 3 nor 5?

[✓ 3.] A computer system requires a case-sensitive, alpha-numeric password containing 4 or 5 characters. How many possible passwords are there?

[✓ 4.] A computer system requires a case-sensitive, alpha-numeric password containing 5 digits. How many possible passwords are there if

   (a) you can repeat characters?

   (b) you cannot repeat characters?

   (c) you can repeat characters but the first character must be a letter and not a digit?

[✓ 5.] A computer system requires an eight-character, case-sensitive, alpha-numeric passwords.

   (a) How many possible passwords are there?

   (b) How many passwords are there that contain at least one digit?

   (c) How many passwords are there that contain at least one letter?

   (d) How many passwords are there that contain at least one digit and one letter?

6. A computer system requires a case-sensitive, alpha-numeric password containing six characters.

(a) How many passwords are there that contain no “A”s?

(b) How many passwords are there that contain no “a”s?

(c) How many passwords are there that contain no “A”s or “a”s?

7. For homework, Peter has assigned reading pages 25-37 inclusive. How many pages has he asked his class to read?

8. Gilles has assigned for homework all the odd questions from 7 to 89. How many homework questions has he assigned?

9. Canadian postal codes are of the form “letter-number-letter number-letter-number”. The first letter shows which province or territory is from, for a total of 13 letters allowable. The remaining letters can be any letter of the alphabet except for O and I. All numbers are allowed. How many possible Canadian postal codes are there?

10. How many days of the week

    (a) contain the letter “t”?

    (b) contain the letter “s”?

    (c) contain the letters “t” and “s”?

    (d) contain the letters “t” or “s”?

11. Pat is writing up systems of equations containing two variables. She will be using lower-case letters for her variables, but doesn’t want to use the letters “e”, “i”, and “o” (for obvious reasons!). How many possible letter combinations does she have to choose from?

12. The mythical Canadian province of Gondor has licence plates of the form “letter-letter number-number-number”. Because of an odd superstition, you cannot repeat a letter on the licence plate, but you can repeat a number. How many possible Gondorian licence plates are there?

8.1 ~~4.1.2~~ Answers

1. First, note that 2-digit numbers run from $10, 11, 12, \dots 99$.

   (a) The even ones are $10, 12, 14, \dots 98$. You can do the really short method to count them: __ __ – the first slot can have the digits 1-9 for 9 choices, and the second can only have 2, 4, 6, 8, or 0 for 5 choices. Then the total number is $9 \times 5 = 45$ numbers.

   (b) Unfortunately, you cannot use the above technique for dividing by 7, since 7 doesn't restrict the last digit. Instead, you have to note that the first 2-digit number that's divisible by 7 is 14, the next is 21, then 28, and so on. To find the last digit, you have to count backwards from 99 to find one that's divisible by 7. 99 does not divide evenly by 7, but with your calculator (sigh) you can quickly find that $98 \div 7 = 14$.

   So our sequence is $14, 21, 28, \dots 98$. This is just $2 \times 7, 3 \times 7, 4 \times 7, \dots 14 \times 7$. So there are $(\text{last} - \text{first} + 1) = 14 - 2 + 1 = 13$ numbers divisible by 7.

   (c) Total number of 2-digit numbers: last - first + 1 = 90 - 10 + 1 = 90. So the total number of 2-digit numbers **not** divisible by 7 is the total number minus the number that **are** divisible by 7. So, we get $90 - 13 = 77$ for our answer.

2. First, note that 4-digit numbers run from $1000, 1001, 1002, \dots 9999$.

   (a) The first number that's divisible by 3 is 1002, the next is 1005, then 1008, and so on up to 9999, which also divides evenly by 3.

   So our sequence is $1002, 1005, 1008, \dots 9999$. This is just $334 \times 3, 335 \times 3, 336 \times 3, \dots 3333 \times 3$. So there are $(\text{last} - \text{first} + 1) = 3333 - 334 + 1 = 3000$ numbers divisible by 3.

   (b) Numbers that divide evenly by 5 end in either 0 or 5. You can do the really short method to count them: the first slot can have the digits 1-9 for 9 choices, the second and third slots can have 0-9 for 10 choices and the second can only have 0 or 5 for 2 choices. Then the total number is $9 \times 10 \times 10 \times 2 = 1800$ numbers.

   (c) Numbers divisible by 3 **and** 5 must be divisible by 15. Looking at our sequence in a), we can see that 1005 must be the first number, then add 15 to get 1020, etc. Starting from 9999 and working downwards, we'll see that the first possibility is 9995, which doesn't divide, but 9990 does.

So our sequence is $1005, 1020, 1035, \dots 9990$. This is just $67 \times 15, 68 \times 15, 69 \times 15, \dots 666$. So the total number is $666 - 67 + 1 = 600$.

(d) $n(3 \text{ or } 5) = n(3) + n(5) - n(3 \text{ and } 5) = 3000 + 1800 - 600 = 4200$

(e) total number with 4-digits: $9999 - 1000 + 1 = 9000$ numbers Then the total divisible by neither 3 nor 5 is $9000 - 4200 = 4800$.

3. Case-sensitive, alpha-numeric passwords have $2 \times 26 + 10$ choices for characters, or 62 different possibilities. The number of passwords containing 4 characters is $62 \times 62 \times 62 \times 62 = 14,776,336$. The number of passwords containing 5 characters is $62 \times 62 \times 62 \times 62 \times 62 = 916,132,832$. The total number of passwords is then the sum of these two (since you can't have four **and** five at the same time), $= 930,909,168$.

4. (a) This is the same as in question #3: $916,132,832$.

(b) If you can't repeat, then you get 62 choices for the first one, 61 for the second, etc., to give $62 \times 61 \times 60 \times 59 \times 58 = 776,520,240$.

(c) If the first number must be a letter, then you only have 52 possibilities for the first slot: $52 \times 62 \times 62 \times 62 \times 62 = 768,369,472$.

5. (a) You have 62 choices for each slot, so result is $62^8 = 2.18 \times 10^{14}$.

(b) The number containing no digits is $52^8 = 5.35 \times 10^{13}$. So the number containing at least one digit is $2.18 \times 10^{14} - 5.35 \times 10^{13} = 1.65 \times 10^{14}$.

(c) The number containing **no** letters is 10^8. So the number containing at least one letter is $2.18 \times 10^{14} - 10^8 = 2.18 \times 10^{14}$ (essentially the same number, since 10^8 is so much smaller).

(d) The number containing at least one digit and one letter must be the total minus (the number containing no digits plus the number containing no letters). So we get $2.18 \times 10^{14} - 5.35 \times 10^{13} - 10^8 = 1.65 \times 10^{14}$ (very close to the answer to (b) – you'd have to write out a few more decimals to see the difference).

6. (a) If no "A"s are allowed, then we are constrained to 61 choices from our original 62. Then we'll get $61^6 = 5.15 \times 10^{10}$ passwords.

(b) This will again give us 61 choices for each character, or $61^6 = 5.15 \times 10^{10}$ passwords.

(c) Now, we're down to 60 choices, since we can't have "A" or "a". We then get $60^6 = 4.67 \times 10^{10}$ passwords.

7. Peter has assigned 25-37 pages, so $25, 26, 27, \ldots 37$. # pages = last − first + 1 = $37 - 25 + 1 = 13$ pages.

8. Gilles has assigned odd questions, so $7, 9, 11, \ldots 89$. It's a bit tricky to do the odd numbers, so I'm going to take all numbers from 7 to 89 and subtract the even numbers.

   total number from 7 to 89: 89 - 7 + 1 = 83

   even numbers: $8, 10, 12, \ldots 88$ is the same as $4 \times 2, 5 \times 2, 6 \times 2, \ldots 44 \times 2$. So we get $44 - 4 + 1 = 41$ even numbers

   odd numbers= $83 - 41 = 42$ odd numbered questions

9. first letter: 13 choices, second and third letters: 24 choices, all numbers: 10 choices

   So we get $\underline{\ 13\ }\ \underline{\ 10\ }\ \underline{\ 24\ }\quad \underline{\ 10\ }\ \underline{\ 24\ }\ \underline{\ 10\ } = 13 \times 10 \times 24 \times 10 \times 24 \times 10 = 7,488,000$ possible postal codes. (Note that since postal codes reference a geographical area and not a group of people, we're not likely to run out any time soon!)

10. (a) Counting on my fingers, I get that Tuesday, Thursday, and Saturday contain the letter "t" for a total of 3.

    (b) Counting on my fingers, I get that all days except for Monday and Friday have "s" in them for a total of 5.

    (c) I see that all of the days containing "t" also contain "s" for a total of 3.

    (d) Using my counting rules, I get that $n(\text{t or s}) = n(\text{t}) + n(\text{s}) - n(\text{t \& s}) = 3 + 5 - 3 = 5$.

11. I will not be using 3 letters, leaving 23 lower-case letters to choose from. But they have to be different, so I'll get 23×22 choices = 506 choices.

12. $\underline{\ 26\ }\ \underline{\ 25\ }\ \underline{\ 10\ }\ \underline{\ 10\ }\ \underline{\ 10\ } = 26 \times 25 \times 10^3 = 650,000$.

8.2 ~~4.2~~ Probability

Suppose you were to roll a six-sided die (one die, two dice – die is the singular of dice). What's the probability of rolling a 1? Well, if the die is fair and all six numbers are equally likely, then the probability of getting any one number is just 1/6. This is the idea of classical probability: if all outcomes are equally likely, then the probability of event E happening is just the number of outcomes in which E happens, $n(E)$, divided by the total number of outcomes n: $P(E) = \frac{n(E)}{n}$

Example: If you roll two four-sided dice, what's the probability of rolling a total of 5?

Answer: The brute force method involves writing out all possible outcomes.

11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44

Then, assuming that the dice are fair, there are sixteen equally likely outcomes, and four of them – 41, 32, 23, and 14 – lead to a total of 5. Then $P(\text{total of } 5) = 4/16 = 1/4$ or 25%.

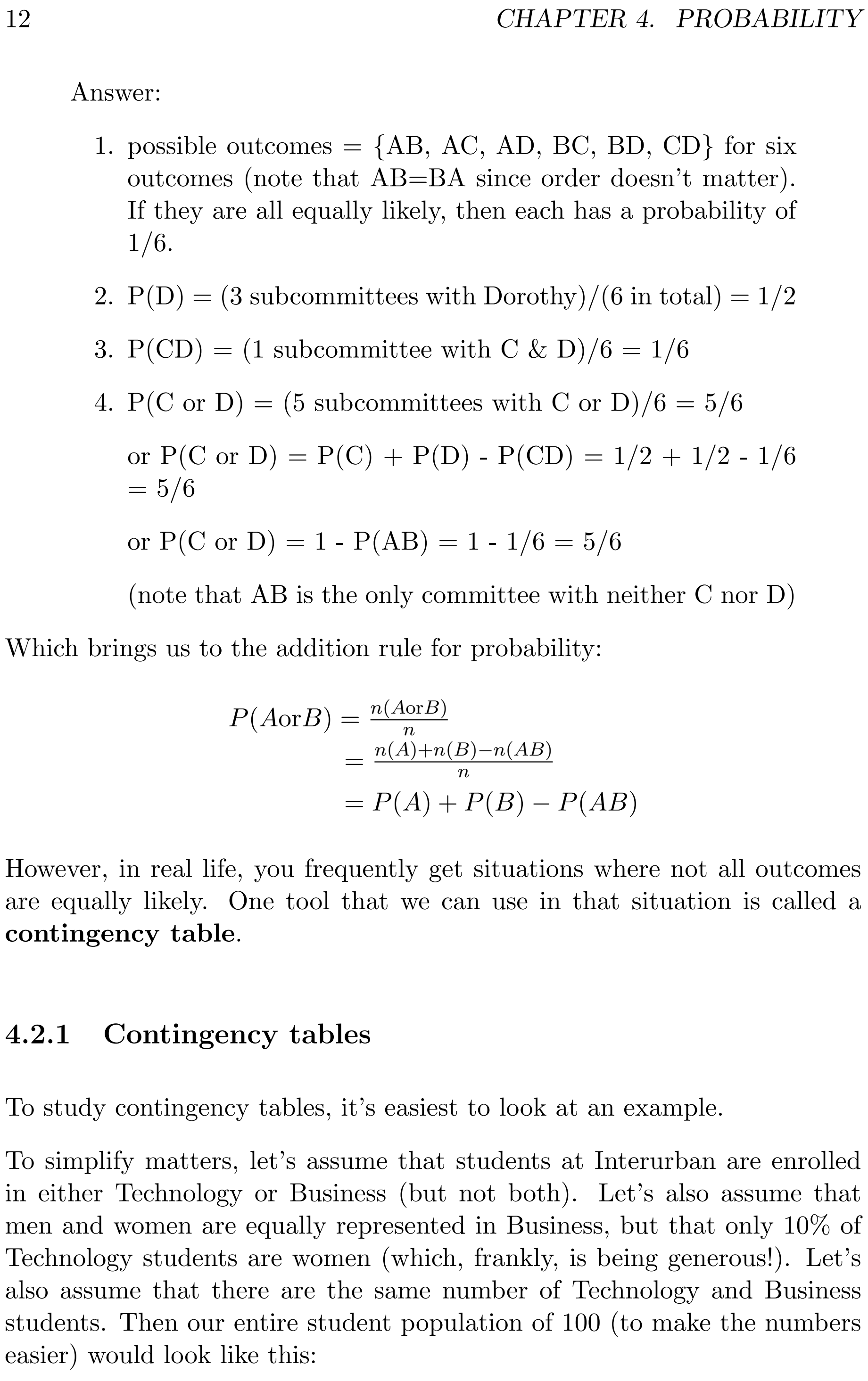
Example: The Saanich city council has four members: Alex, Barbara, Charlie, and Dorothy. Two of these members are to be selected to form a subcommittee to study the city's traffic problems.

1. How many different subcommittees are possible? What probability would you assign to each one if there is an equal chance of selecting each council member?

2. What is the probability that Dorothy is a member of the committee?

3. What is the probability that Charlie and Dorothy are both selected?

4. What is the probability either Charlie or Dorothy or both are selected?



12 CHAPTER 4. PROBABILITY

Answer:

1. possible outcomes = {AB, AC, AD, BC, BD, CD} for six outcomes (note that AB=BA since order doesn't matter). If they are all equally likely, then each has a probability of 1/6.
2. P(D) = (3 subcommittees with Dorothy)/(6 in total) = 1/2
3. P(CD) = (1 subcommittee with C & D)/6 = 1/6
4. P(C or D) = (5 subcommittees with C or D)/6 = 5/6

   or P(C or D) = P(C) + P(D) - P(CD) = 1/2 + 1/2 - 1/6 = 5/6

   or P(C or D) = 1 - P(AB) = 1 - 1/6 = 5/6

   (note that AB is the only committee with neither C nor D)

Which brings us to the addition rule for probability:

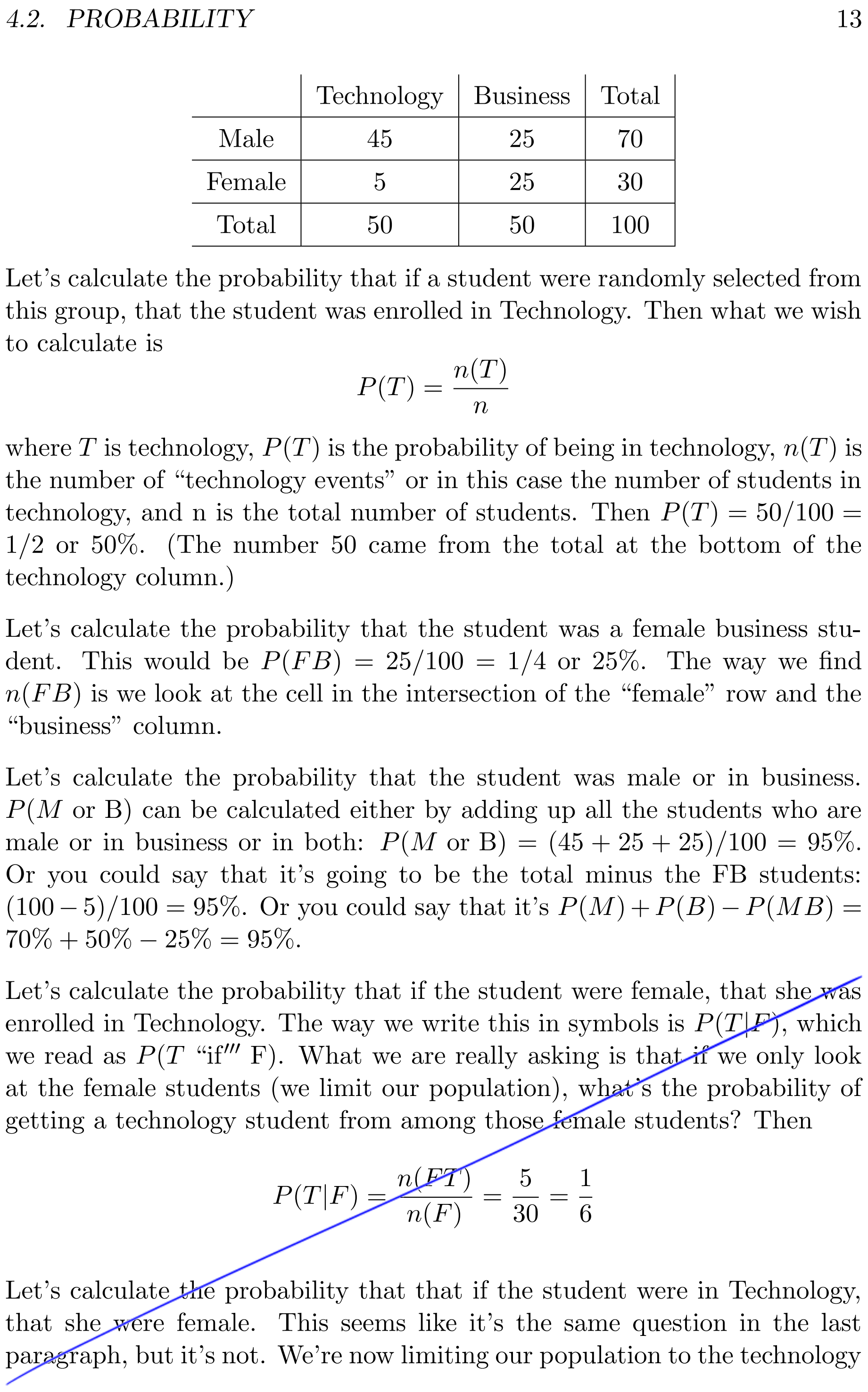
$$\begin{aligned} P(A \text{or} B) &= \tfrac{n(A\text{or}B)}{n} \\ &= \tfrac{n(A)+n(B)-n(AB)}{n} \\ &= P(A)+P(B)-P(AB) \end{aligned}$$

However, in real life, you frequently get situations where not all outcomes are equally likely. One tool that we can use in that situation is called a **contingency table**.

4.2.1 Contingency tables

To study contingency tables, it's easiest to look at an example.

To simplify matters, let's assume that students at Interurban are enrolled in either Technology or Business (but not both). Let's also assume that men and women are equally represented in Business, but that only 10% of Technology students are women (which, frankly, is being generous!). Let's also assume that there are the same number of Technology and Business students. Then our entire student population of 100 (to make the numbers easier) would look like this:



	Technology	Business	Total
Male	45	25	70
Female	5	25	30
Total	50	50	100

Let's calculate the probability that if a student were randomly selected from this group, that the student was enrolled in Technology. Then what we wish to calculate is

$$P(T) = \frac{n(T)}{n}$$

where T is technology, $P(T)$ is the probability of being in technology, $n(T)$ is the number of "technology events" or in this case the number of students in technology, and n is the total number of students. Then $P(T) = 50/100 = 1/2$ or 50%. (The number 50 came from the total at the bottom of the technology column.)

Let's calculate the probability that the student was a female business student. This would be $P(FB) = 25/100 = 1/4$ or 25%. The way we find $n(FB)$ is we look at the cell in the intersection of the "female" row and the "business" column.

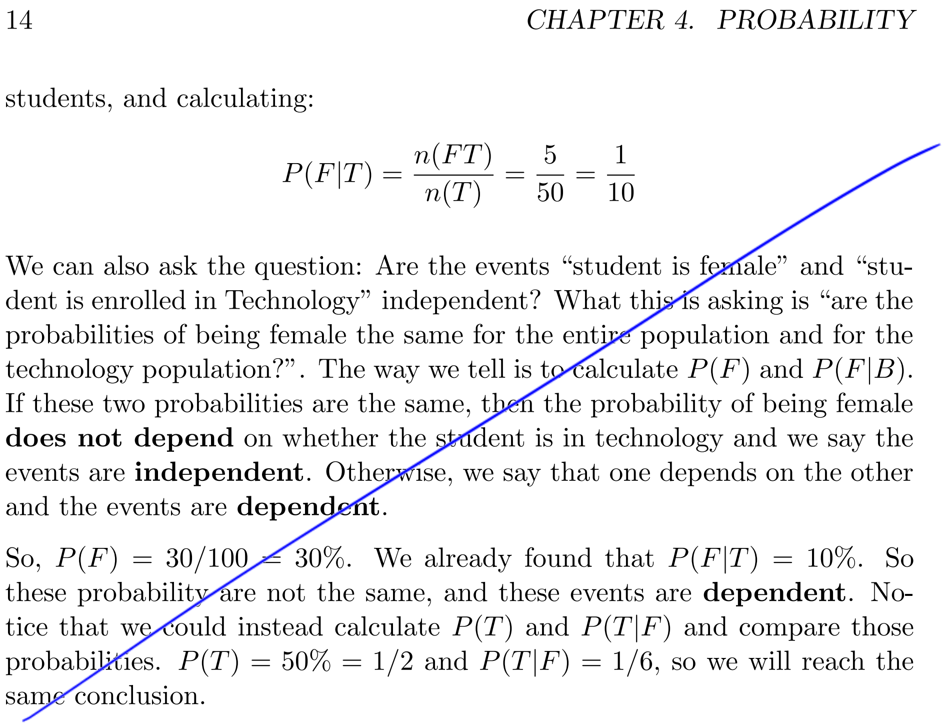
Let's calculate the probability that the student was male or in business. $P(M$ or B) can be calculated either by adding up all the students who are male or in business or in both: $P(M$ or B) $= (45 + 25 + 25)/100 = 95\%$. Or you could say that it's going to be the total minus the FB students: $(100 - 5)/100 = 95\%$. Or you could say that it's $P(M) + P(B) - P(MB) = 70\% + 50\% - 25\% = 95\%$.

~~Let's calculate the probability that if the student were female, that she was enrolled in Technology. The way we write this in symbols is $P(T|F)$, which we read as $P(T$ "if$'''$ F). What we are really asking is that if we only look at the female students (we limit our population), what's the probability of getting a technology student from among those female students? Then~~

~~$P(T|F) = \frac{n(FT)}{n(F)} = \frac{5}{30} = \frac{1}{6}$~~

~~Let's calculate the probability that that if the student were in Technology, that she were female. This seems like it's the same question in the last paragraph, but it's not. We're now limiting our population to the technology~~

we are omitting conditionals/ independence



students, and calculating:

$$P(F|T) = \frac{n(FT)}{n(T)} = \frac{5}{50} = \frac{1}{10}$$

We can also ask the question: Are the events "student is female" and "student is enrolled in Technology" independent? What this is asking is "are the probabilities of being female the same for the entire population and for the technology population?". The way we tell is to calculate $P(F)$ and $P(F|B)$. If these two probabilities are the same, then the probability of being female **does not depend** on whether the student is in technology and we say the events are **independent**. Otherwise, we say that one depends on the other and the events are **dependent**.

So, $P(F) = 30/100 = 30\%$. We already found that $P(F|T) = 10\%$. So these probability are not the same, and these events are **dependent**. Notice that we could instead calculate $P(T)$ and $P(T|F)$ and compare those probabilities. $P(T) = 50\% = 1/2$ and $P(T|F) = 1/6$, so we will reach the same conclusion.

8.2 ~~4.2~~.2 Exercises

1. A fair twelve-sided die is rolled. What is the probability that the roll is

   (a) a 7?

   (b) even?

   (c) greater than 5?

   (d) not a 7?

   (e) a 1 or a 2?

2. Two four-sided dice are rolled. What is the probability that the roll

   (a) results in the same number on both dice?

   (b) results in different numbers on both dice?

   (c) has a sum of 6?

   (d) has at least one die rolling a 3?

3. An individual is presented with three different glasses of soft drink, labeled A, B, and C. He is asked to taste all three and then list them in order of preference. Suppose that the same soft drink has actually been put into all three glasses.

   (a) How many outcomes are there in this experiment? What probability would you assign to each one?

   (b) What is the probability that A is ranked first?

   (c) What is the probability that either B or C is ranked first?

   (d) What is the probability that A is ranked first and B is ranked last?

4. Your ATM/debit card has a four-digit PIN number associated with it. If there are no restrictions on what digits or what order you can pick them, then

   (a) how many PIN numbers are possible?

   (b) what is the probability that someone could guess your PIN randomly?

(c) if that person saw you input the first two digits when you were at the grocery checkout counter, what are their chances of guessing your PIN correctly now?

Complete the following exercises involving contingency tables.

5. One hundred students each from the Computing Systems Technology program and from the English department were asked who is the greatest fictional wizard ever, with the following results.

	Gandalf	Dumbledore	total
CST	90	10	
English	40	60	
total			

(a) Calculate $P(G)$.

~~(b)~~ Calculate $P(C|G)$.

~~(c)~~ Calculate $P(G|C)$.

(d) Calculate $P(E \text{ or } \mathrm{D})$.

~~6.~~ A sampling of CST faculty and students were asked what operating system they used on their home computer, with the following results.

	Windows	Linux
faculty	6	2
students	24	8

(a) What's the probability that a random CST user (faculty or student) will have Linux on their home machine?

(b) What's the probability that a random CST student will have Linux on their home machine?

(c) Are the events "student" and "Linux user' independent?

7. One thousand television watchers from BC and Alberta were asked if they watched the Rick Mercer Report on CBC with the following results.

	Yes	No
BC	500	500
AB	250	750

(a) What's the probability that one of these people, when selected randomly, is from BC or watches the RMR?

(b) What's the probability that one of these people, when selected randomly, is from BC and watches the RMR?

(c) What's the probability that one of these people, when selected randomly, is from Alberta and does not watch the RMR?

~~(d)~~ What's the probability that a Rick Mercer watcher is from BC?

~~(e)~~ What's the probability that a British Columbian watches Rick Mercer?

~~8.~~ A roving reporter surveyed all of the patrons inside the Starbucks and the Moka House coffee houses in Cook Street Village (it was a slow news day). The beverage each patron was drinking was noted and summarized in the following table.

	coffee	tea	other
Starbucks	45	9	6
Moka House	30	8	2

(a) Are the events "drinking coffee" and "Starbucks" independent?

(b) Are the events "tea" and "Moka House" independent?

9. StatsCan surveyed one hundred Canadians and found that 60 of them exercise regularly, 75 of them eat healthy diets, and 45 of them do both. Complete the following contingency table using the above information

	exercise regularly	don't exercise regularly	total
healthy diet			
unhealthy diet			
total			

(a) If one of these Canadians is selected randomly, what is the probability that this person exercises regularly but does not eat a healthy diet?

(b) If one of these Canadians is selected randomly, what is the probability that this person exercises regularly or eats a healthy diet?

~~(c)~~ Is eating a healthy diet independent of exercising regularly for this sample of Canadians?

8.2 ~~4.2.3~~ Answers

1. A fair twelve-sided die is rolled.

   (a) $P(7) = 1/12$ (only one way to get a 7, and there are 12 outcomes)

   (b) Even numbers from 1 to 12: 2, 4, 6, 8, 10, 12, so six possibilities out of 12 outcomes. $P(even) = 6/12 = 1/2$. (Or you could note that exactly half of the outcomes gave an even number to get an even shorter solution.)

   (c) $P(>5) = P(6 \text{ or } 7 \text{ or } 8 \text{ or } 9 \text{ or } 10 \text{ or } 11 \text{ or } 12) = 7/12$

   (d) $P(\text{not } 7) = 1 - \text{P}(7) = 11/12$

   (e) $P(1 \text{ or } 2) = 2/12 = 1/6$

2. I'm going to use the brute force method here and list all possible rolls:

11	12	13	14
21	22	23	24
31	32	33	34
41	42	43	44

   (a) We can see that there are sixteen possibilities in total, and four of them will result in the same number on both dice, so $P(\text{both same}) = 4/16 = 1/4$.

   (b) $P(\text{different}) = 1 - \text{P(same)} = 3/4$.

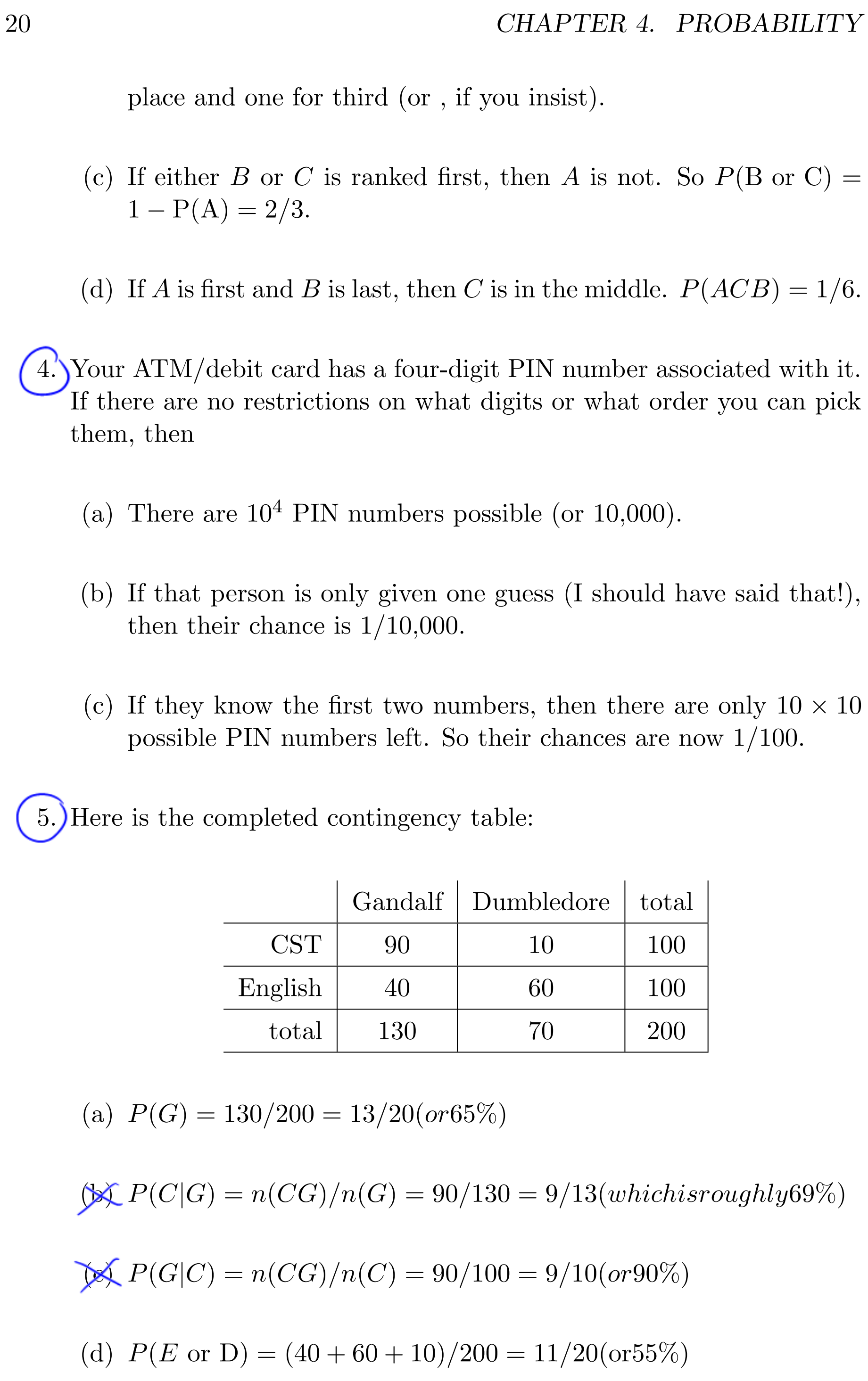
   (c) $P(\text{sum of } 6) = 3/16$ (need 42, 33, or 24).

   (d) You can count them up to find $P(\text{at least one } 3) = 7/16$. [Or you could say that's $P(\text{at least one } 3) = 1 - \text{P(no 3s)}$. And the number of rolls with no 3s is $\underline{3}\ \underline{3} = 9$ possibilities, so then you'd get $1 - 9/16 = 7/16$.]

3. We should note that the order of items matters in this question.

   (a) You can either just calculate or list the possible outcomes: {ABC, ACB, BAC, BCA, CAB, CBA}. Since it's the same soft drink in each glass, the lists should all be equally probable at 1/6 each.

   (b) Only 2 of the 6 outcomes have A first, so $P(\text{A first}) = 2/6 = 1/3$. Or you could say that if A is first, there are two choices for second



place and one for third (or , if you insist).

(c) If either B or C is ranked first, then A is not. So $P(\text{B or C}) = 1 - \text{P(A)} = 2/3$.

(d) If A is first and B is last, then C is in the middle. $P(ACB) = 1/6$.

4. Your ATM/debit card has a four-digit PIN number associated with it. If there are no restrictions on what digits or what order you can pick them, then

(a) There are 10^4 PIN numbers possible (or 10,000).

(b) If that person is only given one guess (I should have said that!), then their chance is 1/10,000.

(c) If they know the first two numbers, then there are only 10×10 possible PIN numbers left. So their chances are now 1/100.

5. Here is the completed contingency table:

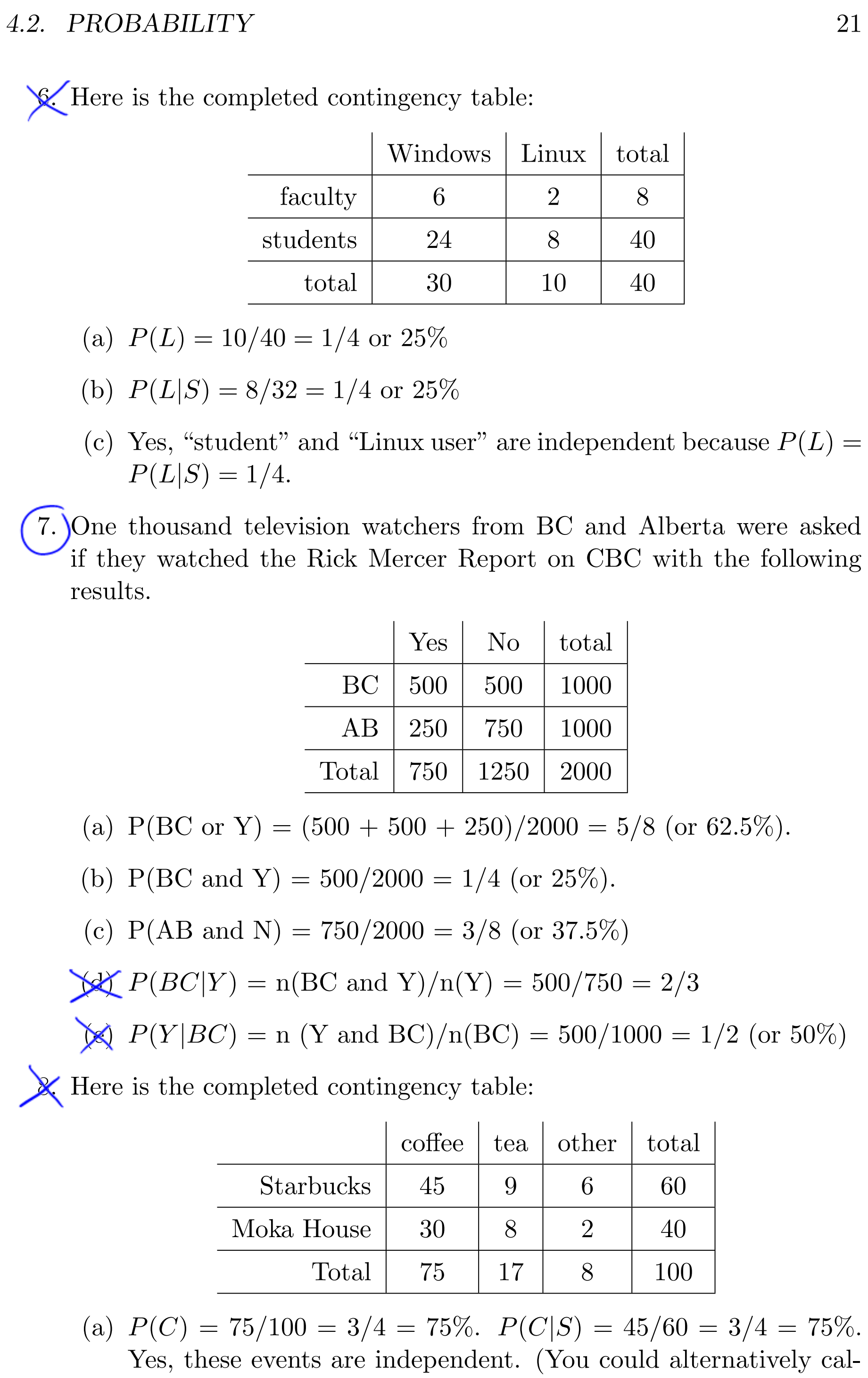
	Gandalf	Dumbledore	total
CST	90	10	100
English	40	60	100
total	130	70	200

(a) $P(G) = 130/200 = 13/20(or65\%)$

~~(b)~~ $P(C|G) = n(CG)/n(G) = 90/130 = 9/13(whichisroughly69\%)$

~~(c)~~ $P(G|C) = n(CG)/n(C) = 90/100 = 9/10(or90\%)$

(d) $P(E \text{ or D}) = (40 + 60 + 10)/200 = 11/20(\text{or}55\%)$



6. Here is the completed contingency table:

	Windows	Linux	total
faculty	6	2	8
students	24	8	40
total	30	10	40

   (a) $P(L) = 10/40 = 1/4$ or 25%

   (b) $P(L|S) = 8/32 = 1/4$ or 25%

   (c) Yes, "student" and "Linux user" are independent because $P(L) = P(L|S) = 1/4$.

7. One thousand television watchers from BC and Alberta were asked if they watched the Rick Mercer Report on CBC with the following results.

	Yes	No	total
BC	500	500	1000
AB	250	750	1000
Total	750	1250	2000

   (a) P(BC or Y) = (500 + 500 + 250)/2000 = 5/8 (or 62.5%).

   (b) P(BC and Y) = 500/2000 = 1/4 (or 25%).

   (c) P(AB and N) = 750/2000 = 3/8 (or 37.5%)

   (d) $P(BC|Y)$ = n(BC and Y)/n(Y) = 500/750 = 2/3

   (e) $P(Y|BC)$ = n (Y and BC)/n(BC) = 500/1000 = 1/2 (or 50%)

8. Here is the completed contingency table:

	coffee	tea	other	total
Starbucks	45	9	6	60
Moka House	30	8	2	40
Total	75	17	8	100

   (a) $P(C) = 75/100 = 3/4 = 75\%$. $P(C|S) = 45/60 = 3/4 = 75\%$. Yes, these events are independent. (You could alternatively calculate $P(S) = 60\%$ and $P(S|C) = 60\%$ to reach the same conclusion.)

(b) $P(T) = 17/100 = 17\%$. $P(T|M) = 8/40 = 1/5 = 20\%$. As these are not the same, the events are dependent.

9. (a) Here is the completed contingency table:

	exercise regularly	don't exercise regularly	total
healthy diet	45	30	75
unhealthy diet	15	10	25
total	60	40	100

(b) $P(E\,\overline{H}) = 15/100 = 15\%$.

(c) $P(E \text{ or H}) = (15 + 45 + 30)/100 = 90/100 = 90\%$.

(or $P(E \text{ or } H) = 1 - P(\overline{E}\,\overline{H}) = 1 - 10/100 = 90\%$)

~~(d)~~ $P(H) = 75/100 = 75\%$. $P(H|E) = 45/60 = 75\%$. Since these probabilities are the same, eating a healthy diet is independent of exercising regularly for this sample of Canadians.

Reading for Section 8.3: The following reading is excerpted from:

> Triola et al. Elementary Statistics. 3rd Canadian edition, Pearson, 2011, pages 174-178, 183, 819.

Section 8.3

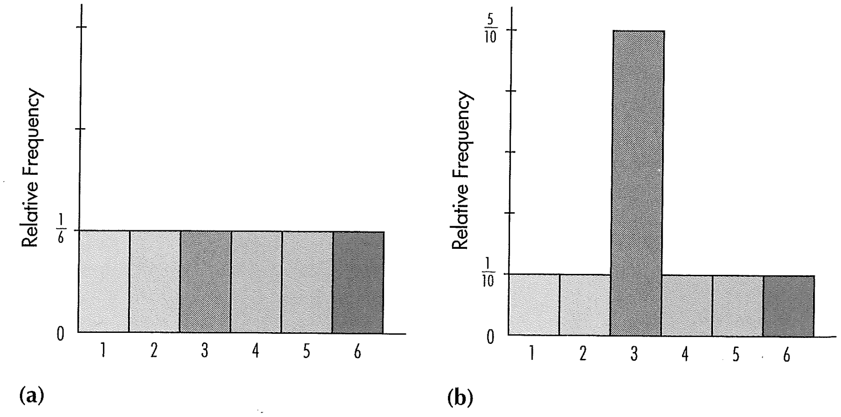
4-1 Overview

In Chapter 2 we saw that we could explore and describe a set of data by using graphs (such as a histogram or boxplot), measures of central tendency (such as the mean), and measures of variation (such as the standard deviation). In Chapter 3 we discussed the basic principles of probability theory. In this chapter we combine those concepts as we develop probability distributions that describe what will *probably* happen instead of what actually did happen. In Chapter 2 we constructed frequency tables and histograms using *observed* real scores, but in this chapter we will construct probability distributions by presenting possible outcomes along with the relative frequencies we *expect*, given an understanding of the relevant circumstances.

Suppose that a casino manager suspects cheating at a dice table. He or she can compare the relative frequency distribution of the actual sample outcomes to a theoretical model that describes the frequency distribution likely to occur with a fair die. A fair die should have a relative frequency histogram similar to the one shown in Figure 4-1(a), so a relative frequency histogram looking like Figure 4-1(b) may arouse suspicion.

In Figure 4-1(a) we see relative frequencies based not on actual outcomes, but on our knowledge of the probabilities for the outcomes of a fair die. In essence, we can describe the frequency table and histogram for a die rolled an infinite number of times. With this knowledge of the population of outcomes, we are able to determine important characteristics, such as the mean and standard deviation. The remainder of this book and the very core of inferential statistics are based on some knowledge of probability distributions. We begin by examining the concept of a random variable, then we consider important distributions that have many real applications.

Figure 4-1
Histograms of Dice Outcomes for (a) a Fair Die and (b) a Loaded Die



4-2 Random Variables

In this section we discuss random variables, probability distributions, and procedures for finding the mean and standard deviation for a probability distribution. We will see that a random variable has a numeric value for each outcome of a procedure, and a probability distribution associates a probability value with each outcome of a procedure.

DEFINITION

A **random variable** is a variable (typically represented by x) that has a single numerical value (determined by chance) for each outcome of a procedure.

Examples of random variables are the following:

x = The number of women among 10 newly hired employees
x = The number of students absent from statistics class today
x = The height (in cm) of a randomly selected adult male

The word *random* is used to remind us that we don't usually know what the value of x is until we observe or perform a procedure.

EXAMPLE

We randomly select 7 vehicles involved in accidents on public roadways and count how many of these are commercial vehicles. If we let the random variable represent the number of commercial vehicles among 7, this procedure has possible outcomes of 0, 1, 2, 3, 4, 5, 6, 7. (Remember, 0 represents no vehicles are commercial vehicles, 1 represents 1 vehicle is a commercial vehicle, and so on.) The variable is random in the sense that we do not know the value until after the 7 vehicles have been selected.

In Section 1–2 we made a distinction between discrete and continuous data. Random variables may also be discrete or continuous, and the following two definitions are consistent with those given in Section 1–2. This chapter deals with discrete random variables, but the following chapters will deal with continuous random variables.

DEFINITIONS

A **discrete random variable** has either a finite number of values or a countable number of values; that is, they result from a counting process.

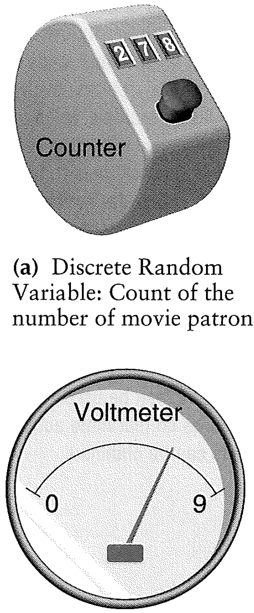
A **continuous random variable** has infinitely many values, and those values can be associated with measurements on a continuous scale in such a way that there are no gaps or interruptions.

STATISTICAL PLANET

Sensitive Surveys

Survey respondents are sometimes reluctant to honestly answer questions on a sensitive topic, such as employee theft or sex. Stanley Warner (York University, Ontario) devised a scheme that leads to more accurate results in such cases. As an example, ask employees if they stole within the past year and also ask them to flip a coin. The employees are instructed to answer no if they didn't steal and the coin turns up heads. Otherwise, they should answer yes. The employees are more likely to be honest because the coin flip helps protect their privacy. Probability theory can then be used to analyze responses so that more accurate results can be obtained.

Figure 4-2
Discrete and Continuous Random Variables



(a) Discrete Random Variable: Count of the number of movie patrons.

(b) Continuous Random Variable: The measured voltage of a smoke detector battery.

EXAMPLE

1. The count of the number of patrons viewing a movie is a whole number and is therefore a discrete random variable. The counting device shown in Figure 4-2(a) is capable of indicating only whole numbers. It can therefore be used to obtain values for a discrete random variable.
2. The measure of voltage for a smoke detector battery can be any value between 0 volts and 9 volts and is therefore a continuous random variable. The voltmeter depicted in Figure 4-2(b) is capable of indicating values on a continuous scale, so it can be used to obtain values for a continuous random variable.

After we identify possible values of a random variable, we can often identify a probability for each of those values. When we know all values of a random variable along with their corresponding probabilities, we have a probability distribution, defined as follows.

DEFINITION

A **probability distribution** gives the probability for each value or range of values of the random variable.

EXAMPLE

Suppose that 20% of all registered motor vehicles are commercial vehicles, and that all vehicles have the same chance of being involved in an accident. If we let the random variable *x* represent the number of commercial vehicles among 7 randomly selected vehicles involved in accidents, then the probability distribution can be described by Table 4-1. (In Section 4-3 we will see how the probabilities in Table 4-1 are found.) In the table, we see that the probability that 0 of 7 vehicles in accidents is a commercial vehicle is 0.210, the probability that 1 of 7 is commercial is 0.367, and so on. The values denoted by 0+ represent positive probabilities that are so small that they become 0.000 when rounded to three decimal places. We avoid using 0.000 because it incorrectly suggests an impossible event with a probability of 0.

Table 4-1 **Probability Distribution for Number of Commercial Vehicles Among Seven Vehicles Involved in Accidents**

x	*P*(*x*)
0	0.210
1	0.367
2	0.275
3	0.115
4	0.029
5	0.004
6	0+
7	0+

Graphs

There are various ways to graph a probability distribution, but we present only the **probability histogram.** Figure 4-3 is a probability histogram that resembles the relative frequency histogram from Chapter 2, but the vertical scale delineates *probabilities* instead of relative frequencies based on actual sample results.

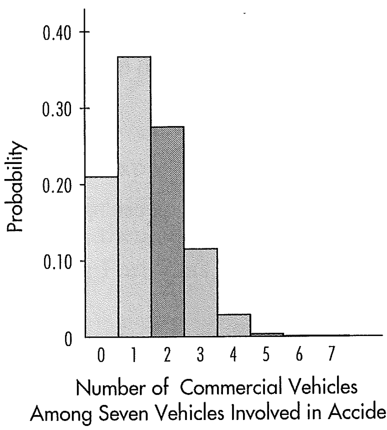


Figure 4-3
Probability Histogram for Number of Commercial Vehicles Among Seven Vehicles Involved in Accidents

In Figure 4-3, note that along the horizontal axis, the values of 0, 1, 2, . . . , 7 are located at the centres of the rectangles. This implies that the rectangles are each 1 unit wide, so the areas of the rectangles are 0.210, 0.367, and so on. When the total area of such a probability histogram is 1, the *probabilities* are equal to the corresponding rectangular *areas*. We will see in Chapter 5 and later chapters that this correspondence between area and probability is very useful in statistics.

For discrete data, every probability distribution must satisfy the following two requirements.

REQUIREMENTS FOR A PROBABILITY DISTRIBUTION (FOR DISCRETE DATA)

1. $\Sigma P(x) = 1$ where x assumes all possible, distinct values
2. $0 \leq P(x) \leq 1$ for every value of x

The first requirement states that the sum of the individual probabilities must equal 1 and is based on the addition rule for mutually exclusive events. The values of the random variable x represent all possible events in the entire sample space, so we are certain (with probability 1) that one of the events will occur. We use simple addition of the values of $P(x)$ because the different values of x correspond to events that are mutually exclusive. In Table 4-1 we can see that the individual probabilities do result

STATISTICAL PLANET

Prophets for Profits

Many books and computer programs claim to be helpful in predicting winning lottery numbers. Some use the theory that particular numbers are "due" (and should be selected) because they haven't been coming up often; others use the theory that some numbers are "cold" (and should be avoided) because they haven't been coming up often; and still others use astrology, numerology, or dreams. Because selections of winning lottery number combinations are independent events, such theories are worthless. A valid approach is to choose numbers that are "rare" in the sense that they are not selected by other people, so that if you win, you will not need to share your jackpot with many others. For this reason, the combination of 1, 2, 3, 4, 5, and 6 is a bad choice because many people use it, whereas 12, 17, 18, 33, 40, 46 is a much better choice (at least until it was published in this book).

in a sum of 1. Also, the probability rule (see Section 3–2) stating that $0 \leq P(A) \leq 1$ for any event A implies that $P(x)$ must be between 0 and 1 for any value of x. Again, refer to Table 4-1 and note that each individual value of $P(x)$ does fall between 0 and 1. Because Table 4-1 does satisfy both of these requirements, it is an example of a probability distribution. A probability distribution may be described by a table, such as Table 4-1, or a graph, such as Figure 4-3, or a formula, as in the following two examples.

EXAMPLE

Can the formula $P(x) = x/5$ (where x can take on the values of 1, 2, 3) determine a probability distribution?

SOLUTION

If a probability distribution is determined, it must conform to the preceding two requirements. But

$$\begin{aligned}\Sigma P(x) &= P(1) + P(2) + P(3)\\ &= \frac{1}{5} + \frac{2}{5} + \frac{3}{5}\\ &= \frac{6}{5} \quad \text{(showing that } \Sigma P(x) \neq 1)\end{aligned}$$

Because the first requirement is not satisfied, we conclude that $P(x)$ given in this example cannot be determining a probability distribution.

EXAMPLE

Can the formula $P(x) = x/3$ (where x can be 1 or 2) determine a probability distribution?

SOLUTION

For the given function, we find that $P(1) = 1/3$ and $P(2) = 2/3$ so that

1. $\Sigma P(x) = \frac{1}{3} + \frac{2}{3} = \frac{3}{3} = 1$
2. Each of the $P(x)$ values is between 0 and 1 inclusive.

Because the two requirements are both satisfied, the $P(x)$ function given in this example is a possible probability distribution.

Section 8.3: Exercises

4-2 Exercises A: Basic Skills and Concepts

In Exercises 1–4, identify the given random variable as being discrete or continuous.

1. The weight of a randomly selected textbook
2. The cost of a randomly selected textbook
3. The number of eggs a hen lays
4. The amount of milk obtained from a cow

In Exercises 5–12, determine whether a probability distribution is given. In those cases where a probability distribution is not described, identify the requirement that is not satisfied. ~~*In those cases where a probability distribution is described, find its mean, variance, and standard deviation.*~~

5. When a household is randomly selected, the probability distribution for the number x of automobiles owned is as described in the accompanying table.

x	$P(x)$
0	0.011
1	0.394
2	0.380
3	0.215

6. If your college hires the next 4 employees without regard to gender, and the pool of applicants is large with an equal number of men and women, then the probability distribution for the number x of women hired is described in the accompanying table.

x	$P(x)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625

7. Statistics Canada found that English is the mother tongue of 62.6% of the residents in the Greater Toronto Area (GTA). A small survey of 8 residents in one area of the GTA will be taken, so the accompanying table describes the probability distribution for the number of residents (among the 8 randomly selected residents) whose mother tongue is English.

x	$P(x)$
0	0.000
1	0.002
2	0.012
3	0.053
4	0.147
5	0.261

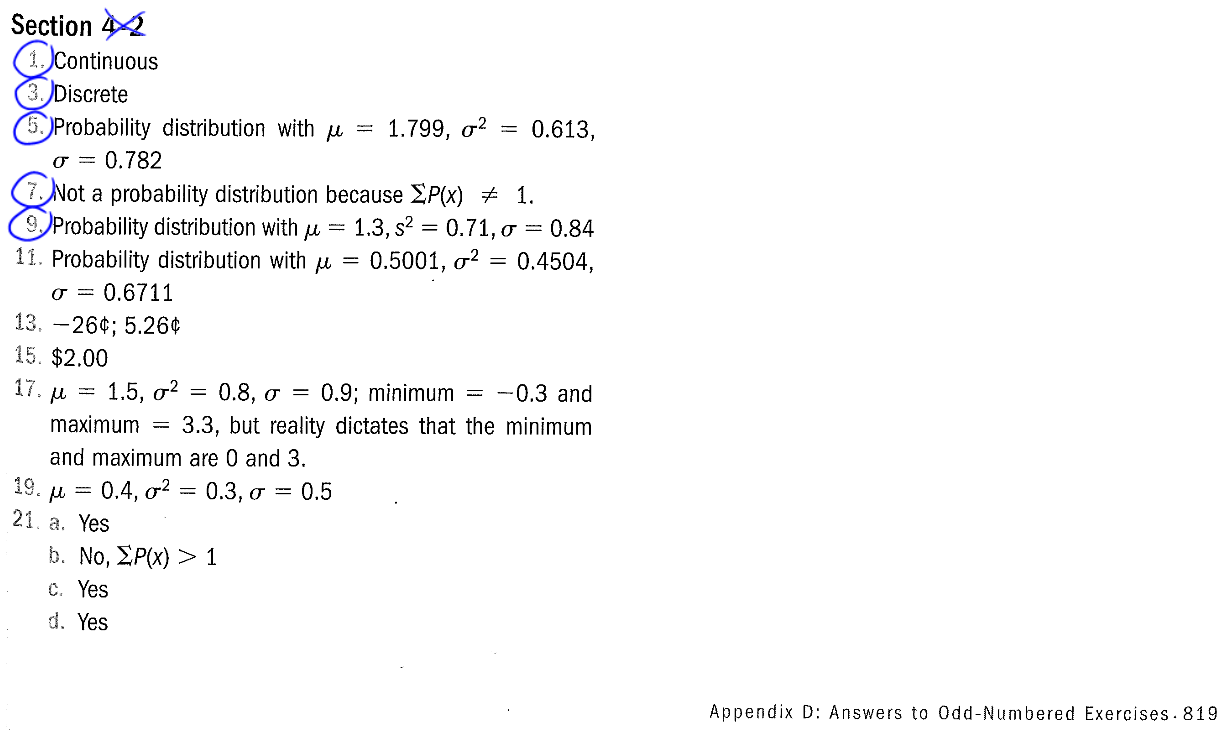
8. In assessing credit risks, a bank investigates the number of credit cards people have. With x representing the number of credit cards adults have, the accompanying table describes the probability distribution for a population of applicants (based on data from Maritz Marketing Research, Inc.).

x	$P(x)$
0	0.26
1	0.16
2	0.12
3	0.09
4	0.07
5	0.09
6	0.07
7	0.14

9. The Willford Printing Company conducted a survey to monitor daily absenteeism at the plant. The accompanying table describes the probability distribution for one department at the plant, where x represents the number of employees absent in the department.

x	$P(x)$
0	0.15
1	0.50
2	0.25
3	0.10

Section 8.3: Answers



Section ~~4-2~~

1. Continuous
3. Discrete
5. Probability distribution with $\mu = 1.799$, $\sigma^2 = 0.613$, $\sigma = 0.782$
7. Not a probability distribution because $\Sigma P(x) \neq 1$.
9. Probability distribution with $\mu = 1.3, s^2 = 0.71, \sigma = 0.84$
11. Probability distribution with $\mu = 0.5001$, $\sigma^2 = 0.4504$, $\sigma = 0.6711$
13. −26¢; 5.26¢
15. $2.00
17. $\mu = 1.5$, $\sigma^2 = 0.8$, $\sigma = 0.9$; minimum $= -0.3$ and maximum $= 3.3$, but reality dictates that the minimum and maximum are 0 and 3.
19. $\mu = 0.4, \sigma^2 = 0.3, \sigma = 0.5$
21. a. Yes
    b. No, $\Sigma P(x) > 1$
    c. Yes
    d. Yes

Chapter 9

Sampling Distributions

Reading for Chapter 9: The following reading is excerpted from:

> Triola et al. Elementary Statistics. 3rd Canadian edition, Pearson, 2011, pages 226-227, 229-239, 241-247, 249-264, 266-269, 821-822.

9.1

~~5-1~~ Overview

In Chapter 4 we introduced the concept of the *random variable* as a variable having a single numerical value (determined by chance) for each outcome of an experiment. We noted that a *probability distribution* gives the probability for each value of the random variable. Chapter 4 was concerned only with *discrete* random variables, such as those in binomial distributions, that have a finite number of possible values. The number of quarters produced by the Royal Canadian Mint each day is an example of a discrete random variables. There are also many different *continuous* probability distributions, such as the weights of the quarters produced at the mint. Distributions can be either discrete or continuous, and they can be described by their *shape*, such as a bell shape.

In Section 4–2 we identified two requirements for a discrete probability distribution: (1) $\sum P(x) = 1$, and (2) $0 \leq P(x) \leq 1$ for all values of x. Also in Section 4–2, we stated that the graph of a discrete probability distribution is called a *probability histogram*. The graph of a continuous probability distribution, such as Figure 5-1, is called a *density curve*, and it must satisfy two properties similar, but not identical, to the requirements for discrete probability distributions, as listed in the folllowing definition.

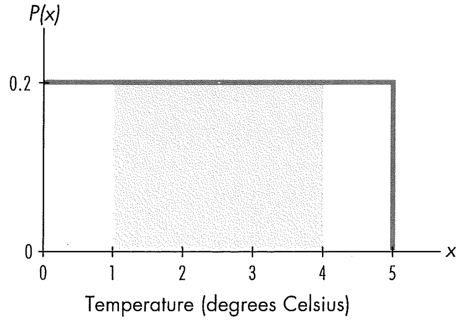
DEFINITION

A **density curve** or **probability density function** is a graph of a continuous probability distribution. It must satisfy the following properties:

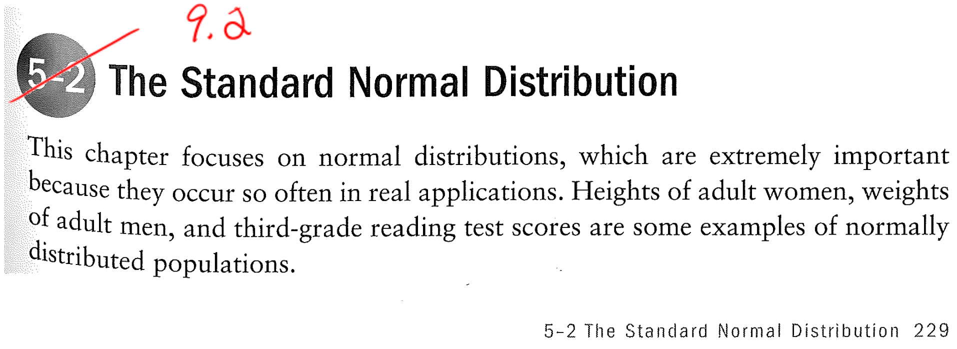
1. The total area under the curve must be 1.
2. Every point on the curve must have a vertical height that is 0 or greater.

Figure 5-1 is an example of a density curve, specifically a uniform distribution curve. By setting the height of the rectangle in Figure 5-1 to be 0.2, we force the enclosed area to be $5 \times 0.2 = 1$, as required. This property (area = 1) makes it very easy to solve probability problems, so the following statement is important:

Figure 5-1
Uniform Distribution of Temperatures



Because the total area under the density curve is equal to 1, there is a correspondence between area and probability.



~~5-2~~ 9.2 The Standard Normal Distribution

This chapter focuses on normal distributions, which are extremely important because they occur so often in real applications. Heights of adult women, weights of adult men, and third-grade reading test scores are some examples of normally distributed populations.

STATISTICAL PLANET

New Data Collection Technology

Measuring or coding products and manually entering results is one way of collecting data, but technology is now providing us with alternatives that are not so susceptible to human error. Supermarkets use barcode readers to set prices and analyze inventory and buying habits. Manufacturers are increasingly using direct data entry devices such as electronic measuring gauges connected directly to a computer that records results. A third possibility is voice data entry, such as the Voice Data Logger, which consists of a headset connected to a small box that attaches to the speaker's belt.

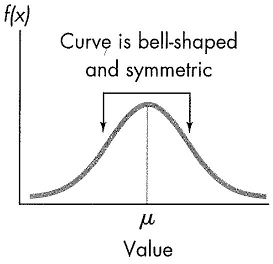
DEFINITION

A continuous random variable has a **normal distribution** if that distribution has a graph that is symmetric and bell-shaped, as in Figure 5-3, and the distribution fits the equation given as Formula 5-1.

Formula 5-1
$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Do not be discouraged by the complexity of Formula 5-1, because it is not really necessary for us actually to use it. What it shows is that any particular normal distribution is determined by two parameters: the mean μ and standard deviation σ. Once specific values are selected for μ and σ, we can graph Formula 5-1 as we would graph any equation relating x and y; the result is a probability distribution with a bell shape. We will see that this normal distribution has many real applications, and we will use it often throughout the remaining chapters.

Figure 5-3
The Normal Distribution

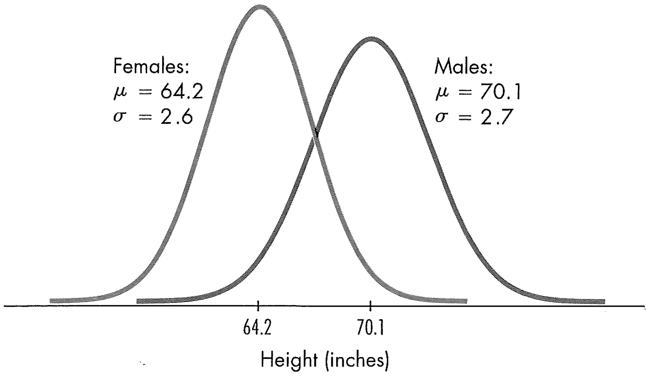


The Standard Normal Distribution

The density curve of a normal distribution has the more complicated bell shape shown in Figure 5-3, so it's more difficult to find areas; but the basic principle is the same: There is a correspondence between area and probability.

There are many different normal distributions, with each one depending on two parameters: the population mean μ and the population standard deviation σ. Figure 5-4 shows density curves for heights of female and male college students. Because males have a larger mean height, the density curve for males is farther to the right. Because males have a slightly larger standard deviation, the density curve for males is slightly wider. Figure 5-4 shows two different possible normal distributions. There are infinite possibilities, but one is of special interest.

Figure 5-4
Heights of Male and Female College Students



DEFINITION

The **standard normal distribution** is a normal probability distribution that has a mean of 0 and a standard deviation of 1. (See Figure 5-5.)

Suppose that somehow we were forced to perform calculations using Formula 5-1. We would quickly see that the most workable values for μ and σ are $\mu = 0$ and $\sigma = 1$. By letting $\mu = 0$ and $\sigma = 1$, mathematicians have calculated areas under the curve. As shown in Figure 5-5, the area under the curve bounded by the mean of 0 and the score of 1 is 0.3413. Remember, the total area under the curve is always 1; this allows us to make the correspondence between area and probability.

Finding Probabilities When Given *z* Scores

Figure 5-5 shows that the area bounded by the curve, the horizontal axis, and the *z scores* of 0 and 1 is an area of 0.3413. Although the figure shows only one area, we can find areas (or probabilities) for many different regions. Such areas can be found by using Table A-2 in Appendix A or by using statistical software. If you are using Table A-2, it is essential to understand the following points.

1. Table A-2 is designed only for the *standard* normal distribution, which has a mean of 0 and a standard deviation of 1.
2. Each value in the body of the table is an area under the curve bounded on the left by a vertical line above the mean of 0 and bounded on the right by a vertical line above a specific positive score denoted by z, as illustrated in Figure 5-6.

Figure 5-5 **Standard Normal Distribution, with Mean $\mu = 0$ and Standard Deviation $\sigma = 1$**

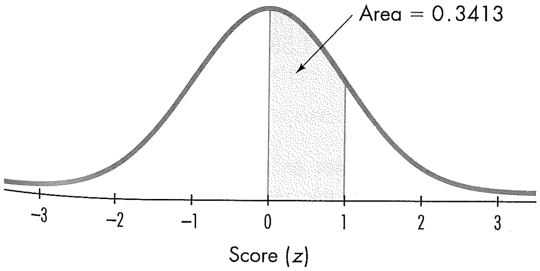
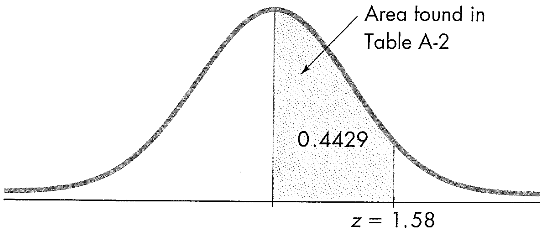


Figure 5-6 **The Standard Normal Distribution**
The area of the shaded region bounded by the mean of 0 and the positive number z can be found in Table A-2.



3. When working with a graph, avoid confusing z scores and areas.

   z score: *distance* along the horizontal scale of the standard normal distribution; refer to the leftmost column and top row of Table A-2

   Area: *region* under the curve; refer to the values in the body of Table A-2

4. The part of the z score denoting hundredths is found across the top row of Table A-2.

The following example requires that we find the probability associated with a score between 0 and 1.58. Begin with the z score of 1.58 by locating 1.5 in the left column; now find the value in the adjoining row of probabilities that is directly below 0.08, as shown in this excerpt from Table A-2.

z	. . .	.08
.		.
.		.
.		.
1.5	. . .	.4429

The area (or probability) value of 0.4429 indicates that there is a probability of 0.4429 of randomly selecting a score between 0 and 1.58. (The following sections will consider cases in which the mean is not 0 or the standard deviation is not 1.)

EXAMPLE

The Precision Scientific Instrument Company manufactures thermometers that are supposed to give readings of 0°C at the freezing point of water. Tests on a large sample of these instruments reveal that at the freezing point of water, some thermometers give readings below 0°C (denoted by negative numbers) and some give readings above 0°C (denoted by positive numbers). Assume that the mean reading is 0°C and the standard deviation of the readings is 1.00°C. Also assume that the frequency distribution of errors closely resembles the normal distribution. If one thermometer is randomly selected, find the probability that, at the freezing point of water, the reading is between 0°C and 1.58°C.

SOLUTION

The probability distribution of the readings is a standard normal distribution because the readings are normally distributed with $\mu = 0$ and $\sigma = 1$. We need to find the area between 0 and z (the shaded region) in Figure 5-6 with $z = 1.58$. From Table A-2 we find that this area is 0.4429.

Interpretation

AS

The probability of randomly selecting a thermometer with an error between 0°C and +1.58°C is therefore 0.4429. Another way to interpret this result is to conclude that 44.29% of the thermometers will have errors between 0°C and +1.58°C.

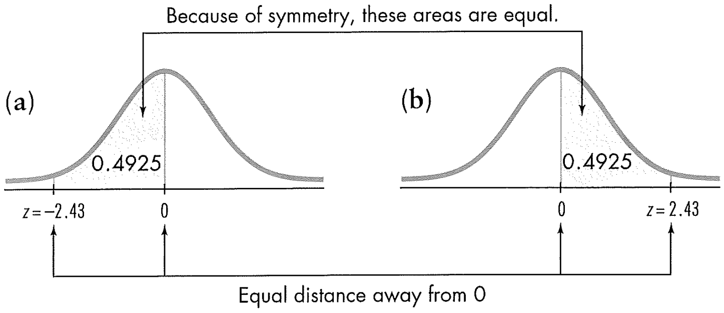

Because of symmetry, these areas are equal.
(a)
(b)
0.4925
0.4925
z=−2.43
0
0
z=2.43
Equal distance away from 0

Figure 5-7
Using Symmetry to Find the Area to the Left of the Mean

EXAMPLE

Using the thermometers from the preceding example, find the probability of randomly selecting one thermometer that reads (at the freezing point of water) between −2.43°C and 0°C.

SOLUTION

We are looking for the region shaded in Figure 5-7(a), but Table A-2 is designed to apply only to regions to the right of the mean (0) as in Figure 5-7(b). By comparing the shaded area in Figure 5-7(a) to the shaded area in Figure 5-7(b), we can see that those two areas are identical because the density curve is symmetric. Referring to Table A-2, we can easily determine that the shaded area of Figure 5-7(b) is 0.4925, so the shaded area of Figure 5-7(a) must also be 0.4925.

Interpretation

The probability of randomly selecting a thermometer with an error between −2.43°C and 0° is 0.4925. In other words, 49.25% of the thermometers have errors between −2.43°C and 0°C.

The above solution illustrates an important principle:

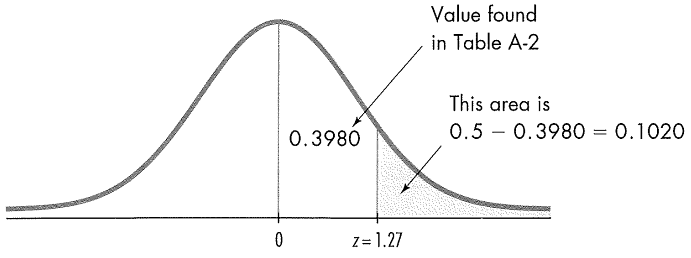
Although a z score can be negative, the area under the curve (or the corresponding probability) can never be negative.

Now recall the empirical rule (presented in Section 2–5) that states that for bell-shaped distributions,

- About 68% of all scores fall within 1 standard deviation of the mean.
- About 95% of all scores fall within 2 standard deviations of the mean.
- About 99.7% of all scores fall within 3 standard deviations of the mean.

If we refer to Figure 5-5 with $z = 1$, Table A-2 shows us that the shaded area is 0.3413. It follows that the proportion of scores between $z = -1$ and $z = 1$ will be $0.3413 + 0.3413 = 0.6826$. That is, about 68% of all scores fall within 1 standard deviation of the mean. A similar calculation with $z = 2$ yields the values of

Figure 5-8
Finding the Area to the Right of $z = 1.27$



0.4772 + 0.4772 = 0.9544 (or about 95%) as the proportion of scores between $z = -2$ and $z = 2$. Similarly, the proportion of scores between $z = -3$ and $z = 3$ is given by 0.4987 + 0.4987 = 0.9974 (or about 99.7%). These exact values correspond very closely to those given in the empirical rule. In fact, the values of the empirical rule were found directly from the probabilities in Table A-2 and have been slightly rounded for convenience. The empirical rule is sometimes called the *68–95–99 rule*; using exact values from Table A-2, it would be called the *68.26–95.44–99.74 rule*, but then it wouldn't sound as snappy.

Because we are dealing with a density curve for a probability distribution, the total area under the curve must be 1. Now refer to Figure 5-8 and see that a vertical line directly above the mean of 0 divides the area under the curve into two equal parts, each containing an area of 0.5. The following example uses this observation.

EXAMPLE

Once again, make a random selection from the same sample of thermometers. Find the probability that the chosen thermometer reads (at the freezing point of water) greater than +1.27°C.

SOLUTION

We are again dealing with normally distributed values having a mean of 0°C and a standard deviation of 1°C. The probability of selecting a thermometer that reads greater than +1.27°C corresponds to the shaded area of Figure 5-8. Table A-2 cannot be used to find that area directly, but we can use the table to find that $z = 1.27$ corresponds to the area of 0.3980, as shown in the figure. We now reason that because the area to the right of zero is one-half of the total area, it has an area of 0.5 and the shaded area is 0.5 − 0.3980, or 0.1020.

Interpretation

We conclude that there is a probability of 0.1020 of randomly selecting one of the thermometers with a reading greater than +1.27°C. Another way to interpret this result is to state that if many thermometers are selected and tested, then 0.1020 (or 10.20%) of them will read greater than +1.27°C.

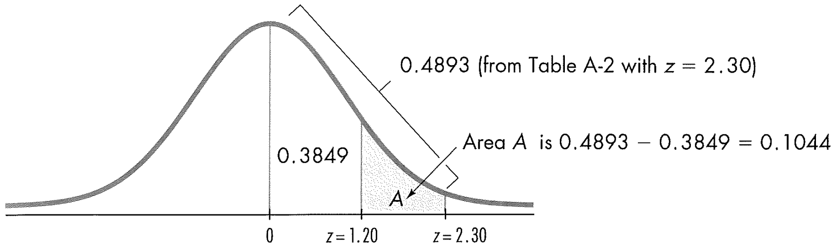


Figure 5-9
Finding the Area Between $z = 1.20$ and $z = 2.30$

We are able to determine the area of the shaded region in Figure 5-8 by an indirect application of Table A-2. The following example illustrates yet another indirect use.

EXAMPLE

Assuming that one thermometer in our sample is randomly selected, find the probability that it reads (at the freezing point of water) between 1.20°C and 2.30°C.

SOLUTION

The probability of selecting a thermometer that reads between 1.20°C and 2.30°C corresponds to the shaded area of Figure 5-9. However, Table A-2 is designed to provide only for regions bounded on the left by the vertical line above 0. We can use the table to find that $z = 1.20$ corresponds to an area of 0.3849 and that $z = 2.30$ corresponds to an area of 0.4893, as shown in the figure. If we denote the area of the shaded region by A, we can see from Figure 5-9 that

$$0.3849 + A = 0.4893$$

so

$$A = 0.4893 - 0.3849 = 0.1044$$

Interpretation

If one thermometer is randomly selected, the probability that it reads (at the freezing point of water) between 1.20°C and 2.30°C is therefore 0.1044.

The above example concluded with the statement that the probability of a reading between 1.20°C and 2.30°C is 0.1044. Such probabilities can also be expressed with the following notation.

NOTATION

$P(a < z < b)$ denotes the probability that the z score is between a and b.

$P(z > a)$ denotes the probability that the z score is greater than a.

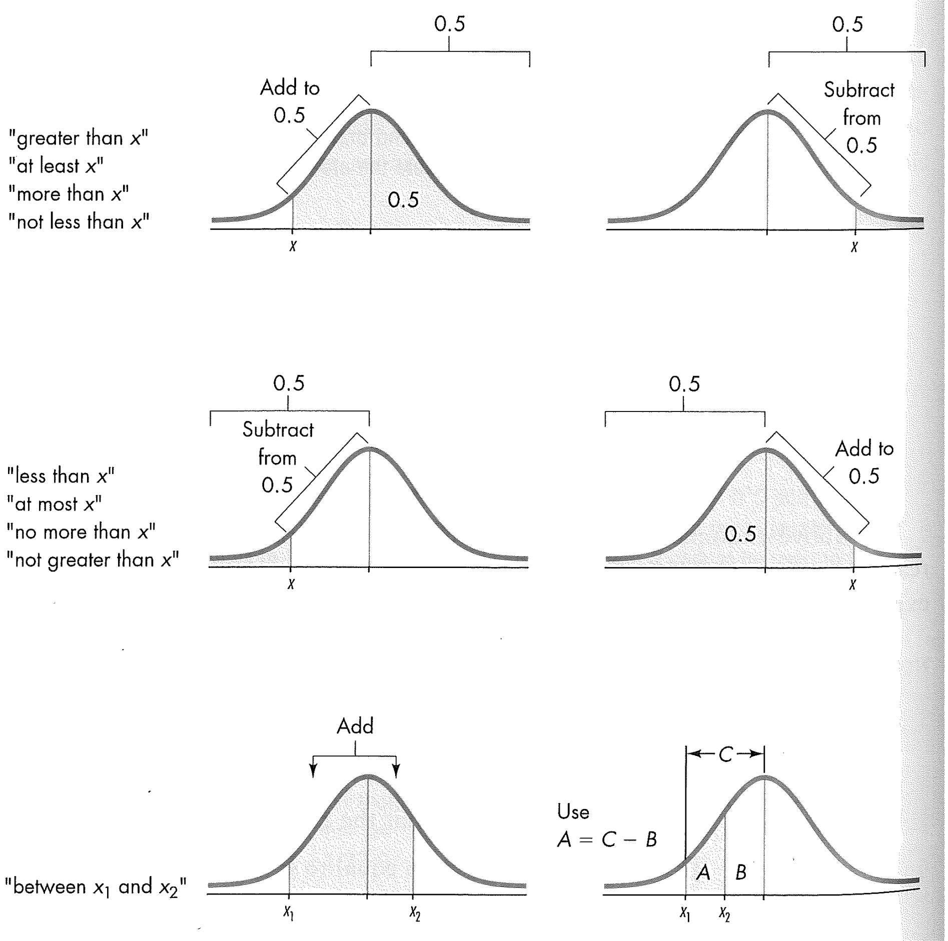
$P(z < a)$ denotes the probability that the z score is less than a.

$P(z = a)$ This probability is always equal to 0.

Using this notation, we can express the result of the last example as $P(1.20 < z < 2.30) = 0.1044$, which states in symbols that the probability of a z score falling between 1.20 and 2.30 is 0.1044. With a continuous probability distribution such as the normal distribution, the probability of getting any single *exact* value is 0. That is, $P(z = a) = 0$.

For example, there is a 0 probability of randomly selecting someone and getting a height of exactly 68.16243357 in. In the normal distribution, any single point on the horizontal scale is represented not by a region under the curve, but by a vertical line above the point. For $P(z = 1.33)$, we have a vertical line above $z = 1.33$, but that vertical line by itself contains no area, so $P(z = 1.33) = 0$. With any continuous random variable, the probability of any one exact value is 0, and it follows that $P(a \le z \le b) = P(a < z < b)$. It also follows that the probability of getting a z score of *at most b* is equal to the probability of getting a z score of *less than b*. It is important to interpret correctly key phrases such as *at most*, *at least*, *more than*, *no more than*, and so on. The illustrations in Figure 5-10

Figure 5-10
Interpreting Areas Correctly



provide an aid to interpreting several of the most common phrases, assuming you will be working with Table A-2.

Finding *z* Scores When Given Probabilities

So far, the examples of this section involving the standard normal distribution have all followed the same format: Given some value(s), we found areas under the curve that represent probabilities. In many other cases, we already know the probability, but we need to find the corresponding *z* score. In such cases, it is very important to avoid confusion between *z* scores and areas. Remember, the numbers Table A-2 shows in the extreme left column and across the top are *z* scores, which are *distances* along the horizontal scale, whereas the numbers in the body of Table A-2 are *areas* (or probabilities). Also, *z* scores to the left of the centre line are always negative (as in Figure 5-7a). If we already know a probability and want to determine the corresponding *z* score using Table A-2, we find it as follows:

1. Draw a bell-shaped curve, draw the centre line, and identify the region under the curve that corresponds to the given probability. If that region is not bounded by the centre line, work with a portion of the curve that *is* bounded, on one side, by the centre line, and, on the other, by a boundary of the area corresponding to the probability.
2. Using the probability representing the area bounded by the centre line, locate the closest probability in the *body* of Table A-2 and identify the corresponding *z* score.
3. If the *z* score is positioned to the left of the centre line, make it negative.

EXAMPLE

Use the same thermometers with temperature readings that are normally distributed with a mean of 0°C and a standard deviation of 1°C. Find the temperature corresponding to P_{95}, the 95th percentile. That is, find the temperature separating the bottom 95% from the top 5%. (See Figure 5-11.)

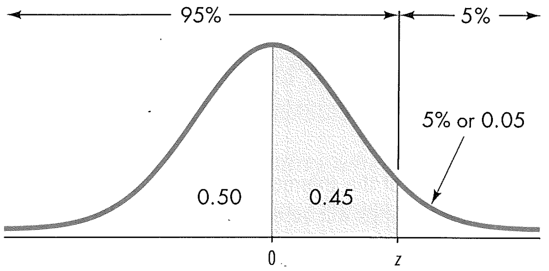


Figure 5-11
Finding the 95th Percentile

SOLUTION

Figure 5-11 shows the z score that is the 95th percentile, separating the top 5% from the bottom 95%. We must refer to Table A-2 to find that z score, and we must use a region bounded by the centre line (where $\mu = 0$) on one side, such as the shaded region of 0.45 in Figure 5-11. (Remember, Table A-2 is designed to directly provide only those areas that are bounded on the left by the centre line and on the right by the z score.) We first search for the area of 0.45 *in the body of the table* and then find the corresponding z score. In Table A-2 the area of 0.45 is between the table values of 0.4495 and 0.4505, but there's an asterisk with a special note indicating that 0.4500 corresponds to a z score of 1.645. We can now conclude that the z score in Figure 5-11 is 1.645, so the 95th percentile is the temperature reading of 1.645°C.

Interpretation

When tested at freezing, 95% of the readings will be less than or equal to 1.645°C, and 5% of them will be greater than or equal to 1.645°C.

Note that in the preceding solution, Table A-2 led to a z score of 1.645, which is midway between 1.64 and 1.65. When using Table A-2, we can usually avoid interpolation by simply selecting the closest value. There are two special cases involving values that are important because they are used so often in a wide variety of applications (see the accompanying table). Except in these two special cases, we can select the closest value in the table. (If a desired value is midway between two table values, select the larger value.) Also, for z scores above 3.09, we can use 0.4999 as an approximation of the corresponding area.

z score	Area
1.645	0.4500
2.575	0.4950

EXAMPLE

Using the same thermometers, find P_{10}, the 10th percentile. That is, find the temperature reading separating the bottom 10% of all temperatures from the top 90%.

SOLUTION

Refer to Figure 5-12, where the 10th percentile is shown as the z score separating the bottom 10% from the top 90%. Table A-2 is designed for areas bounded by the centre line, so we refer to the shaded area of 0.40 (corresponding to 50% − 10%). *In the body of the table*, we select the closest value of 0.3997 and find that it corresponds to $z = 1.28$. However, because the z score is below the mean of 0, it must be negative. The 10th percentile is therefore −1.28°C.

Interpretation

When tested at freezing, 10% of the thermometer readings will be equal to or less than −1.28°C, and 90% of the readings will be equal to or greater than −1.28°C.

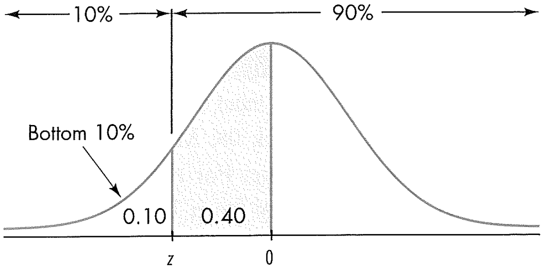
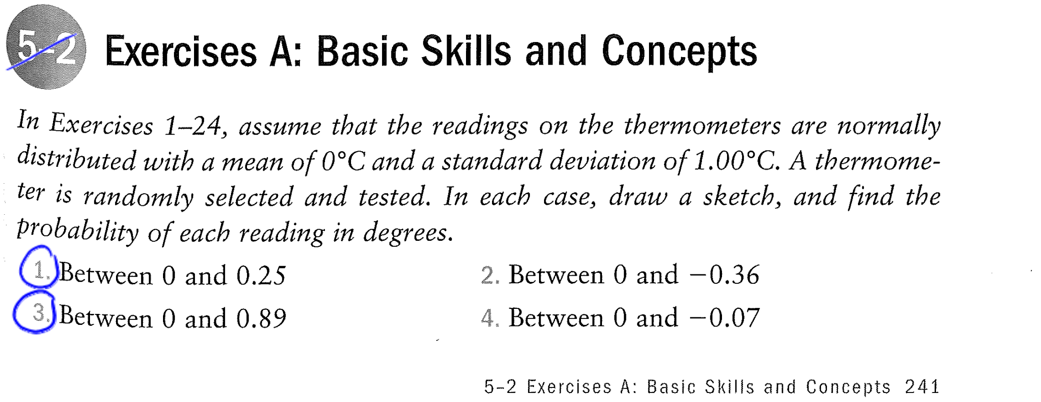


Figure 5-12
Finding the 10th Percentile

The examples in this section were contrived so that the mean of 0 and the standard deviation of 1 coincided exactly with the parameters of the standard normal distribution described in Table A-2. In reality, it is unusual to find such convenient parameters because typical normal distributions involve means different from 0 and standard deviations different from 1. The next section introduces methods for working with such nonstandard normal distributions.

9.2



~~5-2~~ Exercises A: Basic Skills and Concepts

In Exercises 1–24, assume that the readings on the thermometers are normally distributed with a mean of 0°C and a standard deviation of 1.00°C. A thermometer is randomly selected and tested. In each case, draw a sketch, and find the probability of each reading in degrees.

1. Between 0 and 0.25 2. Between 0 and −0.36
3. Between 0 and 0.89 4. Between 0 and −0.07

5. Between 0 and 3.007

6. Between 0 and 1.96

7. Between 0 and -2.331

8. Between 0 and -1.28

9. Greater than 2.58

10. Less than -1.47

11. Less than -2.09

12. Greater than 0.25

13. Between 1.34 and 2.67

14. Between -1.72 and -0.31

15. Between -2.22 and -1.11

16. Between 0.89 and 1.78

17. Less than 0.08

18. Less than 3.01

19. Greater than -2.29

20. Greater than -1.05

21. Between -1.99 and 2.01

22. Between -0.07 and 2.19

23. Between -1.00 and 4.00

24. Between -5.00 and 2.00

In Exercises 25–28, assume that the readings on the thermometers are normally distributed with a mean of 0°C and a standard deviation of 1.00°C. Find the indicated probability, where z is the reading in degrees.

25. $P(z > 2.33)$

26. $P(2.00 < z < 2.50)$

27. $P(-3.00 < z < 2.00)$

28. $P(z < -1.44)$

In Exercises 29–36, assume that the readings on the thermometers are normally distributed with a mean of 0°C and a standard deviation of 1.00°C. A thermometer is randomly selected and tested. In each case, draw a sketch, and find the temperature reading corresponding to the given information.

29. Find P_{90}, the 90th percentile. This is the temperature reading separating the bottom 90% from the top 10%.

30. Find P_{30}, the 30th percentile.

31. Find Q_1, the temperature reading that is the first quartile.

32. Find D_1, the temperature reading that is the first decile.

33. If 4% of the thermometers are rejected because they have readings that are too high, but all other thermometers are acceptable, find the reading that separates the rejected thermometers from the others.

34. If 8% of the thermometers are rejected because they have readings that are too low, but all other thermometers are acceptable, find the reading that separates the rejected thermometers from the others.

35. A quality control analyst wants to examine thermometers that give readings in the bottom 2%. What reading separates the bottom 2% from the others?

36. If 2.5% of the thermometers are rejected because they have readings that are too high and another 2.5% are rejected because they have readings that are too low, find the two readings that are cutoff values separating the rejected thermometers from the others.

9.2

~~5-2~~ Exercises B: Beyond the Basics

37. Assume that z scores are normally distributed with a mean of 0 and a standard deviation of 1.
    a. If $P(0 < z < a) = 0.3212$, find a.
    b. If $P(-b < z < b) = 0.3182$, find b.
    c. If $P(z > c) = 0.2358$, find c.
    d. If $P(z > d) = 0.7517$, find d.
    e. If $P(z < e) = 0.4090$, find e.

38. For a standard normal distribution, find the percentage of data that are
    a. within 1 standard deviation of the mean
    b. within 1.96 standard deviations of the mean
    c. between $\mu - 3\sigma$ and $\mu + 3\sigma$
    d. between 1 standard deviation below the mean and 2 standard deviations above the mean
    e. Suppose it turns out that the distribution is not exactly normal, but is positively skewed. How does this affect your answers to parts (a), (b), and (d) of this exercise?

39. In a manufacturing plant that makes boxes, the width of a certain type of box is normally distributed. The probability that the width is less than 23.9708 cm is 0.0721 and the probability that the width is more than 24.0404 cm is 0.0217. Find the mean and standard deviation of the box width.

40. In a certain region, annual household incomes are normally distributed. The middle 95% of the incomes are between \$72,684 and \$78,564. Find the mean and standard deviation of the annual household incomes for this region.

9.3

~~5-3~~ Normal Distributions: Finding Probabilities

Although Section 5–2 introduced important methods for dealing with normal distributions, the examples and exercises included in that section are generally unrealistic because most normally distributed populations have a nonzero mean, a standard deviation different from 1, or both. In this section we include many real and important nonstandard normal distributions. The basic principle we will be explaining in this section is the following:

If we convert values to standard scores using Formula 5-2, then procedures for working with all normal distributions are the same as for the standard normal distribution.

Formula 5-2

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

If you use certain calculators or software programs to find probabilities under the normal curve, the conversion to *z* scores may not be necessary, because the probabilities can be found directly. Regardless of the method used, however, you need to clearly understand the basic principle of this section, because it is an important foundation for concepts introduced in the following chapters.

See Figure 5-13, where we illustrate the important principle that the area bounded by a score and the population mean is the same as the area bounded by the corresponding *z* score and the mean of 0. Once we convert a nonstandard score to a *z* score, we can use Table A-2 in the same way it was used in Section 5–2. You can use the following procedure for finding probabilities for values of a random variable with a normal probability distribution:

1. Draw a normal curve, label the mean and the specific *x* values, then *shade* the region representing the desired probability.
2. For each relevant score *x* that is a boundary for the shaded region, use Formula 5-2 to find the equivalent *z* score.
3. Refer to Table A-2 to find the area of the shaded region. This area is the desired probability.

The following example uses these three steps, and it illustrates the relationship between a typical nonstandard normal distribution and the standard normal distribution.

EXAMPLE

The mean value of Canadian imports each year from the Middle East and Africa is $2502 million, according to Statistics Canada data for 1980 to 1999. The standard deviation is $905 million. Assuming there is no trend in the data over time, and that the annual figures are normally distributed, what is the probability that in a randomly selected year from that period, the value of imports from the Middle East and Africa is between $2502 million and $4312 million?

SOLUTION

Step 1: See Figure 5-14, where we enter the mean of 2502 and the *x* value of 4312, and we shade the area representing the probability we want.
Step 2: To use Table A-2, we must use Formula 5-2 to convert the nonstandard distribution of import values to the standard normal distribution. The import value of 4312 is converted to a *z* score as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{4312 - 2502}{905} = \frac{1810}{905} = 2.00$$

This result shows that the import value of $4312 million differs from the mean of $2502 million by 2.00 standard deviations.
Step 3: Referring to Table A-2, we find that $z = 2.00$ corresponds to an area of 0.4772.

Figure 5-13
Converting from a Nonstandard Normal Distribution to the Standard Normal Distribution

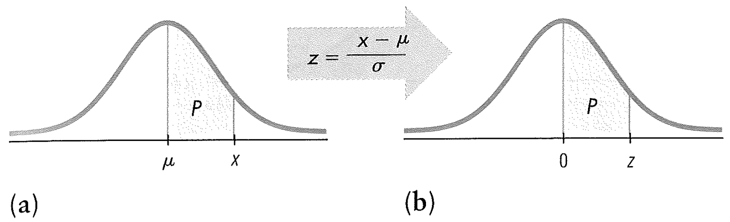
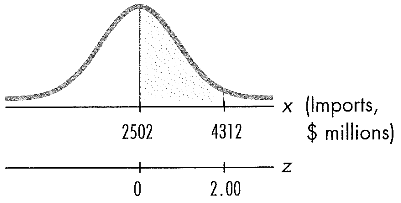


Figure 5-14
Probability of Imports Between $2502 million and $4312 million



Interpretation

There is a probability of 0.4772 of randomly selecting a year with an import value between $2502 million and $4312 million. This can be expressed in symbols as

$$P(2502 < x < 4312) = P(0 < z < 2.00) = 0.4772$$

Another way to interpret this result is to conclude that 47.72% of years have import values from the Middle East and Africa between $2502 million and $4312 million.

EXAMPLE

Assume that the heights of male college students are normally distributed with a mean of 70.1 in. and a standard deviation of 2.7 in.

a. Find the percentage of male students who fall between the Toronto Maple Leafs' average height of 72.4 in. and the Philadelphia Flyers' average height of 74.6 in.

b. Among 500 randomly selected male college students, how many would you expect to fall between the Toronto Maple Leafs' average height and the Philadelphia Flyers' average height?

SOLUTION

a. The shaded region B in Figure 5-15 represents the proportion of male students who fall between the Maple Leafs' average height and the Flyers' average height. We can't find that shaded region directly because Table A-2 isn't designed for such cases, but we can find it indirectly by using the same basic procedures presented in Section 5-2. Find the shaded area B by subtracting region A from the total area of regions A and B combined. That is,

$$B = (A \text{ and } B \text{ combined}) - A$$

For the area of regions A and B combined:

$$z = \frac{x - \mu}{\sigma} = \frac{74.6 - 70.1}{2.7} = 1.67$$

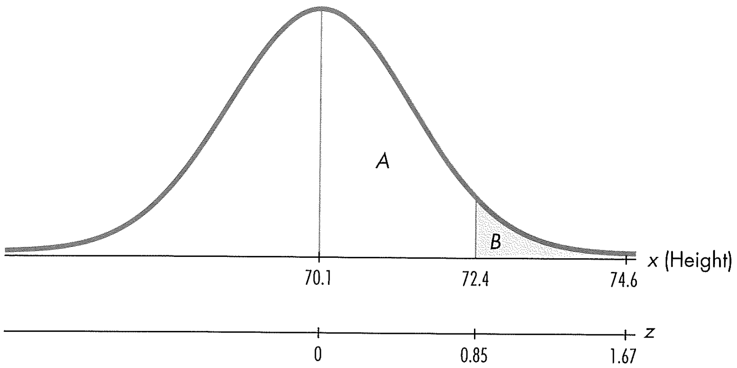
Using Table A-2, we find that $z = 1.67$ corresponds to an area of 0.4525.

STATISTICAL PLANET

Queues

Queuing theory is a branch of mathematics that uses probability and statistics. The study of queues, or waiting lines, is important to businesses such as supermarkets, banks, fast-food restaurants, airlines, and amusement parks. A supermarket in the Netherlands stakes its reputation and revenues on the reliability of queuing theory: A system electronically monitors the number of customers who enter, and cashiers are brought on duty or sent off based on the customer arrival rate. Customers who cannot find a checkout line with fewer than three people don't have to pay for their purchases. (See Nico M. van Dijk in *Chance*, Vol. 10, No. 1.) Bell Laboratories uses queuing theory to optimize telephone network usage, and factories use it to design efficient production lines.

Figure 5-15
Male Students Between Leafs' Average Height and Flyers' Average Height



For the area of region A:

$$z = \frac{x - \mu}{\sigma} = \frac{72.4 - 70.1}{2.7} = 0.85$$

Again using Table A-2, we find that $z = 0.85$ corresponds to an area of 0.3023. Region A has an area of 0.3023.

For the area of region B, the shaded area is the difference between 0.4525 and 0.3023:

$$\begin{aligned}\text{Area } B &= (\text{areas of } A \text{ and } B\text{, combined}) - (\text{area } A)\\ &= 0.4525 - 0.3023 = 0.1502\end{aligned}$$

Interpretation

This number indicates that only 15.02% of male college students have heights between the Maple Leafs' average height and the Flyers' average height.

b. Among 500 randomly selected male college students, we expect that 15.02% of them would have heights between the Maple Leafs' average height and the Flyers' average height:

$$500 \cdot 0.1502 = 75.1 \text{ male students}$$

EXAMPLE

Find the percentage of years in which Canadian imports from the Middle East and Africa are between \$339 million and \$8004 million. Again, assume that the annual import values exhibit no trend, and are normally distributed with a mean of \$2502 million and a standard deviation of \$905 million.

SOLUTION

Figure 5-16 shows the normal distribution of import values, with the shaded region representing values between \$339 million and \$8004 million. The method for finding the area of the shaded region involves breaking it up into parts A and B as shown. We can use

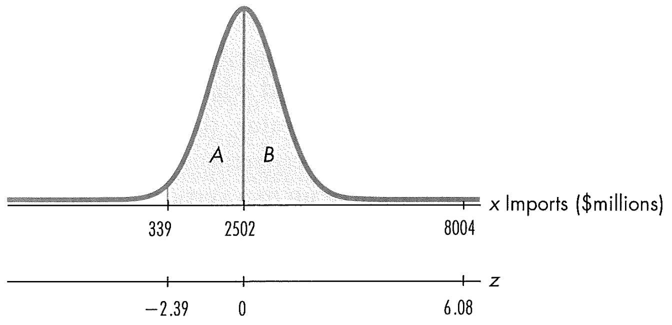


Figure 5-16
Import Values Between $339 Million and $8004 Million

Formula 5-2 and Table A-2 to find the areas of those regions separately; then we can add the results.

For area A only:

$$z = \frac{x - \mu}{\sigma} = \frac{339 - 2502}{905} = -2.39$$

We can use Table A-2 to find that $z = -2.39$ corresponds to 0.4916, so the area A is 0.4916.

For area B only:

$$z = \frac{8004 - 2502}{905} = 6.08$$

Table A-2 does not include z scores above 3.09, but it does include a note that for values of z above 3.09, we should use 0.4999 for the area. (If necessary, more accurate results can be obtained by using special tables or software.) Area B is 0.4999.

For areas of regions A and B combined:

$$0.4916 + 0.4999 = 0.9915$$

Interpretation

The proportion of years in which the value of Canadian imports from the Middle East and Africa is between $339 million and $8004 million is 0.9915. That is to say, imports fall within this range 99.15% of all years.

In this section we have extended the concepts of Section 5–2 to include more realistic nonstandard normal probability distributions. However, all of the examples we have considered so far are of the same general type: We are given specific limit values and we must find an area (or probability, or percentage). In many practical and real cases, the probability (or percentage) is known and we must find the relevant value(s). Problems of this type are discussed in the next section.

9.3

~~5-3~~ Exercises A: Basic Skills and Concepts

In Exercises 1–6, assume that the heights of female students are normally distributed with a mean given by $\mu = 64.2$ in. and a standard deviation given by $\sigma = 2.6$ in. (based on data from a survey of college students). Also assume that a female student is randomly selected. Draw a graph, and find the indicated probability.

1. $P(64.2 \text{ in.} < x < 65.0 \text{ in.})$

2. $P(x < 70.0 \text{ in.})$

3. $P(x > 58.1 \text{ in.})$

4. $P(59.1 \text{ in.} < x < 66.6 \text{ in.})$

5. A fashion agency is looking for females between 65.5 in. and 68.0 in. tall to work as models. Find the probability that a randomly selected female student meets the height requirements to be a model.

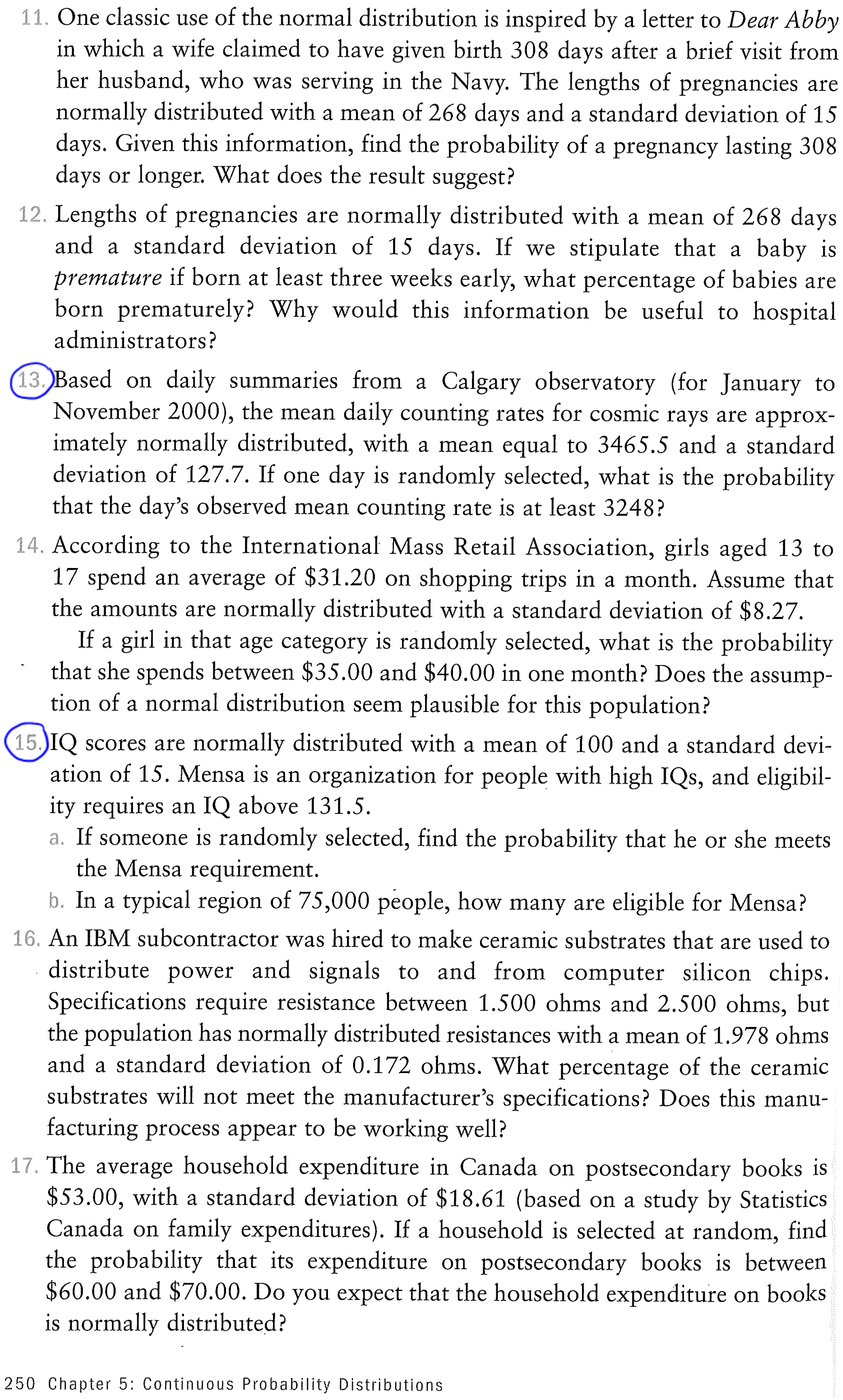
6. The Beanstalk Club, a social organization for tall people, has a requirement that women must be at least 70 in. (or 5 ft 10 in.) tall. Suppose you are trying to decide whether to open a branch of the Beanstalk Club at your college with 500 female students.
   a. Find the percentage of female students who are eligible for membership because they meet the minimum height requirement of 70 in.
   b. Among the 500 female students in your college, how many would be eligible for Beanstalk Club membership?
   c. Will you open a branch of the Beanstalk Club?

7. Replacement times for TV sets are normally distributed with a mean of 8.2 years and a standard deviation of 1.1 years (based on data from "Getting Things Fixed," *Consumer Reports*). Find the probability that a randomly selected TV set will have a replacement time of less than 7.0 years.

8. Replacement times for CD players are normally distributed with a mean of 7.1 years and a standard deviation of 1.4 years (based on data from "Getting Things Fixed," *Consumer Reports*). Find the probability that a randomly selected CD player will have a replacement time of less than 8.0 years.

9. Assume that the heights of soldiers in the Canadian Armed Forces are normally distributed with a mean height of 70.3 in. and a standard deviation of 3.4 in. Find the probability of one soldier who is selected at random having a height of 77.0 in. or greater.

10. Based on the sample results in Data Set 18 of Appendix B, assume that human body temperatures are normally distributed with a mean of 36.4°C and a standard deviation of 0.62°C. If we define a fever to be a body temperature above 37.8°C, what percentage of normal and healthy persons would be considered to have a fever? Does this percentage suggest that a cutoff of 37.8°C is appropriate?



11. One classic use of the normal distribution is inspired by a letter to *Dear Abby* in which a wife claimed to have given birth 308 days after a brief visit from her husband, who was serving in the Navy. The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days. Given this information, find the probability of a pregnancy lasting 308 days or longer. What does the result suggest?

12. Lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days. If we stipulate that a baby is *premature* if born at least three weeks early, what percentage of babies are born prematurely? Why would this information be useful to hospital administrators?

13. Based on daily summaries from a Calgary observatory (for January to November 2000), the mean daily counting rates for cosmic rays are approximately normally distributed, with a mean equal to 3465.5 and a standard deviation of 127.7. If one day is randomly selected, what is the probability that the day's observed mean counting rate is at least 3248?

14. According to the International Mass Retail Association, girls aged 13 to 17 spend an average of $31.20 on shopping trips in a month. Assume that the amounts are normally distributed with a standard deviation of $8.27.

    If a girl in that age category is randomly selected, what is the probability that she spends between $35.00 and $40.00 in one month? Does the assumption of a normal distribution seem plausible for this population?

15. IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. Mensa is an organization for people with high IQs, and eligibility requires an IQ above 131.5.
    a. If someone is randomly selected, find the probability that he or she meets the Mensa requirement.
    b. In a typical region of 75,000 people, how many are eligible for Mensa?

16. An IBM subcontractor was hired to make ceramic substrates that are used to distribute power and signals to and from computer silicon chips. Specifications require resistance between 1.500 ohms and 2.500 ohms, but the population has normally distributed resistances with a mean of 1.978 ohms and a standard deviation of 0.172 ohms. What percentage of the ceramic substrates will not meet the manufacturer's specifications? Does this manufacturing process appear to be working well?

17. The average household expenditure in Canada on postsecondary books is $53.00, with a standard deviation of $18.61 (based on a study by Statistics Canada on family expenditures). If a household is selected at random, find the probability that its expenditure on postsecondary books is between $60.00 and $70.00. Do you expect that the household expenditure on books is normally distributed?

18. Measurements of human skulls from different epochs are analyzed to determine whether they change over time. The maximum breadth is measured for skulls from Egyptian males who lived around 3300 BCE. Results show that those breadths are normally distributed with a mean of 132.6 mm and a standard deviation of 5.4 mm (based on data from *Ancient Races of the Thebaid* by Thomson and Randall-Maciver). An archeologist discovers a male Egyptian skull and a field measurement reveals a maximum breadth of 119 mm. Find the probability of getting a value of 119 or less if a skull is randomly selected from the period around 3300 BCE. Is the newly found skull likely to come from that era?

19. According to a national health survey, the serum cholesterol levels of men aged 18 to 24 are normally distributed with a mean and standard deviation (in mg/100mL) of 178.1 and 40.7, respectively. One criterion for identifying risk of coronary disease is a cholesterol level above 300. If a man aged 18–24 is randomly selected, find the probability that his serum cholesterol level is above 300. Does this probability warrant serious concern?

20. Some vending machines are designed so that their owners can adjust the weights of the quarters that are accepted. If many counterfeit coins are found, adjustments are made to reject more coins, with the effect that most of the counterfeit coins are rejected along with many legal coins. Assume that quarters have weights that are normally distributed with a mean of 5.67 g and a standard deviation of 0.070 g. If a vending machine is adjusted to reject quarters weighing less than 5.50 g or more than 5.80 g, what is the percentage of legal quarters that are rejected?

9.3 ~~5-3~~ Exercises B: Beyond the Basics

In Exercises 21–23, refer to the indicated data set in Appendix B.

a. *Construct a histogram to determine whether the data set has a normal distribution.*

b. *Find the sample mean and sample standard deviation s.*

c. *Use the sample mean as an estimate of the population mean μ, use the sample standard deviation as an estimate of the population standard deviation σ, and use the methods of this section to find the indicated probability.*

21. Use the combined list of 100 weights of M&M plain candies listed in Data Set 11, and estimate the probability of randomly selecting one M&M candy and getting one with a weight greater than 1.000 g.

22. Use the total weights of discarded garbage in Data Set 1, and estimate the probability of randomly selecting a household that discards more than 20.0 lb of garbage in a week.

23. Use the chest measurements of bears in Data Set 3, and estimate the probability of randomly selecting a bear with a chest measuring less than 30 in.

24. When you constructed a histogram for total weights of discarded garbage, for Question 22, was the distribution *exactly* normal? If not, do you believe your probability estimate in Question 22 was likely too low or too high? Explain.

9.4 ~~5-4~~ Normal Distributions: Finding Values

In this section we consider problems such as this: If companies' revenues per employee are normally distributed with μ = \$310,000 and σ = \$147,000, find the revenue per employee separating the bottom 10% from the others. This problem starts from a given probability (0.10). We need to find the appropriate value x. This section reverses the procedure of Section 5–3, where we used a given value to find a probability.

In considering problems of finding values when given probabilities, there are three important cautions to keep in mind.

1. *Don't confuse z scores and areas.* Remember, z scores are *distances* along the horizontal scale, but areas represent *regions* under the normal curve. Table A-2 lists z scores in the left column and across the top row, but areas are found in the body of the table.
2. *Choose the correct (right/left) side of the graph.* A score separating the top 10% from the others will be located on the right side of the graph, but a score separating the bottom 10% will be located on the left side of the graph.
3. *A z score must be negative whenever it is located to the left of the centre line of 0.*

As in Section 5–3, graphs are extremely helpful and they are strongly recommended. Even if you will be using statistical software to find values when given probabilities, you should understand the graphs and procedures that are presented below.

Procedure for Finding Values Using Table A-2 and Formula 5-2

1. Sketch a normal distribution curve; enter the given probability or percentage in the appropriate region of the graph, and identify the x value(s) being sought.
2. Use Table A-2 to find the z score corresponding to the region bounded by x and the centre line of 0. Observe the following cautions:
   - Refer to the *body* of Table A-2 to find the closest area, then identify the corresponding z score.
   - Make the z score *negative* if it is located to the left of the centre line.

3. Using Formula 5-2, enter the values for μ, σ, and the z score found in Step 2, then solve for x. On the basis of the format of Formula 5-2, we can solve for x as follows:

$$x = \mu + (z \cdot \sigma) \qquad \text{(Another form of Formula 5-2)}$$

4. Refer to the sketch of the curve to verify that the solution makes sense in the context of the graph and in the context of the problem.

The following example, which was introduced at the beginning of this section, uses the procedure just outlined. Pay extra attention to Step 2, especially where we make the z score negative because it is to the left of the mean.

EXAMPLE

In its annual lists of the "Top 100 Companies" of Canada, the *Financial Post* uses the companies' revenues as the basis for ranking. There are other company variables which seem, in fact, quite randomly distributed. For example, in 1997, fully two-thirds of the Top 100 companies appeared to have values for the variable "Revenues per employee" that fell into a normal distribution, with a mean and standard deviation (in $1000s) of 310 per employee and 147 per employee, respectively. (The other third of companies truly did distinguish themselves, with values ranging from 3 up to 115 times the mean value for the lower group.) For those lower performing companies, find the value of P_{10}—the revenue per employee separating the bottom 10% from the top 90%.

SOLUTION

Step 1: We begin with the graph shown in Figure 5-17. We have entered the mean of 310, shaded the area representing the bottom 10%, and identified the desired value as x. The area between 310 and x must be 40% of the total area (because the left half of the area must combine to be 50% of the total). Because the total area is 1, the area constituting 40% of the total must be 0.4.

Step 2: We refer to Table A-2, but we look for an area of 0.4000 in the body of the table. (Remember, Table A-2 is designed to list areas only for those regions bounded on the left by the mean and on the right by some value.) The area closest to 0.4000 is 0.3997, and it corresponds to a z score of 1.28. Because the score is to the left of the mean, we make it negative and use $z = -1.28$.

AS

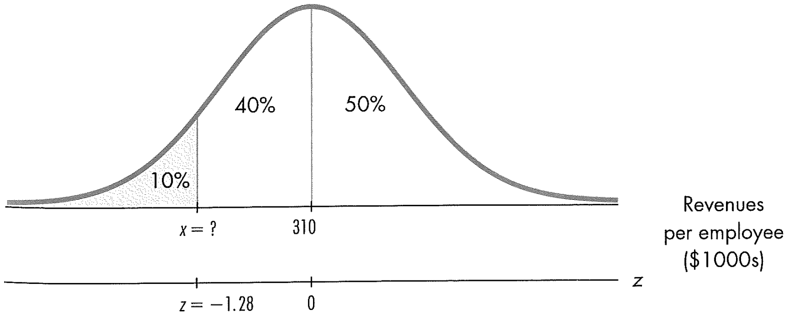


Figure 5-17
Finding P_{10} for Revenues per Employee

STATISTICAL PLANET

Survey Medium Can Affect Results

In a survey of Catholics, the subjects were asked if contraceptives should be made available to unmarried women. In personal interviews, 44% of the respondents said yes. But among a similar group contacted by mail or telephone, 75% of the respondents answered yes to the same question.

Step 3: With $z = -1.28$, $\mu = 310$, and $\sigma = 147$, we solve for x either by using Formula 5-2 directly or by using the following version of Formula 5-2:

$$x = \mu + (z \cdot \sigma) = 310 + (-1.28 \cdot 147) = 121.8$$

Step 4: If we let $x = 121.8$ in Figure 5-17, we see that this solution is reasonable because the 10th percentile should be less than the mean of 310.

Interpretation

A revenue per employee of $121,800 separates the lowest 10% from the highest 90%.

EXAMPLE

Assume that body temperatures of healthy adults are normally distributed with a mean of 36.39°C and a standard deviation of 0.62°C (based on Data Set 18 in Appendix B). If a medical researcher wants to study people in the bottom 2.5% and people in the top 2.5%, find the temperatures separating those limits.

SOLUTION

Step 1: We begin with the graph shown in Figure 5-18. We have shaded the areas representing the bottom 2.5% and top 2.5% (or 0.025). The areas of 0.475 are found by using the fact that the centre line above the mean divides the total area of 1 into two parts, each with area 0.5. We get $0.5 - 0.025 = 0.475$.

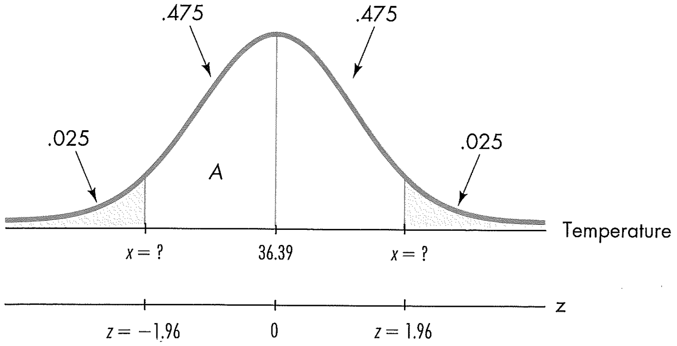
Step 2: We refer to Table A-2, but we look for an area of 0.475 in the body of the table. (Remember, Table A-2 is designed to list areas only for those regions bounded on the left by the mean and on the right by some value.) The area of 0.4750 corresponds to $z = 1.96$. For the x value located on the right in Figure 5-18, we use $z = 1.96$; for the x value located on the left we use $z = -1.96$.

Step 3: With $z = 1.96$, $\mu = 36.39$, and $\sigma = 0.62$, we solve for x using a variation of Formula 5-2:

$$x = \mu + (z \cdot \sigma) = 36.39 + (1.96 \cdot 0.62) = 37.61$$

Figure 5-18

Body Temperatures



With $z = -1.96$, $\mu = 36.39$, and $\sigma = 0.62$, we solve for x using a variation of Formula 5-2:

$$x = \mu + (z \cdot \sigma) = 36.39 + (-1.96 \cdot 0.62) = 35.17$$

Step 4: If we let $x = 35.17$ and 37.61 in Figure 5-18, we see that our solutions are reasonable.

Interpretation

If the researcher's assumptions about the distribution of body temperatures are correct, then he or she should select people with body temperatures below 35.17°C or above 37.61°C.

9.4

~~5-4~~ Exercises A: Basic Skills and Concepts

In Exercises 1–4, assume that female college students have heights that are normally distributed with a mean of 64.2 in. and a standard deviation of 2.6 in. Find the height for the given percentile.

1. P_{85}
2. P_{66}
3. P_{15}
4. P_{35}
5. Replacement times for TV sets are normally distributed with a mean of 8.2 years and a standard deviation of 1.1 years (based on data from "Getting Things Fixed," *Consumer Reports*).
   a. Find the replacement time that separates the top 20% from the bottom 80%.
   b. Find the probability that a randomly selected TV will have a replacement time of less than 5.0 years.
   c. If you want to provide a warranty so that only 1% of the TV sets will be replaced before the warranty expires, what length of time would you recommend for the warranty?
6. Replacement times for CD players are normally distributed with a mean of 7.1 years and a standard deviation of 1.4 years (based on data from "Getting Things Fixed," *Consumer Reports*). Find P_{55}, which is the replacement time separating the top 45% from the bottom 55%.
7. Weights of paper discarded by households each week are normally distributed with a mean of 4.3 kg and a standard deviation of 1.9 kg. Find P_{33}, which is the weight that separates the bottom 33% from the top 67%.
8. Based on the sample results in Data Set 18 of Appendix B, assume that human body temperatures are normally distributed with a mean of 36.4°C and a standard deviation of 0.62°C. What two temperature levels separate the bottom 2% and the top 2%? Could these values serve as reasonable limits that could be used to identify people who are likely to be ill?
9. The durations of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days. If we stipulate that a baby is *premature* if the length of pregnancy is in the lowest 4%, find the duration that separates premature babies from those who are not premature.
10. According to the Opinion Research Corporation, men spend an average of 11.4 min in the shower. Assume that the times are normally distributed with a standard deviation of 1.8 min. Find the values of the quartiles Q_1 and Q_3.
11. IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. If we define a genius to be someone in the top 1% of IQ scores, find the score separating geniuses from the rest of us. Are there any jobs where this score could reasonably be used as one criterion for employment?

12. A subcontractor manufactures ceramic substrates for IBM. These devices have resistances that are normally distributed with a mean of 1.978 ohms and a standard deviation of 0.172 ohms. If the required specifications are to be modified so that 3% of the devices are rejected because their resistances are too low and another 3% are rejected because their resistances are too high, find the cutoff values for the acceptable devices.

13. Scores obtained from the Law School Admission Test (LSAT) are normally distributed with a mean score of 550 and a standard deviation of 110.

a. If a test score is selected at random, find the probability that it is less than 750.

b. If the top 16% of the test scores are usually good enough for admission to law school, find the cutoff score for gaining admission.

14. Measurements of human skulls from different epochs are analyzed to determine whether they change over time. The maximum breadth is measured for skulls from Egyptian males who lived around 3300 BCE. Results show that those breadths are normally distributed with a mean of 132.6 mm and a standard deviation of 5.4 mm (based on data from *Ancient Races of the Thebaid* by Thomson and Randall-Maciver).

a. Find the probability of getting a value greater than 140 mm if a skull is randomly selected from the period of around 3300 BCE.

b. Find the value that is D_2, the second decile.

15. Based on daily summaries from a Calgary observatory over a period of ten months, the mean daily counting rates for cosmic rays are approximately normally distributed, with a mean equal to 3465.5 and a standard deviation of 127.7. Find the sixth decile D_6. What is the mean daily counting rate that separates the lowest 60% from the highest 40%?

16. Quarters have weights that are normally distributed with a mean of 5.67 g and a standard deviation of 0.070 g.

a. If a vending machine is adjusted to reject quarters weighing less than 5.53 g or more than 5.81 g, what is the percentage of legal quarters that are rejected?

b. Find the weights of accepted legal quarters if the machine is readjusted so that the lightest 1.5% are rejected and the heaviest 1.5% are rejected.

c. If your quarter is rejected from a machine that is set to reject the upper 1.5% and the lower 1.5% of coins, by weight, is it a waste of time to reinsert your coin?

9.4

~~5-4~~ Exercises B: Beyond the Basics

17. The construction of a histogram for a data set reveals that the distribution is approximately normal and the boxplot is constructed with these quartiles: $Q_1 = 62$, $Q_2 = 70$, $Q_3 = 78$. Estimate the standard deviation.

18. An instructor informs her physics class that a test is very difficult, but the grades will be curved. Scores for the test are normally distributed with a mean of 25 and a standard deviation of 5.
    a. If she curves by adding 50 to each grade, what is the new mean? What is the new standard deviation?
    b. Is it fair to curve by adding 50 to each grade? Why or why not?
    c. If the grades are curved according to the following scheme (instead of adding 50), find the numerical limits for each letter grade.
       A: Top 10%
       B: Scores above the bottom 70% and below the top 10%
       C: Scores above the bottom 30% and below the top 30%
       D: Scores above the bottom 10% and below the top 70%
       F: Bottom 10%
    d. Which method of curving the grades is fairer: Adding 50 to each grade or using the scheme given in part (c)? Explain.

19. According to data from the College Entrance Examination Board, the mean math SAT score is 475, and 17.0% of the scores are above 600. Find the standard deviation, and then use that result to find the 99th percentile. (Assume that the scores are normally distributed.)

20. The College Entrance Examination Board writes that "for the SAT Achievement Tests, your score would fall in a range [between] about 30 points above [and] below your actual ability about two-thirds of the time. This range is called the standard error of measurement (SEM)." Use that statement to estimate the standard deviation for scores of an individual on an SAT Achievement Test. (Assume that the scores are normally distributed.)

9.5 ~~5-5~~ The Central Limit Theorem

This section presents the central limit theorem, which is one of the most important and useful concepts in statistics. It forms a foundation for estimating population parameters and hypothesis testing—topics discussed at length in the following chapters.

We will not present rigorous proofs in this section, but will instead focus on the concepts and how to apply them. You will need to keep in mind the types of data sets we are considering. Instead of sets of *individual* values—the values of a *random variable* (see Section 4–2)—we will work with data sets in which each value is the *mean* of some other sample. Just as a *probability distribution* describes the probability for each value of a random variable x, the *sampling distribution of sample means* describes the probability for each value of the sample mean, when drawing samples of a given size from a population.

DEFINITION

The **sampling distribution of sample means** is the probability distribution of sample means, with all samples having the sample size *n*.

One property of the sampling distribution of sample means is key to this section:

As the sample size increases, the sampling distribution of sample means approaches a normal distribution.

In other words, if we collect many samples of the same size from the same population, compute their means, and then draw a histogram of those means, that histogram will tend to have the bell shape of a normal distribution. This is true regardless of the shape of the distribution of the original population. Compare Figure 5-19, which shows a probability distribution for a specific variable, and Figure 5-20, which shows the sampling distribution for sample means for the same variable. Figure 5-21 is a general illustration of the same principle, applied to three different shapes of underlying population distribution. Observations exactly like these led to the formulation of the central limit theorem, which we will now discuss.

The **central limit theorem** involves two different distributions: the distribution of the original population and the distribution of the sample means. As in previous chapters, we use the symbols μ and σ to denote the mean and standard deviation of the original population. We now introduce new notation for the mean and standard deviation of the distribution of sample means.

Figure 5-19
Distribution of 200 Digits from Social Insurance Numbers (Last 4 Digits) of 50 Students

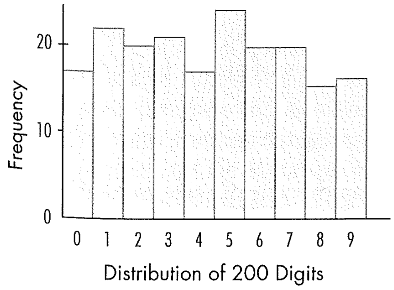


Figure 5-20
Distribution of 50 Sample Means for 50 Students

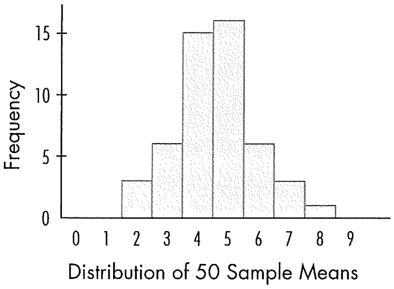
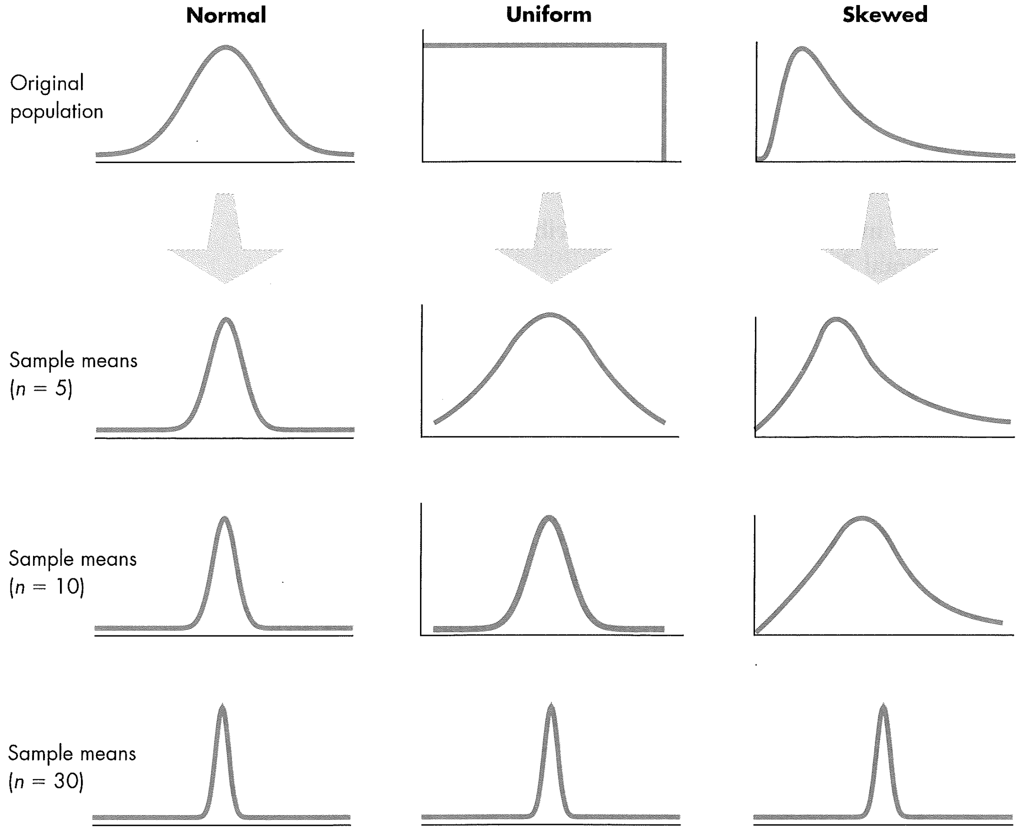


Figure 5-21 **Normal, Uniform, and Skewed Distributions**



CENTRAL LIMIT THEOREM

Given:

1. The random variable x has a distribution (which may *or may not* be normal) with mean μ and standard deviation σ.
2. Samples of size n are randomly selected from this population. (The samples are selected so that all possible samples of size n have the same chance of being selected.)

Conclusions:

1. The distribution of sample means $\overline{x}$ will, as the sample size increases, approach a *normal* distribution.

2. The mean of this distribution of sample means will be the population mean μ.

3. The standard deviation of this distribution of sample means will be $\sigma/\sqrt{n}$.

Practical Rules Commonly Used:

1. For samples of size n larger than 30, the distribution of the sample means can be approximated reasonably well by a normal distribution. The approximation gets better as the sample size n becomes larger.

2. If the original population is itself normally distributed, then the sample means will be normally distributed for *any* sample size n (not just the sample sizes n larger than 30).

NOTATION FOR THE CENTRAL LIMIT THEOREM

If all possible random samples of size n are selected from a population with mean μ and standard deviation σ, the mean of the sample means is denoted by $\mu_{\bar{x}}$, so

$$\mu_{\bar{x}} = \mu$$

Also, the standard deviation of the sample means is denoted by $\sigma_{\bar{x}}$, so

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$\sigma_{\bar{x}}$ is often called the **standard error of the mean.**

EXAMPLE

Table 5-1 illustrates the last four digits of the Social Insurance Numbers of each of 50 students.

The last four digits of Social Insurance Numbers are random. If we combine the four digits from each student into one big collection of 200 numbers, we get a mean of $\bar{x} = 4.5$, a standard deviation of $s = 2.8$, and an approximately uniform distribution with the graph shown in Figure 5-19. Now see what happens when we find the 50 sample means, as shown in Table 5-1. Even though the original collection of data has an approximately *uniform* (that is, not normal) distribution, the sample means have a distribution that is approximately *normal*. This can be a confusing concept, so you should stop right here and study this paragraph until its major point becomes clear: The original set of 200 individual numbers has a uniform distribution (because the digits 0–9 occur with approximately equal frequencies), but the 50 sample means have a normal distribution. It's a truly fascinating and intriguing phenomenon in statistics that by sampling from any distribution, we can create a distribution that is normal or at least approximately normal.

Table 5-1

SIN Digits				$\bar{x}$
1	8	6	4	4.75
5	3	3	6	4.25
9	8	8	8	8.25
5	1	2	5	3.25
9	3	3	5	5.00
4	2	6	2	3.50
7	7	1	6	5.25
9	1	5	4	4.75
5	3	3	9	5.00
7	8	4	1	5.00
0	5	6	1	3.00
9	8	2	2	5.25
6	1	5	7	4.75
8	1	3	0	3.00
5	9	6	9	7.25
6	2	3	4	3.75
7	4	0	7	4.50
5	7	5	6	5.75
4	1	5	7	4.25
1	2	0	6	2.25
4	0	2	8	3.50
3	1	2	5	2.75
0	3	4	0	1.75
1	5	1	0	1.75
9	7	4	0	5.00
7	3	1	1	3.00
9	1	1	3	3.50
8	6	5	9	7.00
5	6	4	1	4.00
9	3	9	5	6.50
6	0	7	3	4.00
8	2	9	6	6.25
0	2	8	6	4.00
2	0	9	7	4.50
5	8	9	0	5.50
6	5	4	9	6.00
4	8	7	6	6.25
7	1	2	0	2.50
2	9	5	0	4.00
8	3	2	2	3.75
2	7	1	6	4.00
6	7	7	1	5.25
2	3	3	9	4.25
2	4	7	5	4.50
5	4	3	7	4.75
0	4	3	8	3.75
2	5	8	6	5.25
7	1	3	4	3.75
8	3	7	0	4.50
5	6	6	7	6.00

Applying the Central Limit Theorem

Many important and practical problems can be solved with the central limit theorem. When working on such problems, remember the following rules.

- **If you are working with a random sample of size $n > 30$, or if the original population is normally distributed, treat the distribution of sample means as a normal distribution.**
- **Treat the mean μ of the original population as the mean of the distribution of sample means.**
- **Treat the calculated value $\sigma/\sqrt{n}$ (based on the original population) as the standard deviation of the distribution of sample means.**

In the following example, part (a) involves an individual value, so we use the methods presented in Section 5–3; those methods apply to the normal distribution of the random variable x. Part (b), however, involves the mean for a *group* of 36 Canadian Football League (CFL) players, so we must use the central limit theorem in working with the random variable $\bar{x}$. Observe the significant difference between the procedures used in part (a) and part (b).

EXAMPLE

In human engineering and product design, it is often important to consider the weights of people so that airplanes or elevators aren't overloaded, chairs don't break, and other such dangerous or embarrassing mishaps do not occur. Given that the population of players in offensive back positions (including quarterback) in the CFL have weights that are approximately normally distributed, with a mean of 197.5 lb and a standard deviation of 14.2 lb (based on sampling from data on the CFL website), find the probability that

a. if one player is randomly selected, his weight is greater than 200 lb

b. if 36 different players are randomly selected, their mean weight is greater than 200 lb

SOLUTION

a. *Approach: Use the methods presented in Section 5-3* (because we are dealing with an *individual* value from a normally distributed population). We seek the area of the shaded region in Figure 5-22(a).

$$z = \frac{x - \mu}{\sigma} = \frac{200 - 197.5}{14.2} = 0.18$$

We now refer to Table A-2 to find that region A is 0.0714. The shaded region is therefore $0.5 - 0.0714 = 0.4286$. The probability of the player weighing more than 200 lb is 0.4286.

b. *Approach: Use the central limit theorem* (because we are dealing with the *mean for a group* of 36 values, not an individual value). Because we are now dealing with a

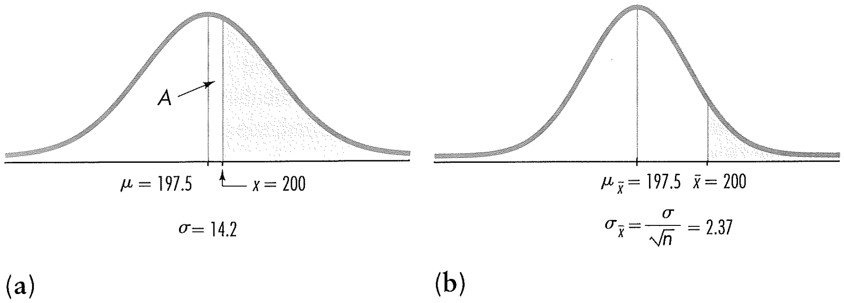


Figure 5-22
Distribution of (a) Individual Offensive Backs' Weights and (b) Means of Samples of 36 Offensive Backs' Weights

distribution of sample means, we must use the parameters $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, which are evaluated as follows:

$$\mu_{\bar{x}} = \mu = 197.5$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14.2}{\sqrt{36}} = 2.37$$

We want to determine the shaded area shown in Figure 5-22(b), and the relevant z score is calculated as follows:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{200 - 197.5}{\frac{14.2}{\sqrt{36}}} = \frac{2.5}{2.37} = 1.05$$

Referring to Table A-2, we find that $z = 1.05$ corresponds to an area of 0.3531, so the shaded region is $0.5 - 0.3531 = 0.1469$. The probability that the 36 players have a mean weight greater than 200 lb is 0.1469.

Interpretation

There is a 0.4286 probability that an offensive back in the CFL will weigh more than 200 lb, but there is only a 0.1469 probability that 36 offensive backs (collectively) will have a mean weight of more than 200 lb. It is much easier for an individual to deviate from the mean than it is for a group of 36. A single extreme weight among the 36 weights will lose its impact when it is averaged in with the other 35 weights.

Also, remember that the calculated probabilities are usually approximate, although they are displayed here to four decimal places. For example, the solutions based on using Table A-2 will differ slightly, due to rounding, from solutions based on statistical software. Moreover, these calculations assume (especially when working with single x values) that the population is normal—which may not be *exactly* true for your data.

The next example illustrates another application of the central limit theorem, but carefully examine the conclusion that is reached. This example shows the type of thinking that is the basis for the important procedure of hypothesis testing (discussed in Chapter 7), and illustrates the **rare event rule** for inferential statistics: If, under a given assumption, the probability of a particular observed event is exceptionally small, we conclude that the assumption is probably not correct.

STATISTICAL PLANET

A Professional Speaks About Sampling Error

Daniel Yankelovich, in an essay for *Time*, commented on the sampling error often reported along with poll results. He stated that sampling error refers only to the inaccuracy created by using random sample data to make an inference about a population; the sampling error does not address issues of poorly stated, biased, or emotional questions. He said, "Most important of all, warning labels about sampling error say nothing about whether or not the public is conflict-ridden or has given a subject much thought. This is the most serious source of opinion poll misinterpretation."

STATISTICAL PLANET

The Fuzzy Central Limit Theorem

In *The Cartoon Guide to Statistics* by Gonick and Smith, the authors describe the *fuzzy central limit theorem* as follows: "Data that are influenced by many small and unrelated random effects are approximately normally distributed. This explains why the normal is everywhere: stock market fluctuations, student weights, yearly temperature averages, [test] scores: All are the result of many different effects." People's heights, for example, are the results of hereditary factors, environmental factors, nutrition, health care, geographic region, and other influences that, when combined, produce normally distributed values.

EXAMPLE

Assume that the population of human body temperatures has a mean of 37.0°C, as is commonly believed. Also assume that the population standard deviation is 0.62°C. If a sample of size $n = 108$ is randomly selected, find the probability of getting a mean of 36.4°C or lower. (The value of 36.4°C was actually obtained; see the 108 temperatures in Data Set 18 of Appendix B.)

SOLUTION

We weren't given the distribution of the population, but because the sample size $n = 108$ exceeds 30, we use the central limit theorem and conclude that the distribution of sample means is a normal distribution with these parameters:

$$\mu_{\bar{x}} = \mu = 37.0$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.62}{\sqrt{108}} = 0.059659$$

Figure 5-23 shows the shaded area (see the left tail of the graph) corresponding to the probability we seek. Having already found the parameters that apply to the distribution shown in Figure 5-23, we can now find the shaded area by using the same procedures developed in the preceding section. We first find the z score:

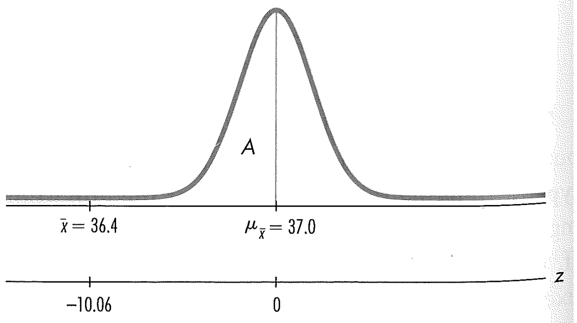
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{36.4 - 37.0}{0.059659} = -10.06$$

Referring to Table A-2, we find that $z = -10.06$ (or its positive equivalent) is off the chart, but for absolute values of z above 3.09, we use an area of 0.4999. We therefore conclude that region A in Figure 5-23 is 0.4999 and the shaded region is $0.5 - 0.4999 = 0.0001$.

Interpretation

The result shows that if the mean of our body temperatures is really 37°C, then there is an extremely small probability of getting a sample mean of 36.4°C or lower when 108 subjects are randomly selected. Either the population mean really is 37.0°C and the sample represents a chance event that is extremely rare, or the population mean is actually lower than 37.0°C so the sample is typical. Because the probability is so low, it seems more reasonable to conclude that the population mean is lower than 37.0°C. This is the type of reasoning used in hypothesis testing, to be introduced in Chapter 7. For now, we should focus on the use of the central limit theorem for finding the probability of 0.0001, but we should also observe that this theorem will be used later in developing some very important concepts in statistics.

Figure 5-23
Distribution of Sample Mean Body Temperatures (n = 108)



No Exercises for 9.5

Ch 9 Answers on next page →

9. $P(\mu - 2\sigma) < X < \mu + 2\sigma) = 4\sigma/(b - a)$
   Since $\sigma = (b - a)/\sqrt{12}$, we get $P(\mu - 2\sigma < X < \mu + 2\sigma) = 4/\sqrt{12} > 1$.
   Since the probability cannot exceed 100%, we can deduce that 100% of a uniform distribution lies within 2 standard deviations of the mean, regardless of the values of a and b.

Section ~~5–~~2 9.2

1. 0.0987
3. 0.3133
5. 0.4987
7. 0.4901
9. 0.0049
11. 0.0183
13. 0.0863
15. 0.1203
17. 0.5319
19. 0.9890
21. 0.9545
23. 0.8412
25. 0.0099
27. 0.9759
29. 1.28°
31. −0.67°
33. 1.75°
35. −2.05°
37. a. 0.92
    b. 0.41
    c. 0.72
    d. −0.68
    e. −0.23
39. $\mu = 24$; $\sigma = 0.02$

Section ~~5–~~3 9.3

1. 0.1217
3. 0.9906
5. 0.2364
7. 0.1379
9. 0.0244
11. 0.0038; either a very rare event has occurred or the husband is not the father.
13. 0.9554
15. a. 0.0179
    b. 1343
17. 0.1706. This answer assumes a normal distribution, but the expenditure distribution is more likely to be bimodal: Many households will spend nothing on postsecondary textbooks (because no one is attending postsecondary institutions); and the minority who are buying textbooks will spend a great deal.
19. 0.0013. No, it is rare for a person in this age group to have a seriously elevated serum cholesterol level.
21. a. Normal distribution
    b. $\bar{x} = 0.9147$, $s = 0.0369$
    c. 0.0104
23. a. Normal distribution
    b. $\bar{x} = 35.66$, $s = 9.35$
    c. 0.271

Section ~~5–~~4 9.4

1. 66.9 in.
3. 61.5 in.
5. a. 9.1 years
   b. 0.0018
   c. 5.6 years.
7. 3.5 kg
9. 242 days
11. 135. Perhaps the employees of a “think tank” company could be held to this standard.
13. a. 0.9656
    b. 658.9
15. 3497.4
17. 11.9
19. 131.6; 782

Chapter 10

Confidence Intervals

Reading for Chapter 10: The following reading is excerpted from:

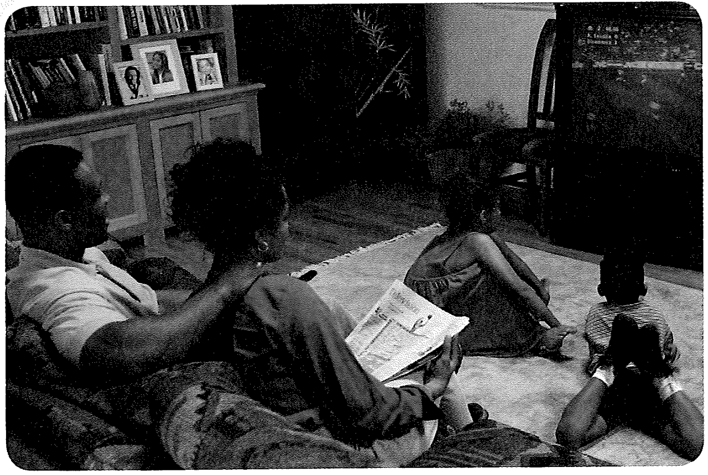
> Bluman and Mayer. Elementary Statistics: A step by step approach. Canadian edition, Mc-Graw Hill Ryerson, 2008, pages 296-307, 748.

Statistics Today

Would You Change the Channel?

A survey by the Roper Organization found that 45% of the people who were offended by a television program would change the channel, while 15% would turn off their television sets. The survey further stated that the margin of error is 3 percentage points, and 4000 adults were interviewed.

Several questions arise:



1. How do these estimates compare with the true population percentages?
2. What is meant by a margin of error of 3 percentage points?
3. Is the sample of 4000 large enough to represent the population of all adults who watch television in Canada?

See Statistics Today—Revisited, page 328, at the end of the chapter for the answers.

After reading this chapter, you will be able to answer these questions, since this chapter explains how statisticians can use statistics to make estimates of parameters.

Source: The Associated Press.

Section 10.1:

~~7–1~~ Introduction

One aspect of inferential statistics is **estimation,** which is the process of estimating the value of a parameter from information obtained from a sample. For example, Statistics Canada and other public as well as private agencies collect economic, societal, and cultural data of the people and resources of the country. Examples of data disseminated through these sources include the following.

> *"The average Canadian male spends $71.73 per week on food compared to female weekly food expenditure of $61.78." (Statistics Canada)*[1]
>
> *"1 in 5 Canadians list talking on mobile phones as their biggest grievance with other drivers on the road, yet 1 in 4 are guilty of it themselves." (Leger Marketing)*
>
> *"Television viewing by Canadians averages 21 hours per week." (Statistics Canada)*[2]
>
> *"44% of Canadian drivers change their own vehicle spark plugs." (DesRosiers Automotive Consultants)*
>
> *"Cannabis dependence is significantly higher among youth (27%) than adults (5%)." (Canadian Community Epidemiology Network on Drug Use)*

Since the populations from which these values were obtained are large, these values are only *estimates* of the true parameters and are derived from data collected from samples.

The statistical procedures for estimating the population mean, proportion, variance, and standard deviation will be explained in this chapter.

[1] Adapted from Statistics Canada publication *Food Expenditure in Canada,* Catalogue 62–554, 2001, Release date: February 21, 2003, page 55.

[2] Adapted from Statistics Canada website http://dissemination.statcan.ca/Daily/English/021202/d021202a.htm. February 28, 2007.

note: no exercises for 10.1

An important question in estimation is that of sample size. How large should the sample be in order to make an accurate estimate? This question is not easy to answer since the size of the sample depends on several factors, such as the accuracy desired and the probability of making a correct estimate. The question of sample size will be explained in this chapter also.

Section 10.2:

~~7-2~~

Confidence Intervals for the Mean (σ Known or $n \geq 30$) and Sample Size

Objective 1

Find the confidence interval for the mean when σ is known or $n \geq 30$.

Suppose a college president wishes to estimate the average age of students attending classes this semester. The president could select a random sample of 100 students and find the average age of these students, say, 22.3 years. From the sample mean, the president could infer that the average age of all the students is 22.3 years. This type of estimate is called a *point estimate.*

A point estimate is a specific numerical value estimate of a parameter. The best point estimate of the population mean μ is the sample mean $\overline{X}$.

One might ask why other measures of central tendency, such as the median and mode, are not used to estimate the population mean. The reason is that the means of samples vary less than other statistics (such as medians and modes) when many samples are selected from the same population. Therefore, the sample mean is the best estimate of the population mean.

Sample measures (i.e., statistics) are used to estimate population measures (i.e., parameters). These statistics are called **estimators.** As previously stated, the sample mean is a better estimator of the population mean than the sample median or sample mode.

A good estimator should satisfy the three properties described now.

Three Properties of a Good Estimator

1. The estimator should be an **unbiased estimator.** That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
2. The estimator should be consistent. For a **consistent estimator,** as sample size increases, the value of the estimator approaches the value of the parameter estimated.
3. The estimator should be a **relatively efficient estimator.** That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.

Confidence Intervals

As stated in Chapter 6, the sample mean will be, for the most part, somewhat different from the population mean due to sampling error. Therefore, one might ask a second question: How good is a point estimate? The answer is that there is no way of knowing how close a particular point estimate is to the population mean.

This answer places some doubt on the accuracy of point estimates. For this reason, statisticians prefer another type of estimate, called an *interval estimate.*

An **interval estimate** of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average age of all students might be $26.9 < \mu < 27.7$, or 27.3 ± 0.4 years.

Either the interval contains the parameter or it does not. A degree of confidence (usually a percent) can be assigned before an interval estimate is made. For instance, one may wish to be 95% confident that the interval contains the true population mean. Another question then arises. Why 95%? Why not 99 or 99.5%?

If one desires to be more confident, such as 99 or 99.5% confident, then the interval must be larger. For example, a 99% confidence interval for the mean age of college students might be $26.7 < \mu < 27.9$, or 27.3 ± 0.6. Hence, a tradeoff occurs. To be more confident that the interval contains the true population mean, one must make the interval wider.

Historical Note

Point and interval estimates were known as long ago as the late 1700s. However, it wasn't until 1937 that a mathematician, Jerzy Neyman, formulated practical applications for them.

The **confidence level** of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated.

A **confidence interval** is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate.

Intervals constructed in this way are called *confidence intervals.* Three common confidence intervals are used: the 90, the 95, and the 99% confidence intervals.

The algebraic derivation of the formula for determining a confidence interval for a mean will be shown later. A brief intuitive explanation will be given first.

The central limit theorem states that when the sample size is large, approximately 95% of the sample means will fall within ±1.96 standard errors of the population mean, that is,

$$\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Now, if a specific sample mean is selected, say, $\overline{X}$, there is a 95% probability that it falls within the range of $\mu \pm 1.96(\sigma/\sqrt{n})$. Likewise, there is a 95% probability that the interval specified by

$$\overline{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

will contain μ, as will be shown later. Stated another way,

$$\overline{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Hence, one can be 95% confident that the population mean is contained within that interval when the values of the variable are normally distributed in the population.

Interesting Fact

Golfers walk, on average, 8 kilometres and burn 1750 calories per round.

For example, in order to target new account advertising, a researcher would like an estimate of the average age and household income of all online stock traders at a confidence level of 95%. Using techniques presented in this chapter, the researcher selects a random sample of new accounts and discovers that on average online stock traders are between 39.5 and 46.5 years old with household incomes between \$89,750 and \$110,250. With a 95% degree of confidence in the estimates of the true population mean, the researcher will target these demographics for new accounts.

The value used for the 95% confidence interval, 1.96, is obtained from Table E–2 in Appendix C. For a 99% confidence interval, the value 2.575 is used instead of 1.96 in the formula. This value is also obtained from Table E–2 and is based on the standard normal distribution. Since other confidence intervals are used in statistics, the symbol $z_{\alpha/2}$ (read "zed sub alpha over two") is used in the general formula for confidence intervals. The Greek letter α (alpha) represents the total area in both tails of the standard normal distribution curve, and $\alpha/2$ represents the area in each one of the tails. More will be said after Examples 7–1 and 7–2 about finding other values for $z_{\alpha/2}$.

The relationship between α and the confidence level is that the stated confidence level is the percentage equivalent to the decimal value of $1 - \alpha$, and vice versa. When the 95% confidence interval is to be found, $\alpha = 0.05$, since $1 - 0.05 = 0.95$, or 95%. When $\alpha = 0.01$, then $1 - \alpha = 1 - 0.01 = 0.99$, and the 99% confidence interval is being calculated.

Formula for the Confidence Interval of the Mean for a Specific α

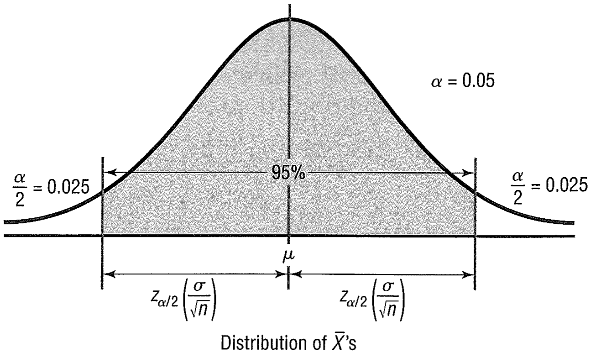
$$\overline{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \overline{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, $z_{\alpha/2} = 1.645$; for a 95% confidence interval, $z_{\alpha/2} = 1.96$; and for a 99% confidence interval, $z_{\alpha/2} = 2.575$.

The term $z_{\alpha/2}(\sigma/\sqrt{n})$ is called the *margin of error,* also referred to as the *maximum error of estimate.* For a specific value, say, $\alpha = 0.05$, 95% of the sample means will fall within this error value on either side of the population mean, as previously explained. See Figure 7–1.

Figure 7–1

95% Confidence Interval



The **margin of error** is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter.

A more detailed explanation of the margin of error follows Examples 7–1 and 7–2, which illustrate the computation of confidence intervals.

Rounding Rule for a Confidence Interval for a Mean When you are computing a confidence interval for a population mean by using *raw data,* round off to one more decimal place than the number of decimal places in the original data. When you are computing a confidence interval for a population mean by using a sample mean and a standard deviation, round off to the same number of decimal places as given for the mean.

Example 7–1

A researcher wishes to estimate the average amount of money a person spends on lottery tickets each month. A sample of 50 people who play the lottery found the mean to be \$19 and the standard deviation to be 6.8. Find the best point estimate of the population mean and the 95% confidence interval of the population mean.

Source: USA TODAY.

Solution

The best point estimate of the mean is \$19. For the 95% confidence interval use $z = 1.96$.

$$19 - 1.96\left(\frac{6.8}{\sqrt{50}}\right) < \mu < 19 + 1.96\left(\frac{6.8}{\sqrt{50}}\right)$$

$$19 - 1.9 < \mu < 19 + 1.9$$

$$17.1 < \mu < 20.9$$

or

$$19 \pm 1.9$$

Hence, one can say with 95% confidence the true mean of the population is between \$17.10 and \$20.90, based on a sample of 50 people who play the lottery.

Example 7–2

A survey of 30 adults found that the mean age of a person's primary vehicle is 5.6 years. Assuming the standard deviation of the population is 0.8 year, find the best point estimate of the population mean and the 99% confidence interval of the population mean.

Source: Based on information in *USA TODAY.*

Solution

The best point estimate of the population mean is 5.6 years.

$$5.6 - 2.575\left(\frac{0.8}{\sqrt{30}}\right) < \mu < 5.6 + 2.575\left(\frac{0.8}{\sqrt{30}}\right)$$

$$5.6 - 0.38 < \mu < 5.6 + 0.38$$

$$5.22 < \mu < 5.98$$

or

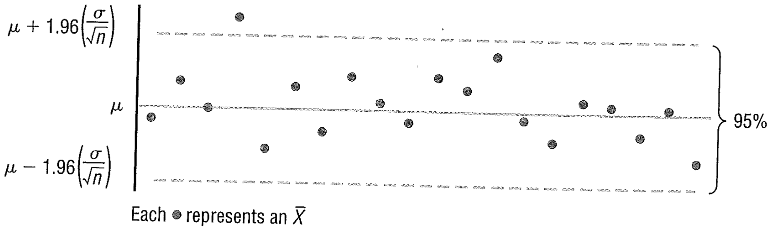
$$5.2 < \mu < 6.0 \text{ (rounded)}$$

Hence, one can be 99% confident that the mean age of all primary vehicles is between 5.2 and 6.0 years, based on 30 vehicles.

Another way of looking at a confidence interval is shown in Figure 7–2. According to the central limit theorem, approximately 95% of the sample means fall within 1.96 standard deviations of the population mean if the sample size is 30 or more or if σ is known when n is less than 30 and the population is normally distributed. If it were possible to build a confidence interval about each sample mean, as was done in Examples 7–1 and 7–2

Figure 7–2

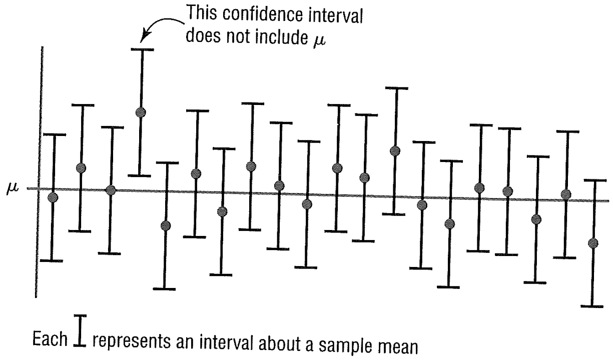
95% Confidence Interval for Sample Means



for μ, 95% of these intervals would contain the population mean, as shown in Figure 7–3. Hence, one can be 95% confident that an interval built around a specific sample mean would contain the population mean. If one desires to be 99% confident, the confidence intervals must be enlarged so that 99 out of every 100 intervals contain the population mean.

Figure 7–3

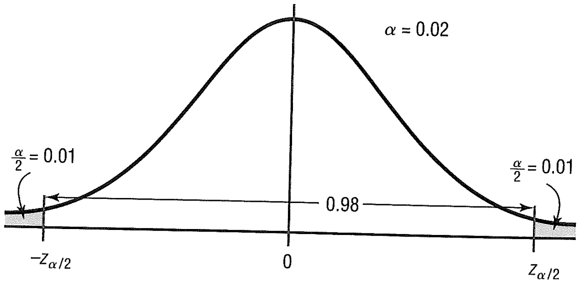
95% Confidence Intervals for Each Sample Mean



Since other confidence intervals (besides 90, 95, and 99%) are sometimes used in statistics, an explanation of how to find the values for $z_{\alpha/2}$ is necessary. As stated previously, the Greek letter α represents the total of the areas in both tails of the normal distribution. The value for α is found by subtracting the decimal equivalent for the desired confidence level from 1. For example, if one wanted to find the 98% confidence interval, one would change 98% to 0.98 and find $\alpha = 1 - 0.98$, or 0.02. Then $\alpha/2$ is obtained by dividing α by 2. So $\alpha/2$ is 0.02/2, or 0.01. Finally, $z_{0.01}$ is the z value that will give an area of 0.01 in both the left and right tails of the standard normal distribution curve. See Figure 7–4.

Figure 7–4

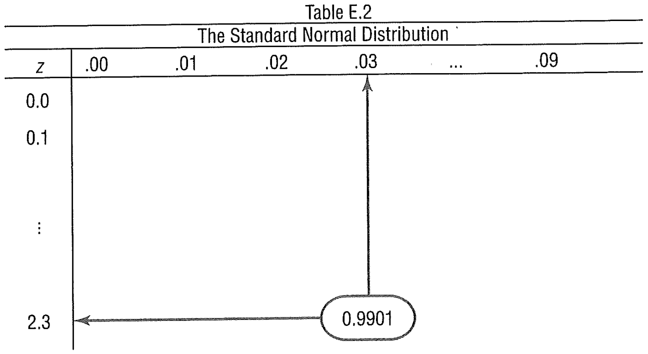
Finding $\alpha/2$ for a 98% Confidence Interval



Once $\alpha/2$ is determined, the corresponding $z_{\alpha/2}$ score can be found by using the procedure shown in Section 6–3 "Finding z Scores When Given Areas". In summary, to obtain the $z_{\alpha/2}$ score for a 98% confidence level, subtract 0.9800 from 1.000 to obtain $\alpha = 0.02$, therefore $\alpha/2 = 0.02/2 = 0.01$. In Table E–2—Positive z Scores the area 0.0100 to the right is the cumulative area 0.9900 (1.0000 − 0.0100) to the left, where $z = 2.33$. Refer to Figure 7–5.

Figure 7–5

Finding $z_{\alpha/2}$ for a 98% Confidence Interval



For confidence intervals, only the positive z score is used in the formula.

When the original variable is normally distributed and σ is known, the standard normal distribution can be used to find confidence intervals regardless of the size of the sample. When $n \geq 30$, the distribution of means will be approximately normal even if the original distribution of the variable departs from normality. Also, if $n \geq 30$ (some authors use $n > 30$), s can be substituted for σ in the formula for confidence intervals; and the standard normal distribution can be used to find confidence intervals for means, as shown in Example 7–3.

Example 7–3

The following data represent a random sample of 30 house prices (in thousands of dollars) in Calgary. Find the 90% confidence interval estimate of the mean price of all Calgary houses.

318.3	507.5	351.9	292.2	487.7	587.2
373.9	261.5	261.5	350.6	381.0	782.3
508.0	262.8	289.3	360.9	447.7	795.8
437.3	276.2	300.0	441.4	545.0	337.0
450.8	317.5	332.5	634.6	945.3	393.6

Source: Calgary Real Estate Board.

Solution

Step 1 Find the mean and standard deviation for the data. Use the formula shown in Chapter 3 or your calculator. The mean $\overline{X} = 434.38$. The standard deviation $s = 171.25$.

Step 2 Find $\alpha/2$. Since the 90% confidence interval is to be used, $\alpha = 1 - 0.90 = 0.10$, and

$$\frac{\alpha}{2} = \frac{0.10}{2} = 0.05$$

Step 3 Find $z_{\alpha/2}$. Subtract 0.05 (area in right tail) from 1.0000 to get 0.9500 (cumulative area to the left). The corresponding z value obtained from Table E–2 for 0.9500 is $z = 1.645$, as this score is halfway between $z = 1.64$ (0.9495) and $z = 1.65$ (0.9505).

Step 4 Substitute into the formula

$$\overline{X} - z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) < \mu < \overline{X} + z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

(Since $n \geq 30$, s is used in place of σ when σ is unknown.)

$$434.38 - 1.645\left(\frac{171.25}{\sqrt{30}}\right) < \mu < 434.38 + 1.645\left(\frac{171.25}{\sqrt{30}}\right)$$

$$434.38 - 51.43 < \mu < 434.38 + 51.43$$

$$382.94 < \mu < 485.81$$

Hence, one can be 90% confident that the population mean of Calgary house prices is between \$382,940 and \$485,810, based on a sample of 30 house sales.

Comment to Computer and Statistical Calculator Users

This chapter and subsequent chapters include examples using raw data. If you are using computer or calculator programs to find the solutions, the answers you get may vary somewhat from the ones given in the textbook. This is so because computers and calculators do not round the answers in the intermediate steps and can use 12 or more decimal places for computation. Also, they use more exact values than those given in the tables in the back of this book. These discrepancies are part and parcel of statistics.

Objective 2

Determine the minimum sample size for finding a confidence interval for the mean.

Sample Size

Sample size determination is closely related to statistical estimation. Quite often, one asks, How large a sample is necessary to make an accurate estimate? The answer is not simple, since it depends on three things: the maximum error of estimate, the population standard deviation, and the degree of confidence. For example, how close to the true mean does one want to be (2 units, 5 units, etc.), and how confident does one wish to be (90, 95, 99%, etc.)? For the purpose of this chapter, it will be assumed that the population standard deviation of the variable is known or has been estimated from a previous study.

The formula for sample size is derived from the margin of error formula

$$E = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

and this formula is solved for n as follows:

$$E\sqrt{n} = z_{\alpha/2}(\sigma)$$

$$\sqrt{n} = \frac{z_{\alpha/2} \cdot \sigma}{E}$$

Hence, $$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

Formula for the Minimum Sample Size Needed for an Interval Estimate of the Population Mean

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$

where E is the margin of error. If necessary, round the answer *up* to obtain a whole number. That is, if there is any fraction or decimal portion in the answer, use the next whole number for sample size n.

Example 7–4

The college president asks the statistics teacher to estimate the average age of the students at their college. How large a sample is necessary? The statistics teacher would like to be 99% confident that the estimate should be accurate within 1 year. From a previous study, the standard deviation of the ages is known to be 3 years.

Solution

Since $\alpha = 0.01$ (or $1 - 0.99$), $z_{\alpha/2} = 2.575$, and $E = 1$, substituting in the formula, one gets

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2 = \left[\frac{(2.575)(3)}{1}\right]^2 \approx 59.7$$

which is rounded up to 60. Therefore, to be 99% confident that the estimate is within 1 year of the true mean age, the teacher needs a sample size of at least 60 students. (Always round n up to the next whole number. For example, if $n = 59.2$, round it up to 60.)

Notice that when one is finding the sample size, the size of the population is irrelevant when the population is large or infinite or when sampling is done with replacement. In other cases, an adjustment is made in the formula for computing sample size. This adjustment is beyond the scope of this book.

Interesting Fact

Canadian households with post-secondary students spend more on tuition fees than food in a year.

The formula for determining sample size requires the use of the population standard deviation. What happens when σ is unknown? In this case, an attempt is made to estimate σ. One such way is to use the standard deviation s obtained from a sample taken previously as an estimate for σ. The standard deviation can also be estimated by dividing the range by 4.

Sometimes, interval estimates rather than point estimates are reported. For instance, one may read a statement: "On the basis of a sample of 200 families, the survey estimates that a Canadian family of two spends an average of \$84 per week for groceries. One can be 95% confident that this estimate is accurate within \$3 of the true mean." This statement means that the 95% confidence interval of the true mean is

$$\$84 - \$3 < \mu < \$84 + \$3$$
$$\$81 < \mu < \$87$$

The algebraic derivation of the formula for a confidence interval is shown next. As explained in Chapter 6, the sampling distribution of the mean is approximately normal when large samples ($n \geq 30$) are taken from a population. Also,

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Furthermore, there is a probability of $1 - \alpha$ that a z will have a value between $-z_{\alpha/2}$ and $+z_{\alpha/2}$. Hence,

$$-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

Using algebra, one finds

$$-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Subtracting $\overline{X}$ from both sides and from the middle, one gets

$$-\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Multiplying by -1, one gets

$$\overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Reversing the inequality, one gets the formula for the confidence interval:

$$\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Applying the Concepts 7–2

Making Decisions with Confidence Intervals

Assume you work for Kimberly Clark Corporation, the makers of Kleenex. The job you are presently working on requires you to decide how many Kleenexes are to be put in the new automobile glove compartment boxes. Complete the following.

1. How will you decide on a reasonable number of Kleenexes to put in the boxes?
2. When do people usually need Kleenexes?
3. What type of data collection technique would you use?
4. Assume you found out that from your sample of 85 people, on average about 57 Kleenexes are used throughout the duration of a cold, with a standard deviation of 15. Use a confidence interval to help you decide how many Kleenexes will go in the boxes.
5. Explain how you decided on how many Kleenexes will go in the boxes.

See page 331 for the answers.

Exercises for Section 10.2

Exercises 7–2

1. What is the difference between a point estimate and an interval estimate of a parameter? Which is better? Why?
2. What information is necessary to calculate a confidence interval?
3. What is the margin of error?
4. What is meant by the 95% confidence interval of the mean?
5. What are three properties of a good estimator?
6. What statistic best estimates μ?
7. What is necessary to determine the sample size?
8. When one is determining the sample size for a confidence interval, is the size of the population relevant?
9. Find each.

   a. $z_{\alpha/2}$ for the 99% confidence interval

   b. $z_{\alpha/2}$ for the 98% confidence interval

   c. $z_{\alpha/2}$ for the 95% confidence interval

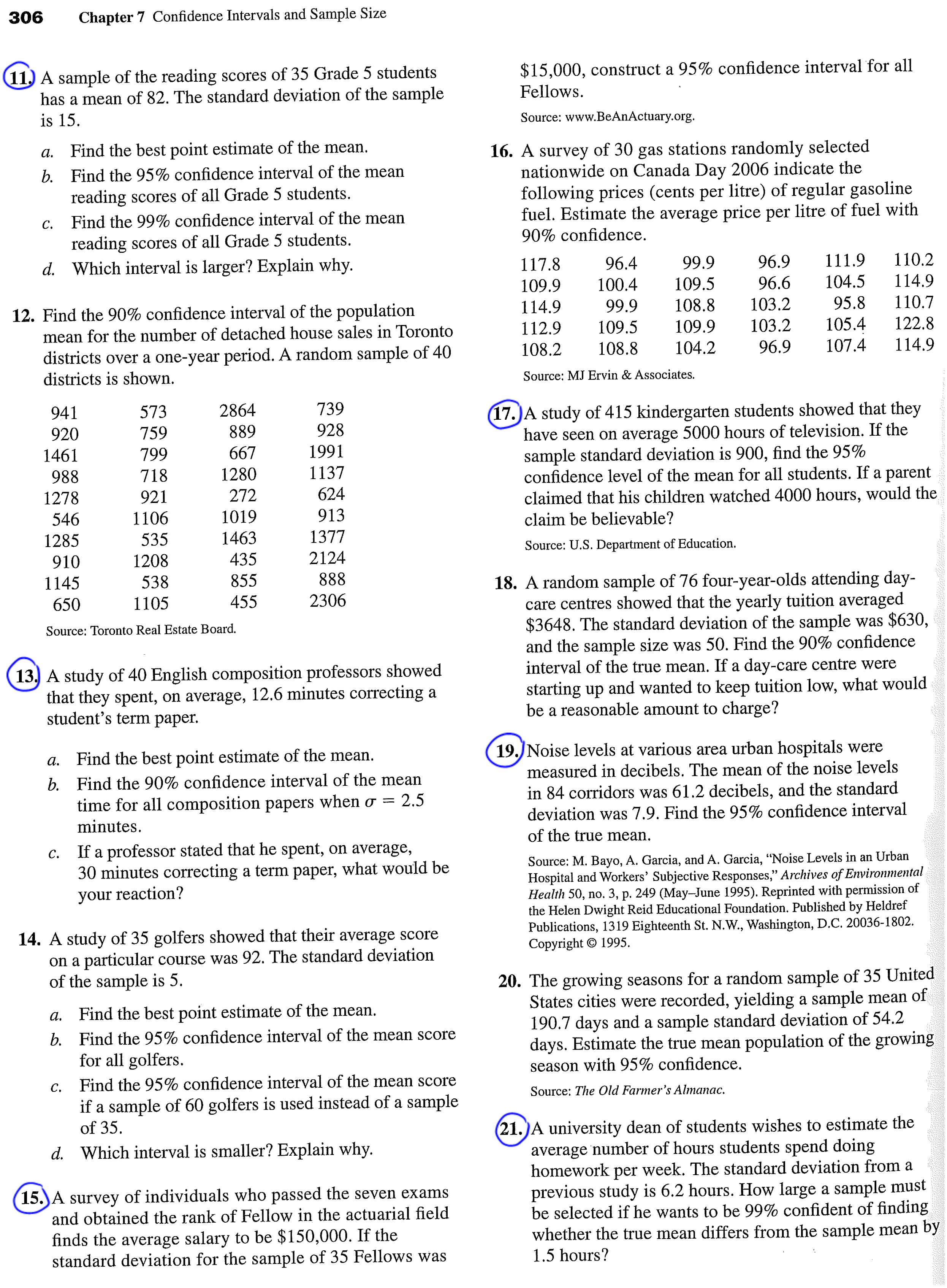
   d. $z_{\alpha/2}$ for the 90% confidence interval

   e. $z_{\alpha/2}$ for the 94% confidence interval

10. Find the 95% confidence interval for the mean paid attendance at the Major League Baseball All-Star games. A random sample of the paid attendances is shown.

47,596	68,751	5,838
69,831	28,843	53,107
31,391	48,829	50,706
62,892	55,105	63,974
56,674	38,362	51,549
31,938	31,851	56,088
34,906	38,359	72,086
34,009	50,850	43,801
46,127	49,926	54,960
32,785	48,321	49,671

Source: *Time Almanac.*



11. A sample of the reading scores of 35 Grade 5 students has a mean of 82. The standard deviation of the sample is 15.

a. Find the best point estimate of the mean.

b. Find the 95% confidence interval of the mean reading scores of all Grade 5 students.

c. Find the 99% confidence interval of the mean reading scores of all Grade 5 students.

d. Which interval is larger? Explain why.

12. Find the 90% confidence interval of the population mean for the number of detached house sales in Toronto districts over a one-year period. A random sample of 40 districts is shown.

941	573	2864	739
920	759	889	928
1461	799	667	1991
988	718	1280	1137
1278	921	272	624
546	1106	1019	913
1285	535	1463	1377
910	1208	435	2124
1145	538	855	888
650	1105	455	2306

Source: Toronto Real Estate Board.

13. A study of 40 English composition professors showed that they spent, on average, 12.6 minutes correcting a student's term paper.

a. Find the best point estimate of the mean.

b. Find the 90% confidence interval of the mean time for all composition papers when $\sigma = 2.5$ minutes.

c. If a professor stated that he spent, on average, 30 minutes correcting a term paper, what would be your reaction?

14. A study of 35 golfers showed that their average score on a particular course was 92. The standard deviation of the sample is 5.

a. Find the best point estimate of the mean.

b. Find the 95% confidence interval of the mean score for all golfers.

c. Find the 95% confidence interval of the mean score if a sample of 60 golfers is used instead of a sample of 35.

d. Which interval is smaller? Explain why.

15. A survey of individuals who passed the seven exams and obtained the rank of Fellow in the actuarial field finds the average salary to be $150,000. If the standard deviation for the sample of 35 Fellows was $15,000, construct a 95% confidence interval for all Fellows.

Source: www.BeAnActuary.org.

16. A survey of 30 gas stations randomly selected nationwide on Canada Day 2006 indicate the following prices (cents per litre) of regular gasoline fuel. Estimate the average price per litre of fuel with 90% confidence.

117.8	96.4	99.9	96.9	111.9	110.2
109.9	100.4	109.5	96.6	104.5	114.9
114.9	99.9	108.8	103.2	95.8	110.7
112.9	109.5	109.9	103.2	105.4	122.8
108.2	108.8	104.2	96.9	107.4	114.9

Source: MJ Ervin & Associates.

17. A study of 415 kindergarten students showed that they have seen on average 5000 hours of television. If the sample standard deviation is 900, find the 95% confidence level of the mean for all students. If a parent claimed that his children watched 4000 hours, would the claim be believable?

Source: U.S. Department of Education.

18. A random sample of 76 four-year-olds attending day-care centres showed that the yearly tuition averaged $3648. The standard deviation of the sample was $630, and the sample size was 50. Find the 90% confidence interval of the true mean. If a day-care centre were starting up and wanted to keep tuition low, what would be a reasonable amount to charge?

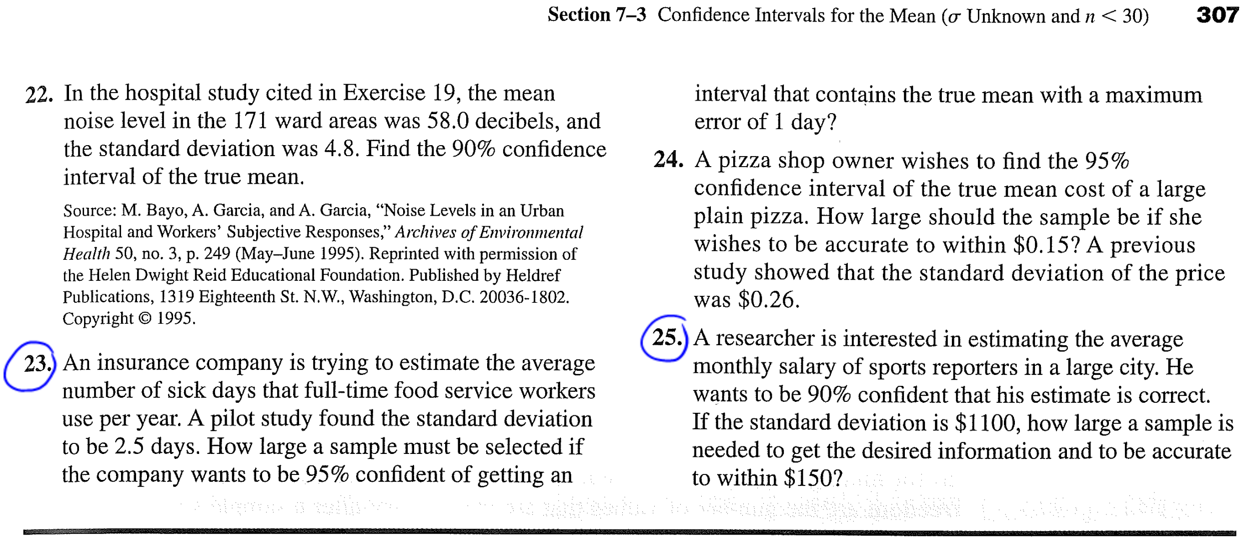
19. Noise levels at various area urban hospitals were measured in decibels. The mean of the noise levels in 84 corridors was 61.2 decibels, and the standard deviation was 7.9. Find the 95% confidence interval of the true mean.

Source: M. Bayo, A. Garcia, and A. Garcia, "Noise Levels in an Urban Hospital and Workers' Subjective Responses," *Archives of Environmental Health* 50, no. 3, p. 249 (May–June 1995). Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 1319 Eighteenth St. N.W., Washington, D.C. 20036-1802. Copyright © 1995.

20. The growing seasons for a random sample of 35 United States cities were recorded, yielding a sample mean of 190.7 days and a sample standard deviation of 54.2 days. Estimate the true mean population of the growing season with 95% confidence.

Source: *The Old Farmer's Almanac.*

21. A university dean of students wishes to estimate the average number of hours students spend doing homework per week. The standard deviation from a previous study is 6.2 hours. How large a sample must be selected if he wants to be 99% confident of finding whether the true mean differs from the sample mean by 1.5 hours?



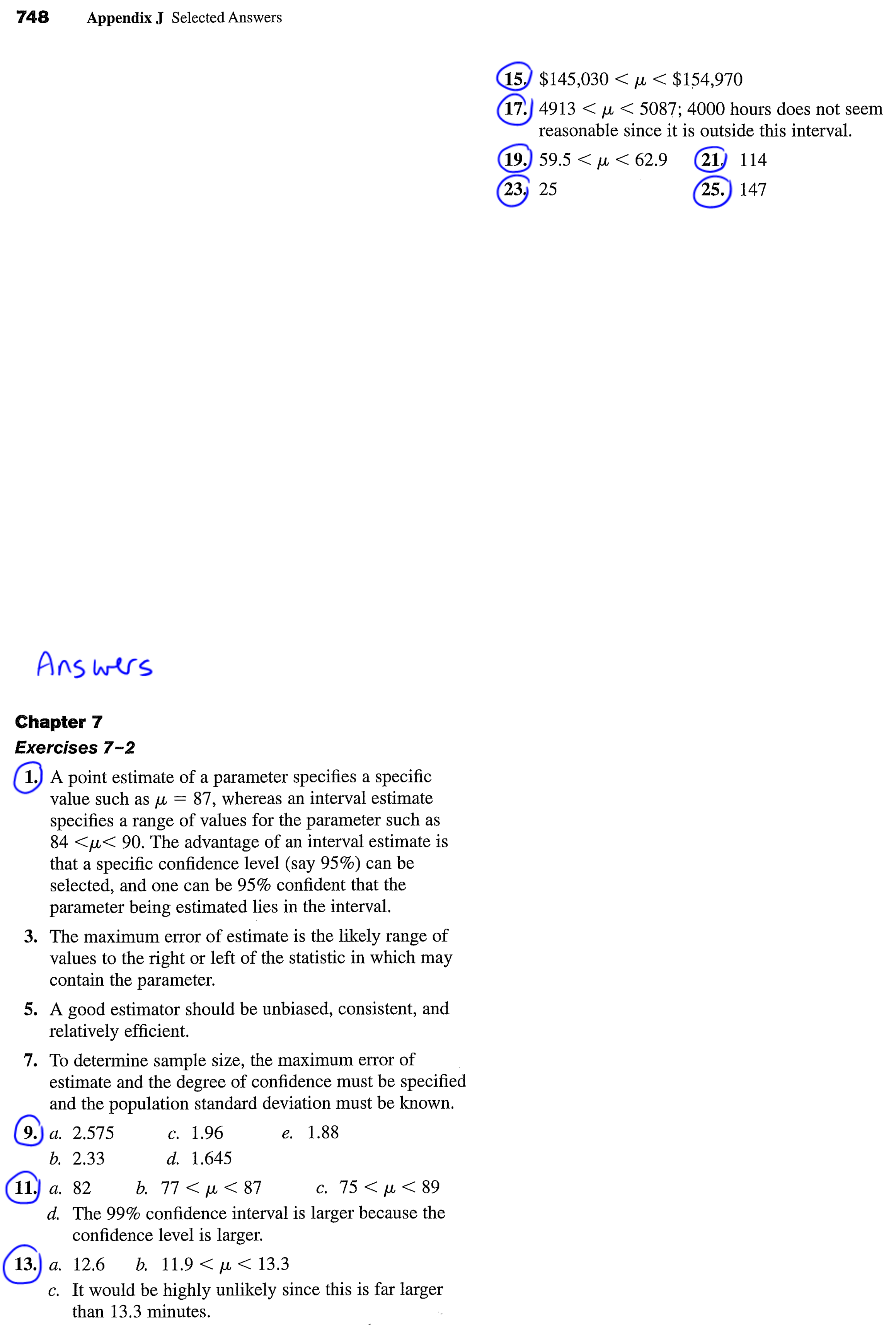
22. In the hospital study cited in Exercise 19, the mean noise level in the 171 ward areas was 58.0 decibels, and the standard deviation was 4.8. Find the 90% confidence interval of the true mean.

Source: M. Bayo, A. Garcia, and A. Garcia, "Noise Levels in an Urban Hospital and Workers' Subjective Responses," *Archives of Environmental Health* 50, no. 3, p. 249 (May–June 1995). Reprinted with permission of the Helen Dwight Reid Educational Foundation. Published by Heldref Publications, 1319 Eighteenth St. N.W., Washington, D.C. 20036-1802. Copyright © 1995.

23. An insurance company is trying to estimate the average number of sick days that full-time food service workers use per year. A pilot study found the standard deviation to be 2.5 days. How large a sample must be selected if the company wants to be 95% confident of getting an interval that contains the true mean with a maximum error of 1 day?

24. A pizza shop owner wishes to find the 95% confidence interval of the true mean cost of a large plain pizza. How large should the sample be if she wishes to be accurate to within \$0.15? A previous study showed that the standard deviation of the price was \$0.26.

25. A researcher is interested in estimating the average monthly salary of sports reporters in a large city. He wants to be 90% confident that his estimate is correct. If the standard deviation is \$1100, how large a sample is needed to get the desired information and to be accurate to within \$150?



Answers

Chapter 7

Exercises 7–2

1. A point estimate of a parameter specifies a specific value such as $\mu = 87$, whereas an interval estimate specifies a range of values for the parameter such as $84 < \mu < 90$. The advantage of an interval estimate is that a specific confidence level (say 95%) can be selected, and one can be 95% confident that the parameter being estimated lies in the interval.

3. The maximum error of estimate is the likely range of values to the right or left of the statistic in which may contain the parameter.

5. A good estimator should be unbiased, consistent, and relatively efficient.

7. To determine sample size, the maximum error of estimate and the degree of confidence must be specified and the population standard deviation must be known.

9. *a.* 2.575 *c.* 1.96 *e.* 1.88
b. 2.33 *d.* 1.645

11. *a.* 82 *b.* $77 < \mu < 87$ *c.* $75 < \mu < 89$
d. The 99% confidence interval is larger because the confidence level is larger.

13. *a.* 12.6 *b.* $11.9 < \mu < 13.3$
c. It would be highly unlikely since this is far larger than 13.3 minutes.

15. $\$145{,}030 < \mu < \$154{,}970$

17. $4913 < \mu < 5087$; 4000 hours does not seem reasonable since it is outside this interval.

19. $59.5 < \mu < 62.9$ **21.** 114

23. 25 **25.** 147