The principles of probability help bridge the worlds of descriptive statistics and inferential statistics. Probability principles are the foundation for the probability distribution, the concept of mathematical expectation, and the binomial and Poisson distributions, topics that are discussed in Chapter 5. In this chapter, you will learn about probability to answer questions such as the following:

- What is the probability that a household is planning to purchase a large-screen HDTV in the next year?
- What is the probability that a household will actually purchase a large-screen HDTV?
- What is the probability that a household is planning to purchase a large-screen HDTV and actually purchases the television?
- Given that the household is planning to purchase a large-screen HDTV, what is the probability that the purchase is made?
- Does knowledge of whether a household *plans* to purchase the television change the likelihood of predicting whether the household *will* purchase the television?
- What is the probability that a household that purchases a large-screen HDTV will purchase a television with a faster refresh rate?
- What is the probability that a household that purchases a large-screen HDTV with a faster refresh rate will also purchase a streaming media box?
- What is the probability that a household that purchases a large-screen HDTV will be satisfied with the purchase?

With answers to questions such as these, you can begin to form a marketing strategy. You can consider whether to target households that have indicated an intent to purchase or to focus on selling televisions that have faster refresh rates or both. You can also explore whether households that purchase large-screen HDTVs with faster refresh rates can be easily persuaded to also purchase streaming media boxes.

4.1 Basic Probability Concepts

Student Tip
Remember, a probability cannot be negative or greater than 1.

What is meant by the word *probability*? A **probability** is the numerical value representing the chance, likelihood, or possibility that a particular event will occur, such as the price of a stock increasing, a rainy day, a defective product, or the outcome five dots in a single toss of a die. In all these instances, the probability involved is a proportion or fraction whose value ranges between 0 and 1, inclusive. An event that has no chance of occurring (the **impossible event**) has a probability of 0. An event that is sure to occur (the **certain event**) has a probability of 1.

There are three types of probability:

- *A priori*
- Empirical
- Subjective

In the simplest case, where each outcome is equally likely, the chance of occurrence of the event is defined in Equation (4.1).

PROBABILITY OF OCCURRENCE

$$\text{Probability of occurrence} = \frac{X}{T} \quad (4.1)$$

where

X = number of ways in which the event occurs
T = total number of possible outcomes

In ***a priori* probability**, the probability of an occurrence is based on prior knowledge of the process involved. Consider a standard deck of cards that has 26 red cards and 26 black cards. The probability of selecting a black card is $26/52 = 0.50$ because there are $X = 26$ black cards and $T = 52$ total cards. What does this probability mean? If each card is replaced after it is selected, does it mean that 1 out of the next 2 cards selected will be black? No, because you cannot say for certain what will happen on the next several selections. However, you can say that in the long run, if this selection process is continually repeated, the proportion of black cards selected will approach 0.50. Example 4.1 shows another example of computing an *a priori* probability.

EXAMPLE 4.1

Finding A Priori Probabilities

A standard six-sided die has six faces. Each face of the die contains either one, two, three, four, five, or six dots. If you roll a die, what is the probability that you will get a face with five dots?

SOLUTION Each face is equally likely to occur. Because there are six faces, the probability of getting a face with five dots is 1/6.

The preceding examples use the *a priori* probability approach because the number of ways the event occurs and the total number of possible outcomes are known from the composition of the deck of cards or the faces of the die.

In the **empirical probability** approach, the probabilities are based on observed data, not on prior knowledge of a process. Surveys are often used to generate empirical probabilities. Examples of this type of probability are the proportion of individuals in the Using Statistics scenario who actually purchase large-screen HDTVs, the proportion of registered voters who prefer a certain political candidate, and the proportion of students who have part-time jobs. For example, if you take a survey of students, and 60% state that they have part-time jobs, then there is a 0.60 probability that an individual student has a part-time job.

The third approach to probability, **subjective probability**, differs from the other two approaches because subjective probability differs from person to person. For example, the development team for a new product may assign a probability of 0.60 to the chance of success for the product, while the president of the company may be less optimistic and assign a probability of 0.30. The assignment of subjective probabilities to various outcomes is usually based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation. Subjective probability is especially useful in making decisions in situations in which you cannot use *a priori* probability or empirical probability.

Events and Sample Spaces

The basic elements of probability theory are the individual outcomes of a variable under study. You need the following definitions to understand probabilities.

Student Tip
Events are represented by letters of the alphabet.

EVENT

Each possible outcome of a variable is referred to as an **event**.
A **simple event** is described by a single characteristic.

For example, when you toss a coin, the two possible outcomes are heads and tails. Each of these represents a simple event. When you roll a standard six-sided die in which the six faces of the die contain either one, two, three, four, five, or six dots, there are six possible simple events. An event can be any one of these simple events, a set of them, or a subset of all of them. For example, the event of an *even number of dots* consists of three simple events (i.e., two, four, or six dots).

Student Tip
The key word when describing a joint event is *and*.

JOINT EVENT

A **joint event** is an event that has two or more characteristics.

Getting two heads when you toss a coin twice is an example of a joint event because it consists of heads on the first toss and heads on the second toss.

COMPLEMENT

The **complement** of event A (represented by the symbol A') includes all events that are not part of A.

The complement of a head is a tail because that is the only event that is not a head. The complement of five dots on a die is not getting five dots. Not getting five dots consists of getting one, two, three, four, or six dots.

SAMPLE SPACE

The collection of all the possible events is called the **sample space**.

The sample space for tossing a coin consists of heads and tails. The sample space when rolling a die consists of one, two, three, four, five, and six dots. Example 4.2 demonstrates events and sample spaces.

EXAMPLE 4.2

Events and Sample Spaces

The Using Statistics scenario on page 149 concerns M&R Electronics World. Table 4.1 presents the results of the sample of 1,000 households in terms of purchase behavior for large-screen HDTVs.

TABLE 4.1

Purchase Behavior for Large-screen HDTVs

PLANNED TO PURCHASE	ACTUALLY PURCHASED		
	Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

What is the sample space? Give examples of simple events and joint events.

SOLUTION The sample space consists of the 1,000 respondents. Simple events are "planned to purchase," "did not plan to purchase," "purchased," and "did not purchase." The complement of the event "planned to purchase" is "did not plan to purchase." The event "planned to purchase and actually purchased" is a joint event because in this joint event, the respondent must plan to purchase the television *and* actually purchase it.

Contingency Tables and Venn Diagrams

There are several ways in which you can view a particular sample space. One way involves using a **contingency table** (see Section 2.1) such as the one displayed in Table 4.1. You get the values in the cells of the table by subdividing the sample space of 1,000 households according to whether someone planned to purchase and actually purchased a large-screen HDTV. For example, 200 of the respondents planned to purchase a large-screen HDTV and subsequently did purchase the large-screen HDTV.

A second way to present the sample space is by using a **Venn diagram**. This diagram graphically represents the various events as "unions" and "intersections" of circles. Figure 4.1 presents a typical Venn diagram for a two-variable situation, with each variable having only two events (A and A', B and B'). The circle on the left (the red one) represents all events that are part of A.

IGURE 4.1

nn diagram for events and B

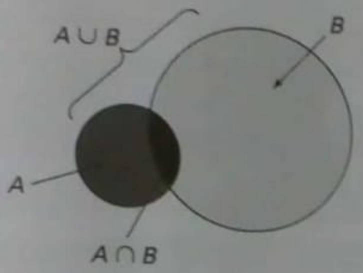
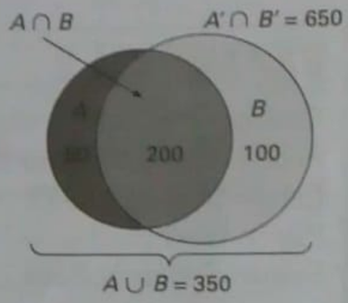


FIGURE 4.2

Venn diagram for the M&R Electronics World example



The circle on the right (the yellow one) represents all events that are part of B. The area contained within circle A and circle B (center area) is the intersection of A and B (written as $A \cap B$), since it is part of A and also part of B. The total area of the two circles is the union of A and B (written as $A \cup B$) and contains all outcomes that are just part of event A, just part of event B, or part of both A and B. The area in the diagram outside of $A \cup B$ contains outcomes that are neither part of A nor part of B.

You must define A and B in order to develop a Venn diagram. You can define either event as A or B, as long as you are consistent in evaluating the various events. For the large-screen HDTV example, you can define the events as follows:

$$A = \text{planned to purchase} \qquad B = \text{actually purchased}$$
$$A' = \text{did not plan to purchase} \qquad B' = \text{did not actually purchase}$$

In drawing the Venn diagram (see Figure 4.2), you must first determine the value of the intersection of A and B so that the sample space can be divided into its parts. $A \cap B$ consists of all 200 households who planned to purchase and actually purchased a large-screen HDTV. The remainder of event A (planned to purchase) consists of the 50 households who planned to purchase a large-screen HDTV but did not actually purchase one. The remainder of event B (actually purchased) consists of the 100 households who did not plan to purchase a large-screen HDTV but actually purchased one. The remaining 650 households represent those who neither planned to purchase nor actually purchased a large-screen HDTV.

Simple Probability

Now you can answer some of the questions posed in the Using Statistics scenario. Because the results are based on data collected in a survey (refer to Table 4.1), you can use the empirical probability approach.

As stated previously, the most fundamental rule for probabilities is that they range in value from 0 to 1. An impossible event has a probability of 0, and an event that is certain to occur has a probability of 1.

Simple probability refers to the probability of occurrence of a simple event, $P(A)$. A simple probability in the Using Statistics scenario is the probability of planning to purchase a large-screen HDTV. How can you determine the probability of selecting a household that planned to purchase a large-screen HDTV? Using Equation (4.1) on page 150:

$$\text{Probability of occurrence} = \frac{X}{T}$$

$$P(\text{Planned to purchase}) = \frac{\text{Number who planned to purchase}}{\text{Total number of households}}$$

$$= \frac{250}{1{,}000} = 0.25$$

Thus, there is a 0.25 (or 25%) chance that a household planned to purchase a large-screen HDTV.

Example 4.3 illustrates another application of simple probability.

EXAMPLE 4.3

Computing the Probability That the Large-Screen HDTV Purchased Had a Faster Refresh Rate

In the Using Statistics follow-up survey, additional questions were asked of the 300 households that actually purchased large-screen HDTVs. Table 4.2 indicates the consumers' responses to whether the television purchased had a faster refresh rate and whether they also purchased a streaming media box in the past 12 months.

Find the probability that if a household that purchased a large-screen HDTV is randomly selected, the television purchased had a faster refresh rate.

TABLE 4.2

Purchase Behavior Regarding Purchasing a Faster Refresh Rate Television and a Streaming Media Box

REFRESH RATE OF TELEVISION PURCHASED	STREAMING MEDIA BOX		
	Yes	No	Total
Faster	38	42	80
Standard	70	150	220
Total	108	192	300

SOLUTION Using the following definitions:

A = purchased a television with a faster refresh rate

A' = purchased a television with a standard refresh rate

B = purchased a streaming media box

B' = did not purchase a streaming media box

$$P(\text{Faster refresh rate}) = \frac{\text{Number of faster refresh rate televisions purchased}}{\text{Total number of televisions}}$$

$$= \frac{80}{300} = 0.267$$

There is a 26.7% chance that a randomly selected large-screen HDTV purchased has a faster refresh rate.

Joint Probability

Whereas simple probability refers to the probability of occurrence of simple events, **joint probability** refers to the probability of an occurrence involving two or more events. An example of joint probability is the probability that you will get heads on the first toss of a coin and heads on the second toss of a coin.

In Table 4.1 on page 152, the group of individuals who planned to purchase and actually purchased a large-screen HDTV consist only of the outcomes in the single cell "yes—planned to purchase *and* yes—actually purchased." Because this group consists of 200 households, the probability of picking a household that planned to purchase *and* actually purchased a large-screen HDTV is

$$P(\text{Planned to purchase } \textit{and} \text{ actually purchased}) = \frac{\text{Planned to purchase } \textit{and} \text{ actually purchased}}{\text{Total number of respondents}}$$

$$= \frac{200}{1{,}000} = 0.20$$

Example 4.4 also demonstrates how to determine joint probability.

XAMPLE 4.4

etermining the int Probability at a Household rchased a Large-reen HDTV with aster Refresh te and Purchased treaming Media x

In Table 4.2 on page 154, the purchases are cross-classified as having a faster refresh rate or having a standard refresh rate and whether the household purchased a streaming media box. Find the probability that a randomly selected household that purchased a large-screen HDTV also purchased a television that had a faster refresh rate and purchased a streaming media box.

SOLUTION Using Equation (4.1) on page 150,

$$P\left(\begin{array}{c}\text{Television with a faster refresh}\\ \text{rate } \textit{and} \text{ streaming media box}\end{array}\right) = \frac{\begin{array}{c}\text{Number that purchased a television with a faster}\\ \text{refresh rate } \textit{and} \text{ purchased a streaming media box}\end{array}}{\text{Total number of large-screen HDTV purchasers}}$$

$$= \frac{38}{300} = 0.127$$

Therefore, there is a 12.7% chance that a randomly selected household that purchased a large-screen HDTV purchased a television that had a faster refresh rate and purchased a streaming media box.

Marginal Probability

The **marginal probability** of an event consists of a set of joint probabilities. You can determine the marginal probability of a particular event by using the concept of joint probability just discussed. For example, if B consists of two events, B_1 and B_2, then $P(A)$, the probability of event A, consists of the joint probability of event A occurring with event B_1 and the joint probability of event A occurring with event B_2. You use Equation (4.2) to compute marginal probabilities.

MARGINAL PROBABILITY

$$P(A) = P(A \textit{ and } B_1) + P(A \textit{ and } B_2) + \cdots + P(A \textit{ and } B_k) \qquad \textbf{(4.2)}$$

where $B_1, B_2, \ldots, B_k$ are k mutually exclusive and collectively exhaustive events, defined as follows:

Two events are **mutually exclusive** if both the events cannot occur simultaneously.
A set of events is **collectively exhaustive** if one of the events must occur.

Heads and tails in a coin toss are mutually exclusive events. The result of a coin toss cannot simultaneously be a head and a tail. Heads and tails in a coin toss are also collectively exhaustive events. One of them must occur. If heads does not occur, tails must occur. If tails does not occur, heads must occur. Being male and being female are mutually exclusive and collectively exhaustive events. No person is both (the two are mutually exclusive), and everyone is one or the other (the two are collectively exhaustive).

You can use Equation (4.2) to compute the marginal probability of "planned to purchase" a large-screen HDTV:

$$\begin{aligned} P(\text{Planned to purchase}) &= P(\text{Planned to purchase } \textit{and} \text{ purchased}) \\ &\quad + P(\text{Planned to purchase } \textit{and} \text{ did not purchase}) \\ &= \frac{200}{1{,}000} + \frac{50}{1{,}000} \\ &= \frac{250}{1{,}000} = 0.25 \end{aligned}$$

You get the same result if you add the number of outcomes that make up the simple event "planned to purchase."

General Addition Rule

Student Tip
The key word when using the addition rule is *or*.

How do you find the probability of event "*A or B*"? You need to consider the occurrence of either event *A* or event *B* or both *A* and *B*. For example, how can you determine the probability that a household planned to purchase *or* actually purchased a large-screen HDTV?

The event "planned to purchase *or* actually purchased" includes all households that planned to purchase and all households that actually purchased a large-screen HDTV. You examine each cell of the contingency table (Table 4.1 on page 152) to determine whether it is part of this event. From Table 4.1, the cell "planned to purchase *and* did not actually purchase" is part of the event because it includes respondents who planned to purchase. The cell "did not plan to purchase and actually purchased" is included because it contains respondents who actually purchased. Finally, the cell "planned to purchase *and* actually purchased" has both characteristics of interest. Therefore, one way to calculate the probability of "planned to purchase *or* actually purchased" is

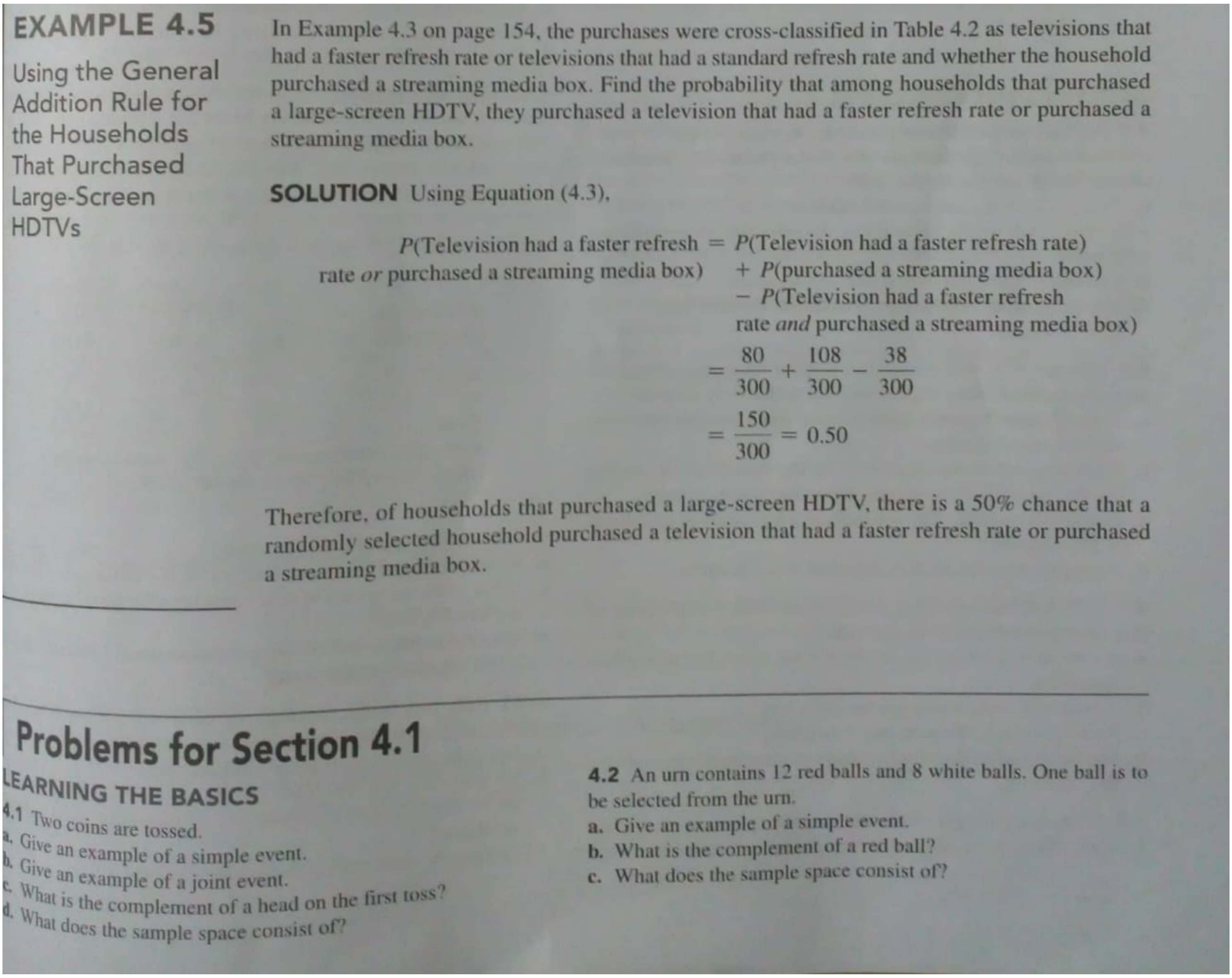
$$\begin{aligned} P(\text{Planned to purchase } \textit{or} \text{ actually purchased}) &= P(\text{Planned to purchase } \textit{and} \text{ did not actually} \\ &\quad \text{purchase}) + P(\text{Did not plan to} \\ &\quad \text{purchase } \textit{and} \text{ actually purchased}) + \\ &\quad P(\text{Planned to purchase } \textit{and} \text{ actually purchased}) \\ &= \frac{50}{1{,}000} + \frac{100}{1{,}000} + \frac{200}{1{,}000} \\ &= \frac{350}{1{,}000} = 0.35 \end{aligned}$$

Often, it is easier to determine $P(A \text{ or } B)$, the probability of the event *A or B*, by using the **general addition rule**, defined in Equation (4.3).

GENERAL ADDITION RULE

The probability of *A or B* is equal to the probability of *A* plus the probability of *B* minus the probability of *A and B*.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{4.3}$$



EXAMPLE 4.5

Using the General Addition Rule for the Households That Purchased Large-Screen HDTVs

In Example 4.3 on page 154, the purchases were cross-classified in Table 4.2 as televisions that had a faster refresh rate or televisions that had a standard refresh rate and whether the household purchased a streaming media box. Find the probability that among households that purchased a large-screen HDTV, they purchased a television that had a faster refresh rate or purchased a streaming media box.

SOLUTION Using Equation (4.3),

$$\begin{aligned} &P(\text{Television had a faster refresh rate } \textit{or} \text{ purchased a streaming media box}) = P(\text{Television had a faster refresh rate}) \\ &\quad + P(\text{purchased a streaming media box}) \\ &\quad - P(\text{Television had a faster refresh rate } \textit{and} \text{ purchased a streaming media box}) \\ &= \frac{80}{300} + \frac{108}{300} - \frac{38}{300} \\ &= \frac{150}{300} = 0.50 \end{aligned}$$

Therefore, of households that purchased a large-screen HDTV, there is a 50% chance that a randomly selected household purchased a television that had a faster refresh rate or purchased a streaming media box.

Problems for Section 4.1

LEARNING THE BASICS

4.1 Two coins are tossed.

a. Give an example of a simple event.

b. Give an example of a joint event.

c. What is the complement of a head on the first toss?

d. What does the sample space consist of?

4.2 An urn contains 12 red balls and 8 white balls. One ball is to be selected from the urn.

a. Give an example of a simple event.

b. What is the complement of a red ball?

c. What does the sample space consist of?

4.3 Consider the following contingency table:

	B	*B'*
A	10	20
A'	20	40

What is the probability of event

a. *A*?
b. *A'*?
c. *A and B*?
d. *A or B*?

4.4 Consider the following contingency table:

	B	*B'*
A	10	30
A'	25	35

What is the probability of event

a. *A'*?
b. *A and B*?
c. *A' and B'*?
d. *A' or B'*?

APPLYING THE CONCEPTS

4.5 For each of the following, indicate whether the type of probability involved is an example of *a priori* probability, empirical probability, or subjective probability.

a. The next toss of a fair coin will land on heads.
b. Italy will win soccer's World Cup the next time the competition is held.
c. The sum of the faces of two dice will be seven.
d. The train taking a commuter to work will be more than 10 minutes late.

4.6 For each of the following, state whether the events created are mutually exclusive and whether they are collectively exhaustive.

a. Undergraduate business students were asked whether they were sophomores or juniors.
b. Each respondent was classified by the type of car he or she drives: sedan, SUV, American, European, Asian, or none.
c. People were asked, "Do you currently live in (i) an apartment or (ii) a house?"
d. A product was classified as defective or not defective.

4.7 Which of the following events occur with a probability of zero? For each, state why or why not.

a. A company is listed on the New York Stock Exchange and NASDAQ.
b. A consumer owns a smartphone and a tablet.
c. A cellphone is a Motorola and a Samsung.
d. An automobile is a Toyota and was manufactured in the United States.

4.8 Do males or females feel more tense or stressed out at work? A survey of employed adults conducted online by Harris Interactive on behalf of the American Psychological Association revealed the following:

	FELT TENSE OR STRESSED OUT AT WORK	
GENDER	**Yes**	**No**
Male	244	495
Female	282	480

Source: Data extracted from "The 2013 Work and Well-Being Survey," American Psychological Association and Harris Interactive, March 2013, p. 5, **bit.ly/11JGcPf**.

a. Give an example of a simple event.
b. Give an example of a joint event.
c. What is the complement of "Felt tense or stressed out at work
d. Why is "Male and felt tense or stressed out at work" a joint eve

4.9 Referring to the contingency table in Problem 4.8, if an employed adult is selected at random, what is the probability that

a. the employed adult felt tense or stressed out at work?
b. the employed adult was a male who felt tense or stressed out work?
c. the employed adult was a male *or* felt tense or stressed out work?
d. Explain the difference in the results in (b) and (c).

4.10 How will marketers change their social media use in the ne future? A survey by Social Media Examiner reported that 78% B2B marketers (marketers that focus primarily on attracting businesses) plan to increase their use of LinkedIn, as compared to 54 of B2C marketers (marketers that primarily target consumer The survey was based on 1,331 B2B marketers and 1,694 B2 marketers. The following table summarizes the results:

INCREASE USE OF LINKEDIN?	BUSINESS FOCUS		
	B2B	**B2C**	**Total**
Yes	1,038	915	1,953
No	293	779	1,072
Total	1,331	1,694	3,025

Source: Data extracted from "2013 Social Media Marketing Industry Report," May 2013, **bit.ly/1g5vMQN**.

a. Give an example of a simple event.
b. Give an example of a joint event.
c. What is the complement of a marketer who plans to increa use of LinkedIn?
d. Why is a marketer who plans to increase use of LinkedIn and a B2C marketer a joint event?

4.11 Referring to the contingency table in Problem 4.10, if marketer is selected at random, what is the probability that

a. he or she plans to increase use of LinkedIn?
b. he or she is a B2C marketer?
c. he or she plans to increase use of LinkedIn *or* is a B2 marketer?
d. Explain the difference in the results in (b) and (c).

SELF Test **4.12** What business and technical skills are critical for today's business intelligence/analytics and information management professionals? As part of InformationWeek's 2013 U.S. IT Salary Survey, business intelligence/analytics and information management professionals, both staff and managers, were asked to indicate what business and technical skills are critical to their job. The list of business and technical skills included *Analyzing Data*. The following table summarizes the responses to this skill:

	PROFESSIONAL POSITION		
ANALYZING DATA	**Staff**	**Management**	**Total**
Critical	4,374	3,633	8,007
Not critical	3,436	2,631	6,067
Total	7,810	6,264	14,074

Source: Data extracted from "IT Salaries Show Slow Growth," *InformationWeek Reports*, April 2013, p. 40, **ubm.io/1ewjKT5**.

If a professional is selected at random, what is the probability that he or she

a. indicates analyzing data as critical to his or her job?

b. is a manager?

c. indicates analyzing data as critical to his or her job *or* is a manager?

d. Explain the difference in the results in (b) and (c).

4.13 Do Americans prefer Coke or Pepsi? A survey was conducted by Public Policy Polling (PPP) in 2013; the results were as follows:

	GENDER		
PREFERENCE	**Female**	**Male**	**Total**
Coke	120	95	215
Pepsi	95	80	175
Neither/Unsure	65	45	110
Total	280	220	500

Source: Data extracted from "Public Policy Polling," Report 2013, **bit.ly/YKXfzN**.

If an American is selected at random, what is the probability that he or she

a. prefers Pepsi?

b. is male *and* prefers Pepsi?

c. is male *or* prefers Pepsi?

d. Explain the difference in the results in (b) and (c).

4.14 A survey of 1,085 adults asked, "Do you enjoy shopping for clothing for yourself?" The results (data extracted from "Split Decision on Clothes Shopping," *USA Today*, January 28, 2011, p. 1B) indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. The sample sizes of males and females were not provided. Suppose that the results indicated that of 542 males, 238 answered yes. Of 543 females, 276 answered yes. Construct a contingency table to evaluate the probabilities. What is the probability that a respondent chosen at random

a. enjoys shopping for clothing for himself or herself?

b. is a female *and* enjoys shopping for clothing for herself?

c. is a female *or* is a person who enjoys shopping for clothing?

d. is a male *or* a female?

4.15 Each year, ratings are compiled concerning the performance of new cars during the first 90 days of use. Suppose that the cars have been categorized according to whether a car needs warranty-related repair (yes or no) and the country in which the company manufacturing a car is based (United States or not United States). Based on the data collected, the probability that the new car needs a warranty repair is 0.04, the probability that the car was manufactured by a U.S.-based company is 0.60, and the probability that the new car needs a warranty repair *and* was manufactured by a U.S.-based company is 0.025. Construct a contingency table to evaluate the probabilities of a warranty-related repair. What is the probability that a new car selected at random

a. needs a warranty repair?

b. needs a warranty repair *and* was manufactured by a U.S.-based company?

c. needs a warranty repair *or* was manufactured by a U.S.-based company?

d. needs a warranty repair *or* was not manufactured by a U.S.-based company?

4.2 Conditional Probability

Each example in Section 4.1 involves finding the probability of an event when sampling from the entire sample space. How do you determine the probability of an event if you know certain information about the events involved?

Computing Conditional Probabilities

Conditional probability refers to the probability of event *A*, given information about the occurrence of another event, *B*.

CONDITIONAL PROBABILITY

The probability of A given B is equal to the probability of A *and* B divided by the probability of B.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad (4.4a)$$

The probability of B given A is equal to the probability of A *and* B divided by the probability of A.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \quad (4.4b)$$

where

$P(A \text{ and } B)$ = joint probability of A *and* B
$P(A)$ = marginal probability of A
$P(B)$ = marginal probability of B

Student Tip
The variable that is *given* goes in the denominator of Equation (4.4). Since you were given planned to purchase, planned to purchase is in the denominator.

Referring to the Using Statistics scenario involving the purchase of large-screen HDTVs, suppose you were told that a household planned to purchase a large-screen HDTV. Now, what is the probability that the household actually purchased the television?

In this example, the objective is to find P(Actually purchased | Planned to purchase). Here you are given the information that the household planned to purchase the large-screen HDTV. Therefore, the sample space does not consist of all 1,000 households in the survey. It consists of only those households that planned to purchase the large-screen HDTV. Of 250 such households, 200 actually purchased the large-screen HDTV. Therefore, based on Table 4.1 on page 152, the probability that a household actually purchased the large-screen HDTV given that they planned to purchase is

$$P(\text{Actually purchased} \mid \text{Planned to purchase}) = \frac{\text{Planned to purchase } \textit{and} \text{ actually purchased}}{\text{Planned to purchase}}$$
$$= \frac{200}{250} = 0.80$$

You can also use Equation (4.4b) to compute this result:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

where

A = planned to purchase

B = actually purchased

then

$$P(\text{Actually purchased} \mid \text{Planned to purchase}) = \frac{200/1{,}000}{250/1{,}000}$$
$$= \frac{200}{250} = 0.80$$

Example 4.6 further illustrates conditional probability.

EXAMPLE 4.6

Finding the Conditional Probability of Purchasing a Streaming Media Box

Table 4.2 on page 154 is a contingency table for whether a household purchased a television with a faster refresh rate and whether the household purchased a streaming media box. If a household purchased a television with a faster refresh rate, what is the probability that it also purchased a streaming media box?

SOLUTION Because you know that the household purchased a television with a faster refresh rate, the sample space is reduced to 80 households. Of these 80 households, 38 also purchased a streaming media box. Therefore, the probability that a household purchased a streaming media box, given that the household purchased a television with a faster refresh rate, is

$$P(\text{Purchased streaming media box} \mid \text{Purchased television with faster refresh rate}) = \frac{\text{Number purchasing television with faster refresh rate } \textit{and} \text{ streaming media box}}{\text{Number purchasing television with faster refresh rate}}$$

$$= \frac{38}{80} = 0.475$$

If you use Equation (4.4b) on page 160:

A = purchased a television with a faster refresh rate

B = purchased a streaming media box

then

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{38/300}{80/300} = 0.475$$

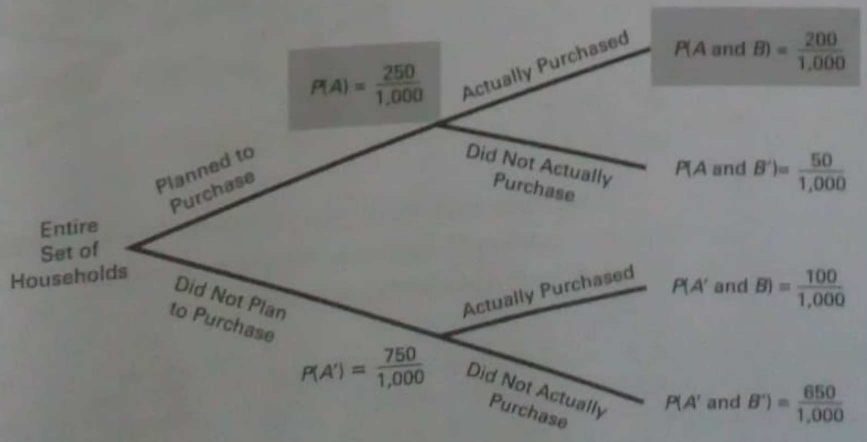
Therefore, given that the household purchased a television with a faster refresh rate, there is a 47.5% chance that the household also purchased a streaming media box. You can compare this conditional probability to the marginal probability of purchasing a streaming media box, which is $108/300 = 0.36$, or 36%. These results tell you that households that purchased televisions with a faster refresh rate are more likely to purchase a streaming media box than are households that purchased large-screen HDTVs that have a standard refresh rate.

Decision Trees

In Table 4.1 on page 152, households are classified according to whether they planned to purchase and whether they actually purchased large-screen HDTVs. A **decision tree** is an alternative to the contingency table. Figure 4.3 represents the decision tree for this example.

FIGURE 4.3

Decision tree for planned to purchase and actually purchased



Entire Set of Households
Planned to Purchase
Did Not Plan to Purchase
$P(A) = \frac{250}{1,000}$
$P(A') = \frac{750}{1,000}$
Actually Purchased
Did Not Actually Purchase
Actually Purchased
Did Not Actually Purchase
$P(A \text{ and } B) = \frac{200}{1,000}$
$P(A \text{ and } B') = \frac{50}{1,000}$
$P(A' \text{ and } B) = \frac{100}{1,000}$
$P(A' \text{ and } B') = \frac{650}{1,000}$

In Figure 4.3, beginning at the left with the entire set of households, there are two "branches" for whether or not the household planned to purchase a large-screen HDTV. Each of these branches has two subbranches, corresponding to whether the household actually purchased or did not actually purchase the large-screen HDTV. The probabilities at the end of the initial branches represent the marginal probabilities of A and A'. The probabilities at the end of each of the four subbranches represent the joint probability for each combination of events A and B. You compute the conditional probability by dividing the joint probability by the appropriate marginal probability.

For example, to compute the probability that the household actually purchased, given that the household planned to purchase the large-screen HDTV, you take P(Planned to purchase *and* actually purchased) and divide by P(Planned to purchase). From Figure 4.3,

$$P(\text{Actually purchased} \mid \text{Planned to purchase}) = \frac{200/1{,}000}{250/1{,}000} = \frac{200}{250} = 0.80$$

Example 4.7 illustrates how to construct a decision tree.

EXAMPLE 4.7

Constructing the Decision Tree for the Households That Purchased Large-Screen HDTVs

Using the cross-classified data in Table 4.2 on page 154, construct the decision tree. Use the decision tree to find the probability that a household purchased a streaming media box, given that the household purchased a television with a faster refresh rate.

SOLUTION The decision tree for purchased a streaming media box and a television with a faster refresh rate is displayed in Figure 4.4.

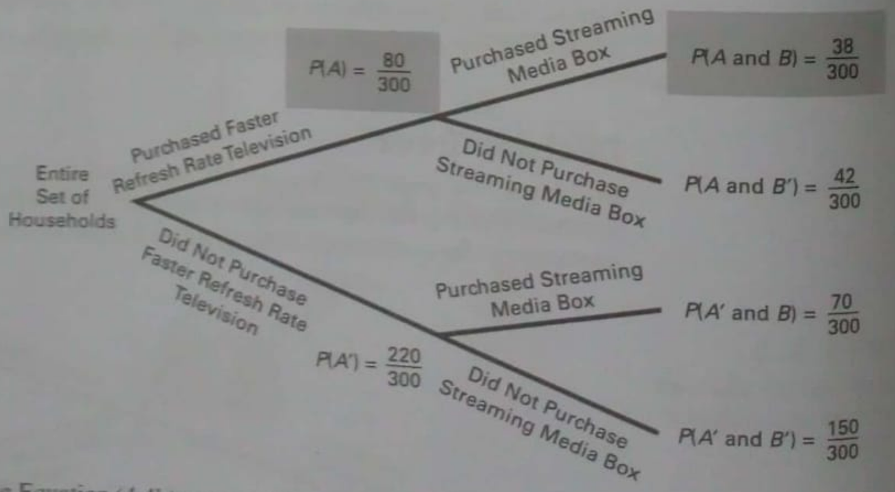

Entire Set of Households
Purchased Faster Refresh Rate Television
$P(A) = \frac{80}{300}$
Did Not Purchase Faster Refresh Rate Television
$P(A') = \frac{220}{300}$
Purchased Streaming Media Box
$P(A \text{ and } B) = \frac{38}{300}$
Did Not Purchase Streaming Media Box
$P(A \text{ and } B') = \frac{42}{300}$
Purchased Streaming Media Box
$P(A' \text{ and } B) = \frac{70}{300}$
Did Not Purchase Streaming Media Box
$P(A' \text{ and } B') = \frac{150}{300}$

FIGURE 4.4
Decision tree for purchased a television with a faster refresh rate and a streaming media box

Using Equation (4.4b) on page 160 and the following definitions,

A = purchased a television with a faster refresh rate
B = purchased a streaming media box

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{38/300}{80/300} = 0.475$$

Independence

In the example concerning the purchase of large-screen HDTVs, the conditional probability is $200/250 = 0.80$ that the selected household actually purchased the large-screen HDTV, given that the household planned to purchase. The simple probability of selecting a household that actually purchased is $300/1{,}000 = 0.30$. This result shows that the prior knowledge that the household planned to purchase affected the probability that the household actually purchased the television. In other words, the outcome of one event is *dependent* on the outcome of a second event.

When the outcome of one event does *not* affect the probability of occurrence of another event, the events are said to be independent. **Independence** can be determined by using Equation (4.5).

INDEPENDENCE

Two events, A and B, are independent if and only if

$$P(A|B) = P(A) \tag{4.5}$$

where

$$P(A|B) = \text{conditional probability of } A \text{ given } B$$
$$P(A) = \text{marginal probability of } A$$

Example 4.8 demonstrates the use of Equation (4.5).

EXAMPLE 4.8 Determining Independence

In the follow-up survey of the 300 households that actually purchased large-screen HDTVs, the households were asked if they were satisfied with their purchases. Table 4.3 cross-classifies the responses to the satisfaction question with the responses to whether the television had a faster refresh rate.

TABLE 4.3 Satisfaction with Purchase of Large-Screen HDTVs

TELEVISION REFRESH RATE	SATISFIED WITH PURCHASE?		
	Yes	No	Total
Faster	64	16	80
Standard	176	44	220
Total	240	60	300

Determine whether being satisfied with the purchase and the refresh rate of the television purchased are independent.

SOLUTION For these data,

$$P(\text{Satisfied}|\text{Faster refresh rate}) = \frac{64/300}{80/300} = \frac{64}{80} = 0.80$$

which is equal to

$$P(\text{Satisfied}) = \frac{240}{300} = 0.80$$

Thus, being satisfied with the purchase and the refresh rate of the television purchased are independent. Knowledge of one event does not affect the probability of the other event.

Example 4.9 demonstrates the use of the general multiplication rule.

EXAMPLE 4.9

Using the General Multiplication Rule

Consider the 80 households that purchased televisions that had a faster refresh rate. In Table 4.3 on page 163, you see that 64 households are satisfied with their purchase, and 16 households are dissatisfied. Suppose 2 households are randomly selected from the 80 households. Find the probability that both households are satisfied with their purchase.

SOLUTION Here you can use the multiplication rule in the following way. If

$$\begin{aligned} A &= \text{second household selected is satisfied} \\ B &= \text{first household selected is satisfied} \end{aligned}$$

then, using Equation (4.6),

$$P(A \text{ and } B) = P(A|B)P(B)$$

The probability that the first household is satisfied with the purchase is 64/80. However, the probability that the second household is also satisfied with the purchase depends on the result of the first selection. If the first household is not returned to the sample after the satisfaction level is determined (i.e., sampling without replacement), the number of households remaining is 79. If the first household is satisfied, the probability that the second is also satisfied is 63/79 because 63 satisfied households remain in the sample. Therefore,

$$P(A \text{ and } B) = \left(\frac{63}{79}\right)\left(\frac{64}{80}\right) = 0.6380$$

There is a 63.80% chance that both of the households sampled will be satisfied with their purchase.

The **multiplication rule for independent events** is derived by substituting $P(A)$ for $P(A|B)$ in Equation (4.6).

If this rule holds for two events, A and B, then A and B are independent. Therefore, there are two ways to determine independence:

1. Events A and B are independent if, and only if, $P(A|B) = P(A)$.
2. Events A and B are independent if, and only if, $P(A \text{ and } B) = P(A)P(B)$.

Marginal Probability Using the General Multiplication Rule

In Section 4.1, marginal probability was defined using Equation (4.2) on page 155. You can state the equation for marginal probability by using the general multiplication rule. If

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k)$$

then, using the general multiplication rule, Equation (4.8) defines the marginal probability.

MARGINAL PROBABILITY USING THE GENERAL MULTIPLICATION RULE

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \qquad \textbf{(4.8)}$$

where $B_1, B_2, \ldots, B_k$ are k mutually exclusive and collectively exhaustive events.

To illustrate Equation (4.8), refer to Table 4.1 on page 152. Let

$$\begin{aligned} P(A) &= \text{probability of planned to purchase} \\ P(B_1) &= \text{probability of actually purchased} \\ P(B_2) &= \text{probability of did not actually purchase} \end{aligned}$$

Then, using Equation (4.8), the probability of planned to purchase is

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) \\ &= \left(\frac{200}{300}\right)\left(\frac{300}{1{,}000}\right) + \left(\frac{50}{700}\right)\left(\frac{700}{1{,}000}\right) \\ &= \frac{200}{1{,}000} + \frac{50}{1{,}000} = \frac{250}{1{,}000} = 0.25 \end{aligned}$$

Problems for Section 4.2

LEARNING THE BASICS

4.16 Consider the following contingency table:

	B	B'
A	10	20
A'	20	40

What is the probability of

a. $A|B$?
b. $A|B'$?
c. $A'|B'$?
d. Are events A and B independent?

4.17 Consider the following contingency table:

	B	B'
A	10	30
A'	25	35

What is the probability of

a. $A|B$?
b. $A'|B'$?
c. $A|B'$?
d. Are events A and B independent?

4.18 If $P(A \text{ and } B) = 0.4$ and $P(B) = 0.8$, find $P(A|B)$.

4.19 If $P(A) = 0.7$, $P(B) = 0.6$, and A and B are independent, find $P(A \textit{ and } B)$.

4.20 If $P(A) = 0.3$, $P(B) = 0.4$, and $P(A \textit{ and } B) = 0.2$, are A and B independent?

APPLYING THE CONCEPTS

4.21 Do males or females feel more tense or stressed out at work? A survey of employed adults conducted online by Harris Interactive on behalf of the American Psychological Association revealed the following:

	FELT TENSE OR STRESSED OUT AT WORK	
GENDER	**Yes**	**No**
Male	244	495
Female	282	480

Source: Data extracted from "The 2013 Work and Well-Being Survey," American Psychological Association and Harris Interactive, March 2013, p. 5, **bit.ly/11JGcPf.**

a. Given that the employed adult felt tense or stressed out at work, what is the probability that the employed adult was a male?
b. Given that the employed adult is male, what is the probability that he felt tense or stressed out at work?
c. Explain the difference in the results in (a) and (b).
d. Is feeling tense or stressed out at work and gender independent?

4.22 How will marketers change their social media use in the near future? A survey by Social Media Examiner of B2B marketers (marketers that focus primarily on attracting businesses) and B2C marketers (marketers that primarily target consumers) was based on 1,331 B2B marketers and 1,694 B2C marketers. The following table summarizes the results:

	BUSINESS FOCUS		
INCREASE USE OF LINKEDIN?	**B2B**	**B2C**	**Total**
Yes	1,038	915	1,953
No	293	779	1,072
Total	1,331	1,694	3,025

Source: Data extracted from "2013 Social Media Marketing Industry Report," May 2013, **bit.ly/1g5vMQN.**

a. Suppose you know that the marketer is a B2B marketer. What is the probability that he or she plans to increase use of LinkedIn?
b. Suppose you know that the marketer is a B2C marketer. What is the probability that he or she plans to increase use of LinkedIn?
c. Are the two events, increase use of LinkedIn and business focus, independent? Explain.

4.23 Do Americans prefer Coke or Pepsi? A survey was conducted by Public Policy Polling (PPP) in 2013; the results were as follows:

	GENDER		
PREFERENCE	**Female**	**Male**	**Total**
Coke	120	95	215
Pepsi	95	80	175
Neither/Unsure	65	45	110
Total	280	220	500

Source: Data extracted from "Public Policy Polling" Report 2013, **bit.ly/YKXfzN.**

a. Given that an American is a male, what is the probability that he prefers Pepsi?
b. Given that an American is a female, what is the probability that she prefers Pepsi?
c. Is preference independent of gender? Explain.

SELF Test **4.24** What business and technical skills are critical for today's business intelligence/analytics and information management professionals? As part of InformationWeek's 2013 U.S. IT Salary Survey, business intelligence/analytics and information management professionals, both staff and managers, were asked to indicate what business and technical skills are critical to their job. The list of business and technical skills included *Analyzing Data*. The following table summarizes the responses to this skill:

	PROFESSIONAL POSITION		
ANALYZING DATA	**Staff**	**Management**	**Total**
Critical	4,374	3,633	8,007
Not critical	3,436	2,631	6,067
Total	7,810	6,264	14,074

Source: Data extracted from "IT Salaries Show Slow Growth," *InformationWeek Reports*, April 2013, p. 40, **ubm.io/1ewjKT5.**

a. Given that a professional is staff, what is the probability that the professional indicates analyzing data as critical to his or her job?
b. Given that a professional is staff, what is the probability that the professional does not indicate analyzing data as critical to his or her job?
c. Given that a professional is a manager, what is the probability that the professional indicates analyzing data as critical to his or her job?
d. Given that a professional is a manager, what is the probability that the professional does not indicate analyzing data as critical to his or her job?

4.25 A survey of 1,085 adults asked, "Do you enjoy shopping for clothing for yourself?" The results (data extracted from "Split Decision on Clothes Shopping," *USA Today*, January 28, 2011, p. 1B) indicated that 51% of the females enjoyed shopping for clothing for themselves as compared to 44% of the males. The sample sizes of males and females were not provided. Suppose that the results were as shown in the following table:

ENJOYS SHOPPING FOR CLOTHING	GENDER		
	Male	Female	Total
Yes	238	276	514
No	304	267	571
Total	542	543	1,085

a. Suppose that the respondent chosen is a female. What is the probability that she does not enjoy shopping for clothing?

b. Suppose that the respondent chosen enjoys shopping for clothing. What is the probability that the individual is a male?

c. Are enjoying shopping for clothing and the gender of the individual independent? Explain.

4.26 Each year, ratings are compiled concerning the performance of new cars during the first 90 days of use. Suppose that the cars have been categorized according to whether a car needs warranty-related repair (yes or no) and the country in which the company manufacturing a car is based (United States or not United States). Based on the data collected, the probability that the new car needs a warranty repair is 0.04, the probability that the car is manufactured by a U.S.-based company is 0.60, and the probability that the new car needs a warranty repair *and* was manufactured by a U.S.-based company is 0.025.

a. Suppose you know that a company based in the United States manufactured a particular car. What is the probability that the car needs a warranty repair?

b. Suppose you know that a company based in the United States did not manufacture a particular car. What is the probability that the car needs a warranty repair?

c. Are need for a warranty repair and location of the company manufacturing the car independent?

4.27 In 41 of the 63 years from 1950 through 2013 (in 2011 there was virtually no change), the S&P 500 finished higher after the first five days of trading. In 36 out of 41 years, the S&P 500 finished higher for the year. Is a good first week a good omen for the upcoming year? The following table gives the first-week and annual performance over this 63-year period:

FIRST WEEK	S&P 500'S ANNUAL PERFORMANCE	
	Higher	Lower
Higher	36	5
Lower	11	11

a. If a year is selected at random, what is the probability that the S&P 500 finished higher for the year?

b. Given that the S&P 500 finished higher after the first five days of trading, what is the probability that it finished higher for the year?

c. Are the two events "first-week performance" and "annual performance" independent? Explain.

d. Look up the performance after the first five days of 2014 and the 2014 annual performance of the S&P 500 at **finance.yahoo.com**. Comment on the results.

4.28 A standard deck of cards is being used to play a game. There are four suits (hearts, diamonds, clubs, and spades), each having 13 faces (ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, jack, queen, and king), making a total of 52 cards. This complete deck is thoroughly mixed, and you will receive the first 2 cards from the deck, without replacement (the first card is not returned to the deck after it is selected).

a. What is the probability that both cards are queens?

b. What is the probability that the first card is a 10 and the second card is a 5 or 6?

c. If you were sampling with replacement (the first card is returned to the deck after it is selected), what would be the answer in (a)?

d. In the game of blackjack, the face cards (jack, queen, king) count as 10 points, and the ace counts as either 1 or 11 points. All other cards are counted at their face value. Blackjack is achieved if 2 cards total 21 points. What is the probability of getting blackjack in this problem?

4.29 A box of nine iPhone 5C cellphones (the iPhone "for the colorful") contains two yellow cellphones and seven green cellphones.

a. If two cellphones are randomly selected from the box, without replacement (the first cellphone is not returned to the box after it is selected), what is the probability that both cellphones selected will be green?

b. If two cellphones are randomly selected from the box, without replacement (the first cellphone is not returned to the box after it is selected), what is the probability that there will be one yellow cellphone and one green cellphone selected?

c. If three cellphones are selected, with replacement (the cellphones are returned to the box after they are selected), what is the probability that all three will be yellow?

d. If you were sampling with replacement (the first cellphone is returned to the box after it is selected), what would be the answers to (a) and (b)?

4.3 Bayes' Theorem

Bayes' theorem is used to revise previously calculated probabilities based on new information. Developed by Thomas Bayes in the eighteenth century (see references 1, 2, 3, and 8), Bayes' theorem is an extension of what you previously learned about conditional probability.

You can apply Bayes' theorem to the situation in which M&R Electronics World is considering marketing a new model of televisions. In the past, 40% of the new-model televisions have been successful, and 60% have been unsuccessful. Before introducing the new-model television, the marketing research department conducts an extensive study and releases a report, either favorable or unfavorable. In the past, 80% of the successful new-model television(s) had received favorable market research reports, and 30% of the unsuccessful new-model television(s) had received favorable reports. For the new model of television under consideration, the marketing research department has issued a favorable report. What is the probability that the television will be successful?

Bayes' theorem is developed from the definition of conditional probability. To find the conditional probability of B given A, consider Equation (4.4b) (originally presented on page 160 and shown again below):

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

Bayes' theorem is derived by substituting Equation (4.8) on page 165 for $P(A)$ in the denominator of Equation (4.4b).

BAYES' THEOREM

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)} \qquad (4.9)$$

where B_i is the ith event out of k mutually exclusive and collectively exhaustive events.

To use Equation (4.9) for the television-marketing example, let

event S = successful television	event F = favorable report
event S' = unsuccessful television	event F' = unfavorable report

and

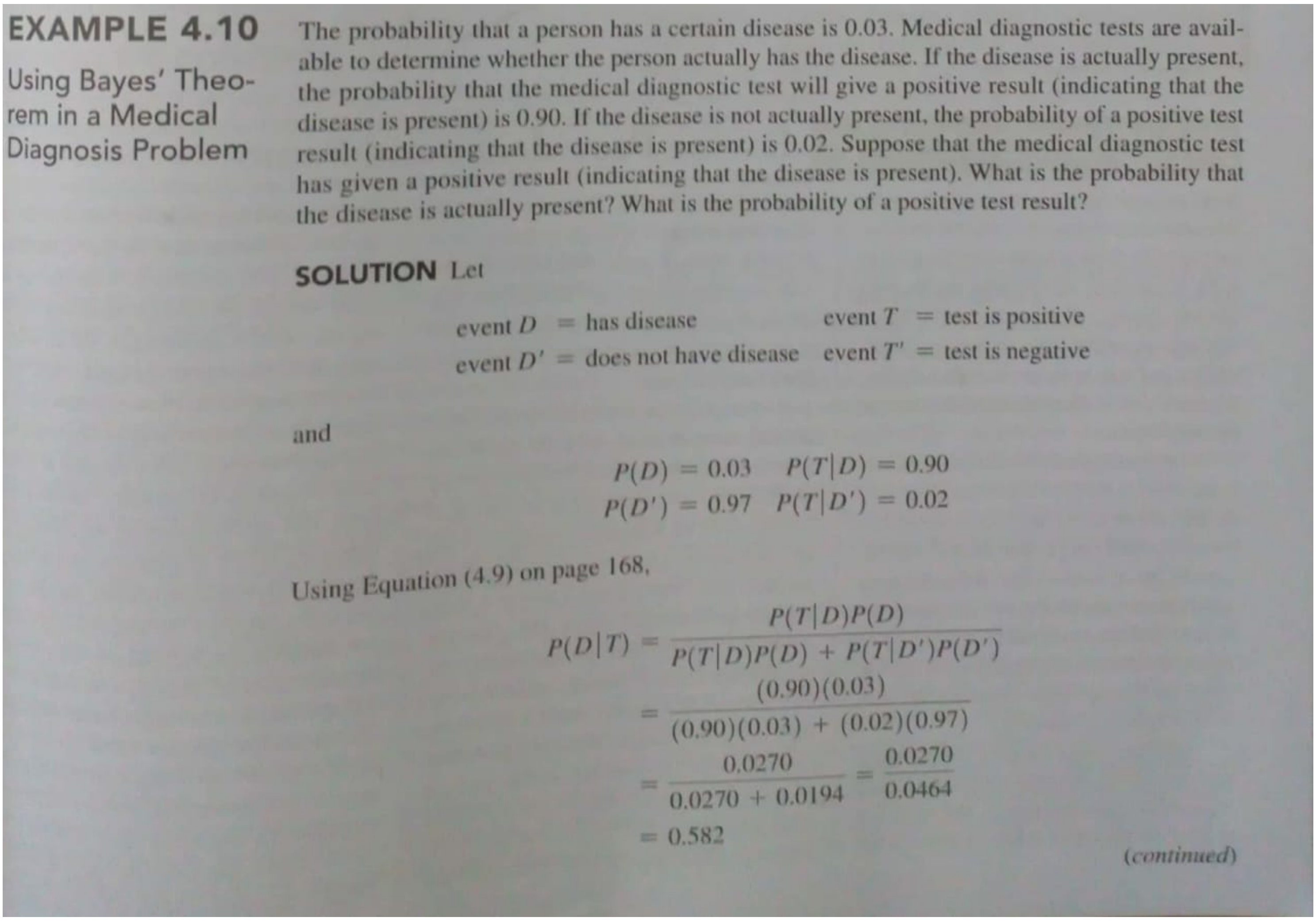
$$P(S) = 0.40 \qquad P(F|S) = 0.80$$
$$P(S') = 0.60 \qquad P(F|S') = 0.30$$

Then, using Equation (4.9),

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|S')P(S')} \\ &= \frac{(0.80)(0.40)}{(0.80)(0.40) + (0.30)(0.60)} \\ &= \frac{0.32}{0.32 + 0.18} = \frac{0.32}{0.50} \\ &= 0.64 \end{aligned}$$

The probability of a successful television, given that a favorable report was received, is 0.64. Thus, the probability of an unsuccessful television, given that a favorable report was received, is 1 − 0.64 = 0.36.

Table 4.4 summarizes the computation of the probabilities, and Figure 4.5 presents the decision tree.



EXAMPLE 4.10

Using Bayes' Theorem in a Medical Diagnosis Problem

The probability that a person has a certain disease is 0.03. Medical diagnostic tests are available to determine whether the person actually has the disease. If the disease is actually present, the probability that the medical diagnostic test will give a positive result (indicating that the disease is present) is 0.90. If the disease is not actually present, the probability of a positive test result (indicating that the disease is present) is 0.02. Suppose that the medical diagnostic test has given a positive result (indicating that the disease is present). What is the probability that the disease is actually present? What is the probability of a positive test result?

SOLUTION Let

event D	= has disease	event T	= test is positive
event D'	= does not have disease	event T'	= test is negative

and

$$P(D) = 0.03 \quad P(T|D) = 0.90$$
$$P(D') = 0.97 \quad P(T|D') = 0.02$$

Using Equation (4.9) on page 168,

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D')P(D')} \\ &= \frac{(0.90)(0.03)}{(0.90)(0.03) + (0.02)(0.97)} \\ &= \frac{0.0270}{0.0270 + 0.0194} = \frac{0.0270}{0.0464} \\ &= 0.582 \end{aligned}$$

(*continued*)

The probability that the disease is actually present, given that a positive result has occurred (indicating that the disease is present), is 0.582. Table 4.5 summarizes the computation of the probabilities, and Figure 4.6 presents the decision tree. The denominator in Bayes' theorem represents $P(T)$, the probability of a positive test result, which in this case is 0.0464, or 4.64%.

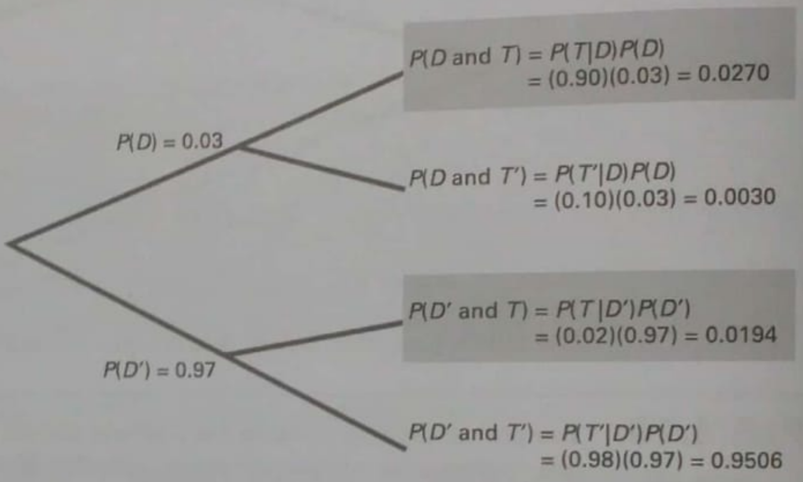
TABLE 4.5

Bayes' Theorem Computations for the Medical Diagnosis Problem

Event D_i	Prior Probability $P(D_i)$	Conditional Probability $P(T\|D_i)$	Joint Probability $P(T\|D_i)P(D_i)$	Revised Probability $P(D_i\|T)$
D = has disease	0.03	0.90	0.0270	$P(D\|T) = 0.0270/0.0464 = 0.582$
D' = does not have disease	0.97	0.02	$\frac{0.0194}{0.0464}$	$P(D'\|T) = 0.0194/0.0464 = 0.418$

FIGURE 4.6

Decision tree for a medical diagnosis problem



THINK ABOUT THIS Divine Providence and Spam

Would you ever guess that the essays *Divine Benevolence: Or, An Attempt to Prove That the Principal End of the Divine Providence and Government Is the Happiness of His Creatures* and *An Essay Towards Solving a Problem in the Doctrine of Chances* were written by the same person? Probably not, and in doing so, you illustrate a modern-day application of Bayesian statistics: spam, or junk mail filters.

In not guessing correctly, you probably looked at the words in the titles of the essays and concluded that they were talking about two different things. An implicit rule you used was that word frequencies vary by subject matter. A statistics essay would very likely contain the word *statistics* as well as words such as *chance*, *problem*, and *solving*. An eighteenth-century essay about theology and religion would be more likely to contain the uppercase forms of *Divine* and *Providence*.

Likewise, there are words you would guess to be very unlikely to appear in either book, such as technical terms from finance, and words that are most likely to appear in both—common words such as *a*, *and*, and *the*. That words would be either likely or unlikely suggests an application of probability theory. Of course, likely and unlikely are fuzzy concepts, and we might occasionally misclassify an essay if we kept things too simple, such as relying solely on the occurrence of the words *Divine* and *Providence*.

For example, a profile of the late Harris Milstead, better known as *Divine*, the star of *Hairspray* and other films, visiting Providence (Rhode Island), would most certainly not be an essay about theology. But if we widened the number of words we examined and found such words as *movie* or the name John Waters (Divine's director in many films), we probably would quickly realize the essay had something to do with twentieth-century cinema and little to do with theology and religion.

We can use a similar process to try to classify a new email message in your in-box as either spam or a legitimate message (called "ham," in this context). We would first need to add to your email program a "spam filter" that has the ability to track word frequencies associated with spam and ham messages as you identify them on a day-to-day basis. This would allow the filter to constantly update the prior probabilities necessary to use Bayes' theorem. With these probabilities, the filter can ask, "What is the probability that an email is spam, given the presence of a certain word?"

Applying the terms of Equation (4.9) on page 168, such a Bayesian spam filter would multiply the probability of finding the word in a spam email, $P(A|B)$, by the probability that the email is spam, $P(B)$, and then divide by the probability of finding the word in an email, the denominator in Equation (4.9). Bayesian spam filters also use shortcuts by focusing on a small set of words that have a high probability of being found in a spam message as well as on a small set of other words that have a low probability of being found in a spam message.

As spammers (people who send junk email) learned of such new filters, they tried to outfox them. Having learned that Bayesian filters might be assigning a high $P(A|B)$ value to words commonly found in spam, such as Viagra, spammers thought they could

USING STATISTICS

Possibilities at M&R Electronics World, Revisited

Shock/Fo

As the marketing manager for M&R Electronics World, you analyzed the survey results of an intent-to-purchase study. This study asked the heads of 1,000 households about their intentions to purchase a large-screen HDTV sometime during the next 12 months, and as a follow-up, M&R surveyed the same people 12 months later to see whether such a television was purchased. In addition, for households purchasing large-screen HDTVs, the survey asked whether the television they purchased had a faster refresh rate, whether they also purchased a streaming media box in the past 12 months, and whether they were satisfied with their purchase of the large-screen HDTV.

By analyzing the results of these surveys, you were able to uncover many pieces of valuable information that will help you plan a marketing strategy to enhance sales and better target those households likely to purchase multiple or more expensive products. Whereas only 30% of the households actually purchased a large-screen HDTV, if a household indicated that it planned to purchase a large-screen HDTV in the next 12 months, there was an 80% chance that the household actually made the purchase. Thus the marketing strategy should target those households that have indicated an intention to purchase.

You determined that for households that purchased a television that had a faster refresh rate, there was a 47.5% chance that the household also purchased a streaming media box. You then compared this conditional probability to the marginal probability of purchasing a streaming media box, which was 36%. Thus, households that purchased televisions that had a faster refresh rate are more likely to purchase a streaming media box than are households that purchased large-screen HDTVs that have a standard refresh rate.

You were also able to apply Bayes' theorem to M&R Electronics World's market research reports. The reports investigate a potential new television model prior to its scheduled release. If a favorable report was received, then there was a 64% chance that the new television model would be successful. However, if an unfavorable report was received, there is only a 16% chance that the model would be successful. Therefore, the marketing strategy of M&R needs to pay close attention to whether a report's conclusion is favorable or unfavorable.

SUMMARY

This chapter began by developing the basic concepts of probability. You learned that probability is a numeric value from 0 to 1 that represents the chance, likelihood, or possibility that a particular event will occur. In addition to simple probability, you learned about conditional probabilities and independent events. Bayes' theorem was used to revise previously calculated probabilities based on new information. Throughout the chapter, contingency tables and decision trees were used to display information. You also learned about several counting rules. In the next chapter, important discrete probability distributions including the binomial and Poisson distributions are developed.

REFERENCES

1. Anderson-Cook, C. M. "Unraveling Bayes' Theorem." *Quality Progress*, March 2014, p. 52–54.
2. Bellhouse, D. R. "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth." *Statistical Science*, 19 (2004), 3–43.
3. Hooper, W. "Probing Probabilities." *Quality Progress*, March 2014, pp. 18–22.
4. Lowd, D., and C. Meek. "Good Word Attacks on Statistical Spam Filters." Presented at the Second Conference on Email and Anti-Spam, 2005.
5. *Microsoft Excel 2013*. Redmond, WA: Microsoft Corp., 2012.
6. *Minitab Release 16*. State College, PA: Minitab, Inc. 2010.
7. Paulos, J. A. *Innumeracy*. New York: Hill and Wang, 1988.
8. Silberman, S. "The Quest for Meaning," *Wired 8.02*, Febru 2000.
9. Zeller, T. "The Fight Against V1@gra (and Other Spam)," Th *New York Times*, May 21, 2006, pp. B1, B6.

Probability of Occurrence

$$\text{Probability of occurrence} = \frac{X}{T} \quad \textbf{(4.1)}$$

Marginal Probability

$$P(A) = P(A \textit{ and } B_1) + P(A \textit{ and } B_2) + \cdots + P(A \textit{ and } B_k) \quad \textbf{(4.2)}$$

General Addition Rule

$$P(A \textit{ or } B) = P(A) + P(B) - P(A \textit{ and } B) \quad \textbf{(4.3)}$$

Conditional Probability

$$P(A|B) = \frac{P(A \textit{ and } B)}{P(B)} \quad \textbf{(4.4a)}$$

$$P(B|A) = \frac{P(A \textit{ and } B)}{P(A)} \quad \textbf{(4.4b)}$$

Independence

$$P(A|B) = P(A) \quad \textbf{(4.5)}$$

General Multiplication Rule

$$P(A \textit{ and } B) = P(A|B)P(B) \quad \textbf{(4.6)}$$

Multiplication Rule for Independent Events

$$P(A \textit{ and } B) = P(A)P(B) \quad \textbf{(4.7)}$$

Marginal Probability Using the General Multiplication Rule

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \quad \textbf{(4.8)}$$

Bayes' Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k)} \quad \textbf{(4.9)}$$

Counting Rule 1

$$k^n \quad \textbf{(4.10)}$$

Counting Rule 2

$$(k_1)(k_2)\ldots(k_n) \quad \textbf{(4.11)}$$

Counting Rule 3

$$n! = (n)(n-1)\ldots(1) \quad \textbf{(4.12)}$$

Counting Rule 4: Permutations

$${}_nP_x = \frac{n!}{(n-x)!} \quad \textbf{(4.13)}$$

Counting Rule 5: Combinations

$${}_nC_x = \frac{n!}{x!(n-x)!} \quad \textbf{(4.14)}$$

KEY TERMS