CHAPTER 1

Probability

1.1 Introduction

The idea of probability, chance, or randomness is quite old, whereas its rigorous axiomatization in mathematical terms occurred relatively recently. Many of the ideas of probability theory originated in the study of games of chance. In this century, the mathematical theory of probability has been applied to a wide variety of phenomena; the following are some representative examples:

- Probability theory has been used in genetics as a model for mutations and ensuing natural variability, and plays a central role in bioinformatics.
- The kinetic theory of gases has an important probabilistic component.
- In designing and analyzing computer operating systems, the lengths of various queues in the system are modeled as random phenomena.
- There are highly developed theories that treat noise in electrical devices and communication systems as random processes.
- Many models of atmospheric turbulence use concepts of probability theory.
- In operations research, the demands on inventories of goods are often modeled as random.
- Actuarial science, which is used by insurance companies, relies heavily on the tools of probability theory.
- Probability theory is used to study complex systems and improve their reliability, such as in modern commercial or military aircraft.
- Probability theory is a cornerstone of the theory of finance.

The list could go on and on.

This book develops the basic ideas of probability and statistics. The first part explores the theory of probability as a mathematical model for chance phenomena. The second part of the book is about statistics, which is essentially concerned with

procedures for analyzing data, especially data that in some vague sense have a random character. To comprehend the theory of statistics, you must have a sound background in probability.

1.2 Sample Spaces

Probability theory is concerned with situations in which the outcomes occur randomly. Generically, such situations are called *experiments*, and the set of all possible outcomes is the **sample space** corresponding to an experiment. The sample space is denoted by Ω, and an element of Ω is denoted by ω. The following are some examples.

EXAMPLE **A** Driving to work, a commuter passes through a sequence of three intersections with traffic lights. At each light, she either stops, s, or continues, c. The sample space is the set of all possible outcomes:

$$\Omega = \{ccc, ccs, css, csc, sss, ssc, scc, scs\}$$

where csc, for example, denotes the outcome that the commuter continues through the first light, stops at the second light, and continues through the third light. ■

EXAMPLE **B** The number of jobs in a print queue of a mainframe computer may be modeled as random. Here the sample space can be taken as

$$\Omega = \{0, 1, 2, 3, \ldots\}$$

that is, all the nonnegative integers. In practice, there is probably an upper limit, N, on how large the print queue can be, so instead the sample space might be defined as

$$\Omega = \{0, 1, 2, \ldots, N\}$$ ■

EXAMPLE **C** Earthquakes exhibit very erratic behavior, which is sometimes modeled as random. For example, the length of time between successive earthquakes in a particular region that are greater in magnitude than a given threshold may be regarded as an experiment. Here Ω is the set of all nonnegative real numbers:

$$\Omega = \{t \mid t \geq 0\}$$ ■

We are often interested in particular subsets of Ω, which in probability language are called **events.** In Example A, the event that the commuter stops at the first light is the subset of Ω denoted by

$$A = \{sss, ssc, scc, scs\}$$

(Events, or subsets, are usually denoted by italic uppercase letters.) In Example B, the event that there are fewer than five jobs in the print queue can be denoted by

$$A = \{0, 1, 2, 3, 4\}$$

The algebra of set theory carries over directly into probability theory. The **union** of two events, A and B, is the event C that either A occurs or B occurs or both occur: $C = A \cup B$. For example, if A is the event that the commuter stops at the first light (listed before), and if B is the event that she stops at the third light,

$$B = \{sss, scs, ccs, css\}$$

then C is the event that she stops at the first light or stops at the third light and consists of the outcomes that are in A or in B or in both:

$$C = \{sss, ssc, scc, scs, ccs, css\}$$

The **intersection** of two events, $C = A \cap B$, is the event that both A and B occur. If A and B are as given previously, then C is the event that the commuter stops at the first light and stops at the third light and thus consists of those outcomes that are common to both A and B:

$$C = \{sss, scs\}$$

The **complement** of an event, A^c, is the event that A does not occur and thus consists of all those elements in the sample space that are not in A. The complement of the event that the commuter stops at the first light is the event that she continues at the first light:

$$A^c = \{ccc, ccs, css, csc\}$$

You may recall from previous exposure to set theory a rather mysterious set called the empty set, usually denoted by $\emptyset$. The **empty set** is the set with no elements; it is the event with no outcomes. For example, if A is the event that the commuter stops at the first light and C is the event that she continues through all three lights, $C = \{ccc\}$, then A and C have no outcomes in common, and we can write

$$A \cap C = \emptyset$$

In such cases, A and C are said to be **disjoint.**

Venn diagrams, such as those in Figure 1.1, are often a useful tool for visualizing set operations.

The following are some laws of set theory.

Commutative Laws:

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associative Laws:

$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive Laws:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

Of these, the distributive laws are the least intuitive, and you may find it instructive to illustrate them with Venn diagrams.

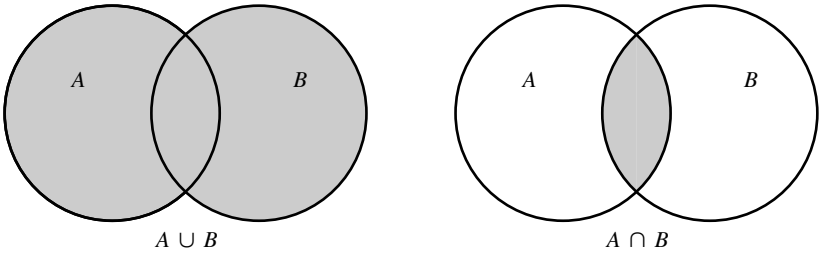

FIGURE **1.1** Venn diagrams of $A \cup B$ and $A \cap B$.

1.3 Probability Measures

A **probability measure** on Ω is a function P from subsets of Ω to the real numbers that satisfies the following axioms:

1. $P(\Omega) = 1$.
2. If $A \subset \Omega$, then $P(A) \geq 0$.
3. If A_1 and A_2 are disjoint, then

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

More generally, if $A_1, A_2, \ldots, A_n, \ldots$ are mutually disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The first two axioms are obviously desirable. Since Ω consists of all possible outcomes, $P(\Omega) = 1$. The second axiom simply states that a probability is nonnegative. The third axiom states that if A and B are disjoint—that is, have no outcomes in common—then $P(A \cup B) = P(A) + P(B)$ and also that this property extends to limits. For example, the probability that the print queue contains either one or three jobs is equal to the probability that it contains one plus the probability that it contains three.

The following properties of probability measures are consequences of the axioms.

Property A $P(A^c) = 1 - P(A)$. This property follows since A and A^c are disjoint with $A \cup A^c = \Omega$ and thus, by the first and third axioms, $P(A) + P(A^c) = 1$. In words, this property says that the probability that an event does not occur equals one minus the probability that it does occur.

Property B $P(\emptyset) = 0$. This property follows from Property A since $\emptyset = \Omega^c$. In words, this says that the probability that there is no outcome at all is zero.

Property C If $A \subset B$, then $P(A) \leq P(B)$. This property states that if B occurs whenever A occurs, then $P(A) \leq P(B)$. For example, if whenever it rains (A) it is cloudy (B), then the probability that it rains is less than or equal to the probability that it is cloudy. Formally, it can be proved as follows: B can be expressed as the union of two disjoint sets:

$$B = A \cup (B \cap A^c)$$

Then, from the third axiom,

$$P(B) = P(A) + P(B \cap A^c)$$

and thus

$$P(A) = P(B) - P(B \cap A^c) \leq P(B)$$

Property D Addition Law $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This property is easy to see from the Venn diagram in Figure 1.2. If $P(A)$ and $P(B)$ are added together, $P(A \cap B)$ is counted twice. To prove it, we decompose $A \cup B$ into three disjoint subsets, as shown in Figure 1.2:

$$C = A \cap B^c$$
$$D = A \cap B$$
$$E = A^c \cap B$$

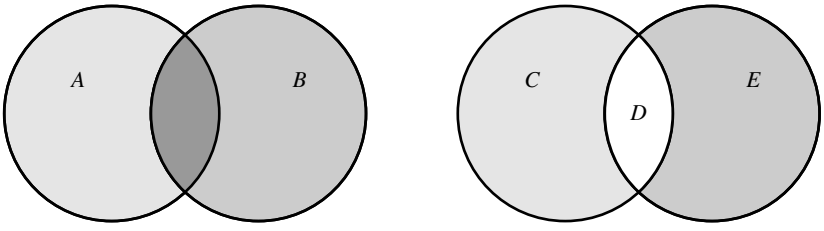

FIGURE **1.2** Venn diagram illustrating the addition law.

We then have, from the third axiom,

$$P(A \cup B) = P(C) + P(D) + P(E)$$

Also, $A = C \cup D$, and C and D are disjoint; so $P(A) = P(C) + P(D)$. Similarly, $P(B) = P(D) + P(E)$. Putting these results together, we see that

$$\begin{aligned} P(A) + P(B) &= P(C) + P(E) + 2P(D) \\ &= P(A \cup B) + P(D) \end{aligned}$$

or

$$P(A \cup B) = P(A) + P(B) - P(D)$$

EXAMPLE A Suppose that a fair coin is thrown twice. Let A denote the event of heads on the first toss, and let B denote the event of heads on the second toss. The sample space is

$$\Omega = \{hh, ht, th, tt\}$$

We assume that each elementary outcome in Ω is equally likely and has probability $\frac{1}{4}$. $C = A \cup B$ is the event that heads comes up on the first toss or on the second toss. Clearly, $P(C) \neq P(A) + P(B) = 1$. Rather, since $A \cap B$ is the event that heads comes up on the first toss and on the second toss,

$$P(C) = P(A) + P(B) - P(A \cap B) = .5 + .5 - .25 = .75 \quad \blacksquare$$

EXAMPLE B An article in the *Los Angeles Times* (August 24, 1987) discussed the statistical risks of AIDS infection:

> Several studies of sexual partners of people infected with the virus show that a single act of unprotected vaginal intercourse has a surprisingly low risk of infecting the uninfected partner—perhaps one in 100 to one in 1000. For an average, consider the risk to be one in 500. If there are 100 acts of intercourse with an infected partner, the odds of infection increase to one in five.
>
> Statistically, 500 acts of intercourse with one infected partner or 100 acts with five partners lead to a 100% probability of infection (statistically, not necessarily in reality).

Following this reasoning, 1000 acts of intercourse with one infected partner would lead to a probability of infection equal to 2 (statistically, but not necessarily in reality). To see the flaw in the reasoning that leads to this conclusion, consider two acts of intercourse. Let A_1 denote the event that infection occurs on the first act and let A_2 denote the event that infection occurs on the second act. Then the event that infection occurs is $B = A_1 \cup A_2$ and

$$P(B) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2) = \frac{2}{500} \quad \blacksquare$$

1.4 Computing Probabilities: Counting Methods

Probabilities are especially easy to compute for finite sample spaces. Suppose that $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ and that $P(\omega_i) = p_i$. To find the probability of an event A, we simply add the probabilities of the ω_i that constitute A.

EXAMPLE A Suppose that a fair coin is thrown twice and the sequence of heads and tails is recorded. The sample space is

$$\Omega = \{hh, ht, th, tt\}$$

As in Example A of the previous section, we assume that each outcome in Ω has probability .25. Let A denote the event that at least one head is thrown. Then $A = \{hh, ht, th\}$, and $P(A) = .75$. ■

This is a simple example of a fairly common situation. The elements of Ω all have equal probability; so if there are N elements in Ω, each of them has probability $1/N$. If A can occur in any of n mutually exclusive ways, then $P(A) = n/N$, or

$$P(A) = \frac{\text{number of ways } A \text{ can occur}}{\text{total number of outcomes}}$$

Note that this formula holds only if all the outcomes are equally likely. In Example A, if only the number of heads were recorded, then Ω would be $\{0, 1, 2\}$. These outcomes are not equally likely, and $P(A)$ is not $\frac{2}{3}$. ■

EXAMPLE **B** *Simpson's Paradox*

A black urn contains 5 red and 6 green balls, and a white urn contains 3 red and 4 green balls. You are allowed to choose an urn and then choose a ball at random from the urn. If you choose a red ball, you get a prize. Which urn should you choose to draw from? If you draw from the black urn, the probability of choosing a red ball is $\frac{5}{11} = .455$ (the number of ways you can draw a red ball divided by the total number of outcomes). If you choose to draw from the white urn, the probability of choosing a red ball is $\frac{3}{7} = .429$, so you should choose to draw from the black urn.

Now consider another game in which a second black urn has 6 red and 3 green balls, and a second white urn has 9 red and 5 green balls. If you draw from the black urn, the probability of a red ball is $\frac{6}{9} = .667$, whereas if you choose to draw from the white urn, the probability is $\frac{9}{14} = .643$. So, again you should choose to draw from the black urn.

In the final game, the contents of the second black urn are added to the first black urn, and the contents of the second white urn are added to the first white urn. Again, you can choose which urn to draw from. Which should you choose? Intuition says choose the black urn, but let's calculate the probabilities. The black urn now contains 11 red and 9 green balls, so the probability of drawing a red ball from it is $\frac{11}{20} = .55$. The white urn now contains 12 red and 9 green balls, so the probability of drawing a red ball from it is $\frac{12}{21} = .571$. So, you should choose the white urn. This counterintuitive result is an example of *Simpson's paradox*. For an example that occurred in real life, see Section 11.4.7. For more amusing examples, see Gardner (1976). ■

In the preceding examples, it was easy to count the number of outcomes and calculate probabilities. To compute probabilities for more complex situations, we must develop systematic ways of counting outcomes, which are the subject of the next two sections.

1.4.1 The Multiplication Principle

The following is a statement of the very useful multiplication principle.

MULTIPLICATION PRINCIPLE

If one experiment has m outcomes and another experiment has n outcomes, then there are mn possible outcomes for the two experiments.

Proof

Denote the outcomes of the first experiment by $a_1, \ldots, a_m$ and the outcomes of the second experiment by $b_1, \ldots, b_n$. The outcomes for the two experiments are the ordered pairs (a_i, b_j). These pairs can be exhibited as the entries of an $m \times n$ rectangular array, in which the pair (a_i, b_j) is in the ith row and the jth column. There are mn entries in this array. ■

EXAMPLE **A** Playing cards have 13 face values and 4 suits. There are thus $4 \times 13 = 52$ face-value/suit combinations. ■

EXAMPLE **B** A class has 12 boys and 18 girls. The teacher selects 1 boy and 1 girl to act as representatives to the student government. She can do this in any of $12 \times 18 = 216$ different ways. ■

EXTENDED MULTIPLICATION PRINCIPLE

If there are p experiments and the first has n_1 possible outcomes, the second $n_2, \ldots,$ and the pth n_p possible outcomes, then there are a total of $n_1 \times n_2 \times \cdots \times n_p$ possible outcomes for the p experiments.

Proof

This principle can be proved from the multiplication principle by induction. We saw that it is true for $p = 2$. Assume that it is true for $p = q$—that is, that there are $n_1 \times n_2 \times \cdots \times n_q$ possible outcomes for the first q experiments. To complete the proof by induction, we must show that it follows that the property holds for $p = q + 1$. We apply the multiplication principle, regarding the first q experiments as a single experiment with $n_1 \times \cdots \times n_q$ outcomes, and conclude that there are $(n_1 \times \cdots \times n_q) \times n_{q+1}$ outcomes for the $q + 1$ experiments. ■

EXAMPLE **C** An 8-bit binary word is a sequence of 8 digits, of which each may be either a 0 or a 1. How many different 8-bit words are there?

There are two choices for the first bit, two for the second, etc., and thus there are

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8 = 256$$

such words. ■

EXAMPLE **D**

A DNA molecule is a sequence of four types of nucleotides, denoted by A, G, C, and T. The molecule can be millions of units long and can thus encode an enormous amount of information. For example, for a molecule 1 million (10^6) units long, there are 4^{10^6} different possible sequences. This is a staggeringly large number having nearly a million digits. An amino acid is coded for by a sequence of three nucleotides; there are $4^3 = 64$ different codes, but there are only 20 amino acids since some of them can be coded for in several ways. A protein molecule is composed of as many as hundreds of amino acid units, and thus there are an incredibly large number of possible proteins. For example, there are 20^{100} different sequences of 100 amino acids. ■

1.4.2 Permutations and Combinations

A **permutation** is an ordered arrangement of objects. Suppose that from the set $C = \{c_1, c_2, \ldots, c_n\}$ we choose r elements and list them in order. How many ways can we do this? The answer depends on whether we are allowed to duplicate items in the list. If no duplication is allowed, we are **sampling without replacement.** If duplication is allowed, we are **sampling with replacement.** We can think of the problem as that of taking labeled balls from an urn. In the first type of sampling, we are not allowed to put a ball back before choosing the next one, but in the second, we are. In either case, when we are done choosing, we have a list of r balls ordered in the sequence in which they were drawn.

The extended multiplication principle can be used to count the number of different ordered samples possible for a set of n elements. First, suppose that sampling is done with replacement. The first ball can be chosen in any of n ways, the second in any of n ways, etc., so that there are $n \times n \times \cdots \times n = n^r$ samples. Next, suppose that sampling is done without replacement. There are n choices for the first ball, $n - 1$ choices for the second ball, $n - 2$ for the third, ..., and $n - r + 1$ for the rth. We have just proved the following proposition.

PROPOSITION **A**

For a set of size n and a sample of size r, there are n^r different ordered samples with replacement and $n(n-1)(n-2)\cdots(n-r+1)$ different ordered samples without replacement. ■

COROLLARY **A**

The number of orderings of n elements is $n(n-1)(n-2)\cdots 1 = n!$. ■

EXAMPLE A How many ways can five children be lined up?

This corresponds to sampling without replacement. According to Corollary A, there are $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$ different lines. ■

EXAMPLE B Suppose that from ten children, five are to be chosen and lined up. How many different lines are possible?

From Proposition A, there are $10 \times 9 \times 8 \times 7 \times 6 = 30{,}240$ different lines. ■

EXAMPLE C In some states, license plates have six characters: three letters followed by three numbers. How many distinct such plates are possible?

This corresponds to sampling with replacement. There are $26^3 = 17{,}576$ different ways to choose the letters and $10^3 = 1000$ ways to choose the numbers. Using the multiplication principle again, we find there are $17{,}576 \times 1000 = 17{,}576{,}000$ different plates. ■

EXAMPLE D If all sequences of six characters are equally likely, what is the probability that the license plate for a new car will contain no duplicate letters or numbers?

Call the desired event A; Ω consists of all 17,576,000 possible sequences. Since these are all equally likely, the probability of A is the ratio of the number of ways that A can occur to the total number of possible outcomes. There are 26 choices for the first letter, 25 for the second, 24 for the third, and hence $26 \times 25 \times 24 = 15{,}600$ ways to choose the letters without duplication (doing so corresponds to sampling without replacement), and $10 \times 9 \times 8 = 720$ ways to choose the numbers without duplication. From the multiplication principle, there are $15{,}600 \times 720 = 11{,}232{,}000$ nonrepeating sequences. The probability of A is thus

$$P(A) = \frac{11{,}232{,}000}{17{,}576{,}000} = .64$$ ■

EXAMPLE E *Birthday Problem*

Suppose that a room contains n people. What is the probability that at least two of them have a common birthday?

This is a famous problem with a counterintuitive answer. Assume that every day of the year is equally likely to be a birthday, disregard leap years, and denote by A the event that at least two people have a common birthday. As is sometimes the case, finding $P(A^c)$ is easier than finding $P(A)$. This is because A can happen in many ways, whereas A^c is much simpler. There are 365^n possible outcomes, and A^c can happen in $365 \times 364 \times \cdots \times (365 - n + 1)$ ways. Thus,

$$P(A^c) = \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$

and

$$P(A) = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$

The following table exhibits the latter probabilities for various values of n:

n	$P(A)$
4	.016
16	.284
23	.507
32	.753
40	.891
56	.988

From the table, we see that if there are only 23 people, the probability of at least one match exceeds .5. The probabilities in the table are larger than one might intuitively guess, showing that the coincidence is not unlikely. Try it in your class. ■

EXAMPLE **F** How many people must you ask to have a 50 : 50 chance of finding someone who shares your birthday?

Suppose that you ask n people; let A denote the event that someone's birthday is the same as yours. Again, working with A^c is easier. The total number of outcomes is 365^n, and the total number of ways that A^c can happen is 364^n. Thus,

$$P(A^c) = \frac{364^n}{365^n}$$

and

$$P(A) = 1 - \frac{364^n}{365^n}$$

For the latter probability to be .5, n should be 253, which may seem counterintuitive. ■

We now shift our attention from counting permutations to counting combinations. Here we are no longer interested in ordered samples, but in the constituents of the samples regardless of the order in which they were obtained. In particular, we ask the following question: If r objects are taken from a set of n objects without replacement and disregarding order, how many different samples are possible? From the multiplication principle, the number of ordered samples equals the number of unordered samples multiplied by the number of ways to order each sample. Since the number of ordered samples is $n(n-1)\cdots(n-r+1)$, and since a sample of size r can be ordered in $r!$ ways (Corollary A), the number of unordered samples is

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{(n-r)!r!}$$

This number is also denoted as $\binom{n}{r}$. We have proved the following proposition.

PROPOSITION **B**

The number of unordered samples of r objects selected from n objects without replacement is $\binom{n}{r}$.

The numbers $\binom{n}{k}$, called the **binomial coefficients,** occur in the expansion

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

In particular,

$$2^n = \sum_{k=0}^{n} \binom{n}{k}$$

This latter result can be interpreted as the number of subsets of a set of n objects. We just add the number of subsets of size 0 (with the usual convention that $0! = 1$), and the number of subsets of size 1, and the number of subsets of size 2, etc. ■

EXAMPLE **G** Up until 1991, a player of the California state lottery could win the jackpot prize by choosing the 6 numbers from 1 to 49 that were subsequently chosen at random by the lottery officials. There are $\binom{49}{6} = 13{,}983{,}816$ possible ways to choose 6 numbers from 49, and so the probability of winning was about 1 in 14 million. If there were no winners, the funds thus accumulated were rolled over (carried over) into the next round of play, producing a bigger jackpot. In 1991, the rules were changed so that a winner had to correctly select 6 numbers from 1 to 53. Since $\binom{53}{6} = 22{,}957{,}480$, the probability of winning decreased to about 1 in 23 million. Because of the ensuing rollover, the jackpot accumulated to a record of about $120 million. This produced a fever of play—people were buying tickets at the rate of between 1 and 2 million per hour and state revenues burgeoned. ■

EXAMPLE **H** In the practice of quality control, only a fraction of the output of a manufacturing process is sampled and examined, since it may be too time-consuming and expensive to examine each item, or because sometimes the testing is destructive. Suppose that n items are in a lot and a sample of size r is taken. There are $\binom{n}{r}$ such samples. Now suppose that the lot contains k defective items. What is the probability that the sample contains exactly m defectives?

Clearly, this question is relevant to the efficacy of the sampling scheme, and the most desirable sample size can be determined by computing such probabilities for various values of r. Call the event in question A. The probability of A is the number of ways A can occur divided by the total number of outcomes. To find the number of ways A can occur, we use the multiplication principle. There are $\binom{k}{m}$ ways to choose the m defective items in the sample from the k defectives in the lot, and there are $\binom{n-k}{r-m}$ ways to choose the $r-m$ nondefective items in the sample from the $n-k$ nondefectives in the lot. Therefore, A can occur in $\binom{k}{m}\binom{n-k}{r-m}$ ways. Thus, $P(A)$ is the

ratio of the number of ways A can occur to the total number of outcomes, or

$$P(A) = \frac{\binom{k}{m}\binom{n-k}{r-m}}{\binom{n}{r}}$$

■

EXAMPLE **I** *Capture/Recapture Method*

The so-called capture/recapture method is sometimes used to estimate the size of a wildlife population. Suppose that 10 animals are captured, tagged, and released. On a later occasion, 20 animals are captured, and it is found that 4 of them are tagged. How large is the population?

We assume that there are n animals in the population, of which 10 are tagged. If the 20 animals captured later are taken in such a way that all $\binom{n}{20}$ possible groups are equally likely (this is a big assumption), then the probability that 4 of them are tagged is (using the technique of the previous example)

$$\frac{\binom{10}{4}\binom{n-10}{16}}{\binom{n}{20}}$$

Clearly, n cannot be precisely determined from the information at hand, but it can be estimated. One method of estimation, called **maximum likelihood,** is to choose that value of n that makes the observed outcome most probable. (The method of maximum likelihood is one of the main subjects of a later chapter in this text.) The probability of the observed outcome as a function of n is called the **likelihood.** Figure 1.3 shows the likelihood as a function of n; the likelihood is maximized at $n = 50$.

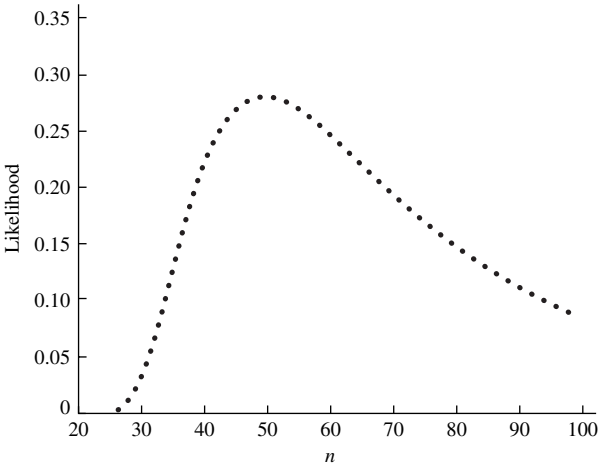

FIGURE **1.3** Likelihood for Example I.

To find the maximum likelihood estimate, suppose that, in general, t animals are tagged. Then, of a second sample of size m, r tagged animals are recaptured. We estimate n by the maximizer of the likelihood

$$L_n = \frac{\binom{t}{r}\binom{n-t}{m-r}}{\binom{n}{m}}$$

To find the value of n that maximizes L_n, consider the ratio of successive terms, which after some algebra is found to be

$$\frac{L_n}{L_{n-1}} = \frac{(n-t)(n-m)}{n(n-t-m+r)}$$

This ratio is greater than 1, i.e., L_n is increasing, if

$$\begin{aligned}(n-t)(n-m) &> n(n-t-m+r)\\ n^2-nm-nt+mt &> n^2-nt-nm-nr\\ mt &> nr\\ \frac{mt}{r} &> n\end{aligned}$$

Thus, L_n increases for $n < mt/r$ and decreases for $n > mt/r$; so the value of n that maximizes L_n is the greatest integer not exceeding mt/r.

Applying this result to the data given previously, we see that the maximum likelihood estimate of n is $\frac{mt}{r} = \frac{20 \cdot 10}{4} = 50$. This estimate has some intuitive appeal, as it equates the proportion of tagged animals in the second sample to the proportion in the population:

$$\frac{4}{20} = \frac{10}{n}$$

■

Proposition B has the following extension.

PROPOSITION **C**

The number of ways that n objects can be grouped into r classes with n_i in the ith class, $i = 1, \ldots, r$, and $\sum_{i=1}^{r} n_i = n$ is

$$\binom{n}{n_1 n_2 \cdots n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}$$

Proof

This can be seen by using Proposition B and the multiplication principle. (Note that Proposition B is the special case for which $r = 2$.) There are $\binom{n}{n_1}$ ways to choose the objects for the first class. Having done that, there are $\binom{n-n_1}{n_2}$ ways of choosing the objects for the second class. Continuing in this manner, there are

$$\frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{(n-n_1-n_2)!n_2!}\cdots\frac{(n-n_1-n_2-\cdots-n_{r-1})!}{0!n_r!}$$

choices in all. After cancellation, this yields the desired result. ■

EXAMPLE **J** A committee of seven members is to be divided into three subcommittees of size three, two, and two. This can be done in

$$\binom{7}{3\,2\,2} = \frac{7!}{3!2!2!} = 210$$

ways. ■

EXAMPLE **K** In how many ways can the set of nucleotides $\{A, A, G, G, G, G, C, C, C\}$ be arranged in a sequence of nine letters? Proposition C can be applied by realizing that this problem can be cast as determining the number of ways that the nine positions in the sequence can be divided into subgroups of sizes two, four, and three (the locations of the letters A, G, and C):

$$\binom{9}{2\,4\,3} = \frac{9!}{2!4!3!} = 1260$$

■

EXAMPLE **L** In how many ways can $n = 2m$ people be paired and assigned to m courts for the first round of a tennis tournament?

In this problem, $n_i = 2, i = 1, \ldots, m$, and, according to Proposition C, there are

$$\frac{(2m)!}{2^m}$$

assignments.

One has to be careful with problems such as this one. Suppose we were asked how many ways $2m$ people could be arranged in pairs without assigning the pairs to courts. Since there are $m!$ ways to assign the m pairs to m courts, the preceding result should be divided by $m!$, giving

$$\frac{(2m)!}{m!2^m}$$

pairs in all. ■

The numbers $\binom{n}{n_1 n_2 \cdots n_r}$ are called **multinomial coefficients.** They occur in the expansion

$$(x_1 + x_2 + \cdots + x_r)^n = \sum \binom{n}{n_1 n_2 \cdots n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

where the sum is over all nonnegative integers $n_1, n_2, \ldots, n_r$ such that $n_1 + n_2 + \cdots + n_r = n$.

1.5 Conditional Probability

We introduce the definition and use of conditional probability with an example. Digitalis therapy is often beneficial to patients who have suffered congestive heart failure, but there is the risk of digitalis intoxication, a serious side effect that is difficult to diagnose. To improve the chances of a correct diagnosis, the concentration of digitalis in the blood can be measured. Bellar et al. (1971) conducted a study of the relation of the concentration of digitalis in the blood to digitalis intoxication in 135 patients. Their results are simplified slightly in the following table, where this notation is used:

$T+$ = high blood concentration (positive test)
$T-$ = low blood concentration (negative test)
$D+$ = toxicity (disease present)
$D-$ = no toxicity (disease absent)

	$D+$	$D-$	*Total*
$T+$	25	14	39
$T-$	18	78	96
Total	43	92	135

Thus, for example, 25 of the 135 patients had a high blood concentration of digitalis and suffered toxicity.

Assume that the relative frequencies in the study roughly hold in some larger population of patients. (Making inferences about the frequencies in a large population from those observed in a small sample is a statistical problem, which will be taken up in a later chapter of this book.) Converting the frequencies in the preceding table to proportions (relative to 135), which we will regard as probabilities, we obtain the following table:

	$D+$	$D-$	*Total*
$T+$	.185	.104	.289
$T-$	.133	.578	.711
Total	.318	.682	1.000

From the table, $P(T+) = .289$ and $P(D+) = .318$, for example. Now if a doctor knows that the test was positive (that there was a high blood concentration), what is the probability of disease (toxicity) given this knowledge? We can restrict our attention to the first row of the table, and we see that of the 39 patients who had positive tests, 25 suffered from toxicity. We denote the probability that a patient shows toxicity given that the test is positive by $P(D+ \mid T+)$, which is called the **conditional probability** of $D+$ given $T+$.

$$P(D+ \mid T+) = \frac{25}{39} = .640$$

Equivalently, we can calculate this probability as

$$P(D+ \mid T+) = \frac{P(D+ \cap T+)}{P(T+)} = \frac{.185}{.289} = .640$$

In summary, we see that the unconditional probability of $D+$ is .318, whereas the conditional probability $D+$ given $T+$ is .640. Therefore, knowing that the test is positive makes toxicity more than twice as likely. What if the test is negative?

$$P(D- \mid T-) = \frac{.578}{.711} = .848$$

For comparison, $P(D-) = .682$. Two other conditional probabilities from this example are of interest: The probability of a false positive is $P(D- \mid T+) = .360$, and the probability of a false negative is $P(D+ \mid T-) = .187$.

In general, we have the following definition.

DEFINITION

Let A and B be two events with $P(B) \neq 0$. The conditional probability of A given B is defined to be

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$ ■

The idea behind this definition is that if we are given that event B occurred, the relevant sample space becomes B rather than Ω, and conditional probability is a probability measure on B. In the digitalis example, to find $P(D+ \mid T+)$, we restricted our attention to the 39 patients who had positive tests. For this new measure to be a probability measure, it must satisfy the axioms, and this can be shown.

In some situations, $P(A \mid B)$ and $P(B)$ can be found rather easily, and we can then find $P(A \cap B)$.

MULTIPLICATION LAW

Let A and B be events and assume $P(B) \neq 0$. Then

$$P(A \cap B) = P(A \mid B)P(B)$$ ■

The multiplication law is often useful in finding the probabilities of intersections, as the following examples illustrate.

EXAMPLE **A** An urn contains three red balls and one blue ball. Two balls are selected without replacement. What is the probability that they are both red?

Let R_1 and R_2 denote the events that a red ball is drawn on the first trial and on the second trial, respectively. From the multiplication law,

$$P(R_1 \cap R_2) = P(R_1)P(R_2 \mid R_1)$$

$P(R_1)$ is clearly $\frac{3}{4}$, and if a red ball has been removed on the first trial, there are two red balls and one blue ball left. Therefore, $P(R_2 \mid R_1) = \frac{2}{3}$. Thus, $P(R_1 \cap R_2) = \frac{1}{2}$. ■

EXAMPLE **B** Suppose that if it is cloudy (B), the probability that it is raining (A) is .3, and that the probability that it is cloudy is $P(B) = .2$ The probability that it is cloudy and raining is

$$P(A \cap B) = P(A \mid B)P(B) = .3 \times .2 = .06$$ ■

Another useful tool for computing probabilities is provided by the following law.

LAW OF TOTAL PROBABILITY

Let $B_1, B_2, \ldots, B_n$ be such that $\bigcup_{i=1}^{n} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, with $P(B_i) > 0$ for all i. Then, for any event A,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

Proof

Before going through a formal proof, it is helpful to state the result in words. The B_i are mutually disjoint events whose union is Ω. To find the probability of an event A, we sum the conditional probabilities of A given B_i, weighted by $P(B_i)$. Now, for the proof, we first observe that

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P\left(A \cap \left(\bigcup_{i=1}^{n} B_i\right)\right) \\ &= P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) \end{aligned}$$

Since the events $A \cap B_i$ are disjoint,

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} (A \cap B_i)\right) &= \sum_{i=1}^{n} P(A \cap B_i) \\ &= \sum_{i=1}^{n} P(A \mid B_i)P(B_i) \end{aligned}$$ ■

The law of total probability is useful in situations where it is not obvious how to calculate $P(A)$ directly but in which $P(A \mid B_i)$ and $P(B_i)$ are more straightforward, such as in the following example.

EXAMPLE **C** Referring to Example A, what is the probability that a red ball is selected on the second draw?

The answer may or may not be intuitively obvious—that depends on your intuition. On the one hand, you could argue that it is "clear from symmetry" that $P(R_2) = P(R_1) = \frac{3}{4}$. On the other hand, you could say that it is obvious that a red ball is likely to be selected on the first draw, leaving fewer red balls for the second draw, so that $P(R_2) < P(R_1)$. The answer can be derived easily by using the law of total probability:

$$\begin{aligned} P(R_2) &= P(R_2 \mid R_1)P(R_1) + P(R_2 \mid B_1)P(B_1) \\ &= \frac{2}{3} \times \frac{3}{4} + 1 \times \frac{1}{4} = \frac{3}{4} \end{aligned}$$

where B_1 denotes the event that a blue ball is drawn on the first trial. ■

As another example of the use of conditional probability, we consider a model that has been used for occupational mobility.

EXAMPLE **D** Suppose that occupations are grouped into upper (U), middle (M), and lower (L) levels. U_1 will denote the event that a father's occupation is upper-level; U_2 will denote the event that a son's occupation is upper-level, etc. (The subscripts index generations.) Glass and Hall (1954) compiled the following statistics on occupational mobility in England and Wales:

	U_2	M_2	L_2
U_1	.45	.48	.07
M_1	.05	.70	.25
L_1	.01	.50	.49

Such a table, which is called a *matrix of transition probabilities*, is to be read in the following way: If a father is in U, the probability that his son is in U is .45, the probability that his son is in M is .48, etc. The table thus gives conditional probabilities: for example, $P(U_2 \mid U_1) = .45$. Examination of the table reveals that there is more upward mobility from L into M than from M into U. Suppose that of the father's generation, 10% are in U, 40% in M, and 50% in L. What is the probability that a son in the next generation is in U?

Applying the law of total probability, we have

$$\begin{aligned} P(U_2) &= P(U_2 \mid U_1)P(U_1) + P(U_2 \mid M_1)P(M_1) + P(U_2 \mid L_1)P(L_1) \\ &= .45 \times .10 + .05 \times .40 + .01 \times .50 = .07 \end{aligned}$$

$P(M_2)$ and $P(L_2)$ can be worked out similarly. ■

Continuing with Example D, suppose we ask a different question: If a son has occupational status U_2, what is the probability that his father had occupational status U_1? Compared to the question asked in Example D, this is an "inverse" problem; we are given an "effect" and are asked to find the probability of a particular "cause." In situations like this, Bayes' rule, which we state shortly, is useful. Before stating the rule, we will see what it amounts to in this particular case.

We wish to find $P(U_1 \mid U_2)$. By definition,

$$P(U_1 \mid U_2) = \frac{P(U_1 \cap U_2)}{P(U_2)}$$
$$= \frac{P(U_2 \mid U_1)P(U_1)}{P(U_2 \mid U_1)P(U_1) + P(U_2 \mid M_1)P(M_1) + P(U_2 \mid L_1)P(L_1)}$$

Here we used the multiplication law to reexpress the numerator and the law of total probability to restate the denominator. The value of the numerator is $P(U_2 \mid U_1)P(U_1) = .45 \times .10 = .045$, and we calculated the denominator in Example D to be .07, so we find that $P(U_1 \mid U_2) = .64$. In other words, 64% of the sons who are in upper-level occupations have fathers who were in upper-level occupations.

We now state Bayes' rule.

BAYES' RULE

Let A and $B_1, \ldots, B_n$ be events where the B_i are disjoint, $\bigcup_{i=1}^{n} B_i = \Omega$, and $P(B_i) > 0$ for all i. Then

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

The proof of Bayes' rule follows exactly as in the preceding discussion. ■

EXAMPLE **E** Diamond and Forrester (1979) applied Bayes' rule to the diagnosis of coronary artery disease. A procedure called cardiac fluoroscopy is used to determine whether there is calcification of coronary arteries and thereby to diagnose coronary artery disease. From the test, it can be determined if 0, 1, 2, or 3 coronary arteries are calcified. Let T_0, T_1, T_2, T_3 denote these events. Let $D+$ or $D-$ denote the event that disease is present or absent, respectively. Diamond and Forrester presented the following table, based on medical studies:

i	$P(T_i \mid D+)$	$P(T_i \mid D-)$
0	.42	.96
1	.24	.02
2	.20	.02
3	.15	.00

According to Bayes' rule,

$$P(D+ \mid T_i) = \frac{P(T_i \mid D+)P(D+)}{P(T_i \mid D+)P(D+) + P(T_i \mid D-)P(D-)}$$

Thus, if the initial probabilities $P(D+)$ and $P(D-)$ are known, the probability that a patient has coronary artery disease can be calculated.

Let us consider two specific cases. For the first, suppose that a male between the ages of 30 and 39 suffers from nonanginal chest pain. For such a patient, it is known from medical statistics that $P(D+) \approx .05$. Suppose that the test shows that no arteries are calcified. From the preceding equation,

$$P(D+ \mid T_0) = \frac{.42 \times .05}{.42 \times .05 + .96 \times .95} = .02$$

It is unlikely that the patient has coronary artery disease. On the other hand, suppose that the test shows that one artery is calcified. Then

$$P(D+ \mid T_1) = \frac{.24 \times .05}{.24 \times .05 + .02 \times .95} = .39$$

Now it is more likely that this patient has coronary artery disease, but by no means certain.

As a second case, suppose that the patient is a male between ages 50 and 59 who suffers typical angina. For such a patient, $P(D+) = .92$. For him, we find that

$$P(D+ \mid T_0) = \frac{.42 \times .92}{.42 \times .92 + .96 \times .08} = .83$$

$$P(D+ \mid T_1) = \frac{.24 \times .92}{.24 \times .92 + .02 \times .08} = .99$$

Comparing the two patients, we see the strong influence of the prior probability, $P(D+)$. ■

EXAMPLE **F** Polygraph tests (lie-detector tests) are often routinely administered to employees or prospective employees in sensitive positions. Let $+$ denote the event that the polygraph reading is positive, indicating that the subject is lying; let T denote the event that the subject is telling the truth; and let L denote the event that the subject is lying. According to studies of polygraph reliability (Gastwirth 1987),

$$P(+ \mid L) = .88$$

from which it follows that $P(- \mid L) = .12$ also

$$P(- \mid T) = .86$$

from which it follows that $P(+ \mid T) = .14$. In words, if a person is lying, the probability that this is detected by the polygraph is .88, whereas if he is telling the truth, the polygraph indicates that he is telling the truth with probability .86. Now suppose that polygraphs are routinely administered to screen employees for security reasons, and that on a particular question the vast majority of subjects have no reason to lie so

that $P(T) = .99$, whereas $P(L) = .01$. A subject produces a positive response on the polygraph. What is the probability that the polygraph is incorrect and that she is in fact telling the truth? We can evaluate this probability with Bayes' rule:

$$\begin{aligned} P(T \mid +) &= \frac{P(+ \mid T)P(T)}{P(+ \mid T)P(T) + P(+ \mid L)P(L)} \\ &= \frac{(.14)(.99)}{(.14)(.99) + (.88)(.01)} \\ &= .94 \end{aligned}$$

Thus, in screening this population of largely innocent people, 94% of the positive polygraph readings will be in error. Most of those placed under suspicion because of the polygraph result will, in fact, be innocent. This example illustrates some of the dangers in using screening procedures on large populations. ■

Bayes' rule is the fundamental mathematical ingredient of a subjective, or "Bayesian," approach to epistemology, theories of evidence, and theories of learning. According to this point of view, an individual's beliefs about the world can be coded in probabilities. For example, an individual's belief that it will hail tomorrow can be represented by a probability $P(H)$. This probability varies from individual to individual. In principle, each individual's probability can be ascertained, or elicited, by offering him or her a series of bets at different odds.

According to Bayesian theory, our beliefs are modified as we are confronted with evidence. If, initially, my probability for a hypothesis is $P(H)$, after seeing evidence E (e.g., a weather forecast), my probability becomes $P(H|E)$. $P(E|H)$ is often easier to evaluate than $P(H|E)$. In this case, the application of Bayes' rule gives

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})}$$

where $\bar{H}$ is the event that H does not hold. This point can be illustrated by the preceding polygraph example. Suppose an investigator is questioning a particular suspect and that the investigator's prior opinion that the suspect is telling the truth is $P(T)$. Then, upon observing a positive polygraph reading, his opinion becomes $P(T|+)$. Note that different investigators will have different prior probabilities $P(T)$ for different suspects, and thus different posterior probabilities.

As appealing as this formulation might be, a long line of research has demonstrated that humans are actually not very good at doing probability calculations in evaluating evidence. For example, Tversky and Kahneman (1974) presented subjects with the following question: "If Linda is a 31-year-old single woman who is outspoken on social issues such as disarmament and equal rights, which of the following statements is more likely to be true?

- Linda is bank teller.
- Linda is a bank teller and active in the feminist movement."

More than 80% of those questioned chose the second statement, despite Property C of Section 1.3.

Even highly trained professionals are not good at doing probability calculations, as illustrated by the following example of Eddy (1982), regarding interpreting the results from mammogram screening. One hundred physicians were presented with the following information:

- In the absence of any special information, the probability that a woman (of the age and health status of this patient) has breast cancer is 1%.
- If the patient has breast cancer, the probability that the radiologist will correctly diagnose it is 80%.
- If the patient has a benign lesion (no breast cancer), the probability that the radiologist will incorrectly diagnose it as cancer is 10%.

They were then asked, "What is the probability that a patient with a positive mammogram actually has breast cancer?"

Ninety-five of the 100 physicians estimated the probability to be about 75%. The correct probability, as given by Bayes' rule, is 7.5%. (You should check this.) So even experts radically overestimate the strength of the evidence provided by a positive outcome on the screening test.

Thus the Bayesian probability calculus does not describe the way people actually assimilate evidence. Advocates for Bayesian learning theory might assert that the theory describes the way people "should think." A softer point of view is that Bayesian learning theory is a model for learning, and it has the merit of being a simple model that can be programmed on computers. Probability theory in general, and Bayesian learning theory in particular, are part of the core of artificial intelligence.

1.6 Independence

Intuitively, we would say that two events, A and B, are independent if knowing that one had occurred gave us no information about whether the other had occurred; that is, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$. Now, if

$$P(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

then

$$P(A \cap B) = P(A)P(B)$$

We will use this last relation as the definition of independence. Note that it is symmetric in A and in B, and does not require the existence of a conditional probability, that is, $P(B)$ can be 0.

DEFINITION

A and B are said to be independent events if $P(A \cap B) = P(A)P(B)$. ■

EXAMPLE **A** A card is selected randomly from a deck. Let A denote the event that it is an ace and D the event that it is a diamond. Knowing that the card is an ace gives no

information about its suit. Checking formally that the events are independent, we have $P(A) = \frac{4}{52} = \frac{1}{13}$ and $P(D) = \frac{1}{4}$. Also, $A \cap D$ is the event that the card is the ace of diamonds and $P(A \cap D) = \frac{1}{52}$. Since $P(A)P(D) = (\frac{1}{4}) \times (\frac{1}{13}) = \frac{1}{52}$, the events are in fact independent. ■

EXAMPLE **B** A system is designed so that it fails only if a unit and a backup unit both fail. Assuming that these failures are independent and that each unit fails with probability p, the system fails with probability p^2. If, for example, the probability that any unit fails during a given year is .1, then the probability that the system fails is .01, which represents a considerable improvement in reliability. ■

Things become more complicated when we consider more than two events. For example, suppose we know that events A, B, and C are **pairwise independent** (any two are independent). We would like to be able to say that they are all independent based on the assumption that knowing something about two of the events does not tell us anything about the third, for example, $P(C \mid A \cap B) = P(C)$. But as the following example shows, pairwise independence does not guarantee mutual independence.

EXAMPLE **C** A fair coin is tossed twice. Let A denote the event of heads on the first toss, B the event of heads on the second toss, and C the event that exactly one head is thrown. A and B are clearly independent, and $P(A) = P(B) = P(C) = .5$. To see that A and C are independent, we observe that $P(C \mid A) = .5$. But

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$$

■

To encompass situations such as that in Example C, we define a collection of events, $A_1, A_2, \ldots, A_n$, to be **mutually independent** if for any subcollection, $A_{i_1}, \ldots, A_{i_m}$,

$$P(A_{i_1} \cap \cdots \cap A_{i_m}) = P(A_{i_1}) \cdots P(A_{i_m})$$

EXAMPLE **D** We return to Example B of Section 1.3 (infectivity of AIDS). Suppose that virus transmissions in 500 acts of intercourse are mutually independent events and that the probability of transmission in any one act is 1/500. Under this model, what is the probability of infection? It is easier to first find the probability of the complement of this event. Let $C_1, C_2, \ldots, C_{500}$ denote the events that virus transmission does not occur during encounters 1, 2, ..., 500. Then the probability of no infection is

$$P(C_1 \cap C_2 \cap \cdots \cap C_{500}) = \left(1 - \frac{1}{500}\right)^{500} = .37$$

so the probability of infection is $1 - .37 = .63$, not 1, which is the answer produced by incorrectly adding probabilities. ■

EXAMPLE **E** Consider a circuit with three relays (Figure 1.4). Let A_i denote the event that the ith relay works, and assume that $P(A_i) = p$ and that the relays are mutually independent. If F denotes the event that current flows through the circuit, then $F = A_3 \cup (A_1 \cap A_2)$ and, from the addition law and the assumption of independence,

$$P(F) = P(A_3) + P(A_1 \cap A_2) - P(A_1 \cap A_2 \cap A_3) = p + p^2 - p^3 \quad \blacksquare$$

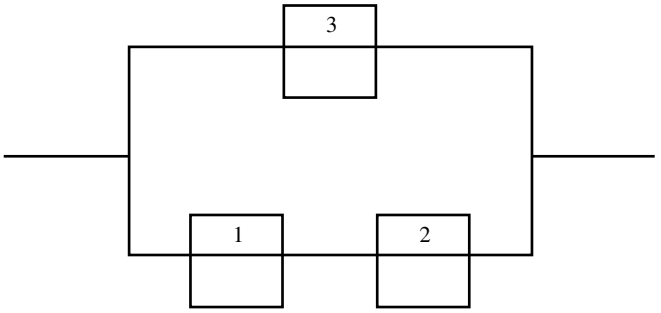

FIGURE **1.4** Circuit with three relays.

EXAMPLE **F** Suppose that a system consists of components connected in a series, so the system fails if any one component fails. If there are n mutually independent components and each fails with probability p, what is the probability that the system will fail?

It is easier to find the probability of the complement of this event; the system works if and only if all the components work, and this situation has probability $(1 - p)^n$. The probability that the system fails is then $1 - (1 - p)^n$. For example, if $n = 10$ and $p = .05$, the probability that the system works is only $.95^{10} = .60$, and the probability that the system fails is .40.

Suppose, instead, that the components are connected in parallel, so the system fails only when all components fail. In this case, the probability that the system fails is only $.05^{10} = 9.8 \times 10^{-14}$. $\blacksquare$

Calculations like those in Example F are made in reliability studies for systems consisting of quite complicated networks of components. The absolutely crucial assumption is that the components are independent of one another. Theoretical studies of the reliability of nuclear power plants have been criticized on the grounds that they incorrectly assume independence of the components.

EXAMPLE **G** *Matching DNA Fragments*

Fragments of DNA are often compared for similarity, for example, across species. A simple way to make a comparison is to count the number of locations, or sites, at which these fragments agree. For example, consider these two sequences, which agree at three sites: fragment 1: AGATCAGT; and fragment 2: TGGATACT.

Many such comparisons are made, and to sort the wheat from the chaff, a probability model is often used. A comparison is deemed interesting if the number of

matches is much larger than would be expected by chance alone. This requires a chance model; a simple one stipulates that the nucleotide at each site of fragment 1 occurs randomly with probabilities p_{A1}, p_{G1}, p_{C1}, p_{T1}, and that the second fragment is similarly composed with probabilities $p_{A2}, \ldots, p_{T2}$. What is the chance that the fragments match at a particular site if in fact the identity of the nucleotide on fragment 1 is independent of that on fragment 2? The match probability can be calculated using the law of total probability:

$$\begin{aligned} P(\text{match}) &= P(\text{match}|\text{A on fragment 1})P(\text{A on fragment 1}) + \\ &\quad \ldots + P(\text{match}|\text{T on fragment 1})P(\text{T on fragment 1}) \\ &= p_{A2}p_{A1} + p_{G2}p_{G1} + p_{C2}p_{C1} + p_{T2}p_{T1} \end{aligned}$$

The problem of determining the probability that they match at k out of a total of n sites is discussed later. ■

1.7 Concluding Remarks

This chapter provides a simple axiomatic development of the mathematical theory of probability. Some subtle issues that arise in a careful analysis of infinite sample spaces have been neglected. Such issues are typically addressed in graduate-level courses in measure theory and probability theory. Certain philosophical questions have also been avoided. One might ask what is meant by the statement "The probability that this coin will land heads up is $\frac{1}{2}$." Two commonly advocated views are the **frequentist approach** and the **Bayesian approach.** According to the frequentist approach, the statement means that if the experiment were repeated many times, the long-run average number of heads would tend to $\frac{1}{2}$. According to the Bayesian approach, the statement is a quantification of the speaker's uncertainty about the outcome of the experiment and thus is a personal or subjective notion; the probability that the coin will land heads up may be different for different speakers, depending on their experience and knowledge of the situation. There has been vigorous and occasionally acrimonious debate among proponents of various versions of these points of view.

In this and ensuing chapters, there are many examples of the use of probability as a model for various phenomena. In any such modeling endeavor, an idealized mathematical theory is hoped to provide an adequate match to characteristics of the phenomenon under study. The standard of adequacy is relative to the field of study and the modeler's goals.

1.8 Problems

1. A coin is tossed three times and the sequence of heads and tails is recorded.

a. List the sample space.

b. List the elements that make up the following events: (1) A = at least two heads, (2) B = the first two tosses are heads, (3) C = the last toss is a tail.

c. List the elements of the following events: (1) A^c, (2) $A \cap B$, (3) $A \cup C$.