conditions. Such a definition allows us to include only specific subsets of Ω, which is especially important for continuous Ω.

Definition: Field $\mathcal{F}$ A *field* is the nonempty set of events of an experiment that satisfy the following conditions:

(i) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

(ii) If $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots, N$, where N is finite, then $\bigcup_{n=1}^{N} A_n \in \mathcal{F}$.

A field is also called an *algebra* of events. Observe that these conditions imply the following properties:

(iii) $\Omega \in \mathcal{F}$.

(iv) $\phi \in \mathcal{F}$.

(v) If $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots, N$, where N is finite, then $\bigcap_{n=1}^{N} A_n \in \mathcal{F}$.

Property (v), which is obtained using one of De Morgan's laws, is sometimes included as a condition in the definition of a field, but it need not be mentioned explicitly because (i) and (ii) are sufficient.

Example 2.19. The simplest (but trivial) field is $\mathcal{F} = \{\phi, \Omega\}$.

Example 2.20. For the experiment of a single coin toss with $\Omega = \{H, T\}$, consider the nontrivial field $\mathcal{F} = \{\phi, \Omega, H, T\}$. The properties above are readily verified: (i) $H \in \mathcal{F} \Longrightarrow T \in \mathcal{F}$ (and vice versa), (ii)–(iii) $H \in \mathcal{F}, T \in \mathcal{F} \Longrightarrow H \cup T \equiv \Omega \in \mathcal{F}$, and (iv)–(v) $H \in \mathcal{F}, T \in \mathcal{F} \Longrightarrow H \cap T \equiv \phi \in \mathcal{F}$. Note that $\mathcal{F}$ is, in fact, the power set $\mathcal{P}(\Omega)$ for this experiment, and no other fields are possible, except the trivial field in Example 2.19.

In some experiments, we might first consider a set of events that do not quite comprise a field, and then include additional elements to *generate* a field. This procedure is illustrated in Example 2.21.

Example 2.21. Suppose that $\Omega = [1, \infty)$ and consider $\mathcal{F} = \{\phi, \Omega, \{[1, a)\}\}$ for every $a > 1$, that is, $\mathcal{F}$ contains all semi-open intervals of the form $[1, a)$. Obviously, $\mathcal{F}$ is not a field because $[1, a)^c = [a, \infty)$ for any $a > 1$ is not in $\mathcal{F}$. Expand the set to be $\mathcal{F} = \{\phi, \Omega, \{[1, a)\}, \{[b, \infty)\}\}$ with $b > 1$. Again, this is not a field because $[1, a) \cap [b, \infty) = [b, a)$ for any $b < a$ is not in $\mathcal{F}$. Expand the set further to be $\mathcal{F} = \{\phi, \Omega, \{[1, a)\}, \{[b, \infty)\}, \{[c, d)\}\}$ for all $a, b, c, d > 1$. We now have a field: all *finite* unions of the elements are events in $\mathcal{F}$, as are all complements of the elements. Later, we demonstrate that $\mathcal{F}$ is *not* a σ-field which is defined for an *infinity* of unions, and is described after the next set of examples.

Example 2.22. For the experiment of tossing two coins, recall that $\Omega = \{HH, TT, HT, TH\}$. The elements of the power set $\mathcal{P}(\Omega)$ are summarized in Table 2.3. Properties (i) and (ii) are easily verified for this case. From this example, it is useful to summarize some characteristics of an event. The four outcomes in the experiment

TABLE 2.3 Elements of Field $\mathcal{F} = \mathcal{P}(\Omega)$ for Example 2.22

Elementary events (the four outcomes):	$\{HH\}, \{TT\}, \{HT\}, \{TH\}$
Events defined by two outcomes:	$\{HH, TT\}, \{HH, HT\}, \{HH, TH\}, \{TT, HT\}, \{TT, TH\}, \{HT, TH\}$
Events defined by three outcomes:	$\{HH, TT, HT\}, \{HH, TT, TH\}, \{HH, HT, TH\}, \{TT, HT, TH\}$
ϕ and Ω:	$\phi = \Omega^c, \Omega = \{HH, TT, HT, TH\}$

are given by the terms in the first row of Table 2.3. These are the elementary events; all other entries in the table are not elementary events. For example, $\{HH, HT\}$ is the event that the first coin is H; $\{HH, TT\}$ is the event that the two coins have the same outcome. These two events are *not* outcomes themselves; instead, they refer to those outcomes that share a common feature. Although event $\{HH, HT, TH\}$ corresponds to those outcomes such that there is at least one H, events with three outcomes together can be more complicated to describe. Event $\{HH, TT, HT\}$ is not so easy to state; it corresponds to the situation where the coins have the same outcome *or* the first coin is H and the second coin is T. Alternatively for this event, we can say that it refers to those outcomes for which the first coin is not T *and* the second coin is not H (thus excluding TH). Since we require all unions of events in $\mathcal{F}$ to be in the field, Ω must be in $\mathcal{F}$. Likewise, $\phi \in \mathcal{F}$ so that complement and intersection are consistent for any event in $\mathcal{F}$. Observe that there are 16 elements in $\mathcal{F}$, as expected because $N = 4$; the power set for this case has cardinality $|\mathcal{P}(\Omega)| = 2^4$.

The power set $\mathcal{P}(\Omega)$ by definition is the largest field possible. Fields are not unique, as mentioned previously and as shown in Example 2.23. Generally, smaller fields are not useful for experiments with finite and even countably infinite outcomes. However, for uncountable experiments, it is necessary that we choose a σ-field that is smaller than the power set.

Example 2.23. From Example 2.19, we know that $\{\phi, \Omega\}$ is a field; it obviously satisfies properties (i)-(ii). Of course, this trivial field contains no useful information about any particular experiment. Consider the experiment of tossing a single die with outcomes $\{1, 2, 3, 4, 5, 6\}$. The power set $\mathcal{P}(\Omega)$ consists of $2^6 = 64$ elements. However, it is possible to define a smaller (nontrivial) field, though as we shall see it is not as useful as the power set. Consider the following two events: $E = \{1, 3, 5\}$ and $F = \{2, 4, 6\}$, and observe that $\mathcal{F} = \{\phi, \Omega, E, F\}$ is a field. Properties (i) and (ii) are satisfied; in particular, $E = F^c \in \mathcal{F}$ and $E \cup F = \Omega \in \mathcal{F}$. Although this is a valid field, it results in a simplification of the experiment such that we are concerned only about the outcome being odd or even; the specific number showing on the die is of no interest. In fact, by defining such a field, this "simplified" experiment is structurally identical to a single toss of a coin where, for example, $H \equiv$ even number and $T \equiv$ odd number showing on the die. Tossing a single die and using the smaller field above *simulates* tossing a single coin.

Consider again the situation where we start with some subset of Ω and then generate a field that includes the subset.

Example 2.24. For the experiment of tossing a single die, we would like to generate a field starting with the subset $\{1, 2\}$. Observe that $\{1, 2\}^c = \{3, 4, 5, 6\}$, $\{1, 2\} \cap \{3, 4, 5, 6\} = \phi$, and $\{1, 2\} \cup \{3, 4, 5, 6\} = \Omega$. Thus the field *generated* by this subset is $\mathcal{F} = \{\phi, \Omega, \{1, 2\}, \{3, 4, 5, 6\}\}$. Obviously, this field fails to capture all outcomes and events of the original experiment; in fact, it is equivalent to a binary experiment such as the toss of a single coin (as in Example 2.23). Consider starting with the subset $\{\{1, 2\}, \{3, 4\}\}$. Then it is easy to show that the generated field is $\mathcal{F} = \{\phi, \Omega, \{1, 2\}, \{3, 4\}, \{5, 6\}, \{3, 4, 5, 6\}, \{1, 2, 5, 6\}, \{1, 2, 3, 4\}\}$. In order to capture all details of the single-die experiment, we would need to choose the power set for the field with $2^6 = 64$ elements.

These last examples illustrate the difference between a "fine" field (the power set) and a "coarse" field (given by some subset of the power set). Later when conditional probability is covered, we use this distinction between fine and coarse fields to derive a useful property in probability theory when conditioning on events.

Next, we describe a property of a field $\mathcal{F}$ when Ω is finite, based on the definition of atoms.

Definition: Atoms of a Field The smallest events in $\mathcal{F}$ (excluding ϕ) are *atoms*. We denote the set of atoms by $\mathcal{A}$.

This definition means that an atom cannot be obtained from the union of other events in $\mathcal{F}$, and they are disjoint. Every other event in $\mathcal{F}$ is obtained by set operations on the elements of $\mathcal{A}$ (in order to satisfy the requirements

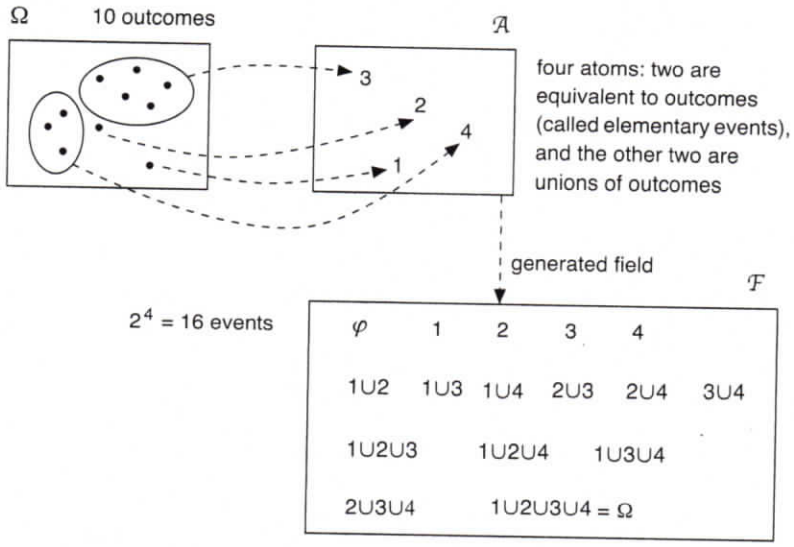

FIGURE 2.4 Example outcomes in Ω, atoms in $\mathcal{A}$, and events in $\mathcal{F}$.

of a field). A simple finite example showing the connection between outcomes in Ω, atoms in $\mathcal{A}$, and events in $\mathcal{F}$ is illustrated in Figure 2.4.

Example 2.25. For the second field in Example 2.24, the three atoms are $\mathcal{A} = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$, which are not outcomes in Ω.

It is clear that the atoms of the power set for an experiment are the elementary events (outcomes). If individual outcomes are excluded from the field, then the atoms are nonelementary events as shown in Example 2.24.

Theorem 2.1. The number of elements in any finite field $\mathcal{F}$ is given by 2^n, where $n = |\mathcal{A}|$.

Proof. From the definition of an atom and the requirements of a field, the elements of $\mathcal{F}$ must be the power set $\mathcal{P}(\mathcal{A})$ of the atoms. We have already shown that the cardinality of the power set is 2^N for a set of size N, and thus $N = n$ completes the proof. □

This theorem is useful because we can immediately determine that $\mathcal{F}$ is *not* a field if the number of elements is not 2, 4, 8, 16, and so on. Note that Ω need not be discrete in order for the field to be finite, as demonstrated in Example 2.26.

Example 2.26. Let $\Omega = [0, \infty) = \mathcal{R}^+$ and $\mathcal{F} = \{\phi, \Omega, [0, c), [c, \infty)\}$ for some $c > 0$. Thus, although the sample space is continuous, the field has only four elements with atoms given by two subintervals: $[0, c)$ and $[c, \infty)$. Suppose instead that the atoms are $[0, c)$, $[c, d)$, and $[d, \infty)$ for some $d > c > 0$. From the theorem above, we know for this continuous sample space that the field has eight elements: $\mathcal{F} = \{\phi, \Omega, [0, c), [c, d), [d, \infty), [0, d), [c, \infty), [0, c) \cup [d, \infty)\}$. Such intervals on $\mathcal{R}$ are important for continuous sample spaces, as discussed later when we introduce Borel sets.

For problems involving a finite sample space, the definition of a field with conditions (i) and (ii) is adequate for us to specify a probability space. However, if the sample space is infinite, either countably infinite or

uncountable, then it is necessary that the definition of a field be extended to include an infinite (but countable) union of events.

Definition: Sigma Field $\mathcal{F}$ A *sigma-field* (σ-*field*) is a field satisfying condition (i), with (ii) extended to a countably infinite number of subsets as follows:

(ii′) If $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

A σ-field is also called a σ-*algebra*.

Of course, this modified definition causes (v) to be extended as follows:

(v′) If $A_n \in \mathcal{F}$ for $n = 1, 2, \ldots$, then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$.

Example 2.27. From Example 2.21, $\mathcal{F} = \{\phi, \Omega, \{[1, a)\}, \{[b, \infty)\}, \{[c, d)\}\}$ for every $a, b, c, d > 1$ is a field for $\Omega = [1, \infty)$, but it is not a σ-field. Suppose that $[c, d)$ is written as $[c + 1/n, d)$, where n is a positive integer. Then we know from the previous discussion on the different types of intervals and Table 2.2 that

$$\bigcup_{n=1}^{\infty} [c + 1/n, d) = (c, d), \tag{2.18}$$

which is a type of interval not in $\mathcal{F}$.

Next, we consider the cardinality of infinite sets.

Definition: Beth Numbers The *beth numbers* $\{\beth_n\}$ (Forster, 1995) are used to represent the cardinality of different types of infinite sets.

In mathematics, the cardinality of the natural numbers $\mathcal{N}$ is denoted by the symbol $\beth_0$ which is called *beth null*. The beth numbers are generated by the following recursion:

$$\beth_{n+1} = 2^{\beth_n}, \tag{2.19}$$

which we see has a form similar to that in Proposition 2.1 for the cardinality of $\mathcal{P}$ for a finite set. It turns out that the cardinality of $\mathcal{R}$ is beth one $\beth_1 = 2^{\beth_0}$ which is the size of the power set for $\mathcal{N}$. Similarly, the cardinality of $\mathcal{P}(\mathcal{R})$ is beth two $\beth_2$. We are not concerned with the theory behind the cardinality of infinite sets and the nature of the different types of infinity; instead, we use beth numbers as a convenient notation to represent infinite sets that have the same cardinality. Several examples are provided in Table 2.4, some of which are defined later.

For $\Omega = \mathcal{N}$, the power set $\mathcal{P}(\mathcal{N})$ is a σ-field. In fact, the power set for *any* Ω is a σ-field, and it is clear from the previous discussions that σ-field $\mathcal{F}$ for any Ω satisfies $\{\Omega, \phi\} \subseteq \mathcal{F} \subseteq \mathcal{P}(\Omega)$.

Definition: Event Space An *event space* is the set of events in Ω that comprise the σ-field $\mathcal{F}$. The event space is written as the double $\{\Omega, \mathcal{F}\}$.

TABLE 2.4 Beth Numbers and Cardinality of Infinite Sets

Beth Number	Infinite Sets with the Same Cardinality
$\beth_0$	$\mathcal{N}$, $\mathcal{Q}$, $\mathcal{Z}$
$\beth_1$	$\mathcal{R}$, $\mathcal{R} - \mathcal{Q}$, $\mathcal{P}(\mathcal{N})$, $\mathcal{B}(\mathcal{R})$, $\mathcal{C}$, $[a, b] \subset \mathcal{R}$, Cantor set
$\beth_2$	$\mathcal{P}(\mathcal{R})$, $\mathcal{P}(\mathcal{P}(\mathcal{N}))$

The event space is designed to include events in Ω that satisfy the conditions of the specific σ-field $\mathcal{F}$. We extend the idea of generating a field described in Examples 2.21 and 2.24 to a σ-field.

Definition: Generated σ-Field For subset $\mathcal{G} \subseteq \Omega$, define the following set of σ-fields:

$$E = \{\mathcal{F} \subseteq \mathcal{P}(\Omega) : \mathcal{G} \subseteq \mathcal{F}\}. \tag{2.20}$$

The σ-field *generated* by subset $\mathcal{G}$ and denoted by $\sigma(\mathcal{G})$ is obtained as the intersection of all elements in E.

Since $\mathcal{G}$ is contained in every σ-field for Ω, $\sigma(\mathcal{G})$ is necessarily the *smallest* σ-field containing $\mathcal{G}$. If $\mathcal{G}$ itself is a σ-field, then $\sigma(\mathcal{G}) = \mathcal{G}$.

Example 2.28. Consider the finite sample space $\Omega = \{-1, 0, 1\}$. The following are all the possible σ-fields:

$$\mathcal{F}_1 = \{\phi, \Omega, -1, 0, 1, \{-1, 0\}, \{-1, 1\}, \{0, 1\}\}, \tag{2.21}$$

$$\mathcal{F}_2 = \{\phi, \Omega, -1, \{0, 1\}\}, \tag{2.22}$$

$$\mathcal{F}_3 = \{\phi, \Omega, 0, \{-1, 1\}\}, \tag{2.23}$$

$$\mathcal{F}_4 = \{\phi, \Omega, 1, \{-1, 0\}\}, \tag{2.24}$$

$$\mathcal{F}_5 = \{\phi, \Omega\}. \tag{2.25}$$

If $\mathcal{G} = \Omega$, then $\sigma(\mathcal{G}) = \mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_3 \cap \mathcal{F}_4 \cap \mathcal{F}_5 = \mathcal{F}_5$ which is the trivial σ-field. If $\mathcal{G} = \{1\}$, then $\sigma(\mathcal{G}) = \mathcal{F}_1 \cap \mathcal{F}_4 = \mathcal{F}_4$. If $\mathcal{G} = \{0, 1\}$, then $\sigma(\mathcal{G}) = \mathcal{F}_1$, which is the power set $\mathcal{P}(\Omega)$.

We are interested in σ-fields that are *measurable* so that probabilities can be assigned in a consistent manner to the event space $\{\Omega, \mathcal{F}\}$. For finite and countably infinite experiments, there is generally no need to consider σ-fields smaller than the power set $\mathcal{P}(\Omega)$. However, for continuous sample spaces, the power set is "too large." The Vitali set described at the end of this chapter for $\Omega = [0, 1]$ is an example of a subset that is not measurable. It will be convenient from a practical sense, as well as for most applications, to consider *intervals* for continuous experiments. Also, since sample spaces are mapped to the real line $\mathcal{R}$ for the random variables in Chapter 3, we are interested in a specific type of σ-field for intervals on $\mathcal{R}$ called the Borel σ-field.

Definition: Borel σ-Field The *Borel σ-field* $\mathcal{B}(\mathcal{R})$ on the real line is the smallest σ-field generated by all open intervals in $\Omega = \mathcal{R}$. Elements of $\mathcal{B}(\mathcal{R})$ are called *Borel sets*.

The Borel σ-field is generated by starting with all open intervals in $\mathcal{R}$ of the form (a, b). Individual points (singletons) are included by observing that

$$a = \lim_{n \to \infty} (a - 1/n, a + 1/n). \tag{2.26}$$

This result allows us to include all semi-open intervals:

$$(a, b] = (a, b) \cup b, \qquad [a, b) = (a, b) \cup a. \tag{2.27}$$

Since $(-\infty, b] = (b, \infty)^c$ and $[a, \infty) = (-\infty, a)^c$, all closed intervals are also included:

$$[a, b] = [a, \infty) \cap (-\infty, b]. \tag{2.28}$$

Thus, $\mathcal{B}(\mathcal{R})$ consists of all types of intervals on $\mathcal{R}$ as well as all individual points. It is not easy to visualize subsets of $\mathcal{R}$ that lie outside of $\mathcal{B}(\mathcal{R})$; we present an example of such a subset when measure theory is discussed

later. However, it turns out that these subsets do not arise in practical applications and so they are not of concern. Since we will also consider σ-fields in $\Omega = \mathcal{R}^N$ when *random vectors* are covered in Chapter 4, the above definition extends to the smallest σ-field of subsets of $\mathcal{R}^N$ denoted by $\mathcal{B}(\mathcal{R}^N)$, which is defined to be open *rectangles* on $\mathcal{R}^2$ and open hyper-rectangles on $\mathcal{R}^N$ (for $N > 2$).

Example 2.29. Examples of *Borel sets* include the following: (i) closed interval $[a, b]$, (ii) open interval (a, b), (iii) set of rational numbers $\mathcal{Q}$, (iv) set of irrational numbers $\mathcal{R} - \mathcal{Q}$, (v) $\{1, 2, 3, 4, 5, 6\}$, and (vi) the Cantor set described later.

2.5 SUMMARY OF A RANDOM EXPERIMENT

It will be useful at this point to summarize our description of a random experiment for which we want to assign a probability measure:

- The *sample space* Ω is the collection of all outcomes of an experiment. Depending on the type of experiment, it may contain a finite, countably infinite, or uncountable number of elements.
- An *event* of an experiment is a subset of Ω consisting of one or more outcomes, and which usually share some feature. Since the goal is to assign a probability measure to events that is consistent, we may not want to include all possible subsets of Ω.
- A *σ-field* $\mathcal{F}$ is a collection of subsets of Ω that satisfy the algebraic conditions in (i) and (ii′), as well as (iii), (iv), and (v′). One can envision constructing a σ-field by starting with some subset of Ω and expanding the number of subsets until the collection satisfies (i) and (ii′).
- If Ω is finite or countably infinite, then all subsets described by the power set $\mathcal{P}(\Omega)$ comprise a useful σ-field $\mathcal{F}$ for the experiment.
- If Ω is uncountable, then the Borel σ-field $\mathcal{B}(\mathcal{R})$ is a useful σ-field for the real line that consists of all intervals (open, semi-open, and closed) as well as individual points.

With these definitions of the sample space Ω and the event space $\{\Omega, \mathcal{F}\}$, we have a collection of events that satisfy the additivity property in (ii′). It is now possible to formulate a *probability space* which we denote by the triple $\{\Omega, \mathcal{F}, P\}$. We begin with a brief discussion of measure theory, of which the probability measure is a particular case.

2.6 MEASURE THEORY

Measure theory is concerned with assigning a "size" called a *measure* to subsets of a sample space Ω. Depending on the problem, this size might be a length, an area, or a volume in Euclidean space; this is known as the Lebesgue measure which is defined below. In our case, we are interested in assigning a *probability measure* to subsets of Ω that comprise the σ-field $\mathcal{F}$. For uncountable experiments, it is not possible to assign probabilities in a consistent manner to $\mathcal{P}(\mathcal{R})$. The power set of $\Omega = \mathcal{R}$ is too large (it has cardinality beth two $\beth_2$), so instead we choose $\mathcal{F}$ to be $\mathcal{B}(\mathcal{R})$, which is the Borel σ-field on the real line and has cardinality beth one $\beth_1$ (the same as $\mathcal{R}$). The Borel σ-field is useful for essentially all practical applications because we are generally interested in events described by *intervals* on the real line.

Let μ be the notation for a measure that we want to define for the event space, and denote the *measure space* by the triple $\{\Omega, \mathcal{F}, \mu\}$. (Note that μ should not be confused with the mean of a random variable covered later in Chapter 5.)

Definition: Measure A *measure* is a mapping of events $\{E_n\} \in \mathcal{F}$ to $\mathcal{R}$ that has the following properties:

- $\mu(E_n) \geq 0$ for all events $E_n \in \mathcal{F}$.
- $\mu(\phi) = 0$.

- For all events $\{E_n\}$ such that $E_n \cap E_m = \phi$ for $n \neq m$ (mutually exclusive):

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n). \tag{2.29}$$

Similar properties are shown later for the probability measure where they are known as the *three axioms of probability*.

Example 2.30. The *trivial measure* is $\mu(E_n) = 0$ for all $E_n \in \mathcal{F}$ in the event space $\{\Omega, \mathcal{F}\}$.

Example 2.31. For $\Omega = \mathcal{R}$ and $\mathcal{F} = \mathcal{B}(\mathcal{R})$, the *standard Gaussian measure* is

$$\mu(E) = \frac{1}{\sqrt{2\pi}} \int_E \exp(-x^2/2)\mathrm{d}x, \tag{2.30}$$

which turns out to be the probability $P(E)$ of event E for the standard Gaussian random variable X with zero mean and unit variance.

Example 2.32. The *Dirac measure* for event E in the event space $\{\Omega, \mathcal{F}\}$ is given by

$$\delta_\omega(E) \triangleq \begin{cases} 1, & \omega \in E \\ 0, & \text{else.} \end{cases} \tag{2.31}$$

It can be shown that $\delta_\omega(E)$ is a probability measure, and thus E is the almost sure event in Ω (see Problem 2.23).

Consider two important measures that will be useful later.

Definition: Counting Measure Let Ω be a countable sample space and $\mathcal{P}(\Omega)$ its power set. The *counting measure* for any set $E \in \mathcal{P}(\Omega)$ is

$$C(E) = \begin{cases} |E|, & E \text{ is a finite subset} \\ \infty, & \text{otherwise,} \end{cases} \tag{2.32}$$

where $|E|$ is the cardinality of E. The infinity in this definition is actually beth null $\beth_0$ mentioned above, the size of a countably infinite set.

Obviously, this measure is useful only for finite sets; it certainly provides no information about the size of uncountable events, such as intervals on $\mathcal{R}$. For this latter case, we consider the Lebesgue measure.

Definition: Lebesgue Measure on $\mathcal{R}$ The *Lebesgue measure* of interval $[a, b] \in \mathcal{B}(\mathcal{R})$ is

$$L([a, b]) \triangleq b - a, \tag{2.33}$$

which is the *length* of the interval. Since the Lebesgue measure of a single point (singleton) is defined to be zero, we also have $L((a, b)) = L((a, b]) = L([a, b)) = b - a$.

In order to extend the Lebesgue measure to N-dimensional Euclidean space, it will be convenient to define the Cartesian product.

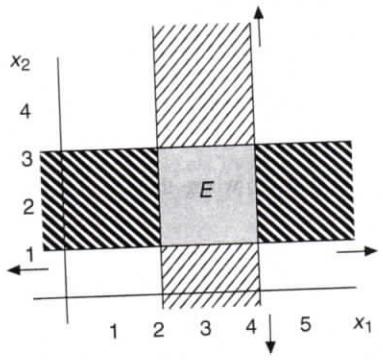

FIGURE 2.5 Cartesian product $E = [2, 4] \times [1, 3]$ for Example 2.33.

Definition: Cartesian Product The *Cartesian product* of a set of intervals $\{[a_n, b_n]\}$, $n = 1, \ldots, N$, is

$$E \triangleq \{(x_1, \ldots, x_N) : x_1 \in [a_1, b_1], \ldots, x_N \in [a_N, b_N]\}. \tag{2.34}$$

E corresponds to an ordered N-tuple, which can be represented using the following product notation:

$$E = [a_1, b_1] \times \cdots \times [a_N, b_N] = \prod_{n=1}^{N} [a_n, b_n]. \tag{2.35}$$

The Cartesian product of intervals on $\mathcal{R}$ is a rectangle on $\mathcal{R}^2$ and a hyper-rectangle on $\mathcal{R}^N$ (for $N > 2$).

Example 2.33. For intervals $[1, 3]$ and $[2, 4]$, the Cartesian product $E = \{(x_1, x_2) : x_1 \in [2, 4], x_2 \in [1, 3]\}$ is a rectangle on $\mathcal{R}^2$ described by all points on the abscissa in the interval $[2, 4]$ *intersected* with all points on the ordinate in the interval $[1, 3]$. This is depicted in Figure 2.5.

For the real line $\mathcal{R}$, we can write $\mathcal{R}^2 = \mathcal{R} \times \mathcal{R}$ (a plane), $\mathcal{R}^3 = \mathcal{R}^2 \times \mathcal{R} = \mathcal{R} \times \mathcal{R} \times \mathcal{R}$ (a hyperplane), and so on. With this geometric viewpoint, we readily see in general that the Lebesgue measure of E corresponds to the area of a rectangle, and in three dimensions it is a *volume*. For the N-dimensions, we have the following definition.

Definition: Lebesgue Measure on $\mathcal{R}^N$ The *Lebesgue measure* of the Cartesian product $E = \prod_{n=1}^{N} [a_n, b_n] \in \mathcal{B}(\mathcal{R}^N)$ is

$$L(E) \triangleq \prod_{n=1}^{N} (b_n - a_n), \tag{2.36}$$

which is the hyper-volume of the corresponding hyper-rectangle on $\mathcal{R}^N$.

Example 2.34. For the Cartesian product in Example 2.33, $L([1, 3] \times [2, 4]) = 4$ is the area of a rectangle, and $L([8, 10] \times [1, 2] \times [4, 8]) = 8$ is the volume of a rectangular cuboid.

The Cantor set described in Example 2.35 is interesting because of the way it is constructed, and because it has Lebesgue measure zero.

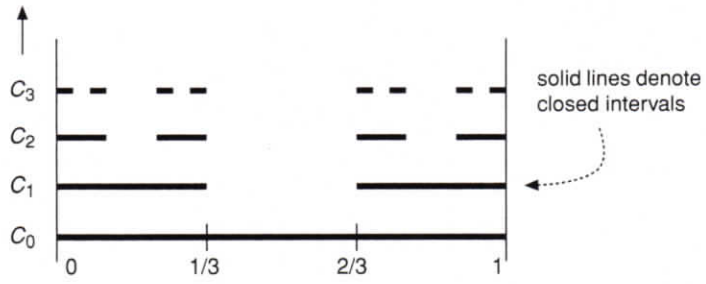

FIGURE 2.6 Sequence of C_m for $m = 1, 2, 3$, leading to the Cantor set C as $m \to \infty$.

Example 2.35 (Cantor set). The Cantor set C (Kingman and Taylor, 1966) on $[0, 1]$ is obtained by removing an infinite set of open intervals as follows:

$$C = [0, 1] - \bigcup_{m=1}^{\infty} \bigcup_{n=0}^{2^{m-1}-1} ((3n + 1)/3^m, (3n + 2)/3^m). \tag{2.37}$$

Let $C_0 = [0, 1]$ and denote successive sets C_m for $m \in \mathcal{N}$ by removing open intervals in the middle third of the remaining closed intervals. Thus

$$C_1 = [0, 1] - (1/3, 2/3) = [0, 1/3] \cup [2/3, 1] \tag{2.38}$$

such that $[0, 1]$ is divided into three regions and the middle open interval is removed. This procedure is performed again for each closed interval remaining in C_1, yielding

$$\begin{aligned} C_2 &= C_1 - (1/9, 2/9) - (7/9, 8/9), \\ &= [1, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1]. \end{aligned} \tag{2.39}$$

Examples of C_m are shown in Figure 2.6. From (2.37), we see that each successive set is obtained from the previous set by scaling the intervals down by a factor of three, and then performing a union of the scaled set with a version of itself shifted by $2/3$, which is described by the general formula

$$C_m = [C_{m-1}/3] \cup [C_{m-1}/3 + 2/3]. \tag{2.40}$$

This process of removing the set of middle-third open intervals is repeated ad infinitum to generate C. Note that instead of subtracting open intervals as in (2.37), the Cantor set can also be defined in terms of intersections as follows:

$$C = \bigcap_{m=0}^{\infty} C_m. \tag{2.41}$$

The Cantor set C has the following interesting properties:

- From (2.37), the length of each open interval removed from $[0, 1]$ is

$$(3n + 2)/3^m - (3n + 1)/3^m = 1/3^m, \tag{2.42}$$

which is the Lebesgue measure of $((3n+1)/3^m, (3n+2)/3)$. For each m, 2^{m-1} open intervals are removed. Thus, by changing variables and recognizing that the sum below is geometric (see Appendix E), the total length removed is

$$\sum_{m=1}^{\infty} 2^{m-1}/3^m = \sum_{m=0}^{\infty} 2^m/3^{m-1} = 3/(1-2/3) = 1, \tag{2.43}$$

which is a surprising result because the length (Lebesgue measure) of $[0, 1]$ is also one. This occurs because at each iteration in the construction of $\{C_m\}$, the end points of the middle interval are not removed, and the fact that the Lebesgue measure of an interval is the same whether or not the end points are included:

- From the preceding result, it follows that C has Lebesgue measure zero.
- C is uncountable with the same cardinality as $[0, 1]$ (which has the same cardinality as $\mathcal{R}$, denoted by beth one $\beth_1$).
- C is a *fractal* because each iteration in its construction (removing the set of middle-third open intervals) results in a reduced-size version C_{m+1} of the previous C_m. This behavior is evident from the first few iterations in Figure 2.6.

The Cantor function based on the Cantor set is described later in Example 3.10.

The Lebesgue measure is useful in Chapter 3 where events in Ω are mapped to the real line $\mathcal{R}$ to form a *random variable*, and to $\mathcal{R}^N$ to form a *random vector*. Since the mapping of an event in Ω to a random variable is well defined (measurable), the probability measure for events in Ω carries over in a consistent manner to events given by intervals on $\mathcal{R}$ (and hyper-rectangles on $\mathcal{R}^N$).

Finally, we define the probability measure for events in the sample space Ω.

Definition: Probability Measure P The *probability measure* P for event space $\{\Omega, \mathcal{F}\}$ is a function that maps events in $\mathcal{F}$ to real numbers in $[0, 1]$, and satisfies the *three axioms of probability*. The resulting *probability space* is denoted by the triple $\{\Omega, \mathcal{F}, P\}$.

It is important to emphasize that P is defined for events in the σ-field $\mathcal{F}$, which may not include all subsets of Ω depending on the experiment. This was, of course, the purpose of constructing $\mathcal{F}$, so that a probability measure could be specified for events that is consistent with the three axioms of probability. In particular, like any measure, the probability measure must satisfy a countable additivity condition in order to "make sense," and which appeals to our intuition about how probabilities should be computed when performing set operations on events in $\mathcal{F}$.

2.7 AXIOMS OF PROBABILITY

The probability measure P of the probability space $\{\Omega, \mathcal{F}, P\}$ satisfies the following axioms:

- *Axiom 1.* $P(E) \geq 0$.
- *Axiom 2.* $P(\Omega) = 1$.
- *Axiom 3.* For $E_n \cap E_m = \phi$ and $n \neq m$, where $\{E_m\} \in \mathcal{F}$:

$$P\left(\bigcup_{m=1}^{\infty} E_m\right) = \sum_{m=1}^{\infty} P(E_m). \tag{2.44}$$

All three axioms are consistent with intuition and the frequency interpretation of probability, as described in Chapter 1. Axiom 1 obviously states that the probability measure is nonnegative. Axiom 2 states that the

probability of some event in $\mathcal{F}$ occurring is 1: "something must happen." From Axioms 1 and 2, the probability of any event in $\mathcal{F}$ must lie in the range [0, 1]; events are assigned probabilities that do not exceed 1. P is a special case of the general measure μ, where $P(\Omega) = 1$ is used instead of $\mu(\phi) = 0$. This difference in Axiom 2 is the reason for $P \in [0, 1]$; all measures μ must be nonnegative but they are *not* necessarily bounded by 1 as is the probability measure.

$P(E) = 0$ does not imply that E cannot occur: there is a difference between an event with probability zero (which can occur) and the *impossible event*, which can never occur. Likewise, $P(E) = 1$ does not imply that E must occur: there is a difference between an event with probability one and the *sure event* which must happen. Events that have probability one are known as *almost sure events*. These differences are illustrated by the following example.

Example 2.36. Consider an experiment where the possible outcomes are *uniformly distributed* on $\Omega = [-1, 1]$. As discussed later for continuous sample spaces, the probability of any point in this interval is necessarily zero, for example, $P(E = 0) = 0$. However, $E = 0$ could happen. Events corresponding to numbers outside of $[-1, 1]$ are impossible events: they not only have probability zero, but they can never occur. The empty set ϕ also can never occur, because some event in $[-1, 1]$ must happen. Similarly, $P(F = [-1, 1)) = 1$ does not mean this event must happen: it is possible (though with probability zero) that event $E = 1$ happens. We are *almost sure* that $F = [-1, 1)$ will occur. On the other hand, event $F = [-1, 1] = \Omega$ is a sure event; some subset of this interval must occur, because no numbers outside this interval are possible.

Axiom 3 is the most important of the axioms in terms of ensuring a consistent probability measure; it specifies how the probabilities of several events are combined. If events are *pairwise* mutually exclusive (disjoint), then the probability of their union is the sum of their probabilities. This includes unions for an infinity of events, which is necessary for countably infinite and uncountable experiments for which a σ-field (not just a field) is required. In the next few sections, we cover some basic results in probability, as well as two important topics: *conditional probability* and *independent events*. Later, we consider probability assignments for discrete and continuous samples spaces, in preparation for the random variables discussed in Chapter 3.

2.8 BASIC PROBABILITY RESULTS

The following propositions follow from the three axioms of probability. They also illustrate how to rearrange a problem statement in terms of disjoint sets for which one can use Axiom 3 to derive the result.

Proposition 2.2. $P(\phi) = 0$.

Proof. Since $\Omega \cup \phi = \Omega$ and $\Omega \cap \phi = \phi$:

$$P(\Omega) = P(\Omega \cup \phi) = P(\Omega) + P(\phi), \tag{2.45}$$

which gives $1 = 1 + P(\phi)$ and $P(\phi) = 0$. □

Proposition 2.3. *For event* $E \in \Omega$, $P(E^c) = 1 - P(E)$.

Proof. Since $\Omega = E \cup E^c$ and $E \cap E^c = \phi$:

$$P(\Omega) = P(E^c + E) = P(E^c) + P(E) = 1, \tag{2.46}$$

which can be rewritten as $P(E^c) = 1 - P(E)$. □

Proposition 2.4. *If* $E \subset F$, *then* $P(E) \leq P(F)$.

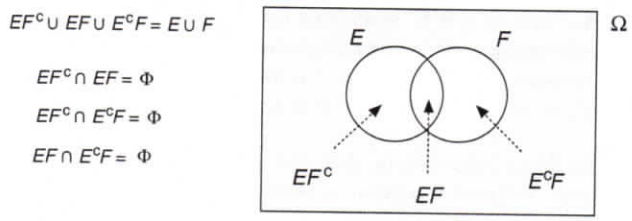

FIGURE 2.7 Partition of $E \cup F$ into three disjoint events.

Proof. Since F contains E, it can be partitioned in terms of E as follows:

$$F = E \cup E^c F. \tag{2.47}$$

Thus

$$P(F) = P(E) + P(E^c F) \geq P(E) \tag{2.48}$$

because all probabilities are nonnegative. □

Proposition 2.5, which is the most important one in this section, can be viewed as an extension of the third axiom to events that are not disjoint.

Proposition 2.5. *For events $E, F \in \Omega$:*

$$P(E \cup F) = P(E) + P(F) - P(EF). \tag{2.49}$$

Proof. Although E and F may not be disjoint, $E \cup F$ can always be written as the union of three mutually exclusive events:

$$E \cup F = EF^c \cup E^c F \cup EF \tag{2.50}$$

as illustrated in Figure 2.7. From the third axiom of probability:

$$P(E \cup F) = P(EF^c) + P(E^c F) + P(EF). \tag{2.51}$$

Observe that E and F can be expressed as the union of disjoint events: $E = EF^c \cup EF$ and $F = E^c F \cup EF$, yielding

$$P(E) = P(EF^c) + P(EF), \qquad P(F) = P(E^c F) + P(EF). \tag{2.52}$$

Substituting these expressions into (2.51) gives

$$\begin{aligned} P(E \cup F) &= P(E) - P(EF) + P(F) - P(EF) + P(EF), \\ &= P(E) + P(F) - P(EF), \end{aligned} \tag{2.53}$$

which completes the proof. □

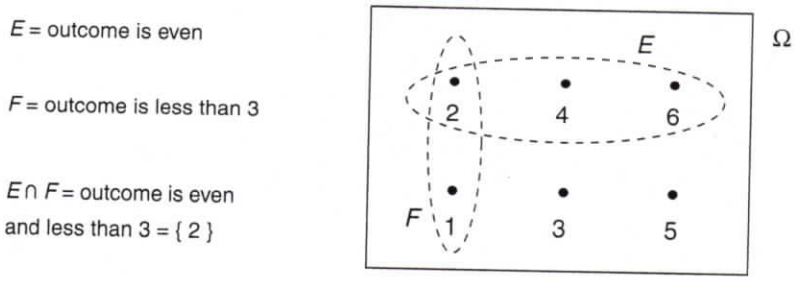

FIGURE 2.8 Venn diagram for calculating the probability of events that are not disjoint in Example 2.37.

Example 2.37. Consider the experiment of tossing a single fair die. Let event $E = \{\text{even number}\}$ and $F = \{\text{number} < 3\}$. Obviously these are not disjoint events. Since $P(E) = 1/2$, $P(F) = 1/3$, and $P(EF) = 1/6$, we have from (2.49):

$$P(E \cup F) = 1/2 + 2/3 - 1/6 = 2/3. \tag{2.54}$$

This result is easily seen from the Venn diagram in Figure 2.8.

2.9 CONDITIONAL PROBABILITY

In many applications, information is often known about some events of an experiment that can influence the probability of other events of interest. It is possible to incorporate this prior information into the calculation of probabilities. Consider two events A and B in the probability space $\{\Omega, \mathcal{F}, P\}$. If it is known that B has occurred, then the probability of A is usually different from that without this knowledge (though not necessarily). The notation $P(A|B)$ is used to denote the probability of A *given* B (also stated as given B is *true*). The following is a simple example for which it easy to compute conditional probabilities.

Example 2.38. For the experiment of tossing two fair coins, let A be the event that both coins are heads, and B the event that *at least one* coin is heads. It is easy to interpret the conditioning event (i.e., knowing this additional information) as causing the sample space to be *reduced* from four elements to three elements. Event B causes the "new" sample space to be $\{HH, HT, TH\}$ because TT is excluded, and thus we immediately have $P(A|B) = 1/3$. We can also compute $P(B|A)$, though this probability turns out to be trivial. The reduced sample space is simply $\{HH\}$, from which it is clear that B (at least one H) is always true: $P(B|A) = 1$.

$P(A|B)$ is a *conditional probability*. Although conditional probabilities were easy to calculate in Example 2.38 by examining the reduced sample space, in general it is preferable to use the following definition based on the intersection of events.

Definition: Conditional Probability The *conditional probability* of event A given event B is

$$P(A|B) \triangleq \frac{P(AB)}{P(B)} \tag{2.55}$$

assuming $P(B) \neq 0$.

We can interpret this definition using the Venn diagram for events A and B shown in Figure 2.9. By conditioning on B, the sample space is reduced from the original Ω to include only elements in B (because B has occurred). Thus, the numerator in (2.55) refers only to those elements in A that are also in B: since B is true, we are

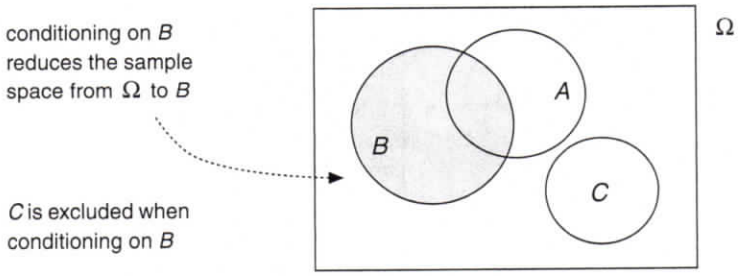

FIGURE 2.9 Conditional probability: elements in A that are not in B have zero probability in the measure $P(\cdot|B)$ (and, in fact, are impossible events).

interested in the probability of the elements in the *intersection* of A and B. Because $P(AB)$ is calculated using the original sample space based on all the original outcomes, it needs to be *normalized* to account for the fact that the sample space is now reduced. This is achieved by dividing $P(AB)$ by $P(B)$, which allows $P(\cdot|B)$ to be a consistent probability measure that satisfies the three axioms of probability, as proved in the following proposition.

Proposition 2.6. *Conditional probability $P(A|B)$ is a valid probability measure with respect to event* A.

Proof. For Axiom 1 and from the definition of conditional probability, we require that $P(AB)/P(B) \geq 0$. Since $P(AB)$ and $P(B)$ are valid probabilities in the original sample space Ω, it immediately follows that $P(A|B) \geq 0$. For Axiom 2,

$$P(\Omega|B) = \frac{P(\Omega B)}{P(B)} = \frac{P(B)}{P(B)} = 1. \tag{2.56}$$

Finally, for Axiom 3 and countably infinite disjoint events $\{A_n\}$,

$$P\left(\bigcup_{n=1}^{\infty} A_n \middle| B\right) = \frac{1}{P(B)} P\left(\bigcup_{n=1}^{\infty} A_n B\right) = \frac{1}{P(B)} \sum_{n=1}^{\infty} P(A_n B) = \sum_{n=1}^{\infty} P(A_n|B), \tag{2.57}$$

where Axiom 3 has been applied to the union of joint events $\{A_n B\}$ to complete the proof. □

We can also verify that the definition is intuitively correct using the frequency interpretation of probability for *finite* experiments where $P(B) = |B|/|\Omega|$. From the reduced sample space,

$$P(A|B) = \frac{|AB|}{|B|}. \tag{2.58}$$

This equation describes how the conditional probability was computed in Example 2.38. But note that this can be rewritten as

$$P(A|B) = \frac{|AB|/|\Omega|}{|B|/|\Omega|} = \frac{P(AB)}{P(B)}, \tag{2.59}$$

which is the same as the definition. Although we cannot apply the frequency interpretation to continuous experiments, the intuition is still the same, and the definition in (2.55) applies to any well-defined random experiment: continuous, discrete, or mixed.

It is important to note that $P(A|B)$ is a valid probability measure only with respect to event A and *not* with respect to the conditioning event B. Consider three events A, B, and C. The following expression:

$$P(A \cup B|C) = P(A|C) + P(B|C) - P(AB|C) \tag{2.60}$$

is a valid extension of the result in Proposition 2.6. The conditioning event C has been carried over to the right-hand side for each of the three terms. If the conditioning event is written in terms of a set operation such as $P(A|B \cup C)$, then in general

$$P(A|B \cup C) \neq P(A|B) + P(A|C) - P(A|BC). \tag{2.61}$$

Set operations on the conditioning event do not satisfy the axioms of probability, and in general we cannot rewrite expressions such as $P(A|B \cup C)$ in terms of probabilities conditioned on the individual events B and C.

Consider again the original probability space $\{\Omega, \mathcal{F}, P\}$. By conditioning on event B, the sample space is reduced to $B \subset \Omega$, the σ-field is reduced to the sub-σ-field $\mathcal{G} \subset \mathcal{F}$, and the probability measure is $P(\cdot|B)$ as described above. Thus, the *conditional probability space* is $\{B, \mathcal{G}, P(\cdot|B)\}$, which we illustrate by the following simple example.

Example 2.39. The sample space for the experiment of tossing two coins is $\Omega = \{HH, TT, TH, HT\}$. Let $\mathcal{F}$ be the power set with the elements summarized previously in Table 2.3. Define the conditioning event $B = \{\text{at least one } H\}$. This reduces the sample space to $B = \{HH, TH, HT\}$, and the corresponding sub-σ-field (the power set for B) is $\mathcal{G} = \{\phi, B, HH, TH, HT, \{HH, TH\}, \{HH, HT\}, \{TH, HT\}\}$.

We conclude this section with another example showing how conditioning reduces the sample space.

Example 2.40. When tossing a single die, suppose we are interested in the probability of $F = \{\text{outcome is } < 3\}$ given $E = \{\text{outcome is even}\}$. It is easily shown that $P(F|E) = 1/3$. Only outcome 2 is less than three in the reduced sample space $\{2, 4, 6\}$. For $P(E|F)$, the sample space is reduced to $\{1, 2\}$, and we immediately have $P(E|F) = 1/2$. In both cases, the conditional probabilities are the same as the corresponding probabilities without conditioning: $P(F|E) = P(F)$ and $P(E|F) = P(E)$. This is not a coincidence: it turns out that events E and F are *independent* in this experiment.

2.10 INDEPENDENCE

Independence in probability is an important property that is often assumed when modeling an experiment, which usually leads to simplified analyses. Intuitively, we can say that two events are independent if they have no influence on the outcomes of each other. It is defined as follows.

Definition: Independent Events Events A and B are *independent* if and only if

$$P(AB) = P(A)P(B). \tag{2.62}$$

Probability $P(AB)$ is known as the *joint* probability of A and B (i.e., they occur jointly). The individual probabilities $P(A)$ and $P(B)$ in the context of a joint probability are known as *marginal* probabilities. The definition of independence is clear for the following simple example.

Example 2.41. The probability of observing two heads when simultaneously tossing two fair coins is $P(HH) = P(H)P(H) = (1/2)(1/2) = 1/4$. It is obvious from a physical view of the dynamics of tossing a coin, there is no reason why the outcome of one coin should influence the other. Intuitively, and from the frequency interpretation of probability, the probabilities should split as in (2.62) for independent events. Likewise, the

probability of any pair of outcomes $\{\omega_m, \omega_n\}$ when tossing two fair dice is $P(\omega_m, \omega_n) = P(\omega_m)P(\omega_n) = 1/36$, where $\omega_m \in \{1, 2, 3, 4, 5, 6\}$.

Observe from the definition of conditional probability that

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A). \tag{2.63}$$

Likewise

$$P(B|A) = P(B), \tag{2.64}$$

demonstrating that independence is a symmetrical property of two events: they are *mutually independent*. Although (2.62) is the standard definition for independence, the result in (2.63) can also be taken as the definition; in fact, it is more appealing because it explicitly shows that conditioning on B has no influence on the probability of A when they are independent. In order to test whether or not two events are independent for specific numerical examples, one need only show that the left-hand side in (2.55) equals the right-hand side. That is, if both sides of the equation have the same numerical value, then the events are independent; dependent events *cannot* coincidentally have numerical equality in (2.62).

The definition of two independent events can be extended to several events.

Definition: Mutually Independent Events Events $\{A_1, \ldots, A_N\}$ are *mutually independent* if

$$P(A_m \cdots A_n) = P(A_m) \cdots P(A_n) \tag{2.65}$$

for all possible subsets of $\{A_1, \ldots, A_N\}$. This includes combinations of two events, three events, and so on, as well as all N events:

$$P(A_m A_n) = P(A_m)P(A_m), \qquad m \neq n, \tag{2.66}$$

$$P(A_m A_n A_p) = P(A_m)P(A_n)P(A_p), \qquad m \neq n \neq p, \tag{2.67}$$

$$\vdots$$

$$P(A_1 \cdots A_N) = P(A_1) \cdots P(A_N). \tag{2.68}$$

If one or more combinations above do not hold, then the events are dependent by definition. Thus, it is usually easier to demonstrate that multiple events are *dependent*. One might begin by checking all probabilities of two events, then of three events, and so on, until one of the conditions above does not hold. To demonstrate that all events are mutually independent, however, *all* combinations above must be examined.

Example 2.42. From the previous die-toss example with events $E = \{\text{even}\}$ and $F = \{\text{less than three}\}$, we know that $P(EF) = P(E)P(F)$. Consider another event $G = \{\text{divisible by three}\}$, corresponding to $G = \{3, 6\}$ for which $P(G) = 1/3$. From the Venn diagram in Figure 2.8, we find that $EG = \{6\}$ and thus $P(EG) = 1/6$. Observe that $P(FG) = 0$ (F and G are disjoint), but $P(F)P(G) = 1/9$. Events $\{E, F, G\}$ are not mutually independent.

2.11 BAYES' FORMULA

Bayes' formula is a useful equation that is used to compute conditional probabilities, or derive marginal probabilities by conditioning on events.

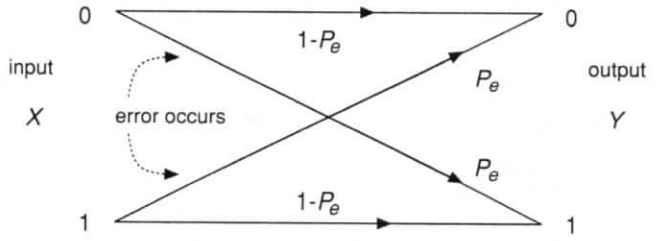

FIGURE 2.10 Binary symmetric channel (BSC) with input X and output Y. Binary $\Longrightarrow |X| = |Y| = 2$, and symmetric means $P(Y = 0|X = 0) = P(Y = 1|X = 0) \triangleq P_e$.

Theorem 2.2 (Bayes' formula). For events $\{A, B\} \in \Omega$ with $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2.69}$$

Likewise, for $P(A) \neq 0$:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{2.70}$$

Proof. From the definition of conditional probability, we can write the following two expressions:

$$P(AB) = P(B|A)P(A) = P(A|B)P(B). \tag{2.71}$$

Combining these results and solving for $P(A|B)$ or $P(B|A)$ yields Bayes' formula, which is also called *Bayes' rule*. □

Example 2.43 (Binary symmetric channel). An important example of Bayes' formula is the binary symmetric channel (BSC) shown in Figure 2.10, which is a model for bit errors that occur in a digital communication system. For a binary system, the transmitted symbol has two outcomes: we will use $\{0, 1\}$ to represent the two *binary* elementary events. This model is equivalent to the coin-toss experiment with $H \equiv 1$ and $T \equiv 0$. As shown in Chapter 3, it is convenient to represent the transmitted symbol by *random variable* X, which "randomly" takes on one of the two values. X itself is not random: it is well defined by the two possible outcomes (called the *alphabet*) and the specific probability for each value. The randomness of a random variable is due to the fact that it is not known beforehand which value will occur; X is not deterministic. Typically in a real system, $P(X = 0) = P(X = 1) = 1/2$ (analogous to a fair coin).

The received symbol can likewise be represented by random variable Y, which also has alphabet $\{0, 1\}$. We are interested in the probability of a transmission error across the channel, of which there are two types represented by conditional probabilities $P(Y = 1|X = 0)$ and $P(Y = 0|X = 1)$. The channel is *symmetric* if these two conditional probabilities are equal. The average probability of error is defined by conditioning on the two values for X:

$$P_e \triangleq P(Y = 1|X = 0)P(X = 0) + P(Y = 0|X = 1)P(X = 1), \tag{2.72}$$

which is an example of the *law of total probability* discussed later. Since the channel is symmetric, only one of the conditional probabilities needs to be examined:

$$P_e = P(Y = 1|X = 0)[P(X = 0) + P(X = 1)] = P(Y = 1|X = 0), \tag{2.73}$$

which does not require equally likely outcomes for this simplification. In a real communication system, P_e can usually be estimated by performing measurements over many trials of the experiment, by sending a stream of 0s and 1s that are known at the receiver and then counting the errors.

The receiver in a communication system requires the reverse conditional probabilities: $P(X = 0|Y = 0)$, $P(X = 0|Y = 1)$, $P(X = 1|Y = 0)$, and $P(X = 1|Y = 1)$. For example, if $Y = 0$ is received, the goal of the receiver is to decide whether $X = 0$ or $X = 1$ was transmitted. Since the input is random, the receiver cannot decide exactly which symbol was transmitted. The detection criterion, known as the maximum a posteriori (MAP) decision rule (see Chapter 10), chooses the symbol maximizing $P(X = x|Y = y)$ for $x, y \in \{0, 1\}$. Using Bayes' formula, this can be rewritten as

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}, \tag{2.74}$$

where $P(Y = y|X = x)$ is generally known for the specific communication channel, as mentioned above. The prior probability $P(X = x)$ (before conditioning on Y) is also known by the receiver so that $P(Y = y)$ is readily calculated using the law of total probability as follows:

$$P(Y = y) = P(Y = y|X = 0)P(X = 0) + P(Y = y|X = 1)P(X = 1). \tag{2.75}$$

This example is continued in Section 2.12.

2.12 TOTAL PROBABILITY

The law of total probability is an extension of Bayes' formula to a *partition* of events in the sample space Ω. It is a convenient means for computing the probability of an event in terms of conditional probabilities that might be easier to compute for some problems, such as the BSC discussed in Example 2.43.

Theorem 2.3 (Total probability). Let $\{B_n\}$, $n = 1, \ldots, N$, be a partition of Ω. Then

$$P(A) = \sum_{n=1}^{N} P(A|B_n)P(B_n). \tag{2.76}$$

Proof. Since $\{B_n\}$ form a partition, $B_n \cap B_m = \phi$ and $AB_n \cap AB_m = \phi$ for $n \neq m$. We also have

$$\bigcup_{n=1}^{N} AB_n = \Omega \tag{2.77}$$

so that

$$P(A) = \sum_{n=1}^{N} P(AB_n). \tag{2.78}$$

Applying conditional probability to each term in the sum gives $P(AB_n) = P(A|B_n)P(B_n)$, which completes the proof. □

FIGURE 2.1
$p = P(X =$

Example 2.
by conditio

Since the p
known, $\{P($
$P(Y = 1|X$
and (2.80) g

Figure 2.11
a communi
$P(X = 1))$.

Example 2
probability
disease. Co
be made: P
can be esti
population
patient has

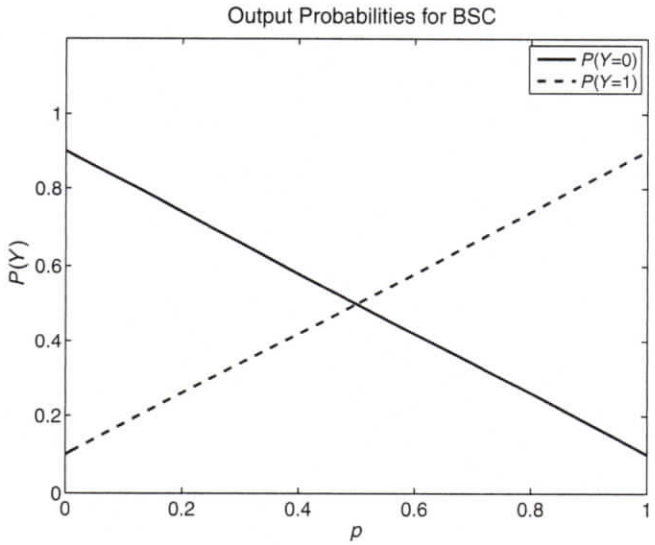

FIGURE 2.11 Output probabilities for the binary symmetric channel (BSC) in Example 2.44 as a function of $p = P(X = 1)$ for probability of error $P_e = 0.1$.

Example 2.44. Continuing with the BSC in Example 2.43 with input X and output Y, (2.75) was generated by conditioning on a partition for X; the results for each value of Y are

$$P(Y = 0) = P(Y = 0|X = 0)P(X = 0) + P(Y = 0|X = 1)P(X = 1), \tag{2.79}$$

$$P(Y = 1) = P(Y = 1|X = 0)P(X = 0) + P(Y = 1|X = 1)P(X = 1). \tag{2.80}$$

Since the prior probabilities $\{P(X = x)\}$ and the channel error probabilities $\{P(Y = y|X = x)\}$ are usually known, $\{P(Y = y)\}$ can be computed. Let $P(X = 1) = p$. Because the channel is symmetric, substituting $P(Y = 1|X = 0) = P(Y = 0|X = 1) = P_e$ and $P(Y = 1|X = 1) = P(Y = 0|X = 0) = 1 - P_e$ into (2.79) and (2.80) gives

$$P(Y = 0) = (1 - P_e)(1 - p) + P_e p, \tag{2.81}$$

$$P(Y = 1) = P_e(1 - p) + (1 - P_e)p. \tag{2.82}$$

Figure 2.11 shows plots of these two functions of p for $P_e = 0.1$ (which is a high error probability for a communication system). Clearly, $P(Y = 0) \neq P(Y = 1)$ except for $p = 1/2$ (i.e., when $P(X = 0) = P(X = 1)$).

Example 2.45 (Medical test). Another well-known application of Bayes' formula and the law of total probability is in medicine where a test has been developed to determine whether a patient has a particular disease. Consider two events: $A = \{\text{test is positive}\}$ and $B = \{\text{patient has the disease}\}$. Two types of errors can be made: $P(A|B^c)$ is known as a *false positive* and $P(A^c|B)$ is known as a *false negative*. These probabilities can be estimated from experimental trials, perhaps when the medicine is developed. Also, based on general population data, we assume that $P(B)$ is known with a high degree of accuracy. Thus, the probability that the patient has the disease given that the test is positive can be computed using Bayes' formula as follows:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{2.83}$$

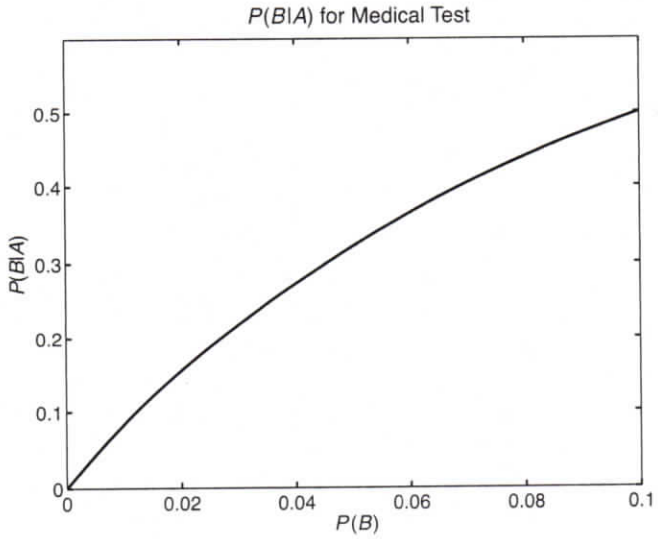

FIGURE 2.12 $P(B|A)$ in Example 2.45 as a function of $P(B)$ for $P(A^c|B) = P(A|B^c) = 0.1$.

Substituting the law of total probability gives

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}. \tag{2.84}$$

All quantities on the right-hand side are known because $P(A|B) + P(A^c|B) = 1$ and $P(A|B^c) + P(A^c|B^c) = 1$. For example, if $P(A^c|B) = P(A|B^c) = 0.1$ and $P(B) = 0.02$, then $P(B|A) \approx 0.16$. Figure 2.12 shows a plot of $P(B|A)$ versus the probability that the general population has the disease, as it varies from 0 to 0.1.

Finally, there is a multiplicative expression for a joint probability that involves conditioning similar to that used in the law of total probability, but does not require that the events form a partition. The conditioning is performed on an increasing number of events as described in the following theorem.

Theorem 2.4. Let N events $\{B_n\}$ in Ω be such that their joint probability $P(B_1 \cdots B_N) \neq 0$. Successively conditioning on these events gives

$$P(B_1 \cdots B_N) = P(B_1)P(B_2|B_1)P(B_3|B_1B_2) \cdots P(B_N|B_1 \cdots B_{N-1}). \tag{2.85}$$

Proof. The proof is by induction. Conditioning on one event as in Bayes' formula gives

$$P(B_1B_2) = P(B_1)P(B_2|B_1). \tag{2.86}$$

Assume the following expression is true for $N - 1$ events:

$$P(B_1 \cdots B_{N-1}) = P(B_1)P(B_2|B_1)P(B_3|B_1B_2) \cdots P(B_{N-1}|B_1 \cdots B_{N-2}). \tag{2.87}$$

Then

$$P(B_1 \cdots B_N) = P(B_N|B_1 \cdots B_{N-1})P(B_1 \cdots B_{N-1}) \tag{2.88}$$

is also a form of Bayes' formula because the intersection $B_1 \cdots B_{N-1}$ can be viewed as a single event. Substituting (2.87) into the last expression verifies the result for N events, thus completing the proof. □

2.13 DISCRETE SAMPLE SPACES

For discrete sample spaces (finite or countably infinite), it is straightforward to assign probabilities because events are expressed in terms of the individual sample points in Ω, which in turn are assigned probabilities satisfying the axioms of probability. As mentioned earlier, we generally assume $\mathcal{F} = \mathcal{P}(\Omega)$ for discrete Ω.

Example 2.46. Consider an experiment where the countable outcomes $\{x_n\}$ are the natural numbers $\mathcal{N}$. Suppose we assign probabilities to these outcomes as follows:

$$P(x_n) = (1 - p)^{n-1}p, \tag{2.89}$$

where $0 < p < 1$. This probability assignment yields the *geometric random variable* described in Chapter 3. In fact, it corresponds to the experiment of successively tossing a single coin with $P(H) = p$ until the *first* H appears. Even though an infinite number of outcomes are possible, these probabilities sum to one:

$$\sum_{n=1}^{\infty}(1 - p)^{n-1}p = p\sum_{n=0}^{\infty}(1 - p)^n = \frac{p}{1 - (1 - p)} = 1, \tag{2.90}$$

where the closed-form expression for an infinite series in Appendix E has been used. This is an example of a convergent series that has been scaled by p so that it converges to 1, and is due to the fact that there is a countable number of terms. In fact, the terms of any convergent series can be assigned probabilities as long as they are scaled to sum to 1.

Next, consider the special case of a *finite* sample space Ω with *equally likely* outcomes $\{\omega_n\}$. (Obviously, a countably infinite sample space *cannot* have equally likely outcomes). The probability of any event $E \in \Omega$ for this situation is determined by *counting* the number of outcomes in E and computing the ratio

$$P(E) = \frac{|E|}{|\Omega|}, \tag{2.91}$$

where $|E|$ is the cardinality of E. For such equally likely cases, finding probabilities reduces to a problem in *combinatorics*. Computing $|\Omega|$ is usually easy, whereas finding $|E|$ often requires more work because it involves a subset of outcomes. In order to handle such problems, we need to be clear about the difference between a combination and a permutation.

Definition: Combination and Permutation A *combination* is an unordered set of distinct elements. A *permutation* is one of the many ordered sets of distinct elements.

Given N distinct elements, there is only one combination: the order of these elements does not matter, because each ordering is still the same combination. On the other hand, the various arrangements of those elements are different permutations.

Example 2.47. Consider the set $E = \{a_1, a_2, a_3\}$. There is only one combination of those elements: $\{a_2, a_1, a_3\}$ is the same combination. On the other hand, there are six permutations: in addition to the two above, they are $\{a_1, a_3, a_2\}$, $\{a_2, a_3, a_1\}$, $\{a_3, a_1, a_2\}$, and $\{a_3, a_2, a_1\}$.

It is clear from this simple example that for a set of N distinct elements, there can be only one combination, but there are $N!$ permutations. This is readily seen for the general case as follows. Since a permutation is an ordered set, meaning that the location of an element relative to the other elements must be considered, then for the first entry we can "choose" any of the N elements. For the second entry, we can choose only from the $N - 1$ remaining elements (since the elements in the permutation must be distinct, we cannot choose the element already in the first entry). Similarly for the third entry, we can choose from $N - 2$ elements and so on until the last entry, for which there is only one element remaining. Thus, the number of permutations of N distinct elements is

$$N(N-1)\cdots 2\cdot 1 = N!. \tag{2.92}$$

When counting the number of outcomes in event E, we must first determine whether or not their order is important. The following counting examples demonstrate some cases.

Example 2.48. Suppose we are dealt a hand of five distinct cards from a standard deck of 52 cards. There are 5! ways that we can arrange those cards, corresponding to 120 permutations. However, as we know in card games such as poker, the various permutations do not affect the strength of the hand. Whether or not we have a winning hand depends only on the combination of the cards; their order is not relevant. (This is true even for a straight such as {5, 6, 7, 8, 9}: the strength of this hand does not require that we hold the five cards in increasing or decreasing order.)

Example 2.49. Generally, when an arrangement is specified in the problem statement, permutations should be considered. Consider counting the number of social security numbers (which have nine digits) such that a digit is *not* repeated, and which does not include a zero. Then obviously the order is important: there are $9! = 362{,}880$ such social security numbers. These are all the permutations of the numbers 1 through 9 without repeated digits. This type of "experiment" is called *sampling without replacement*: once a number has been selected for a digit, it cannot be used again. Suppose now that we allow digits to be repeated. In this case, there are 9^9 such numbers (again, excluding zeros). This problem is not handled by the formula in (2.92) because the digits are no longer distinct. This scenario corresponds to *sampling with replacement*. The number of elements is readily determined by an approach similar to that used to derive (2.92): allowing digits to be repeated, there are $9 \cdot 9 \cdots 9 = 9^9 = 387{,}420{,}489$ numbers.

For problems involving N distinct equally likely outcomes, we are often interested in a subset of $M < N$ elements corresponding to some event. This case yields more than one combination because there are many ways to choose M elements from N without regard to order. The number of permutations for this case is easily seen to be

$$N(N-1)\cdots(N-M+1) \triangleq (N)_M, \tag{2.93}$$

where $(N)_M$ is the notation for a *falling factorial* (see Appendix B). It can be rewritten as

$$(N)_M = \frac{N(N-1)\cdots(N-M+1)}{(N-M)\cdots 2\cdot 1}[(N-M)\cdots 2\cdot 1] = \frac{N!}{(N-M)!}. \tag{2.94}$$

Since the set has M elements, there are $M!$ permutations from (2.92). The number of combinations obtained by selecting M elements from N is computed by dividing (2.94) by the number of permutations. This causes us to count all permutations of the same M elements as a single combination:

$$\frac{N!}{(N-M)!M!} = \frac{(N)_M}{M!} \triangleq \binom{N}{M}. \tag{2.95}$$

This last expression, which we state as "N choose M," is called a *binomial coefficient*. Equations (2.94) and (2.95) summarize the number of permutations and combinations, respectively, for a size-M subset chosen from N distinct elements. Some factorials and binomial coefficients are provided in Tables E.1 and E.2 in Appendix E, along with results for Stirling's formula which approximates $N!$ for large N.

Example 2.50. Suppose we are interested in the number of six-digit license plates (no letters allowed) without repeated digits (i.e., without replacement) and excluding any zeros. This problem is handled by the formula in (2.94) with $N = 9$ and $M = 6$. Clearly, order is important; the number of such license plates is

$$\frac{9!}{(9-6)!} = 504 \tag{2.96}$$

permutations. If we are concerned only with the number of license plates regardless of the order of the digits, then (2.95) is used, yielding

$$\binom{9}{6} = 84 \tag{2.97}$$

combinations. When determining the number of combinations, license plate 123456 is "equivalent" to 213456, and thus only one of them above is counted. This is handled in (2.95) by dividing by $M! = 6! = 720$ permutations of the six numbers.

These results can be generalized to an experiment where the N elements are no longer distinct, but instead can be grouped into M subsets, each of which has identical elements. The first group has N_1 indistinguishable elements, the second group has N_2 indistinguishable elements, and so on until the Mth group with N_M indistinguishable elements such that

$$\sum_{m=1}^{M} N_m = N. \tag{2.98}$$

This case reduces to the previous experiment when each group has only one element: $N_m = 1$ for $m = 1, \ldots, M$. (Note, of course, that we assume no groups have the same type of elements; otherwise, such groups should be combined into larger groups.)

Since there are no longer N distinct elements, the overall number of permutations given by $N!$ is necessarily reduced by the permutations for all groups. Since the first group has N_1 identical elements, it is clear that we must divide $N!$ by the amount $N_1!$, so that permutations in the first group are not included in the overall number. This is done for each of the groups so that the number of permutations is

$$\frac{N!}{N_1!N_2!\cdots N_M!} \triangleq \binom{N}{N_1, N_2, \ldots, N_M}. \tag{2.99}$$

The right-hand side is known as a *multinomial coefficient*, which reduces to (2.92) when all the groups have only one element.

Example 2.51. Consider an experiment where four fair coins are tossed, and let E be the event of three tails. The elements of E are $\{T, T, T, H\}$, $\{T, T, H, T\}$, $\{T, H, T, T\}$, and $\{H, T, T, T\}$, which correspond to the four possible permutations of three tails and one head. Thus $|E| = 4$ and since $|\Omega| = 2^4 = 16$, the probability is $P(E) = |E|/|\Omega| = 1/4$. In this experiment, the order of elements is important: we must take into consideration the four ways that a single H can occur. The formula in (2.92) does not apply directly because the elements are not distinct: the three tails are indistinguishable, and their order within the four elements of E is not relevant. Using (2.99), we find that $N = 4$, $N_1 = 1$, and $N_2 = 3$, which gives

$$|E| = \binom{4}{3, 1} = 4. \tag{2.100}$$

Next, we examine how the problem changes by considering event F consisting of two tails. The elements of F are $\{T, T, H, H\}$, $\{H, H, T, T\}$, $\{T, H, T, H\}$, $\{H, T, H, T\}$, $\{H, T, T, H\}$, and $\{T, H, H, T\}$; these are all the possible permutations. From (2.99),

$$|F| = \binom{4}{2, 2} = 6 \tag{2.101}$$

and the probability is $P(F) = |F|/|\Omega| = 3/8$.

Finally, we discuss further the process of *sampling* when dealing with finite sample spaces and equally likely outcomes. Consider again two previous examples: (i) drawing five cards for a hand in a game of poker and (ii) tossing a single fair die. In the first case as the cards are drawn from the deck, they are not replaced: this experiment is an example of *sampling without replacement*. The significance of this case is that when five additional cards are drawn from the same deck (for another player's hand), the sample space is changed: there are now only 47 cards. This, of course, must be considered when computing probabilities. The second example corresponds to *sampling with replacement*: each time the die is tossed, the sample space is unchanged. The same six outcomes are always possible. If the die is tossed M times, the number of permutations is

$$\underbrace{N \cdots N}_{M} = N^M. \tag{2.102}$$

Table 2.5 summarizes the formulas for determining the number of samples in a finite random experiment. The fourth case gives the number of combinations with replacement; it is derived in Problem 2.49.

Example 2.52. Suppose a fair die is tossed twice. The number of combinations with replacement from the formula in Table 2.5 is

$$\binom{6+2-1}{2} = \frac{7!}{2!5!} = 21. \tag{2.103}$$

TABLE 2.5 Number of Permutations and Combinations

Type of Sample (N Elements, Choose M)	Formula
Combinations without replacement	$\binom{N}{M} = \dfrac{N!}{M!(N-M)!}$
Permutations without replacement	$(N)_M = \dfrac{N!}{(N-M)!}$
Combinations with replacement	$\binom{N+M-1}{M} = \dfrac{(N+M-1)!}{M!(N-1)!}$
Permutations with replacement	N^M