Chapter 2

Probability Basics

This chapter reviews the probability that you need to read the remainder of this book. Section 2.1 presents the basics of probability theory; Section 2.2 reviews random variables; Section 2.3 briefly discusses the meaning of probability; and Section 2.4 shows how random variables are used in practice.

2.1 Probability Basics

After defining probability spaces, we discuss conditional probability, independence and conditional independence, and Bayes' Theorem.

2.1.1 Probability Spaces

You may recall using probability in situations such as drawing the top card from a deck of playing cards, tossing a coin, or drawing a ball from an urn. We call the process of drawing the top card or tossing a coin an **experiment**. Probability theory has to do with experiments that have a set of distinct **outcomes**. The set of all outcomes is called the **sample space** or **population**. Mathematicians ordinarily say sample space, while social scientists ordinarily say population. We will say sample space. In this simple review, we assume that the sample space is finite. Any subset of a sample space is called an **event**. A subset containing exactly one element is called an **elementary event**.

Example 2.1 *Suppose we have the experiment of drawing the top card from an ordinary deck of cards. Then the set*

$$\mathsf{E} = \{\textit{jack of hearts, jack of clubs, jack of spades, jack of diamonds}\}$$

is an event, and the set

$$\mathsf{F} = \{\textit{jack of hearts}\}$$

is an elementary event.

The meaning of an event is that one of the elements of the subset is the outcome of the experiment. In the preceding example, the meaning of the event E is that the card drawn is one of the four jacks, and the meaning of the elementary event F is that the card is the jack of hearts.

We articulate our certainty that an event contains the outcome of the experiment with a real number between 0 and 1. This number is called the **probability** of the event. When the sample space is finite, a probability of 0 means we are certain the event does not contain the outcome, whereas a probability of 1 means we are certain it does. Values in between represent varying degrees of belief. The following definition formally defines probability for a finite sample space.

Definition 2.1 *Suppose we have a sample space Ω containing n distinct elements; that is,*

$$\Omega = \{e_1, e_2, \ldots, e_n\}.$$

A function that assigns a real number $P(\mathsf{E})$ to each event $\mathsf{E} \subseteq \Omega$ is called a ***probability function*** *on the set of subsets of Ω if it satisfies the following conditions:*

1. $0 \leq P(e_i) \leq 1 \qquad$ *for* $1 \leq i \leq n$.
2. $P(e_1) + P(e_2) + \ldots + P(e_n) = 1$.
3. *For each event that is not an elementary event, $P(\mathsf{E})$ is the sum of the probabilities of the elementary events whose outcomes are in E. For example, if*

$$\mathsf{E} = \{e_3, e_6, e_8\}$$

then

$$P(\mathsf{E}) = P(e_3) + P(e_6) + P(e_8).$$

The pair (Ω, P) *is called a* ***probability space****.*

Because probability is defined as a function whose domain is a set of sets, we should write $P(\{e_i\})$ instead of $P(e_i)$ when denoting the probability of an elementary event. However, for the sake of simplicity, we do not do this. In the same way, we write $P(e_3, e_6, e_8)$ instead of $P(\{e_3, e_6, e_8\})$.

The most straightforward way to assign probabilities is to use the **Principle of Indifference**, which says that outcomes are to be considered equiprobable if we have no reason to expect one over the other. According to this principle, when there are n elementary events, each has probability equal to $1/n$.

Example 2.2 *Let the experiment be tossing a coin. Then the sample space is*

$$\Omega = \{heads, tails\},$$

and, according to the Principle of Indifference, we assign

$$P(heads) = P(tails) = .5.$$

We stress that there is nothing in the definition of a probability space that says we must assign the value of .5 to the probabilities of heads and tails. We could assign $P(heads) = .7$ and $P(tails) = .3$. However, if we have no reason to expect one outcome over the other, we give them the same probability.

Example 2.3 *Let the experiment be drawing the top card from a deck of* 52 *cards. Then* Ω *contains the faces of the* 52 *cards, and, according to the Principle of Indifference, we assign* $P(e) = 1/52$ *for each* $e \in \Omega$*. For example,*

$$P(\textit{jack of hearts}) = \frac{1}{52}.$$

The event

$$E = \{\textit{jack of hearts, jack of clubs, jack of spades, jack of diamonds}\}$$

means that the card drawn is a jack. Its probability is

$$\begin{aligned} P(E) &= P(\textit{jack of hearts}) + P(\textit{jack of clubs}) + \\ &\quad P(\textit{jack of spades}) + P(\textit{jack of diamonds}) \\ &= \frac{1}{52} + \frac{1}{52} + \frac{1}{52} + \frac{1}{52} = \frac{1}{13}. \end{aligned}$$

We have Theorem 2.1 concerning probability spaces. Its proof is left as an exercise.

Theorem 2.1 *Let (Ω, P) be a probability space. Then*

1. $P(\Omega) = 1$.
2. $0 \leq P(\mathsf{E}) \leq 1 \qquad$ *for every* $\mathsf{E} \subseteq \Omega$.
3. *For every two subsets* E *and* F *of* Ω *such that* $\mathsf{E} \cap \mathsf{F} = \varnothing$,
$$P(\mathsf{E} \cup \mathsf{F}) = P(\mathsf{E}) + P(\mathsf{F}),$$
where $\varnothing$ *denotes the empty set.*

Example 2.4 *Suppose we draw the top card from a deck of cards. Denote by* Queen *the set containing the* 4 *queens and by* King *the set containing the* 4 *kings. Then*
$$P(\mathsf{Queen} \cup \mathsf{King}) = P(\mathsf{Queen}) + P(\mathsf{King}) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$
because $\mathsf{Queen} \cap \mathsf{King} = \varnothing$. *Next denote by* Spade *the set containing the* 13 *spades. The sets* Queen *and* Spade *are not disjoint, so their probabilities are not additive. However, it is not hard to prove that, in general,*
$$P(\mathsf{E} \cup \mathsf{F}) = P(\mathsf{E}) + P(\mathsf{F}) - P(\mathsf{E} \cap \mathsf{F}).$$
So
$$\begin{aligned} P(\mathsf{Queen} \cup \mathsf{Spade}) &= P(\mathsf{Queen}) + P(\mathsf{Spade}) - P(\mathsf{Queen} \cap \mathsf{Spade}) \\ &= \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}. \end{aligned}$$

2.1.2 Conditional Probability and Independence

We start with a definition.

Definition 2.2 *Let* E *and* F *be events such that* $P(\mathsf{F}) \neq 0$. *Then the* ***conditional probability*** *of* E *given* F, *denoted* $P(\mathsf{E}|\mathsf{F})$, *is given by*
$$P(\mathsf{E}|\mathsf{F}) = \frac{P(\mathsf{E} \cap \mathsf{F})}{P(\mathsf{F})}.$$

We can gain intuition for this definition by considering probabilities that are assigned using the Principle of Indifference. In this case, $P(\mathsf{E}|\mathsf{F})$, as defined previously, is the ratio of the number of items in $\mathsf{E} \cap \mathsf{F}$ to the number of items in F. We show this as follows: Let n be the number of items in the sample space, n_{F} be the number of items in F, and n_{EF} be the number of items in $\mathsf{E} \cap \mathsf{F}$. Then
$$\frac{P(\mathsf{E} \cap \mathsf{F})}{P(\mathsf{F})} = \frac{n_{\mathsf{EF}}/n}{n_{\mathsf{F}}/n} = \frac{n_{\mathsf{EF}}}{n_{\mathsf{F}}},$$
which is the ratio of the number of items in $\mathsf{E} \cap \mathsf{F}$ to the number of items in F. As far as the meaning is concerned, $P(\mathsf{E}|\mathsf{F})$ is our belief that event E contains the outcome (i.e. E occurs) when we already know that event F contains the outcome (i.e. F occurred).

Example 2.5 *Again, consider drawing the top card from a deck of cards. Let* Jack *be the set of the* 4 *jacks,* RedRoyalCard *be the set of the* 6 *red royal cards,*[1] *and* Club *be the set of the* 13 *clubs. Then*

$$P(\mathsf{Jack}) = \frac{4}{52} = \frac{1}{13}$$

$$P(\mathsf{Jack}|\mathsf{RedRoyalCard}) = \frac{P(\mathsf{Jack} \cap \mathsf{RedRoyalCard})}{P(\mathsf{RedRoyalCard})} = \frac{2/52}{6/52} = \frac{1}{3}$$

$$P(\mathsf{Jack}|\mathsf{Club}) = \frac{P(\mathsf{Jack} \cap \mathsf{Club})}{P(\mathsf{Club})} = \frac{1/52}{13/52} = \frac{1}{13}.$$

Notice in the previous example that $P(\mathsf{Jack}|\mathsf{Club}) = P(\mathsf{Jack})$. This means that finding out the card is a club does not change the likelihood that it is a jack. We say that the two events are independent in this case, which is formalized in the following definition.

Definition 2.3 *Two events* E *and* F *are* ***independent*** *if one of the following holds:*

1. $P(\mathsf{E}|\mathsf{F}) = P(\mathsf{E})$ *and* $P(\mathsf{E}) \neq 0$, $P(\mathsf{F}) \neq 0$.
2. $P(\mathsf{E}) = 0$ *or* $P(\mathsf{F}) = 0$.

Notice that the definition states that the two events are independent even though it is in terms of the conditional probability of E given F. The reason is that independence is symmetric. That is, if $P(\mathsf{E}) \neq 0$ and $P(\mathsf{F}) \neq 0$, then $P(\mathsf{E}|\mathsf{F}) = P(\mathsf{E})$ if and only if $P(\mathsf{F}|\mathsf{E}) = P(\mathsf{F})$. It is straightforward to prove that E and F are independent if and only if $P(\mathsf{E} \cap \mathsf{F}) = P(\mathsf{E})P(\mathsf{F})$.

If you've previously studied probability, you should have already been introduced to the concept of independence. However, a generalization of independence, called **conditional independence**, is not covered in many introductory texts. This concept is important to the applications discussed in this book. We discuss it next.

Definition 2.4 *Two events* E *and* F *are* ***conditionally independent*** *given* G *if* $P(\mathsf{G}) \neq 0$ *and one of the following holds:*

1. $P(\mathsf{E}|\mathsf{F} \cap \mathsf{G}) = P(\mathsf{E}|\mathsf{G})$ *and* $P(\mathsf{E}|\mathsf{G}) \neq 0$, $P(\mathsf{F}|\mathsf{G}) \neq 0$.
2. $P(\mathsf{E}|\mathsf{G}) = 0$ *or* $P(\mathsf{F}|\mathsf{G}) = 0$.

Notice that this definition is identical to the definition of independence except that everything is conditional on G. The definition entails that E and F are independent once we know that the outcome is in G. The next example illustrates this.

[1] A royal card is a jack, queen, or king.

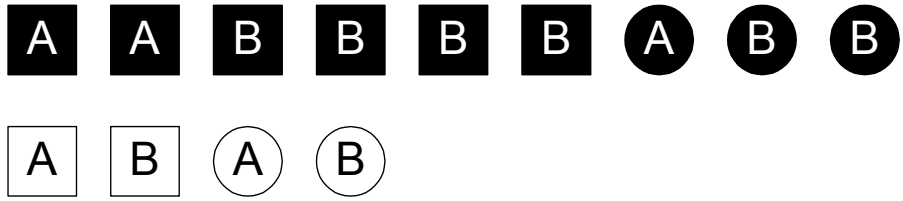

Figure 2.1: Using the Principle of Indifference, we assign a probability of 1/13 to each object.

Example 2.6 *Let* Ω *be the set of all objects in Figure 2.1. Using the Principle of Indifference, we assign a probability of* 1/13 *to each object. Let* Black *be the set of all black objects,* White *be the set of all white objects,* Square *be the set of all square objects, and* A *be the set of all objects containing an A. We then have that*

$$\begin{aligned} P(\mathsf{A}) &= \frac{5}{13} \\ P(\mathsf{A}|\mathsf{Square}) &= \frac{3}{8}. \end{aligned}$$

So A *and* Square *are not independent. However,*

$$\begin{aligned} P(\mathsf{A}|\mathsf{Black}) &= \frac{3}{9} = \frac{1}{3} \\ P(\mathsf{A}|\mathsf{Square} \cap \mathsf{Black}) &= \frac{2}{6} = \frac{1}{3}. \end{aligned}$$

We see that A *and* Square *are conditionally independent given* Black. *Furthermore,*

$$\begin{aligned} P(\mathsf{A}|\mathsf{White}) &= \frac{2}{4} = \frac{1}{2} \\ P(\mathsf{A}|\mathsf{Square} \cap \mathsf{White}) &= \frac{1}{2}. \end{aligned}$$

So A *and* Square *are also conditionally independent given* White.

Next, we discuss an important rule involving conditional probabilities. Suppose we have n events $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_n$ such that

$$\mathsf{E}_i \cap \mathsf{E}_j = \varnothing \qquad \text{for } i \neq j$$

and

$$\mathsf{E}_1 \cup \mathsf{E}_2 \cup \ldots \cup \mathsf{E}_n = \Omega.$$

Such events are called **mutually exclusive and exhaustive**. Then the **Law of Total Probability** says that for any other event F,

$$P(\mathsf{F}) = P(\mathsf{F} \cap \mathsf{E}_1) + P(\mathsf{F} \cap \mathsf{E}_2) + \cdots + P(\mathsf{F} \cap \mathsf{E}_n). \tag{2.1}$$

You are asked to prove this rule in the exercises. If $P(\mathsf{E}_i) \neq 0$, then

$$P(\mathsf{F} \cap \mathsf{E}_i) = P(\mathsf{F}|\mathsf{E}_i)P(\mathsf{E}_i).$$

Therefore, if $P(\mathsf{E}_i) \neq 0$ for all i, the law is often applied in the following form:

$$P(\mathsf{F}) = P(\mathsf{F}|\mathsf{E}_1)P(\mathsf{E}_1) + P(\mathsf{F}|\mathsf{E}_2)P(\mathsf{E}_2) + \cdots + P(\mathsf{F}|\mathsf{E}_n)P(\mathsf{E}_n). \tag{2.2}$$

Example 2.7 *Suppose we have the objects discussed in Example 2.6. Then, according to the Law of Total Probability,*

$$\begin{aligned} P(\mathsf{A}) &= P(\mathsf{A}|\mathsf{Black})P(\mathsf{Black}) + P(\mathsf{A}|\mathsf{White})P(\mathsf{White}) \\ &= \left(\frac{1}{3}\right)\left(\frac{9}{13}\right) + \left(\frac{1}{2}\right)\left(\frac{4}{13}\right) = \frac{5}{13}. \end{aligned}$$

2.1.3 Bayes' Theorem

We can compute conditional probabilities of events of interest from known probabilities using the following theorem.

Theorem 2.2 ***(Bayes)*** *Given two events* E *and* F *such that* $P(\mathsf{E}) \neq 0$ *and* $P(\mathsf{F}) \neq 0$*, we have*

$$P(\mathsf{E}|\mathsf{F}) = \frac{P(\mathsf{F}|\mathsf{E})P(\mathsf{E})}{P(\mathsf{F})}. \tag{2.3}$$

Furthermore, given n *mutually exclusive and exhaustive events* $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_n$ *such that* $P(\mathsf{E}_i) \neq 0$ *for all* i*, we have for* $1 \leq i \leq n$*,*

$$P(\mathsf{E}_i|\mathsf{F}) = \frac{P(\mathsf{F}|\mathsf{E}_i)P(\mathsf{E}_i)}{P(\mathsf{F}|\mathsf{E}_1)P(\mathsf{E}_1) + P(\mathsf{F}|\mathsf{E}_2)P(\mathsf{E}_2) + \cdots P(\mathsf{F}|\mathsf{E}_n)P(\mathsf{E}_n)}. \tag{2.4}$$

Proof. *To obtain Equality 2.3, we first use the definition of conditional probability as follows:*

$$P(\mathsf{E}|\mathsf{F}) = \frac{P(\mathsf{E} \cap \mathsf{F})}{P(\mathsf{F})} \qquad \textit{and} \qquad P(\mathsf{F}|\mathsf{E}) = \frac{P(\mathsf{F} \cap \mathsf{E})}{P(\mathsf{E})}.$$

Next we multiply each of these equalities by the denominator on its right side to show that

$$P(\mathsf{E}|\mathsf{F})P(\mathsf{F}) = P(\mathsf{F}|\mathsf{E})P(\mathsf{E})$$

because they both equal $P(\mathsf{E} \cap \mathsf{F})$*. Finally, we divide this last equality by* $P(\mathsf{F})$ *to obtain our result.*

To obtain Equality 2.4, we place the expression for F*, obtained using the Law of Total Probability (Equality 2.2), in the denominator of Equality 2.3.* ■

Both of the formulas in the preceding theorem are called **Bayes' Theorem** because the original version was developed by Thomas Bayes, published in 1763. The first enables us to compute $P(\mathsf{E}|\mathsf{F})$ if we know $P(\mathsf{F}|\mathsf{E})$, $P(\mathsf{E})$, and $P(\mathsf{F})$; the second enables us to compute $P(\mathsf{E}_i|\mathsf{F})$ if we know $P(\mathsf{F}|\mathsf{E}_j)$ and $P(\mathsf{E}_j)$ for $1 \leq j \leq n$. The next example illustrates the use of Bayes' Theorem.

Example 2.8 *Let Ω be the set of all objects in Figure 2.1, and assign each object a probability of 1/13. Let* A *be the set of all objects containing an A,* B *be the set of all objects containing a B, and* Black *be the set of all black objects. Then, according to Bayes' Theorem,*

$$\begin{aligned} P(\mathsf{Black}|\mathsf{A}) &= \frac{P(\mathsf{A}|\mathsf{Black})P(\mathsf{Black})}{P(\mathsf{A}|\mathsf{Black})P(\mathsf{Black}) + P(\mathsf{A}|\mathsf{White})P(\mathsf{White})} \\ &= \frac{\left(\frac{1}{3}\right)\left(\frac{9}{13}\right)}{\left(\frac{1}{3}\right)\left(\frac{9}{13}\right) + \left(\frac{1}{2}\right)\left(\frac{4}{13}\right)} = \frac{3}{5}, \end{aligned}$$

which is the same value we get by computing $P(\mathsf{Black}|\mathsf{A})$ directly.

In the previous example we can just as easily compute $P(\mathsf{Black}|\mathsf{A})$ directly. We will see a useful application of Bayes' Theorem in Section 2.4.

2.2 Random Variables

In this section we present the formal definition and mathematical properties of a random variable. In Section 2.4 we show how they are developed in practice.

2.2.1 Probability Distributions of Random Variables

Definition 2.5 *Given a probability space (Ω, P), a* ***random variable*** *X is a function whose domain is Ω.*

The range of X is called the **space** of X.

Example 2.9 *Let Ω contain all outcomes of a throw of a pair of six-sided dice, and let P assign 1/36 to each outcome. Then Ω is the following set of ordered pairs:*

$$\Omega = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),\ldots(6,5),(6,6)\}.$$

Let the random variable X assign the sum of each ordered pair to that pair, and let the random variable Y assign odd to each pair of odd numbers and even to a pair if at least one number in that pair is an even number. The following table shows some of the values of X and Y.

e	$X(e)$	$Y(e)$
$(1,1)$	2	odd
$(1,2)$	3	even
$\cdots$	$\cdots$	$\cdots$
$(2,1)$	3	even
$\cdots$	$\cdots$	$\cdots$
$(6,6)$	12	even

The space of X is $\{2,3,4,5,6,7,8,9,10,11,12\}$, and that of Y is $\{$odd,even$\}$.

For a random variable X, we use $X = x$ to denote the subset containing all elements $e \in \Omega$ that X maps to the value of x. That is,

$$X = x \quad \text{represents the event} \quad \{e \text{ such that } X(e) = x\}.$$

Note the difference between X and x. Small x denotes any element in the space of X, whereas X is a function.

Example 2.10 *Let Ω, P, and X be as in Example 2.9. Then*

$$X = 3 \quad \textit{represents the event} \quad \{(1,2),(2,1)\} \textit{ and}$$

$$P(X = 3) = \frac{1}{18}.$$

Notice that

$$\sum_{x \in space(X)} P(X = x) = 1.$$

Example 2.11 *Let Ω, P, and Y be as in Example 2.9. Then*

$$\begin{aligned} \sum_{y \in space(Y)} P(Y = y) &= P(Y = odd) + P(Y = even) \\ &= \frac{9}{36} + \frac{27}{36} = 1. \end{aligned}$$

We call the values of $P(X = x)$ for all values x of X the **probability distribution** of the random variable X. When we are referring to the probability distribution of X, we write $P(X)$.

We often use x alone to represent the event $X = x$, and so we write $P(x)$ instead of $P(X = x)$ when we are referring to the probability that X has value x.

Example 2.12 *Let Ω, P, and X be as in Example 2.9. Then if $x = 3$,*

$$P(x) = P(X = x) = \frac{1}{18}.$$

If we want to refer to all values of, for example, the random variables X, we sometimes write $P(X)$ instead of $P(X = x)$ or $P(x)$.

Example 2.13 *Let Ω, P, and X be as in Example 2.9. Then for all values of X*

$$P(X) > 1.$$

Given two random variables X and Y, defined on the same sample space Ω, we use $X = x, Y = y$ to denote the subset containing all elements $e \in \Omega$ that are mapped both by X to x and by Y to y. That is,

$X = x, Y = y$ represents the event

$$\{e \text{ such that } X(e) = x\} \cap \{e \text{ such that } Y(e) = y\}.$$

Example 2.14 *Let Ω, P, X, and Y be as in Example 2.9. Then*

$$X = 4, Y = odd \quad \text{represents the event} \quad \{(1,3),(3,1)\},$$

and so

$$P(X = 4, Y = odd) = 1/18.$$

We call $P(X = x, Y = y)$ the **joint probability distribution** of X and Y. If $\mathsf{A} = \{X, Y\}$, we also call this the joint probability distribution of A. Furthermore, we often just say *joint distribution* or *probability distribution*.

For brevity, we often use x, y to represent the event $X = x, Y = y$, and so we write $P(x, y)$ instead of $P(X = x, Y = y)$. This concept extends to three or more random variables. For example, $P(X = x, Y = y, Z = z)$ is the joint probability distribution function of the random variables X, Y, and Z, and we often write $P(x, y, z)$.

Example 2.15 *Let Ω, P, X, and Y be as in Example 2.9. Then, if $x = 4$ and $y = odd$,*

$$P(x, y) = P(X = x, Y = y) = 1/18.$$

Similar to the case of a single random variable, if we want to refer to all values of, for example, the random variables X and Y, we sometimes write $P(X, Y)$ instead of $P(X = x, Y = y)$ or $P(x, y)$.

Example 2.16 *Let Ω, P, X, and Y be as in Example 2.9. It is left as an exercise to show that for all values of x and y we have*

$$P(X = x, Y = y) < 1/2.$$

For example, as shown in Example 2.14

$$P(X = 4, Y = odd) = 1/18 < 1/2.$$

We can restate this fact as follows: for all values of X and Y we have that

$$P(X, Y) < 1/2.$$

If, for example, we let $\mathsf{A} = \{X, Y\}$ and $\mathsf{a} = \{x, y\}$, we use

$$\mathsf{A} = \mathsf{a} \qquad \text{to represent} \qquad X = x, Y = y,$$

and we often write $P(\mathsf{a})$ instead of $P(\mathsf{A} = \mathsf{a})$.

Example 2.17 *Let Ω, P, X, and Y be as in Example 2.9. If $\mathsf{A} = \{X, Y\}$, $\mathsf{a} = \{x, y\}$, $x = 4$, and $y = odd$, then*

$$P(\mathsf{A} = \mathsf{a}) = P(X = x, Y = y) = 1/18.$$

Recall the Law of Total Probability (Equalities 2.1 and 2.2). For two random variables X and Y, these equalities are as follows:

$$P(X = x) = \sum_y P(X = x, Y = y). \tag{2.5}$$

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y). \tag{2.6}$$

It is left as an exercise to show this.

Example 2.18 *Let Ω, P, X, and Y be as in Example 2.9. Then, owing to Equality 2.5,*

$$\begin{aligned} P(X = 4) &= \sum_y P(X = 4, Y = y) \\ &= P(X = 4, Y = odd) + P(X = 4, Y = even) = \frac{1}{18} + \frac{1}{36} = \frac{1}{12}. \end{aligned}$$

Example 2.19 *Again, let Ω, P, X, and Y be as in Example 2.9. Then, due to Equality 2.6,*

$$\begin{aligned} P(X = 4) &= \sum_y P(X = x|Y = y)P(Y = y) \\ &= P(X = 4|Y = odd)P(Y = odd) + \\ &\quad P(X = 4|Y = even)P(Y = even) \\ &= \frac{2}{9} \times \frac{9}{36} + \frac{1}{27} \times \frac{27}{36} = \frac{1}{12}. \end{aligned}$$

In Equality 2.5 the probability distribution $P(X = x)$ is called the **marginal probability distribution** of X relative to the joint distribution $P(X = x, Y = y)$ because it is obtained using a process similar to adding across a row or column in a table of numbers. This concept also extends in a straightforward way to three or more random variables. For example, if we have a joint distribution $P(X = x, Y = y, Z = z)$ of X, Y, and Z, the marginal distribution $P(X = x, Y = y)$ of X and Y is obtained by summing over all values of Z. If $\mathsf{A} = \{X, Y\}$, we also call this the **marginal probability distribution** of A.

The next example reviews the concepts covered so far concerning random variables.

Example 2.20 *Let Ω be a set of 12 individuals, and let P assign 1/12 to each individual. Suppose the sexes, heights, and wages of the individuals are as follows:*

Case	Sex	Height (inches)	Wage ($)
1	female	64	30,000
2	female	64	30,000
3	female	64	40,000
4	female	64	40,000
5	female	68	30,000
6	female	68	40,000
7	male	64	40,000
8	male	64	50,000
9	male	68	40,000
10	male	68	50,000
11	male	70	40,000
12	male	70	50,000

Let the random variables S, H, and W, respectively, assign the sex, height, and wage of an individual to that individual. Then the probability distributions of the three random variables are as follows (recall that, for example, $P(s)$ represents $P(S = s)$).

s	$P(s)$
female	1/2
male	1/2

h	$P(h)$
64	1/2
68	1/3
70	1/6

w	$P(w)$
30,000	1/4
40,000	1/2
50,000	1/4

The joint distribution of S and H is as follows:

s	h	$P(s, h)$
female	64	1/3
female	68	1/6
female	70	0
male	64	1/6
male	68	1/6
male	70	1/6

The following table also shows the joint distribution of S and H and illustrates that the individual distributions can be obtained by summing the joint distribution over all values of the other variable.

h / s	64	68	70	Distribution of S
female	1/3	1/6	0	1/2
male	1/6	1/6	1/6	1/2
Distribution of H	1/2	1/3	1/6	

The table that follows shows the first few values in the joint distribution of S, H, and W. There are 18 values in all, many of which are 0.

s	h	w	$P(s,h,w)$
female	64	30,000	1/6
female	64	40,000	1/6
female	64	50,000	0
female	68	30,000	1/12
...	...	...	...

We close with the **chain rule** for random variables, which says that given n random variables $X_1, X_2, \ldots, X_n$, defined on the same sample space Ω,

$$P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \cdots \times P(x_2|x_1) \times P(x_1)$$

whenever $P(x_1, x_2, \ldots, x_n) \neq 0$. It is straightforward to prove this rule using the rule for conditional probability.

Example 2.21 *Suppose we have the random variables in Example 2.20. Then, according to the chain rule for all values s, h, and w of S, H, and W,*

$$P(s,h,w) = P(w|h,s)P(h|s)P(s).$$

There are eight combinations of values of the three random variables. The table that follows shows that the equality holds for two of the combination.

s	h	w	$P(s,h,w)$	$P(w\|h,s)P(h\|s)P(s)$
female	64	30,000	$\frac{1}{6}$	$\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{6}$
female	64	40,000	$\frac{1}{12}$	$\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{1}{12}$

It is left as an exercise to show that the equality holds for the other six combinations.

2.2.2 Independence of Random Variables

The notion of independence extends naturally to random variables.

Definition 2.6 *Suppose we have a probability space (Ω, P) and two random variables X and Y defined on Ω. Then X and Y are **independent** if, for all values x of X and y of Y, the events $X = x$ and $Y = y$ are independent. When this is the case, we write*

$$I_P(X, Y),$$

where I_P stands for independent in P.

Example 2.22 *Let Ω be the set of all cards in an ordinary deck, and let P assign 1/52 to each card. Define random variables as follows:*

Variable	Value	Outcomes Mapped to This Value
R	r_1	All royal cards
	r_2	All nonroyal cards
S	s_1	All spades
	s_2	All nonspades

Then the random variables R and S are independent. That is,

$$I_P(R, S).$$

To show this, we need show for all values of r and s that

$$P(r|s) = P(r).$$

The following table shows that this is the case.

s	r	$P(r)$	$P(r\|s)$
s_1	r_1	$\frac{12}{52} = \frac{3}{13}$	$\frac{3}{13}$
s_1	r_2	$\frac{40}{52} = \frac{10}{13}$	$\frac{10}{13}$
s_2	r_1	$\frac{12}{52} = \frac{3}{13}$	$\frac{9}{39} = \frac{3}{13}$
s_2	r_2	$\frac{40}{52} = \frac{10}{13}$	$\frac{30}{39} = \frac{10}{13}$

The concept of conditional independence also extends naturally to random variables.

Definition 2.7 *Suppose we have a probability space (Ω, P), and three random variables X, Y, and Z defined on Ω. Then X and Y are **conditionally independent** given Z if for all values x of X, y of Y, and z of Z, whenever $P(z) \neq 0$, the events $X = x$ and $Y = y$ are conditionally independent given the event $Z = z$. When this is the case, we write*

$$I_P(X, Y|Z).$$

Example 2.23 *Let Ω be the set of all objects in Figure 2.1, and let P assign $1/13$ to each object. Define random variables S (for shape), L (for letter), and C (for color) as follows:*

Variable	Value	Outcomes Mapped to This Value
L	l_1	All objects containing an A
	l_2	All objects containing a B
S	s_1	All square objects
	s_2	All circular objects
C	c_1	All black objects
	c_2	All white objects

Then L and S are conditionally independent given C. That is,

$$I_P(L, S|C).$$

To show this, we need to show for all values of l, s, and c that

$$P(l|s, c) = P(l|c).$$

There is a total of eight combinations of the three variables. The table that follows shows that the equality holds for two of the combinations:

c	s	l	$P(l\|s,c)$	$P(l\|c)$
c_1	s_1	l_1	$\frac{2}{6} = \frac{1}{3}$	$\frac{3}{9} = \frac{1}{3}$
c_1	s_1	l_2	$\frac{4}{6} = \frac{2}{3}$	$\frac{6}{9} = \frac{2}{3}$

It is left as an exercise to show that it holds for the other combinations.

Independence and conditional independence can also be defined for sets of random variables.

Definition 2.8 *Suppose we have a probability space* (Ω, P) *and two sets* A *and* B *containing random variables defined on* Ω*. Let* a *and* b *be sets of values of the random variables in* A *and* B*, respectively. The sets* A *and* B *are said to be* ***independent*** *if, for all values of the variables in the sets* a *and* b*, the events* $\mathsf{A} = \mathsf{a}$ *and* $\mathsf{B} = \mathsf{b}$ *are independent. When this is the case, we write*

$$I_P(\mathsf{A}, \mathsf{B}),$$

where I_P *stands for* independent in P*.*

Example 2.24 *Let* Ω *be the set of all cards in an ordinary deck, and let* P *assign* $1/52$ *to each card. Define random variables as follows:*

Variable	Value	Outcomes Mapped to This Value
R	r_1	All royal cards
	r_2	All nonroyal cards
T	t_1	All tens and jacks
	t_2	All cards that are neither tens nor jacks
S	s_1	All spades
	s_2	All nonspades

Then the sets $\{R, T\}$ *and* $\{S\}$ *are independent. That is,*

$$I_P(\{R, T\}, \{S\}). \tag{2.7}$$

To show this, we need to show for all values of r*,* t*, and* s *that*

$$P(r, t|s) = P(r, t).$$

There are eight combinations of values of the three random variables. The table that follows shows that the equality holds for two of the combinations.

s	r	t	$P(r,t\|s)$	$P(r,t)$
s_1	r_1	t_1	$\frac{1}{13}$	$\frac{4}{52} = \frac{1}{13}$
s_1	r_1	t_2	$\frac{2}{13}$	$\frac{8}{52} = \frac{2}{13}$

It is left as an exercise to show that it holds for the other combinations.

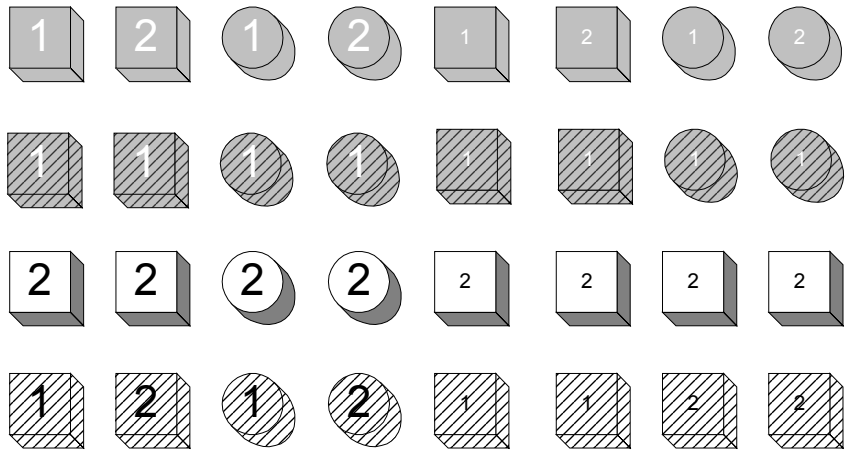

Figure 2.2: Objects with five properties.

When a set contains a single variable, we do not ordinarily show the braces. For example, we write Independency 2.7 as

$$I_P(\{R, T\}, S).$$

Definition 2.9 *Suppose we have a probability space* (Ω, P) *and three sets* A, B, *and* C *containing random variables defined on* Ω. *Let* a, b, *and* c *be sets of values of the random variables in* A, B, *and* C, *respectively. Then the sets* A *and* B *are said to be* ***conditionally independent*** *given the set* C *if, for all values of the variables in the sets* a, b, *and* c, *whenever* $P(\mathsf{c}) \neq 0$, *the events* $\mathsf{A} = \mathsf{a}$ *and* $\mathsf{B} = \mathsf{b}$ *are conditionally independent given the event* $\mathsf{C} = \mathsf{c}$. *When this is the case, we write*

$$I_P(\mathsf{A}, \mathsf{B}|\mathsf{C}).$$

Example 2.25 *Suppose we use the Principle of Indifference to assign probabilities to the objects in Figure 2.2, and we define random variables as follows:*

Variable	Value	Outcomes Mapped to This Value
V	v_1	All objects containing a 1
	v_2	All objects containing a 2
L	l_1	All objects covered with lines
	l_2	All objects not covered with lines
C	c_1	All gray objects
	c_2	All white objects
S	s_1	All square objects
	s_2	All circular objects
F	f_1	All objects containing a number in a large font
	f_2	All objects containing a number in a small font

It is left as an exercise to show for all values of v, l, c, s, and f that

$$P(v, l|s, f, c) = P(v, l|c).$$

So we have

$$I_P(\{V, L\}, \{S, F\}|C).$$

2.3 The Meaning of Probability

When one does not have the opportunity to study probability theory in depth, one is often left with the impression that all probabilities are computed using ratios. Next, we discuss the meaning of probability in more depth and show that this is not how probabilities are ordinarily determined.

2.3.1 Relative Frequency Approach to Probability

A classic textbook example of probability concerns tossing a coin. Because the coin is symmetrical, we use the Principle of Indifference to assign

$$P(\mathsf{Heads}) = P(\mathsf{Tails}) = .5.$$

Suppose that instead we toss a thumbtack. It can also land one of two ways. That is, it could land on its flat end, which we will call "heads," or it could land with the edge of the flat end and the point touching the ground, which we will call "tails." Because the thumbtack is not symmetrical, we have no reason to apply the Principle of Indifference and assign probabilities of .5 to both outcomes. How then should we assign the probabilities? In the case of the coin, when we assign $P(heads) = .5$, we are implicitly assuming that if we tossed the coin a large number of times it would land heads about half the time. That is, if we tossed the coin 1000 times, we would expect it to land heads about 500 times. This notion of repeatedly performing the experiment gives us a method for computing (or at least estimating) the probability. That is, if we repeat an experiment many times, we are fairly certain that the probability of an outcome is about equal to the fraction of times the outcome occurs. For example, a student tossed a thumbtack 10,000 times and it landed heads 3761 times. So

$$P(\mathsf{Heads}) \approx \frac{3761}{10{,}000} = .3761.$$

Indeed, in 1919 Richard von Mises used the limit of this fraction as the definition of probability. That is, if n is the number of tosses and S_n is the number of times the thumbtack lands heads, then

$$P(\mathsf{Heads}) \equiv \lim_{n\to\infty} \frac{S_n}{n}.$$

This definition assumes that a limit actually is approached. That is, it assumes that the ratio does not fluctuate. For example, there is no reason *a priori* to assume that the ratio is not .5 after 100 tosses, .1 after 1000 tosses, .5 after

10,000 tosses, .1 after 100,000 tosses, and so on. Only experiments in the real world can substantiate that a limit is approached. In 1946 J. E. Kerrich conducted many such experiments using games of chance in which the Principle of Indifference seemed to apply (e.g., drawing a card from a deck). His results indicated that the relative frequency does appear to approach a limit and that this limit is the value obtained using the Principle of Indifference.

This approach to probability is called the **relative frequency approach to probability**, and probabilities obtained using this approach are called **relative frequencies**. A **frequentist** is someone who feels this is the only way we can obtain probabilities. Note that, according to this approach, we can never know a probability for certain. For example, if we tossed a coin 10,000 times and it landed heads 4991 times, we would estimate

$$P(\mathsf{Heads}) \approx \frac{4991}{10{,}000} = .4991.$$

On the other hand, if we used the Principle of Indifference, we would assign $P(\mathsf{Heads}) = .5$. In the case of the coin, the probability might not actually be .5 because the coin might not be perfectly symmetrical. For example, Kerrich [1946] found that the six came up the most in the toss of a die and that one came up the least. This makes sense because, at that time, the spots on the die were hollowed out of the die, so the die was lightest on the side with a six. On the other hand, in experiments involving cards or urns, it seems we can be certain of probabilities obtained using the Principle of Indifference.

Example 2.26 *Suppose we toss an asymmetrical six-sided die, and in* 1000 *tosses we observe that each of the six sides comes up the following number of times.*

Side	Number of Times
1	250
2	150
3	200
4	70
5	280
6	50

So we estimate $P(1) \approx .25$, $P(2) \approx .15$, $P(3) \approx .2$, $P(4) \approx .07$, $P(5) \approx .28$, *and* $P(6) \approx .05$.

Repeatedly performing an experiment (so as to estimate a relative frequency) is called **sampling**, and the set of outcomes is called a **random sample** (or simply a sample). The set from which we sample is called a **population**.

Example 2.27 *Suppose our population is all males in the United States between the ages of* 31 *and* 85, *and we are interested in the probability of such*

males having high blood pressure. Then, if we sample 10,000 *males, this set of males is our sample. Furthermore, if* 3210 *have high blood pressure, we estimate*

$$P(\textit{High Blood Pressure}) \approx \frac{3210}{10{,}000} = .321.$$

Technically, we should not call the set of all current males in this age group the population. Rather, the theory says that there is a propensity for a male in this group to have high blood pressure and that this propensity is the probability. This propensity might not be equal to the fraction of current males in the group who have high blood pressure. In theory, we would have to have an infinite number of males to determine the probability exactly. The current set of males in this age group is called a **finite population**. The fraction of them with high blood pressure is the probability of obtaining a male with high blood pressure when we sample him from the set of all males in the age group. This latter probability is simply the ratio of males with high blood pressure.

When doing statistical inference, we sometimes want to estimate the ratio in a finite population from a sample of the population, and at other times we want to estimate a propensity from a finite sequence of observations. For example, TV raters ordinarily want to estimate the actual fraction of people in a nation watching a show from a sample of those people. On the other hand, medical scientists want to estimate the propensity with which males tend to have high blood pressure from a finite sequence of males. One can create an infinite sequence from a finite population by returning a sampled item back to the population before sampling the next item. This is called **sampling with replacement**. In practice, it is rarely done, but ordinarily the finite population is so large that statisticians make the simplifying assumption that it is done. That is, they do not replace the item but still assume that the ratio is unchanged for the next item sampled.

In sampling, the observed relative frequency is called the **maximum likelihood estimate (MLE)** of the probability (limit of the relative frequency) because it is the estimate of the probability that makes the observed sequence most probable when we assume that the trials (repetitions of the experiment) are probabilistically independent. See Section 3.1.3 for further discussion of this estimate.

Another facet of von Mises' relative frequency approach is that a random process is generating the sequence of outcomes. According to von Mises' theory, a **random process** is defined as a repeatable experiment for which the infinite sequence of outcomes is assumed to be a random sequence. Intuitively, a **random sequence** is one that shows no regularity or pattern. For example, the finite binary sequence 1011101100 appears random, whereas the sequence 1010101010 does not, because it has the pattern 10 repeated five times. There is evidence that experiments such as coin tossing and dice throwing are indeed random processes. In 1971 Iversen et al. ran many experiments with dice indicating the sequence of outcomes is random. It is believed that unbiased sampling also yields a random sequence and is therefore a random process. See [van Lambalgen, 1987] for a formal treatment of random sequences.

2.3.2 Subjective Approach to Probability

If we tossed a thumbtack 10,000 times and it landed heads 6000 times, we would estimate $P(heads)$ to be .6. Exactly what does this number approximate? Is there some probability, accurate to an arbitrary number of digits, of the thumbtack landing heads? It seems not. Indeed, as we toss the thumbtack, its shape will slowly be altered, changing any propensity for landing heads. As another example, is there really an exact propensity for a male in a certain age group to have high blood pressure? Again, it seems not. So it seems that, outside of games of chance involving cards and urns, the relative frequency notion of probability is only an idealization. Regardless, we obtain useful insights concerning our beliefs from this notion. For example, after the thumbtack lands heads 6000 times out of 10,000 tosses, we believe it has about a .6 chance of landing heads on the next toss, and we bet accordingly. That is, we would consider it fair to win \$0.40 if the thumbtack landed heads and to lose $\$1 - \$0.40 = \$0.60$ if the thumbtack landed tails. Since the bet is considered fair, the opposite position, namely, to lose \$0.40 if the thumbtack landed heads and to win \$0.60 if it landed tails, would also be considered fair. Hence, we would take either side of the bet. This notion of a probability as a value that determines a fair bet is called a **subjective approach to probability**, and probabilities assigned within this frame are called **subjective probabilities** or **beliefs**. A **subjectivist** is someone who feels we can assign probabilities within this framework. More concretely, in this approach the **subjective probability** of an uncertain event is the fraction p of units of money we would agree it is fair to give (lose) if the event does not occur in exchange for the promise to receive (win) $1 - p$ units if it does occur.

Example 2.28 *Suppose we estimate that* $P(\mathsf{Heads}) = .6$. *This means that we would agree it is fair to give* \$0.60 *if heads does not occur for the promise to receive* \$0.40 *if it does occur. Notice that if we repeated the experiment* 100 *times and heads did occur* 60% *of the time (as we expected), we would win* $60(\$0.40) = \$24$ *and lose* $40(\$0.60) = \$24$. *That is, we would break even.*

Unlike the relative frequency approach to probability, the subjective approach allows us to compute probabilities of events that are not repeatable. A classic example concerns betting at the racetrack. To decide how to bet, we must first determine how likely we feel it is that each horse will win. A particular race has never run before and never will be run again, so we cannot look at previous occurrences of the race to obtain our belief. Rather, we obtain this belief from a careful analysis of the horses' overall previous performance, of track conditions, of jockeys, and so on. Clearly, not everyone will arrive at the same probabilities based on their analyses. This is why these probabilities are called *subjective*. They are particular to individuals. In general, they do not have objective values in nature on which we all must agree. Of course, if we did do an experiment such as tossing a thumbtack 10,000 times and it landed heads 6000 times, most would agree that the probability of heads is about .6. Indeed, de Finetti [1937] showed that if we make certain reasonable assumptions about

your beliefs, this would have to be your probability.

Before pursuing this matter further, we discuss a concept related to probability, namely, odds. Mathematically, if $P(E)$ is the probability of event E, then the **odds** $O(E)$ are defined by

$$O(E) = \frac{P(E)}{1 - P(E)}.$$

As far as betting, $O(E)$ is the amount of money we would consider it fair to lose if E did not occur in return for gaining \$1 if E did occur.

Example 2.29 *Let E be the event that the horse Oldnag wins the Kentucky Derby. If we feel $P(E) = .2$, then*

$$O(E) = \frac{P(E)}{1 - P(E)} = \frac{.2}{1 - .2} = .25.$$

This means we would consider it fair to lose \$0.25 *if the horse did not win in return for gaining* \$1 *if it did win.*

If we state the fair bet in terms of probability (as discussed previously), we would consider it fair to lose \$0.20 *if the horse did not win in return for gaining* \$0.80 *if it did win. Notice that with both methods the ratio of the amount won to the amount lost is* 4*, so they are consistent in the way they determine betting behavior.*

At the racetrack, the betting odds shown are the odds against the event. That is, they are the odds of the event not occurring. If $P(E) = .2$ and $\neg E$ denotes that E does not occur, then

$$O(\neg E) = \frac{P(\neg E)}{1 - P(\neg E)} = \frac{.8}{1 - .8} = 4,$$

and the odds shown at the racetrack are 4 to 1 against E. If you bet on E, you will lose \$1 if E does not occur and win \$4 if E does occur. Note that these are the track odds based on the betting behavior of all participants. If you believe $P(E) = .5$, for you the odds against E are 1 to 1 (even money), and you should jump at the chance to get 4 to 1.

Some individuals are uncomfortable at being forced to consider wagering to assess a subjective probability. There are other methods for ascertaining these probabilities. One of the most popular is the following, which was suggested by Lindley in 1985. This method says an individual should liken the uncertain outcome to a game of chance by considering an urn containing white and black balls. The individual should determine for what fraction of white balls the individual would be indifferent between receiving a small prize if the uncertain event E happened (or turned out to be true) and receiving the same small prize if a white ball was drawn from the urn. That fraction is the individual's probability of the outcome. Such a probability can be constructed using binary cuts. If, for example, you were indifferent when the fraction was .8, for you

$P(E) = .8$. If someone else were indifferent when the fraction was .6, for that individual $P(E) = .6$. Again, neither individual is right or wrong.

It would be a mistake to assume that subjective probabilities are only important in gambling situations. Actually, they are important in all the applications discussed in this book. In the next section we illustrate interesting uses of subjective probabilities.

See [Neapolitan, 1990] for more on the two approaches to probability presented here.

2.4 Random Variables in Applications

Although it is mathematically elegant to first specify a sample space and then define random variables on the space, in practice this is not what we ordinarily do. In practice some single entity or set of entities has features, the states of which we want to determine but that we cannot determine for certain. So we settle for determining how likely it is that a particular feature is in a particular state. An example of a single entity is a jurisdiction in which we are considering introducing an economically beneficial chemical that might be carcinogenic. We would want to determine the relative risk of the chemical versus its benefits. An example of a set of entities is a set of patients with similar diseases and symptoms. In this case, we would want to diagnose diseases based on symptoms. As mentioned in Section 2.3.1, this set of entities is called a *population*, and technically it is usually not the set of all currently existing entities, but rather is, in theory, an infinite set of entities.

In these applications, a random variable represents some feature of the entity being modeled, and we are uncertain as to the value of this feature. In the case of a single entity, we are uncertain as to the value of the feature for that entity, whereas in the case of a set of entities, we are uncertain as to the value of the feature for some members of the set. To help resolve this uncertainty, we develop probabilistic relationships among the variables. When there is a set of entities, we assume the entities in the set all have the same probabilistic relationships concerning the variables used in the model. When this is not the case, our analysis is not applicable. In the case of the scenario concerning introducing a chemical, features may include the amount of human exposure and the carcinogenic potential. If these are our features of interest, we identify the random variables $HumanExposure$ and $CarcinogenicPotential$. (For simplicity, our illustrations include only a few variables. An actual application ordinarily includes many more than this.) In the case of a set of patients, features of interest might include whether or not diseases such as lung cancer are present, whether or not manifestations of diseases such as a chest X-ray are present, and whether or not causes of diseases such as smoking are present. Given these features, we would identify the random variables $ChestXray$, $LungCancer$, and $SmokingHistory$, respectively.

After identifying the random variables, we distinguish a set of mutually exclusive and exhaustive values for each of them. The possible values of a random variable are the different states that the feature can take. For example,

the state of $LungCancer$ could be *present* or *absent*, the state of $ChestXray$ could be *positive* or *negative*, and the state of $SmokingHistory$ could be *yes* or *no*, where *yes* might mean the patient has smoked one or more packs of cigarettes every day during the past 10 years.

After distinguishing the possible values of the random variables (i.e., their spaces), we judge the probabilities of the random variables having their values. However, in general, we do not directly determine values in a joint probability distribution of the random variables. Rather, we ascertain probabilities concerning relationships among random variables that are accessible to us. We can then reason with these variables using Bayes' Theorem to obtain probabilities of events of interest. The next example illustrates this idea.

Example 2.30 *Suppose Sam plans to marry, and to obtain a marriage licence in the state in which he resides, one must take the blood test enzyme-linked immunosorbent assay (ELISA), which tests for the presence of human immunodeficiency virus (HIV). Sam takes the test and it comes back positive for HIV. How likely is it that Sam is infected with HIV? Without knowing the accuracy of the test, Sam really has no way of knowing how probable it is that he is infected with HIV.*

The data we ordinarily have on such tests are the true positive rate (sensitivity) and the true negative rate (specificity). The true positive rate is the number of people who both have the infection and test positive divided by the total number of people who have the infection. For example, to obtain this number for ELISA, 10,000 *people who were known to be infected with HIV were identified. This was done using the Western Blot, which is the gold standard test for HIV. These people were then tested with ELISA, and* 9990 *tested positive. Therefore, the true positive rate is* .999*. The true negative rate is the number of people who both do not have the infection and test negative divided by the total number of people who do not have the infection. To obtain this number for ELISA* 10,000 *nuns who denied risk factors for HIV infection were tested. Of these,* 9980 *tested negative using the ELISA test. Furthermore, the* 20 *positive-testing nuns tested negative using the Western Blot test. So, the true negative rate is* .998*, which means that the false positive rate is* .002*. We therefore formulate the following random variables and subjective probabilities:*

$$P(ELISA = positive|HIV = present) = .999 \tag{2.8}$$

$$P(ELISA = positive|HIV = absent) = .002. \tag{2.9}$$

You might wonder why we called these subjective probabilities *when we obtained them from data. Recall that the frequentist approach says that we can never know the actual relative frequencies (objective probabilities); we can only estimate them from data. However, within the subjective approach, we can make our beliefs (subjective probabilities) equal to the fractions obtained from the data.*

It might seem that Sam almost certainly is infected with HIV, since the test is so accurate. However, notice that neither the probability in Equality 2.8 nor

the one in Equality 2.9 is the probability of Sam being infected with HIV. Since we know that Sam tested positive on ELISA, that probability is

$$P(HIV = present|ELISA = positive).$$

We can compute this probability using Bayes' Theorem if we know $P(HIV = present)$. *Recall that Sam took the blood test simply because the state required it. He did not take it because he thought for any reason he was infected with HIV. So, the only other information we have about Sam is that he is a male in the state in which he resides. Therefore if* 1 *in* 100,000 *men in Sam's state is infected with HIV, we assign the following subjective probability:*

$$P(HIV = present) = .00001.$$

We now employ Bayes' Theorem to compute

$$\begin{aligned} & P(present|positive) \\ = & \frac{P(positive|present)P(present)}{P(positive|present)P(present) + P(positive|absent)P(absent)} \\ = & \frac{(.999)(.00001)}{(.999)(.00001) + (.002)(.99999)} \\ = & .00497. \end{aligned}$$

Surprisingly, we are fairly confident that Sam is not infected with HIV.

A probability such as $P(HIV = present)$ is called a **prior probability** because, in a particular model, it is the probability of some event prior to updating the probability of that event, within the framework of that model, using new information. Do not mistakenly think it means a probability prior to any information. A probability such as $P(HIV = present|ELISA = positive)$ is called a **posterior probability** because it is the probability of an event after its prior probability has been updated, within the framework of some model, based on new information. In the previous example the reason the posterior probability is small, even though the test is fairly accurate, is that the prior probability is extremely low. The next example shows how dramatically a different prior probability can change things.

Example 2.31 *Suppose Mary and her husband have been trying to have a baby and she suspects she is pregnant. She takes a pregnancy test that has a true positive rate of* .99 *and a false positive rate of* .02. *Suppose further that* 20% *of all women who take this pregnancy test are indeed pregnant. Using Bayes' Theorem we then have*

$$\begin{aligned} & P(present|positive) \\ = & \frac{P(positive|present)P(present)}{P(positive|present)P(present) + P(positive|absent)P(absent)} \\ = & \frac{(.99)(.2)}{(.99)(.2) + (.02)(.8)} \\ = & .92523. \end{aligned}$$

Even though Mary's test was far less accurate than Sam's test, she probably is pregnant, whereas he probably is not infected with HIV. This is due to the prior information. There was a significant prior probability (.2) that Mary was pregnant, because only women who suspect they are pregnant on other grounds take pregnancy tests. Sam, on the other hand, took his test simply because he wanted to get married. We had no previous information indicating he could be infected with HIV.

In the previous two examples we obtained our beliefs (subjective probabilities) directly from the observed fractions in the data. Although this is often done, it is not necessary. In general, we obtain our beliefs from our information about the past, which means that these beliefs are a composite of all our experience rather than merely observed relative frequencies. We will see examples of this throughout the book. The following example illustrates such a case.

Example 2.32 *Suppose you feel there is a .4 probability the NASDAQ will go up at least 1% today. This is based on your knowledge that, after trading closed yesterday, excellent earnings were reported by several big companies in the technology sector, and that U.S. crude oil supplies unexpectedly increased. Furthermore, if the NASDAQ does go up at least 1% today, you feel there is a .1 probability that your favorite stock NTPA will go up at least 10% today. If the NASDAQ does not go up at least 1% today, you feel there is only a .02 probability NTPA will go up at least 10% today. You have these beliefs because you know from the past that NTPA's performance is linked to overall performance in the technology sector. You checked NTPA after the close of trading, and you noticed it went up over 10%. What is the probability that the NASDAQ went up at least 1%? Using Bayes' Theorem we have*

$$
\begin{aligned}
& P(NASDAQ = up\ 1\% | NTPA = up\ 10\%) \\
= \ & \frac{P(up\ 10\% | up\ 1\%) P(up\ 1\%)}{P(up\ 10\% | up\ 1\%) P(up\ 1\%) + P(up\ 10\% | not\ up\ 1\%) P(not\ up\ 1\%)} \\
= \ & \frac{(.1)\,(.4)}{(.1)\,(.4) + (.02)(.6)} = .769.
\end{aligned}
$$

In the previous three examples we used Bayes' Theorem to compute posterior subjective probabilities from known subjective probabilities. In a rigorous sense, we can only do this within the subjective framework. That is, since strict frequentists say we can never know probabilities for certain, they cannot use Bayes' Theorem. They can only do analyses such as the computation of a confidence interval for the value of an unknown probability based on the data. These techniques are discussed in any classic statistics text such as [Hogg and Craig, 1972]. Since subjectivists are the ones who use Bayes' Theorem, they are often called **Bayesians**.

EXERCISES

Section 2.1

Exercise 2.1 Let the experiment be drawing the top card from a deck of 52 cards. Let Heart be the event a heart is drawn, and RoyalCard be the event a royal card is drawn.

1. Compute $P(\mathsf{Heart})$.
2. Compute $P(\mathsf{RoyalCard})$.
3. Compute $P(\mathsf{Heart} \cup \mathsf{RoyalCard})$.

Exercise 2.2 Prove Theorem 2.1.

Exercise 2.3 Example 2.5 showed that, in the draw of the top card from a deck, the event Jack is independent of the event Club. That is, it showed $P(\mathsf{Jack}|\ \mathsf{Club}) = P(\mathsf{Jack})$.

1. Show directly that the event Club is independent of the event Jack. That is, show $P(\mathsf{Club}|\mathsf{Jack}) = P(\mathsf{Club})$. Show also that $P(\mathsf{Jack} \cap \mathsf{Club}) = P(\mathsf{Jack})P(\mathsf{Club})$.
2. Show, in general, that if $P(\mathsf{E}) \neq 0$ and $P(\mathsf{F}) \neq 0$, then $P(\mathsf{E}|\mathsf{F}) = P(\mathsf{E})$ if and only if $P(\mathsf{F}|\mathsf{E}) = P(\mathsf{F})$, and each of these holds if and only if $P(\mathsf{E} \cap \mathsf{F}) = P(\mathsf{E})P(\mathsf{F})$.

Exercise 2.4 The complement of a set E consists of all the elements in Ω that are not in E and is denoted by $\overline{\mathsf{E}}$.

1. Show that E is independent of F if and only if $\overline{\mathsf{E}}$ is independent of F, which is true if and only if $\overline{\mathsf{E}}$ is independent of $\overline{\mathsf{F}}$.
2. Example 2.6 showed that, for the objects in Figure 2.1, A and Square are conditionally independent given Black and given White. Let B be the set of all objects containing a B, and Circle be the set of all circular objects. Use the result just obtained to conclude that A and Circle, B and Square, and B and Circle are each conditionally independent given either Black or White.

Exercise 2.5 Show that in the draw of the top card from a deck, the event $\mathsf{E} = \{kh, ks, qh\}$ and the event $\mathsf{F} = \{kh, kc, qh\}$ are conditionally independent given the event $\mathsf{G} = \{kh, ks, kc, kd\}$. Determine whether E and F are conditionally independent given $\overline{\mathsf{G}}$.

Exercise 2.6 Prove the Law of Total Probability, which says that if we have n mutually exclusive and exhaustive events $\mathsf{E}_1, \mathsf{E}_2, \ldots, \mathsf{E}_n$, then for any other event F,

$$P(\mathsf{F}) = P(\mathsf{F} \cap \mathsf{E}_1) + P(\mathsf{F} \cap \mathsf{E}_2) + \cdots + P(\mathsf{F} \cap \mathsf{E}_n).$$

Exercise 2.7 Let Ω be the set of all objects in Figure 2.1, and assign each object a probability of 1/13. Let A be the set of all objects containing an A, and Square be the set of all square objects. Compute $P(\mathsf{A}|\mathsf{Square})$ directly and using Bayes' Theorem.

Section 2.2

Exercise 2.8 Consider the probability space and random variables given in Example 2.20.

1. Determine the joint distribution of S and W, the joint distribution of W and H, and the remaining values in the joint distribution of S, H, and W.

2. Show that the joint distribution of S and H can be obtained by summing the joint distribution of S, H, and W over all values of W.

Exercise 2.9 Let a joint probability distribution be given. Using the law of total probability, show that, in general, the probability distribution of any one of the random variables is obtained by summing over all values of the other variables.

Exercise 2.10 The chain rule says that for n random variables $X_1, X_2, \ldots, X_n$, defined on the same sample space Ω,

$$P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1}, x_{n-2}, \ldots x_1) \cdots \times P(x_2|x_1) \times P(x_1)$$

whenever $P(x_1, x_2, \ldots, x_n) \neq 0$. Prove this rule.

Exercise 2.11 Use the results in Exercise 2.4 (1) to conclude that it was only necessary in Example 2.22 to show that $P(r, t) = P(r, t|s_1)$ for all values of r and t.

Exercise 2.12 Suppose we have two random variables X and Y with spaces $\{x_1, x_2\}$ and $\{y_1, y_2\}$, respectively.

1. Use the results in Exercise 2.4 (1) to conclude that we need only show $P(y_1|x_1) = P(y_1)$ to conclude $I_P(X, Y)$.

2. Develop an example showing that if X and Y both have spaces containing more than two values, then we need to check whether $P(y|x) = P(y)$ for all values of x and y to conclude $I_P(X, Y)$.

Exercise 2.13 Consider the probability space and random variables given in Example 2.20.

1. Are H and W independent?
2. Are H and W conditionally independent given S?
3. If this small sample is indicative of the probabilistic relationships among the variables in some population, what causal relationships might account for this dependency and conditional independency?

Exercise 2.14 In Example 2.25, it was left as an exercise to show for all values v of V, l of L, c of C, s of S, and f of F that

$$P(v, l|s, f, c) = P(v, l|c).$$

Show this.

Section 2.3

Exercise 2.15 Kerrich [1946] performed experiments such as tossing a coin many times, and he found that the relative frequency did appear to approach a limit. That is, for example, he found that after 100 tosses the relative frequency may have been .51, after 1000 tosses it may have been .508, after 10,000 tosses it may have been .5003, and after 100,000 tosses it may have been .50006. The pattern is that the 5 in the first place to the right of the decimal point remains in all relative frequencies after the first 100 tosses, the 0 in the second place remains in all relative frequencies after the first 1000 tosses, and so on. Toss a thumbtack at least 1000 times and see if you obtain similar results.

Exercise 2.16 Pick some upcoming event. It could be a sporting event or it could be the event that you will get an A in this course. Determine your probability of the event using Lindley's [1985] method of comparing the uncertain event to a draw of a ball from an urn. (See the discussion following Example 2.29.)

Section 2.4

Exercise 2.17 A forgetful nurse is supposed to give Mr. Nguyen a pill each day. The probability that the nurse will forget to give the pill on a given day is .3. If Mr. Nguyen receives the pill, the probability he will die is .1. If he does not receive the pill, the probability he will die is .8. Mr. Nguyen died today. Use Bayes' Theorem to compute the probability that the nurse forgot to give him the pill.

Exercise 2.18 An oil well might be drilled on Professor Neapolitan's farm in Texas. Based on what has happened on similar farms, we judge the probability of oil being present to be .5, the probability of only natural gas being present to be .2, and the probability of neither being present to be .3. If oil is present, a geological test will give a positive result with probability .9; if only natural gas is present, it will give a positive result with probability .3; and if neither is present, the test will be positive with probability .1. Suppose the test comes back positive. Use Bayes' Theorem to compute the probability that oil is present.