1

Sample Space and Probability

Contents

"Probability" is a very useful concept, but can be interpreted in a number of ways. As an illustration, consider the following.

A patient is admitted to the hospital and a potentially life-saving drug is administered. The following dialog takes place between the nurse and a concerned relative.

RELATIVE: Nurse, what is the probability that the drug will work?
NURSE: I hope it works, we'll know tomorrow.
RELATIVE: Yes, but what is the probability that it will?
NURSE: Each case is different, we have to wait.
RELATIVE: But let's see, out of a hundred patients that are treated under similar conditions, how many times would you expect it to work?
NURSE (somewhat annoyed): I told you, every person is different, for some it works, for some it doesn't.
RELATIVE (insisting): Then tell me, if you had to bet whether it will work or not, which side of the bet would you take?
NURSE (cheering up for a moment): I'd bet it will work.
RELATIVE (somewhat relieved): OK, now, would you be willing to lose two dollars if it doesn't work, and gain one dollar if it does?
NURSE (exasperated): What a sick thought! You are wasting my time!

In this conversation, the relative attempts to use the concept of probability to discuss an **uncertain** situation. The nurse's initial response indicates that the meaning of "probability" is not uniformly shared or understood, and the relative tries to make it more concrete. The first approach is to define probability in terms of **frequency of occurrence**, as a percentage of successes in a moderately large number of similar situations. Such an interpretation is often natural. For example, when we say that a perfectly manufactured coin lands on heads "with probability 50%," we typically mean "roughly half of the time." But the nurse may not be entirely wrong in refusing to discuss in such terms. What if this was an experimental drug that was administered for the very first time in this hospital or in the nurse's experience?

While there are many situations involving uncertainty in which the frequency interpretation is appropriate, there are other situations in which it is not. Consider, for example, a scholar who asserts that the Iliad and the Odyssey were composed by the same person, with probability 90%. Such an assertion conveys some information, but not in terms of frequencies, since the subject is a one-time event. Rather, it is an expression of the scholar's **subjective belief**. One might think that subjective beliefs are not interesting, at least from a mathematical or scientific point of view. On the other hand, people often have to make choices in the presence of uncertainty, and a systematic way of making use of their beliefs is a prerequisite for successful, or at least consistent, decision

making.

In fact, the choices and actions of a rational person, can reveal a lot about the inner-held subjective probabilities, even if the person does not make conscious use of probabilistic reasoning. Indeed, the last part of the earlier dialog was an attempt to infer the nurse's beliefs in an indirect manner. Since the nurse was willing to accept a one-for-one bet that the drug would work, we may infer that the probability of success was judged to be at least 50%. And had the nurse accepted the last proposed bet (two-for-one), that would have indicated a success probability of at least 2/3.

Rather than dwelling further into philosophical issues about the appropriateness of probabilistic reasoning, we will simply take it as a given that the theory of probability is useful in a broad variety of contexts, including some where the assumed probabilities only reflect subjective beliefs. There is a large body of successful applications in science, engineering, medicine, management, etc., and on the basis of this empirical evidence, probability theory is an extremely useful tool.

Our main objective in this book is to develop the art of describing uncertainty in terms of probabilistic models, as well as the skill of probabilistic reasoning. The first step, which is the subject of this chapter, is to describe the generic structure of such models, and their basic properties. The models we consider assign probabilities to collections (sets) of possible outcomes. For this reason, we must begin with a short review of set theory.

1.1 SETS

Probability makes extensive use of set operations, so let us introduce at the outset the relevant notation and terminology.

A **set** is a collection of objects, which are the **elements** of the set. If S is a set and x is an element of S, we write $x \in S$. If x is not an element of S, we write $x \notin S$. A set can have no elements, in which case it is called the **empty set**, denoted by $\emptyset$.

Sets can be specified in a variety of ways. If S contains a finite number of elements, say $x_1, x_2, \ldots, x_n$, we write it as a list of the elements, in braces:

$$S = \{x_1, x_2, \ldots, x_n\}.$$

For example, the set of possible outcomes of a die roll is $\{1, 2, 3, 4, 5, 6\}$, and the set of possible outcomes of a coin toss is $\{H, T\}$, where H stands for "heads" and T stands for "tails."

If S contains infinitely many elements $x_1, x_2, \ldots$, which can be enumerated in a list (so that there are as many elements as there are positive integers) we write

$$S = \{x_1, x_2, \ldots\},$$

and we say that S is **countably infinite**. For example, the set of even integers can be written as $\{0, 2, -2, 4, -4, \ldots\}$, and is countably infinite.

Alternatively, we can consider the set of all x that have a certain property P, and denote it by

$$\{x \mid x \text{ satisfies } P\}.$$

(The symbol "|" is to be read as "such that.") For example the set of even integers can be written as $\{k \mid k/2 \text{ is integer}\}$. Similarly, the set of all scalars x in the interval $[0, 1]$ can be written as $\{x \mid 0 \leq x \leq 1\}$. Note that the elements x of the latter set take a continuous range of values, and cannot be written down in a list (a proof is sketched in the theoretical problems); such a set is said to be **uncountable**.

If every element of a set S is also an element of a set T, we say that S is a **subset** of T, and we write $S \subset T$ or $T \supset S$. If $S \subset T$ and $T \subset S$, the two sets are **equal**, and we write $S = T$. It is also expedient to introduce a **universal set**, denoted by Ω, which contains all objects that could conceivably be of interest in a particular context. Having specified the context in terms of a universal set Ω, we only consider sets S that are subsets of Ω.

Set Operations

The **complement** of a set S, with respect to the universe Ω, is the set $\{x \in \Omega \mid x \notin S\}$ of all elements of Ω that do not belong to S, and is denoted by S^c. Note that $\Omega^c = \emptyset$.

The **union** of two sets S and T is the set of all elements that belong to S or T (or both), and is denoted by $S \cup T$. The **intersection** of two sets S and T is the set of all elements that belong to both S and T, and is denoted by $S \cap T$. Thus,

$$S \cup T = \{x \mid x \in S \text{ or } x \in T\},$$
$$S \cap T = \{x \mid x \in S \text{ and } x \in T\}.$$

In some cases, we will have to consider the union or the intersection of several, even infinitely many sets, defined in the obvious way. For example, if for every positive integer n, we are given a set S_n, then

$$\bigcup_{n=1}^{\infty} S_n = S_1 \cup S_2 \cup \cdots = \{x \mid x \in S_n \text{ for some } n\},$$

and

$$\bigcap_{n=1}^{\infty} S_n = S_1 \cap S_2 \cap \cdots = \{x \mid x \in S_n \text{ for all } n\}.$$

Two sets are said to be disjoint if their intersection is empty. More generally, several sets are said to be **disjoint** if no two of them have a common element. A collection of sets is said to be a **partition** of a set S if the sets in the collection are disjoint and their union is S.

If x and y are two objects, we use (x, y) to denote the **ordered pair** of x and y. The set of scalars (real numbers) is denoted by $\Re$; the set of pairs (or triplets) of scalars, i.e., the two-dimensional plane (or three-dimensional space, respectively) is denoted by $\Re^2$ (or $\Re^3$, respectively).

Sets and the associated operations are easy to visualize in terms of **Venn diagrams**, as illustrated in Fig. 1.1.

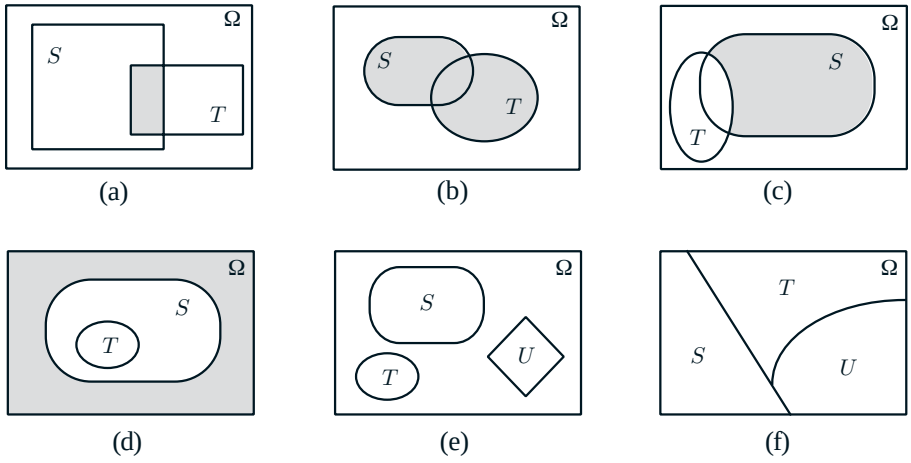

Figure 1.1: Examples of Venn diagrams. (a) The shaded region is $S \cap T$. (b) The shaded region is $S \cup T$. (c) The shaded region is $S \cap T^c$. (d) Here, $T \subset S$. The shaded region is the complement of S. (e) The sets S, T, and U are disjoint. (f) The sets S, T, and U form a partition of the set Ω.

The Algebra of Sets

Set operations have several properties, which are elementary consequences of the definitions. Some examples are:

$$\begin{aligned} S \cup T &= T \cup S, & S \cup (T \cup U) &= (S \cup T) \cup U, \\ S \cap (T \cup U) &= (S \cap T) \cup (S \cap U), & S \cup (T \cap U) &= (S \cup T) \cap (S \cup U), \\ (S^c)^c &= S, & S \cap S^c &= \emptyset, \\ S \cup \Omega &= \Omega, & S \cap \Omega &= S. \end{aligned}$$

Two particularly useful properties are given by **de Morgan's laws** which state that

$$\left(\bigcup_n S_n \right)^c = \bigcap_n S_n^c, \qquad \left(\bigcap_n S_n \right)^c = \bigcup_n S_n^c.$$

To establish the first law, suppose that $x \in (\cup_n S_n)^c$. Then, $x \notin \cup_n S_n$, which implies that for every n, we have $x \notin S_n$. Thus, x belongs to the complement

of every S_n, and $x_n \in \cap_n S_n^c$. This shows that $(\cup_n S_n)^c \subset \cap_n S_n^c$. The converse inclusion is established by reversing the above argument, and the first law follows. The argument for the second law is similar.

1.2 PROBABILISTIC MODELS

A probabilistic model is a mathematical description of an uncertain situation. It must be in accordance with a fundamental framework that we discuss in this section. Its two main ingredients are listed below and are visualized in Fig. 1.2.

Elements of a Probabilistic Model

- The **sample space** Ω, which is the set of all possible outcomes of an experiment.
- The **probability law**, which assigns to a set A of possible outcomes (also called an **event**) a nonnegative number $\mathbf{P}(A)$ (called the **probability** of A) that encodes our knowledge or belief about the collective "likelihood" of the elements of A. The probability law must satisfy certain properties to be introduced shortly.

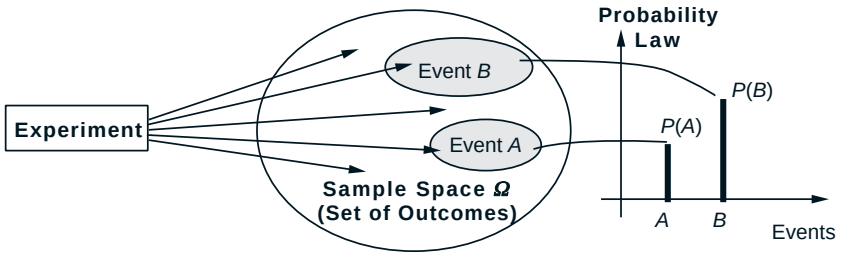

Figure 1.2: The main ingredients of a probabilistic model.

Sample Spaces and Events

Every probabilistic model involves an underlying process, called the **experiment**, that will produce exactly one out of several possible **outcomes**. The set of all possible outcomes is called the **sample space** of the experiment, and is denoted by Ω. A subset of the sample space, that is, a collection of possible

outcomes, is called an **event**.† There is no restriction on what constitutes an experiment. For example, it could be a single toss of a coin, or three tosses, or an infinite sequence of tosses. However, it is important to note that in our formulation of a probabilistic model, there is only one experiment. So, three tosses of a coin constitute a single experiment, rather than three experiments.

The sample space of an experiment may consist of a finite or an infinite number of possible outcomes. Finite sample spaces are conceptually and mathematically simpler. Still, sample spaces with an infinite number of elements are quite common. For an example, consider throwing a dart on a square target and viewing the point of impact as the outcome.

Choosing an Appropriate Sample Space

Regardless of their number, different elements of the sample space should be distinct and **mutually exclusive** so that when the experiment is carried out, there is a unique outcome. For example, the sample space associated with the roll of a die cannot contain "1 or 3" as a possible outcome and also "1 or 4" as another possible outcome. When the roll is a 1, the outcome of the experiment would not be unique.

A given physical situation may be modeled in several different ways, depending on the kind of questions that we are interested in. Generally, the sample space chosen for a probabilistic model must be **collectively exhaustive**, in the sense that no matter what happens in the experiment, we always obtain an outcome that has been included in the sample space. In addition, the sample space should have enough detail to distinguish between all outcomes of interest to the modeler, while avoiding irrelevant details.

Example 1.1. Consider two alternative games, both involving ten successive coin tosses:

Game 1: We receive \$1 each time a head comes up.

Game 2: We receive \$1 for every coin toss, up to and including the first time a head comes up. Then, we receive \$2 for every coin toss, up to the second time a head comes up. More generally, the dollar amount per toss is doubled each time a head comes up.

---

† Any collection of possible outcomes, including the entire sample space Ω and its complement, the empty set $\emptyset$, may qualify as an event. Strictly speaking, however, some sets have to be excluded. In particular, when dealing with probabilistic models involving an uncountably infinite sample space, there are certain unusual subsets for which one cannot associate meaningful probabilities. This is an intricate technical issue, involving the mathematics of measure theory. Fortunately, such pathological subsets do not arise in the problems considered in this text or in practice, and the issue can be safely ignored.

In game 1, it is only the total number of heads in the ten-toss sequence that matters, while in game 2, the order of heads and tails is also important. Thus, in a probabilistic model for game 1, we can work with a sample space consisting of eleven possible outcomes, namely, $0, 1, \ldots, 10$. In game 2, a finer grain description of the experiment is called for, and it is more appropriate to let the sample space consist of every possible ten-long sequence of heads and tails.

Sequential Models

Many experiments have an inherently sequential character, such as for example tossing a coin three times, or observing the value of a stock on five successive days, or receiving eight successive digits at a communication receiver. It is then often useful to describe the experiment and the associated sample space by means of a **tree-based sequential description**, as in Fig. 1.3.

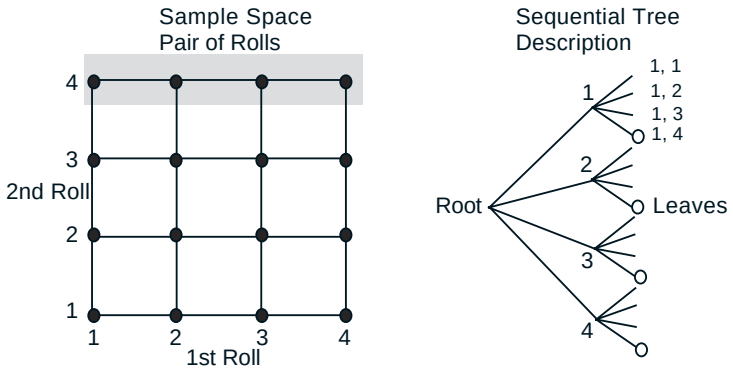

Figure 1.3: Two equivalent descriptions of the sample space of an experiment involving two rolls of a 4-sided die. The possible outcomes are all the ordered pairs of the form (i, j), where i is the result of the first roll, and j is the result of the second. These outcomes can be arranged in a 2-dimensional grid as in the figure on the left, or they can be described by the tree on the right, which reflects the sequential character of the experiment. Here, each possible outcome corresponds to a leaf of the tree and is associated with the unique path from the root to that leaf. The shaded area on the left is the event $\{(1, 4),\ (2, 4),\ (3, 4),\ (4, 4)\}$ that the result of the second roll is 4. That same event can be described as a set of leaves, as shown on the right. Note also that every node of the tree can be identified with an event, namely, the set of all leaves downstream from that node. For example, the node labeled by a 1 can be identified with the event $\{(1, 1), (1, 2), (1, 3), (1, 4)\}$ that the result of the first roll is 1.

Probability Laws

Suppose we have settled on the sample space Ω associated with an experiment.

Then, to complete the probabilistic model, we must introduce a **probability law**. Intuitively, this specifies the "likelihood" of any outcome, or of any set of possible outcomes (an event, as we have called it earlier). More precisely, the probability law assigns to every event A, a number $\mathbf{P}(A)$, called the **probability** of A, satisfying the following axioms.

Probability Axioms

1. **(Nonnegativity)** $\mathbf{P}(A) \geq 0$, for every event A.
2. **(Additivity)** If A and B are two disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

   Furthermore, if the sample space has an infinite number of elements and $A_1, A_2, \ldots$ is a sequence of disjoint events, then the probability of their union satisfies

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots$$

3. **(Normalization)** The probability of the entire sample space Ω is equal to 1, that is, $\mathbf{P}(\Omega) = 1$.

In order to visualize a probability law, consider a unit of mass which is to be "spread" over the sample space. Then, $\mathbf{P}(A)$ is simply the total mass that was assigned collectively to the elements of A. In terms of this analogy, the additivity axiom becomes quite intuitive: the total mass in a sequence of disjoint events is the sum of their individual masses.

A more concrete interpretation of probabilities is in terms of relative frequencies: a statement such as $\mathbf{P}(A) = 2/3$ often represents a belief that event A will materialize in about two thirds out of a large number of repetitions of the experiment. Such an interpretation, though not always appropriate, can sometimes facilitate our intuitive understanding. It will be revisited in Chapter 7, in our study of limit theorems.

There are many natural properties of a probability law which have not been included in the above axioms for the simple reason that they can be **derived** from them. For example, note that the normalization and additivity axioms imply that

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\Omega \cup \emptyset) = \mathbf{P}(\Omega) + \mathbf{P}(\emptyset) = 1 + \mathbf{P}(\emptyset),$$

and this shows that the probability of the empty event is 0:

$$\mathbf{P}(\emptyset) = 0.$$

As another example, consider three disjoint events A_1, A_2, and A_3. We can use the additivity axiom for two disjoint events repeatedly, to obtain

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \cup A_3) &= \mathbf{P}\big(A_1 \cup (A_2 \cup A_2)\big) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup A_3) \\ &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3).\end{aligned}$$

Proceeding similarly, we obtain that the probability of the union of finitely many disjoint events is always equal to the sum of the probabilities of these events. More such properties will be considered shortly.

Discrete Models

Here is an illustration of how to construct a probability law starting from some common sense assumptions about a model.

Example 1.2. Coin tosses. Consider an experiment involving a single coin toss. There are two possible outcomes, heads (H) and tails (T). The sample space is $\Omega = \{H, T\}$, and the events are

$$\{H, T\},\ \{H\},\ \{T\},\ \emptyset.$$

If the coin is fair, i.e., if we believe that heads and tails are "equally likely," we should assign equal probabilities to the two possible outcomes and specify that $\mathbf{P}\big(\{H\}\big) = \mathbf{P}\big(\{T\}\big) = 0.5$. The additivity axiom implies that

$$\mathbf{P}\big(\{H, T\}\big) = \mathbf{P}\big(\{H\}\big) + \mathbf{P}\big(\{T\}\big) = 1,$$

which is consistent with the normalization axiom. Thus, the probability law is given by

$$\mathbf{P}\big(\{H, T\}\big) = 1, \qquad \mathbf{P}\big(\{H\}\big) = 0.5, \qquad \mathbf{P}\big(\{T\}\big) = 0.5, \qquad \mathbf{P}(\emptyset) = 0,$$

and satisfies all three axioms.

Consider another experiment involving three coin tosses. The outcome will now be a 3-long string of heads or tails. The sample space is

$$\Omega = \{HHH,\ HHT,\ HTH,\ HTT,\ THH,\ THT,\ TTH,\ TTT\}.$$

We assume that each possible outcome has the same probability of $1/8$. Let us construct a probability law that satisfies the three axioms. Consider, as an example, the event

$$A = \{\text{exactly 2 heads occur}\} = \{HHT,\ HTH,\ THH\}.$$

Using additivity, the probability of A is the sum of the probabilities of its elements:

$$\begin{aligned}\mathbf{P}\big(\{HHT,\, HTH,\, THH\}\big) &= \mathbf{P}\big(\{HHT\}\big) + \mathbf{P}\big(\{HTH\}\big) + \mathbf{P}\big(\{THH\}\big) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \\ &= \frac{3}{8}.\end{aligned}$$

Similarly, the probability of any event is equal to 1/8 times the number of possible outcomes contained in the event. This defines a probability law that satisfies the three axioms.

By using the additivity axiom and by generalizing the reasoning in the preceding example, we reach the following conclusion.

Discrete Probability Law

If the sample space consists of a finite number of possible outcomes, then the probability law is specified by the probabilities of the events that consist of a single element. In particular, the probability of any event $\{s_1, s_2, \ldots, s_n\}$ is the sum of the probabilities of its elements:

$$\mathbf{P}\big(\{s_1, s_2, \ldots, s_n\}\big) = \mathbf{P}\big(\{s_1\}\big) + \mathbf{P}\big(\{s_2\}\big) + \cdots + \mathbf{P}\big(\{s_n\}\big).$$

In the special case where the probabilities $\mathbf{P}\big(\{s_1\}\big), \ldots, \mathbf{P}\big(\{s_n\}\big)$ are all the same (by necessity equal to $1/n$, in view of the normalization axiom), we obtain the following.

Discrete Uniform Probability Law

If the sample space consists of n possible outcomes which are equally likely (i.e., all single-element events have the same probability), then the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{Number of elements of } A}{n}.$$

Let us provide a few more examples of sample spaces and probability laws.

Example 1.3. Dice. Consider the experiment of rolling a pair of 4-sided dice (cf. Fig. 1.4). We assume the dice are fair, and we interpret this assumption to mean

that each of the sixteen possible outcomes [ordered pairs (i, j), with $i, j = 1, 2, 3, 4$], has the same probability of 1/16. To calculate the probability of an event, we must count the number of elements of event and divide by 16 (the total number of possible outcomes). Here are some event probabilities calculated in this way:

$$\begin{aligned}
\mathbf{P}\big(\{\text{the sum of the rolls is even}\}\big) &= 8/16 = 1/2,\\
\mathbf{P}\big(\{\text{the sum of the rolls is odd}\}\big) &= 8/16 = 1/2,\\
\mathbf{P}\big(\{\text{the first roll is equal to the second}\}\big) &= 4/16 = 1/4,\\
\mathbf{P}\big(\{\text{the first roll is larger than the second}\}\big) &= 6/16 = 3/8,\\
\mathbf{P}\big(\{\text{at least one roll is equal to 4}\}\big) &= 7/16.
\end{aligned}$$

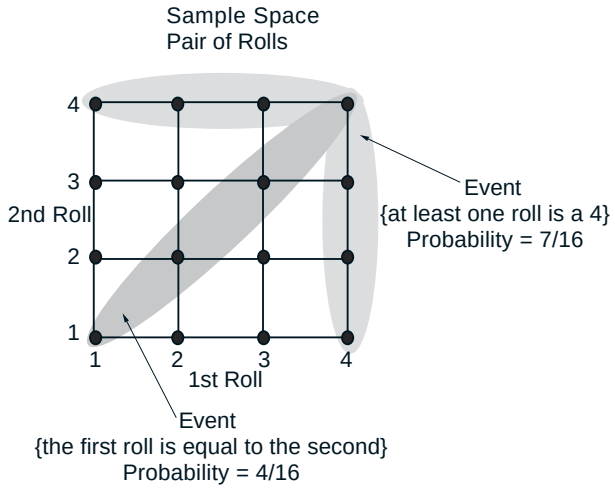

Figure 1.4: Various events in the experiment of rolling a pair of 4-sided dice, and their probabilities, calculated according to the discrete uniform law.

Continuous Models

Probabilistic models with continuous sample spaces differ from their discrete counterparts in that the probabilities of the single-element events may not be sufficient to characterize the probability law. This is illustrated in the following examples, which also illustrate how to generalize the uniform probability law to the case of a continuous sample space.

Example 1.4. A wheel of fortune is continuously calibrated from 0 to 1, so the possible outcomes of an experiment consisting of a single spin are the numbers in the interval $\Omega = [0, 1]$. Assuming a fair wheel, it is appropriate to consider all outcomes equally likely, but what is the probability of the event consisting of a single element? It cannot be positive, because then, using the additivity axiom, it would follow that events with a sufficiently large number of elements would have probability larger than 1. Therefore, the probability of any event that consists of a single element must be 0.

In this example, it makes sense to assign probability $b - a$ to any subinterval $[a, b]$ of $[0, 1]$, and to calculate the probability of a more complicated set by evaluating its "length."† This assignment satisfies the three probability axioms and qualifies as a legitimate probability law.

Example 1.5. Romeo and Juliet have a date at a given time, and each will arrive at the meeting place with a delay between 0 and 1 hour, with all pairs of delays being equally likely. The first to arrive will wait for 15 minutes and will leave if the other has not yet arrived. What is the probability that they will meet?

Let us use as sample space the square $\Omega = [0, 1] \times [0, 1]$, whose elements are the possible pairs of delays for the two of them. Our interpretation of "equally likely" pairs of delays is to let the probability of a subset of Ω be equal to its area. This probability law satisfies the three probability axioms. The event that Romeo and Juliet will meet is the shaded region in Fig. 1.5, and its probability is calculated to be 7/16.

Properties of Probability Laws

Probability laws have a number of properties, which can be deduced from the axioms. Some of them are summarized below.

Some Properties of Probability Laws

Consider a probability law, and let A, B, and C be events.

(a) If $A \subset B$, then $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(b) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(c) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

(d) $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A^c \cap B^c \cap C)$.

† The "length" of a subset S of $[0, 1]$ is the integral $\int_S dt$, which is defined, for "nice" sets S, in the usual calculus sense. For unusual sets, this integral may not be well defined mathematically, but such issues belong to a more advanced treatment of the subject.

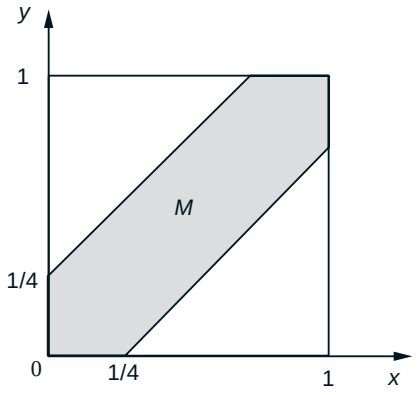

Figure 1.5: The event M that Romeo and Juliet will arrive within 15 minutes of each other (cf. Example 1.5) is

$$M = \Big\{(x, y) \;\Big|\; |x - y| \leq 1/4,\ 0 \leq x \leq 1,\ 0 \leq y \leq 1\Big\},$$

and is shaded in the figure. The area of M is 1 minus the area of the two unshaded triangles, or $1 - (3/4) \cdot (3/4) = 7/16$. Thus, the probability of meeting is $7/16$.

These properties, and other similar ones, can be visualized and verified graphically using Venn diagrams, as in Fig. 1.6. For a further example, note that we can apply property (c) repeatedly and obtain the inequality

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbf{P}(A_i).$$

In more detail, let us apply property (c) to the sets A_1 and $A_2 \cup \cdots \cup A_n$, to obtain

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup \cdots \cup A_n).$$

We also apply property (c) to the sets A_2 and $A_3 \cup \cdots \cup A_n$ to obtain

$$\mathbf{P}(A_2 \cup \cdots \cup A_n) \leq \mathbf{P}(A_2) + \mathbf{P}(A_3 \cup \cdots \cup A_n),$$

continue similarly, and finally add.

Models and Reality

Using the framework of probability theory to analyze a physical but uncertain situation, involves two distinct stages.

(a) In the first stage, we construct a probabilistic model, by specifying a probability law on a suitably defined sample space. There are no hard rules to

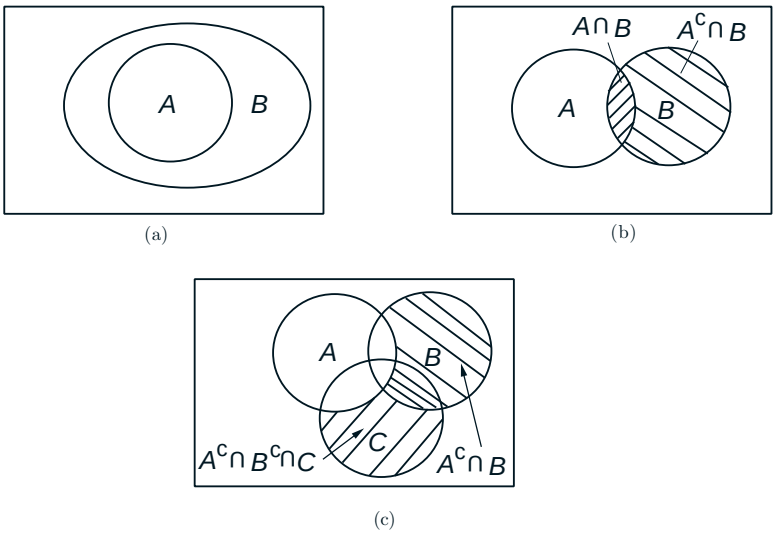

Figure 1.6: Visualization and verification of various properties of probability laws using Venn diagrams. If $A \subset B$, then B is the union of the two disjoint events A and $A^c \cap B$; see diagram (a). Therefore, by the additivity axiom, we have

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B) \geq \mathbf{P}(A),$$

where the inequality follows from the nonnegativity axiom, and verifies property (a).

From diagram (b), we can express the events $A \cup B$ and B as unions of disjoint events:

$$A \cup B = A \cup (A^c \cap B), \qquad B = (A \cap B) \cup (A^c \cap B).$$

The additivity axiom yields

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(A^c \cap B), \qquad \mathbf{P}(B) = \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B).$$

Subtracting the second equality from the first and rearranging terms, we obtain $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, verifying property (b). Using also the fact $\mathbf{P}(A \cap B) \geq 0$ (the nonnegativity axiom), we obtain $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$, verifying property (c)

From diagram (c), we see that the event $A \cup B \cup C$ can be expressed as a union of three disjoint events:

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

so property (d) follows as a consequence of the additivity axiom.

guide this step, other than the requirement that the probability law conform to the three axioms. Reasonable people may disagree on which model best represents reality. In many cases, one may even want to use a somewhat "incorrect" model, if it is simpler than the "correct" one or allows for tractable calculations. This is consistent with common practice in science and engineering, where the choice of a model often involves a tradeoff between accuracy, simplicity, and tractability. Sometimes, a model is chosen on the basis of historical data or past outcomes of similar experiments. Systematic methods for doing so belong to the field of **statistics**, a topic that we will touch upon in the last chapter of this book.

(b) In the second stage, we work within a fully specified probabilistic model and derive the probabilities of certain events, or deduce some interesting properties. While the first stage entails the often open-ended task of connecting the real world with mathematics, the second one is tightly regulated by the rules of ordinary logic and the axioms of probability. Difficulties may arise in the latter if some required calculations are complex, or if a probability law is specified in an indirect fashion. Even so, there is no room for ambiguity: all conceivable questions have precise answers and it is only a matter of developing the skill to arrive at them.

Probability theory is full of "paradoxes" in which different calculation methods seem to give different answers to the same question. Invariably though, these apparent inconsistencies turn out to reflect poorly specified or ambiguous probabilistic models.

1.3 CONDITIONAL PROBABILITY

Conditional probability provides us with a way to reason about the outcome of an experiment, based on **partial information**. Here are some examples of situations we have in mind:

(a) In an experiment involving two successive rolls of a die, you are told that the sum of the two rolls is 9. How likely is it that the first roll was a 6?

(b) In a word guessing game, the first letter of the word is a "t". What is the likelihood that the second letter is an "h"?

(c) How likely is it that a person has a disease given that a medical test was negative?

(d) A spot shows up on a radar screen. How likely is it that it corresponds to an aircraft?

In more precise terms, given an experiment, a corresponding sample space, and a probability law, suppose that we know that the outcome is within some given event B. We wish to quantify the likelihood that the outcome also belongs

to some other given event A. We thus seek to construct a new probability law, which takes into account this knowledge and which, for any event A, gives us the **conditional probability of A given B**, denoted by $\mathbf{P}(A \mid B)$.

We would like the conditional probabilities $\mathbf{P}(A \mid B)$ of different events A to constitute a legitimate probability law, that satisfies the probability axioms. They should also be consistent with our intuition in important special cases, e.g., when all possible outcomes of the experiment are equally likely. For example, suppose that all six possible outcomes of a fair die roll are equally likely. If we are told that the outcome is even, we are left with only three possible outcomes, namely, 2, 4, and 6. These three outcomes were equally likely to start with, and so they should remain equally likely given the additional knowledge that the outcome was even. Thus, it is reasonable to let

$$\mathbf{P}(\text{the outcome is } 6 \mid \text{the outcome is even}) = \frac{1}{3}.$$

This argument suggests that an appropriate definition of conditional probability when all outcomes are equally likely, is given by

$$\mathbf{P}(A \mid B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Generalizing the argument, we introduce the following definition of conditional probability:

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

where we assume that $\mathbf{P}(B) > 0$; the conditional probability is undefined if the conditioning event has zero probability. In words, out of the total probability of the elements of B, $\mathbf{P}(A \mid B)$ is the fraction that is assigned to possible outcomes that also belong to A.

Conditional Probabilities Specify a Probability Law

For a fixed event B, it can be verified that the conditional probabilities $\mathbf{P}(A \mid B)$ form a legitimate probability law that satisfies the three axioms. Indeed, nonnegativity is clear. Furthermore,

$$\mathbf{P}(\Omega \mid B) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1,$$

and the normalization axiom is also satisfied. In fact, since we have $\mathbf{P}(B \mid B) = \mathbf{P}(B)/\mathbf{P}(B) = 1$, all of the conditional probability is concentrated on B. Thus, we might as well discard all possible outcomes outside B and treat the conditional probabilities as a probability law defined on the new universe B.

To verify the additivity axiom, we write for any two disjoint events A_1 and A_2,

$$\begin{aligned}\mathbf{P}(A_1 \cup A_2 \,|\, B) &= \frac{\mathbf{P}\big((A_1 \cup A_2) \cap B\big)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}((A_1 \cap B) \cup (A_2 \cap B))}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1 \cap B) + \mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} + \frac{\mathbf{P}(A_2 \cap B)}{\mathbf{P}(B)} \\ &= \mathbf{P}(A_1 \,|\, B) + \mathbf{P}(A_2 \,|\, B),\end{aligned}$$

where for the second equality, we used the fact that $A_1 \cap B$ and $A_2 \cap B$ are disjoint sets, and for the third equality we used the additivity axiom for the (unconditional) probability law. The argument for a countable collection of disjoint sets is similar.

Since conditional probabilities constitute a legitimate probability law, all general properties of probability laws remain valid. For example, a fact such as $\mathbf{P}(A \cup C) \leq \mathbf{P}(A) + \mathbf{P}(C)$ translates to the new fact

$$\mathbf{P}(A \cup C \,|\, B) \leq \mathbf{P}(A \,|\, B) + \mathbf{P}(C \,|\, B).$$

Let us summarize the conclusions reached so far.

Properties of Conditional Probability

- The conditional probability of an event A, given an event B with $\mathbf{P}(B) > 0$, is defined by

$$\mathbf{P}(A \,|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)},$$

  and specifies a new (conditional) probability law on the same sample space Ω. In particular, all known properties of probability laws remain valid for conditional probability laws.

- Conditional probabilities can also be viewed as a probability law on a new universe B, because all of the conditional probability is concentrated on B.

- In the case where the possible outcomes are finitely many and equally likely, we have

$$\mathbf{P}(A \,|\, B) = \frac{\text{number of elements of } A \cap B}{\text{number of elements of } B}.$$

Example 1.6. We toss a fair coin three successive times. We wish to find the conditional probability $\mathbf{P}(A \mid B)$ when A and B are the events

$$A = \{\text{more heads than tails come up}\}, \qquad B = \{\text{1st toss is a head}\}.$$

The sample space consists of eight sequences,

$$\Omega = \{HHH,\, HHT,\, HTH,\, HTT,\, THH,\, THT,\, TTH,\, TTT\},$$

which we assume to be equally likely. The event B consists of the four elements HHH, HHT, HTH, HTT, so its probability is

$$\mathbf{P}(B) = \frac{4}{8}.$$

The event $A \cap B$ consists of the three elements outcomes HHH, HHT, HTH, so its probability is

$$\mathbf{P}(A \cap B) = \frac{3}{8}.$$

Thus, the conditional probability $\mathbf{P}(A \mid B)$ is

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{3/8}{4/8} = \frac{3}{4}.$$

Because all possible outcomes are equally likely here, we can also compute $\mathbf{P}(A \mid B)$ using a shortcut. We can bypass the calculation of $\mathbf{P}(B)$ and $\mathbf{P}(A \cap B)$, and simply divide the number of elements shared by A and B (which is 3) with the number of elements of B (which is 4), to obtain the same result 3/4.

Example 1.7. A fair 4-sided die is rolled twice and we assume that all sixteen possible outcomes are equally likely. Let X and Y be the result of the 1st and the 2nd roll, respectively. We wish to determine the conditional probability $\mathbf{P}(A \mid B)$ where

$$A = \big\{\max(X, Y) = m\big\}, \qquad B = \big\{\min(X, Y) = 2\big\},$$

and m takes each of the values 1, 2, 3, 4.

As in the preceding example, we can first determine the probabilities $\mathbf{P}(A \cap B)$ and $\mathbf{P}(B)$ by counting the number of elements of $A \cap B$ and B, respectively, and dividing by 16. Alternatively, we can directly divide the number of elements of $A \cap B$ with the number of elements of B; see Fig. 1.7.

Example 1.8. A conservative design team, call it C, and an innovative design team, call it N, are asked to separately design a new product within a month. From past experience we know that:

(a) The probability that team C is successful is 2/3.

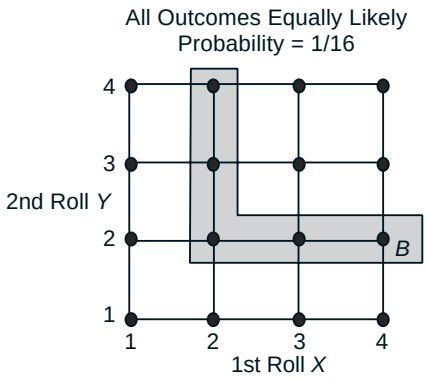

Figure 1.7: Sample space of an experiment involving two rolls of a 4-sided die. (cf. Example 1.7). The conditioning event $B = \{\min(X, Y) = 2\}$ consists of the 5-element shaded set. The set $A = \{\max(X, Y) = m\}$ shares with B two elements if $m = 3$ or $m = 4$, one element if $m = 2$, and no element if $m = 1$. Thus, we have

$$\mathbf{P}\big(\{\max(X, Y) = m\} \,|\, B\big) = \begin{cases} 2/5 & \text{if } m = 3 \text{ or } m = 4, \\ 1/5 & \text{if } m = 2, \\ 0 & \text{if } m = 1. \end{cases}$$

(b) The probability that team N is successful is 1/2.

(c) The probability that at least one team is successful is 3/4.

If both teams are successful, the design of team N is adopted. Assuming that exactly one successful design is produced, what is the probability that it was designed by team N?

There are four possible outcomes here, corresponding to the four combinations of success and failure of the two teams:

SS: both succeed, FF: both fail,
SF: C succeeds, N fails, FS: C fails, N succeeds.

We are given that the probabilities of these outcomes satisfy

$$\mathbf{P}(SS) + \mathbf{P}(SF) = \frac{2}{3}, \quad \mathbf{P}(SS) + \mathbf{P}(FS) = \frac{1}{2}, \quad \mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) = \frac{3}{4}.$$

From these relations, together with the normalization equation $\mathbf{P}(SS) + \mathbf{P}(SF) + \mathbf{P}(FS) + \mathbf{P}(FF) = 1$, we can obtain the probabilities of all the outcomes:

$$\mathbf{P}(SS) = \frac{5}{12}, \qquad \mathbf{P}(SF) = \frac{1}{4}, \qquad \mathbf{P}(FS) = \frac{1}{12}, \qquad \mathbf{P}(FF) = \frac{1}{4}.$$

The desired conditional probability is

$$\mathbf{P}\big(\{FS\} \,|\, \{SF, FS\}\big) = \frac{\dfrac{1}{12}}{\dfrac{1}{4} + \dfrac{1}{12}} = \frac{1}{4}.$$

Using Conditional Probability for Modeling

When constructing probabilistic models for experiments that have a sequential character, it is often natural and convenient to first specify conditional probabilities and then use them to determine unconditional probabilities. The rule $\mathbf{P}(A \cap B) = \mathbf{P}(B)\mathbf{P}(A \mid B)$, which is a restatement of the definition of conditional probability, is often helpful in this process.

Example 1.9. Radar detection. If an aircraft is present in a certain area, a radar correctly registers its presence with probability 0.99. If it is not present, the radar falsely registers an aircraft presence with probability 0.10. We assume that an aircraft is present with probability 0.05. What is the probability of false alarm (a false indication of aircraft presence), and the probability of missed detection (nothing registers, even though an aircraft is present)?

A sequential representation of the sample space is appropriate here, as shown in Fig. 1.8. Let A and B be the events

$$A = \{\text{an aircraft is present}\},$$

$$B = \{\text{the radar registers an aircraft presence}\},$$

and consider also their complements

$$A^c = \{\text{an aircraft is not present}\},$$

$$B^c = \{\text{the radar does not register an aircraft presence}\}.$$

The given probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.8. Each event of interest corresponds to a leaf of the tree and its probability is equal to the product of the probabilities associated with the branches in a path from the root to the corresponding leaf. The desired probabilities of false alarm and missed detection are

$$\mathbf{P}(\text{false alarm}) = \mathbf{P}(A^c \cap B) = \mathbf{P}(A^c)\mathbf{P}(B \mid A^c) = 0.95 \cdot 0.10 = 0.095,$$

$$\mathbf{P}(\text{missed detection}) = \mathbf{P}(A \cap B^c) = \mathbf{P}(A)\mathbf{P}(B^c \mid A) = 0.05 \cdot 0.01 = 0.0005.$$

Extending the preceding example, we have a general rule for calculating various probabilities in conjunction with a tree-based sequential description of an experiment. In particular:

(a) We set up the tree so that an event of interest is associated with a leaf. We view the occurrence of the event as a sequence of steps, namely, the traversals of the branches along the path from the root to the leaf.

(b) We record the conditional probabilities associated with the branches of the tree.

(c) We obtain the probability of a leaf by multiplying the probabilities recorded along the corresponding path of the tree.

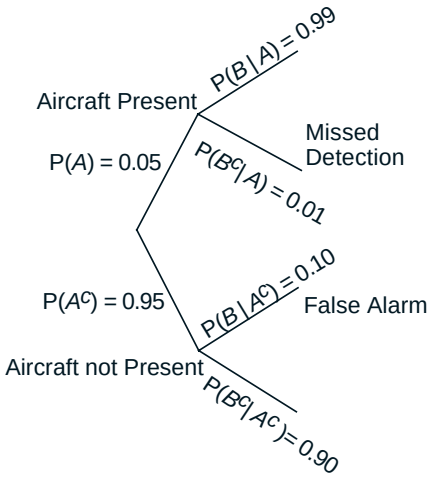

Figure 1.8: Sequential description of the sample space for the radar detection problem in Example 1.9

In mathematical terms, we are dealing with an event A which occurs if and only if each one of several events $A_1, \ldots, A_n$ has occurred, i.e., $A = A_1 \cap A_2 \cap \cdots \cap A_n$. The occurrence of A is viewed as an occurrence of A_1, followed by the occurrence of A_2, then of A_3, etc, and it is visualized as a path on the tree with n branches, corresponding to the events $A_1, \ldots, A_n$. The probability of A is given by the following rule (see also Fig. 1.9).

Multiplication Rule

Assuming that all of the conditioning events have positive probability, we have

$$\mathbf{P}\big(\cap_{i=1}^{n} A_i\big) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2) \cdots \mathbf{P}\big(A_n \,|\, \cap_{i=1}^{n-1} A_i\big).$$

The multiplication rule can be verified by writing

$$\mathbf{P}\big(\cap_{i=1}^{n} A_i\big) = \mathbf{P}(A_1)\frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)}\frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)} \cdots \frac{\mathbf{P}\big(\cap_{i=1}^{n} A_i\big)}{\mathbf{P}\big(\cap_{i=1}^{n-1} A_i\big)},$$

and by using the definition of conditional probability to rewrite the right-hand side above as

$$\mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2) \cdots \mathbf{P}\big(A_n \,|\, \cap_{i=1}^{n-1} A_i\big).$$

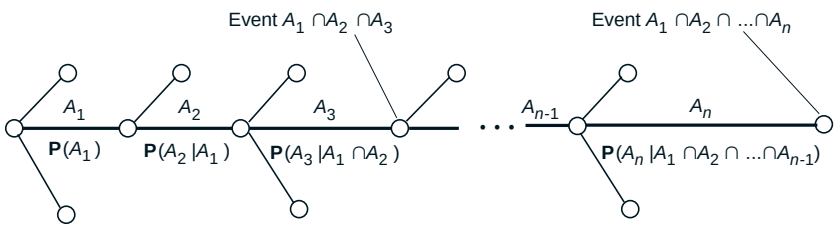

Figure 1.9: Visualization of the total probability theorem. The intersection event $A = A_1 \cap A_2 \cap \cdots \cap A_n$ is associated with a path on the tree of a sequential description of the experiment. We associate the branches of this path with the events $A_1, \ldots, A_n$, and we record next to the branches the corresponding conditional probabilities.

The final node of the path corresponds to the intersection event A, and its probability is obtained by multiplying the conditional probabilities recorded along the branches of the path

$$\mathbf{P}(A_1 \cap A_2 \cap \cdots \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1) \cdots \mathbf{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Note that any intermediate node along the path also corresponds to some intersection event and its probability is obtained by multiplying the corresponding conditional probabilities up to that node. For example, the event $A_1 \cap A_2 \cap A_3$ corresponds to the node shown in the figure, and its probability is

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2).$$

For the case of just two events, A_1 and A_2, the multiplication rule is simply the definition of conditional probability.

Example 1.10. Three cards are drawn from an ordinary 52-card deck without replacement (drawn cards are not placed back in the deck). We wish to find the probability that none of the three cards is a heart. We assume that at each step, each one of the remaining cards is equally likely to be picked. By symmetry, this implies that every triplet of cards is equally likely to be drawn. A cumbersome approach, that we will not use, is to count the number of all card triplets that do not include a heart, and divide it with the number of all possible card triplets. Instead, we use a sequential description of the sample space in conjunction with the multiplication rule (cf. Fig. 1.10).

Define the events

$$A_i = \{\text{the } i\text{th card is not a heart}\}, \qquad i = 1, 2, 3.$$

We will calculate $\mathbf{P}(A_1 \cap A_2 \cap A_3)$, the probability that none of the three cards is a heart, using the multiplication rule,

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \mid A_1)\mathbf{P}(A_3 \mid A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{39}{52},$$

since there are 39 cards that are not hearts in the 52-card deck. Given that the first card is not a heart, we are left with 51 cards, 38 of which are not hearts, and

$$\mathbf{P}(A_2 \,|\, A_1) = \frac{38}{51}.$$

Finally, given that the first two cards drawn are not hearts, there are 37 cards which are not hearts in the remaining 50-card deck, and

$$\mathbf{P}(A_3 \,|\, A_1 \cap A_2) = \frac{37}{50}.$$

These probabilities are recorded along the corresponding branches of the tree describing the sample space, as shown in Fig. 1.10. The desired probability is now obtained by multiplying the probabilities recorded along the corresponding path of the tree:

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{37}{50}.$$

Note that once the probabilities are recorded along the tree, the probability of several other events can be similarly calculated. For example,

$$\begin{aligned}
\mathbf{P}(\text{1st is not a heart and 2nd is a heart}) &= \frac{39}{52} \cdot \frac{13}{51}, \\
\mathbf{P}(\text{1st two are not hearts and 3rd is a heart}) &= \frac{39}{52} \cdot \frac{38}{51} \cdot \frac{13}{50}.
\end{aligned}$$

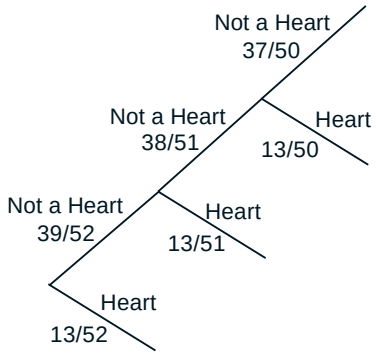

Figure 1.10: Sequential description of the sample space of the 3-card selection problem in Example 1.10.

Example 1.11. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into 4 groups of 4. What is the probability that each group includes a graduate student? We interpret randomly to mean that given the assignment of some students to certain slots, any of the remaining students is equally likely to be assigned to any of the remaining slots. We then calculate the desired probability using the multiplication rule, based on the sequential description shown in Fig. 1.11. Let us denote the four graduate students by 1, 2, 3, 4, and consider the events

$$A_1 = \{\text{students 1 and 2 are in different groups}\},$$

$$A_2 = \{\text{students 1, 2, and 3 are in different groups}\},$$

$$A_3 = \{\text{students 1, 2, 3, and 4 are in different groups}\}.$$

We will calculate $\mathbf{P}(A_3)$ using the multiplication rule:

$$\mathbf{P}(A_3) = \mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2 \,|\, A_1)\mathbf{P}(A_3 \,|\, A_1 \cap A_2).$$

We have

$$\mathbf{P}(A_1) = \frac{12}{15},$$

since there are 12 student slots in groups other than the one of student 1, and there are 15 student slots overall, excluding student 1. Similarly,

$$\mathbf{P}(A_2 \,|\, A_1) = \frac{8}{14},$$

since there are 8 student slots in groups other than the one of students 1 and 2, and there are 14 student slots, excluding students 1 and 2. Also,

$$\mathbf{P}(A_3 \,|\, A_1 \cap A_2) = \frac{4}{13},$$

since there are 4 student slots in groups other than the one of students 1, 2, and 3, and there are 13 student slots, excluding students 1, 2, and 3. Thus, the desired probability is

$$\frac{12}{15} \cdot \frac{8}{14} \cdot \frac{4}{13},$$

and is obtained by multiplying the conditional probabilities along the corresponding path of the tree of Fig. 1.11.

1.4 TOTAL PROBABILITY THEOREM AND BAYES' RULE

In this section, we explore some applications of conditional probability. We start with the following theorem, which is often useful for computing the probabilities of various events, using a "divide-and-conquer" approach.

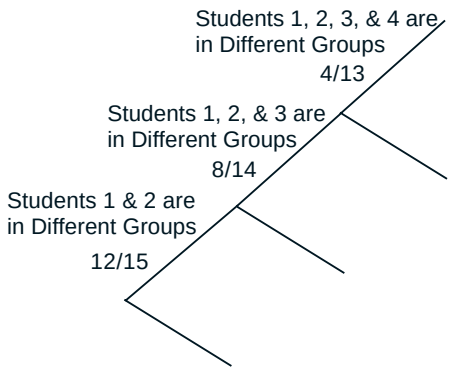

Figure 1.11: Sequential description of the sample space of the student problem in Example 1.11.

Total Probability Theorem

Let $A_1, \ldots, A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in one and only one of the events $A_1, \ldots, A_n$) and assume that $\mathbf{P}(A_i) > 0$, for all $i = 1, \ldots, n$. Then, for any event B, we have

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B) \\ &= \mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \mid A_n).\end{aligned}$$

The theorem is visualized and proved in Fig. 1.12. Intuitively, we are partitioning the sample space into a number of scenarios (events) A_i. Then, the probability that B occurs is a weighted average of its conditional probability under each scenario, where each scenario is weighted according to its (unconditional) probability. One of the uses of the theorem is to compute the probability of various events B for which the conditional probabilities $\mathbf{P}(B \mid A_i)$ are known or easy to derive. The key is to choose appropriately the partition $A_1, \ldots, A_n$, and this choice is often suggested by the problem structure. Here are some examples.

Example 1.12. You enter a chess tournament where your probability of winning a game is 0.3 against half the players (call them type 1), 0.4 against a quarter of the players (call them type 2), and 0.5 against the remaining quarter of the players (call them type 3). You play a game against a randomly chosen opponent. What is the probability of winning?

Let A_i be the event of playing with an opponent of type i. We have

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

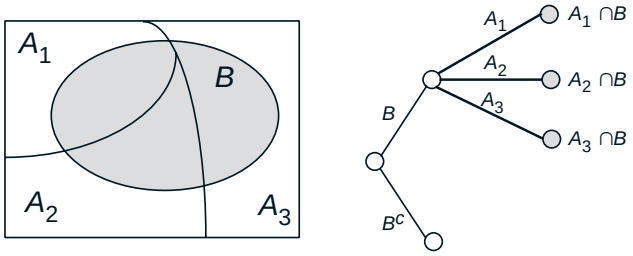

Figure 1.12: Visualization and verification of the total probability theorem. The events $A_1, \ldots, A_n$ form a partition of the sample space, so the event B can be decomposed into the disjoint union of its intersections $A_i \cap B$ with the sets A_i, i.e.,

$$B = (A_1 \cap B) \cup \cdots \cup (A_n \cap B).$$

Using the additivity axiom, it follows that

$$\mathbf{P}(B) = \mathbf{P}(A_1 \cap B) + \cdots + \mathbf{P}(A_n \cap B).$$

Since, by the definition of conditional probability, we have

$$\mathbf{P}(A_i \cap B) = \mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i),$$

the preceding equality yields

$$\mathbf{P}(B) = \mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \,|\, A_n).$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability of the leaf $A_i \cap B$ is the product $\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)$ of the probabilities along the path leading to that leaf. The event B consists of the three highlighted leaves and $\mathbf{P}(B)$ is obtained by adding their probabilities.

Let also B be the event of winning. We have

$$\mathbf{P}(B \,|\, A_1) = 0.3, \qquad \mathbf{P}(B \,|\, A_2) = 0.4, \qquad \mathbf{P}(B \,|\, A_3) = 0.5.$$

Thus, by the total probability theorem, the probability of winning is

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \mathbf{P}(A_2)\mathbf{P}(B \,|\, A_2) + \mathbf{P}(A_3)\mathbf{P}(B \,|\, A_3) \\ &= 0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5 \\ &= 0.375. \end{aligned}$$

Example 1.13. We roll a fair four-sided die. If the result is 1 or 2, we roll once more but otherwise, we stop. What is the probability that the sum total of our rolls is at least 4?

Let A_i be the event that the result of first roll is i, and note that $\mathbf{P}(A_i) = 1/4$ for each i. Let B be the event that the sum total is at least 4. Given the event A_1, the sum total will be at least 4 if the second roll results in 3 or 4, which happens with probability 1/2. Similarly, given the event A_2, the sum total will be at least 4 if the second roll results in 2, 3, or 4, which happens with probability 3/4. Also, given the event A_3, we stop and the sum total remains below 4. Therefore,

$$\mathbf{P}(B \,|\, A_1) = \frac{1}{2}, \qquad \mathbf{P}(B \,|\, A_2) = \frac{3}{4}, \qquad \mathbf{P}(B \,|\, A_3) = 0, \qquad \mathbf{P}(B \,|\, A_4) = 1.$$

By the total probability theorem,

$$\mathbf{P}(B) = \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 = \frac{9}{16}.$$

The total probability theorem can be applied repeatedly to calculate probabilities in experiments that have a sequential character, as shown in the following example.

Example 1.14. Alice is taking a probability class and at the end of each week she can be either up-to-date or she may have fallen behind. If she is up-to-date in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.8 (or 0.2, respectively). If she is behind in a given week, the probability that she will be up-to-date (or behind) in the next week is 0.6 (or 0.4, respectively). Alice is (by default) up-to-date when she starts the class. What is the probability that she is up-to-date after three weeks?

Let U_i and B_i be the events that Alice is up-to-date or behind, respectively, after i weeks. According to the total probability theorem, the desired probability $\mathbf{P}(U_3)$ is given by

$$\mathbf{P}(U_3) = \mathbf{P}(U_2)\mathbf{P}(U_3 \,|\, U_2) + \mathbf{P}(B_2)\mathbf{P}(U_3 \,|\, B_2) = \mathbf{P}(U_2) \cdot 0.8 + \mathbf{P}(B_2) \cdot 0.4.$$

The probabilities $\mathbf{P}(U_2)$ and $\mathbf{P}(B_2)$ can also be calculated using the total probability theorem:

$$\mathbf{P}(U_2) = \mathbf{P}(U_1)\mathbf{P}(U_2 \,|\, U_1) + \mathbf{P}(B_1)\mathbf{P}(U_2 \,|\, B_1) = \mathbf{P}(U_1) \cdot 0.8 + \mathbf{P}(B_1) \cdot 0.4,$$

$$\mathbf{P}(B_2) = \mathbf{P}(U_1)\mathbf{P}(B_2 \,|\, U_1) + \mathbf{P}(B_1)\mathbf{P}(B_2 \,|\, B_1) = \mathbf{P}(U_1) \cdot 0.2 + \mathbf{P}(B_1) \cdot 0.6.$$

Finally, since Alice starts her class up-to-date, we have

$$\mathbf{P}(U_1) = 0.8, \qquad \mathbf{P}(B_1) = 0.2.$$

We can now combine the preceding three equations to obtain

$$\mathbf{P}(U_2) = 0.8 \cdot 0.8 + 0.2 \cdot 0.4 = 0.72,$$

$$\mathbf{P}(B_2) = 0.8 \cdot 0.2 + 0.2 \cdot 0.6 = 0.28.$$

and by using the above probabilities in the formula for $\mathbf{P}(U_3)$:

$$\mathbf{P}(U_3) = 0.72 \cdot 0.8 + 0.28 \cdot 0.4 = 0.688.$$

Note that we could have calculated the desired probability $\mathbf{P}(U_3)$ by constructing a tree description of the experiment, by calculating the probability of every element of U_3 using the multiplication rule on the tree, and by adding. In experiments with a sequential character one may often choose between using the multiplication rule or the total probability theorem for calculation of various probabilities. However, there are cases where the calculation based on the total probability theorem is more convenient. For example, suppose we are interested in the probability $\mathbf{P}(U_{20})$ that Alice is up-to-date after 20 weeks. Calculating this probability using the multiplication rule is very cumbersome, because the tree representing the experiment is 20-stages deep and has 2^{20} leaves. On the other hand, with a computer, a sequential caclulation using the total probability formulas

$$\mathbf{P}(U_{i+1}) = \mathbf{P}(U_i) \cdot 0.8 + \mathbf{P}(B_i) \cdot 0.4,$$

$$\mathbf{P}(B_{i+1}) = \mathbf{P}(U_i) \cdot 0.2 + \mathbf{P}(B_i) \cdot 0.6,$$

and the initial conditions $\mathbf{P}(U_1) = 0.8$, $\mathbf{P}(B_1) = 0.2$ is very simple.

The total probability theorem is often used in conjunction with the following celebrated theorem, which relates conditional probabilities of the form $\mathbf{P}(A \,|\, B)$ with conditional probabilities of the form $\mathbf{P}(B \,|\, A)$, in which the order of the conditioning is reversed.

Bayes' Rule

Let $A_1, A_2, \ldots, A_n$ be disjoint events that form a partition of the sample space, and assume that $\mathbf{P}(A_i) > 0$, for all i. Then, for any event B such that $\mathbf{P}(B) > 0$, we have

$$\begin{aligned}\mathbf{P}(A_i \,|\, B) &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \cdots + \mathbf{P}(A_n)\mathbf{P}(B \,|\, A_n)}.\end{aligned}$$

To verify Bayes' rule, note that $\mathbf{P}(A_i)\mathbf{P}(B \,|\, A_i)$ and $\mathbf{P}(A_i \,|\, B)\mathbf{P}(B)$ are equal, because they are both equal to $\mathbf{P}(A_i \cap B)$. This yields the first equality. The second equality follows from the first by using the total probability theorem to rewrite $\mathbf{P}(B)$.

Bayes' rule is often used for **inference**. There are a number of "causes" that may result in a certain "effect." We observe the effect, and we wish to infer

the cause. The events $A_1, \ldots, A_n$ are associated with the causes and the event B represents the effect. The probability $\mathbf{P}(B \mid A_i)$ that the effect will be observed when the cause A_i is present amounts to a probabilistic model of the cause-effect relation (cf. Fig. 1.13). Given that the effect B has been observed, we wish to evaluate the (conditional) probability $\mathbf{P}(A_i \mid B)$ that the cause A_i is present.

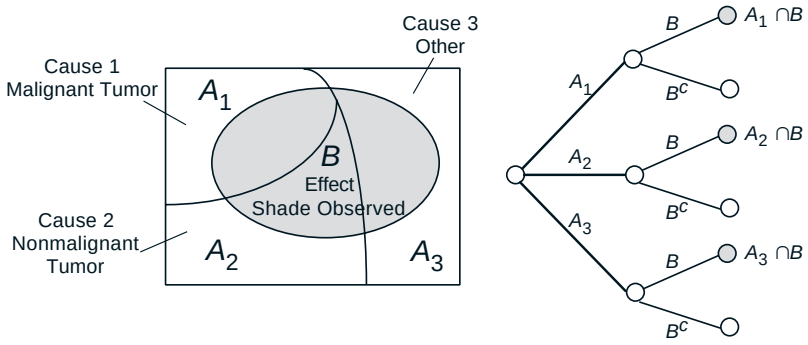

Figure 1.13: An example of the inference context that is implicit in Bayes' rule. We observe a shade in a person's X-ray (this is event B, the "effect") and we want to estimate the likelihood of three mutually exclusive and collectively exhaustive potential causes: cause 1 (event A_1) is that there is a malignant tumor, cause 2 (event A_2) is that there is a nonmalignant tumor, and cause 3 (event A_3) corresponds to reasons other than a tumor. We assume that we know the probabilities $\mathbf{P}(A_i)$ and $\mathbf{P}(B \mid A_i)$, $i = 1, 2, 3$. Given that we see a shade (event B occurs), Bayes' rule gives the conditional probabilities of the various causes as

$$\mathbf{P}(A_i \mid B) = \frac{\mathbf{P}(A_i)\mathbf{P}(B \mid A_i)}{\mathbf{P}(A_1)\mathbf{P}(B \mid A_1) + \mathbf{P}(A_2)\mathbf{P}(B \mid A_2) + \mathbf{P}(A_3)\mathbf{P}(B \mid A_3)}, \quad i = 1, 2, 3.$$

For an alternative view, consider an equivalent sequential model, as shown on the right. The probability $\mathbf{P}(A_1 \mid B)$ of a malignant tumor is the probability of the first highlighted leaf, which is $\mathbf{P}(A_1 \cap B)$, divided by the total probability of the highlighted leaves, which is $\mathbf{P}(B)$.

Example 1.15. Let us return to the radar detection problem of Example 1.9 and Fig. 1.8. Let

$$A = \{\text{an aircraft is present}\},$$
$$B = \{\text{the radar registers an aircraft presence}\}.$$

We are given that

$$\mathbf{P}(A) = 0.05, \qquad \mathbf{P}(B \mid A) = 0.99, \qquad \mathbf{P}(B \mid A^c) = 0.1.$$

Applying Bayes' rule, with $A_1 = A$ and $A_2 = A^c$, we obtain

$$\begin{aligned}\mathbf{P}(\text{aircraft present} \,|\, \text{radar registers}) &= \mathbf{P}(A \,|\, B) \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(B)} \\ &= \frac{\mathbf{P}(A)\mathbf{P}(B \,|\, A)}{\mathbf{P}(A)\mathbf{P}(B \,|\, A) + \mathbf{P}(A^c)\mathbf{P}(B \,|\, A^c)} \\ &= \frac{0.05 \cdot 0.99}{0.05 \cdot 0.99 + 0.95 \cdot 0.1} \\ &\approx 0.3426.\end{aligned}$$

Example 1.16. Let us return to the chess problem of Example 1.12. Here A_i is the event of getting an opponent of type i, and

$$\mathbf{P}(A_1) = 0.5, \qquad \mathbf{P}(A_2) = 0.25, \qquad \mathbf{P}(A_3) = 0.25.$$

Also, B is the event of winning, and

$$\mathbf{P}(B \,|\, A_1) = 0.3, \qquad \mathbf{P}(B \,|\, A_2) = 0.4, \qquad \mathbf{P}(B \,|\, A_3) = 0.5.$$

Suppose that you win. What is the probability $\mathbf{P}(A_1 \,|\, B)$ that you had an opponent of type 1?

Using Bayes' rule, we have

$$\begin{aligned}\mathbf{P}(A_1 \,|\, B) &= \frac{\mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1)}{\mathbf{P}(A_1)\mathbf{P}(B \,|\, A_1) + \mathbf{P}(A_2)\mathbf{P}(B \,|\, A_2) + \mathbf{P}(A_3)\mathbf{P}(B \,|\, A_3)} \\ &= \frac{0.5 \cdot 0.3}{0.5 \cdot 0.3 + 0.25 \cdot 0.4 + 0.25 \cdot 0.5} \\ &= 0.4.\end{aligned}$$

1.5 INDEPENDENCE

We have introduced the conditional probability $\mathbf{P}(A \,|\, B)$ to capture the partial information that event B provides about event A. An interesting and important special case arises when the occurrence of B provides no information and does not alter the probability that A has occurred, i.e.,

$$\mathbf{P}(A \,|\, B) = \mathbf{P}(A).$$

When the above equality holds, we say that A is **independent** of B. Note that by the definition $\mathbf{P}(A \mid B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, this is equivalent to

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

We adopt this latter relation as the definition of independence because it can be used even if $\mathbf{P}(B) = 0$, in which case $\mathbf{P}(A \mid B)$ is undefined. The symmetry of this relation also implies that independence is a symmetric property; that is, if A is independent of B, then B is independent of A, and we can unambiguously say that A and B are **independent events**.

Independence is often easy to grasp intuitively. For example, if the occurrence of two events is governed by distinct and noninteracting physical processes, such events will turn out to be independent. On the other hand, independence is not easily visualized in terms of the sample space. A common first thought is that two events are independent if they are disjoint, but in fact the opposite is true: two disjoint events A and B with $\mathbf{P}(A) > 0$ and $\mathbf{P}(B) > 0$ are never independent, since their intersection $A \cap B$ is empty and has probability 0.

Example 1.17. Consider an experiment involving two successive rolls of a 4-sided die in which all 16 possible outcomes are equally likely and have probability 1/16.

(a) Are the events

$$A_i = \{\text{1st roll results in } i\}, \qquad B_j = \{\text{2nd roll results in } j\},$$

independent? We have

$$\mathbf{P}(A \cap B) = \mathbf{P}\big(\text{the result of the two rolls is } (i,j)\big) = \frac{1}{16},$$
$$\mathbf{P}(A_i) = \frac{\text{number of elements of } A_i}{\text{total number of possible outcomes}} = \frac{4}{16},$$
$$\mathbf{P}(B_j) = \frac{\text{number of elements of } B_j}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

We observe that $\mathbf{P}(A_i \cap B_j) = \mathbf{P}(A_i)\mathbf{P}(B_j)$, and the independence of A_i and B_j is verified. Thus, our choice of the discrete uniform probability law (which might have seemed arbitrary) models the independence of the two rolls.

(b) Are the events

$$A = \{\text{1st roll is a 1}\}, \qquad B = \{\text{sum of the two rolls is a 5}\},$$

independent? The answer here is not quite obvious. We have

$$\mathbf{P}(A \cap B) = \mathbf{P}\big(\text{the result of the two rolls is (1,4)}\big) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

The event B consists of the outcomes (1,4), (2,3), (3,2), and (4,1), and

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{4}{16}.$$

Thus, we see that $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, and the events A and B are independent.

(c) Are the events

$$A = \{\text{maximum of the two rolls is 2}\}, \quad B = \{\text{minimum of the two rolls is 2}\},$$

independent? Intuitively, the answer is "no" because the minimum of the two rolls tells us something about the maximum. For example, if the minimum is 2, the maximum cannot be 1. More precisely, to verify that A and B are not independent, we calculate

$$\mathbf{P}(A \cap B) = \mathbf{P}\big(\text{the result of the two rolls is (2,2)}\big) = \frac{1}{16},$$

and also

$$\mathbf{P}(A) = \frac{\text{number of elements of } A}{\text{total number of possible outcomes}} = \frac{3}{16},$$

$$\mathbf{P}(B) = \frac{\text{number of elements of } B}{\text{total number of possible outcomes}} = \frac{5}{16}.$$

We have $\mathbf{P}(A)\mathbf{P}(B) = 15/(16)^2$, so that $\mathbf{P}(A \cap B) \neq \mathbf{P}(A)\mathbf{P}(B)$, and A and B are not independent.

Conditional Independence

We noted earlier that the conditional probabilities of events, conditioned on a particular event, form a legitimate probability law. We can thus talk about independence of various events with respect to this conditional law. In particular, given an event C, the events A and B are called **conditionally independent** if

$$\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).$$

The definition of the conditional probability and the multiplication rule yield

$$\begin{aligned}\mathbf{P}(A \cap B \,|\, C) &= \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(C)} \\ &= \frac{\mathbf{P}(C)\mathbf{P}(B \,|\, C)\mathbf{P}(A \,|\, B \cap C)}{\mathbf{P}(C)} \\ &= \mathbf{P}(B \,|\, C)\mathbf{P}(A \,|\, B \cap C).\end{aligned}$$

After canceling the factor $\mathbf{P}(B \mid C)$, assumed nonzero, we see that conditional independence is the same as the condition

$$\mathbf{P}(A \mid B \cap C) = \mathbf{P}(A \mid C).$$

In words, this relation states that if C is known to have occurred, the additional knowledge that B also occurred does not change the probability of A.

Interestingly, independence of two events A and B with respect to the unconditional probability law, does not imply conditional independence, and vice versa, as illustrated by the next two examples.

Example 1.18. Consider two independent fair coin tosses, in which all four possible outcomes are equally likely. Let

$$\begin{aligned} H_1 &= \{\text{1st toss is a head}\}, \\ H_2 &= \{\text{2nd toss is a head}\}, \\ D &= \{\text{the two tosses have different results}\}. \end{aligned}$$

The events H_1 and H_2 are (unconditionally) independent. But

$$\mathbf{P}(H_1 \mid D) = \frac{1}{2}, \qquad \mathbf{P}(H_2 \mid D) = \frac{1}{2}, \qquad \mathbf{P}(H_1 \cap H_2 \mid D) = 0,$$

so that $\mathbf{P}(H_1 \cap H_2 \mid D) \neq \mathbf{P}(H_1 \mid D)\mathbf{P}(H_2 \mid D)$, and H_1, H_2 are not conditionally independent.

Example 1.19. There are two coins, a blue and a red one. We choose one of the two at random, each being chosen with probability 1/2, and proceed with two independent tosses. The coins are biased: with the blue coin, the probability of heads in any given toss is 0.99, whereas for the red coin it is 0.01.

Let B be the event that the blue coin was selected. Let also H_i be the event that the ith toss resulted in heads. Given the choice of a coin, the events H_1 and H_2 are independent, because of our assumption of independent tosses. Thus,

$$\mathbf{P}(H_1 \cap H_2 \mid B) = \mathbf{P}(H_1 \mid B)\mathbf{P}(H_2 \mid B) = 0.99 \cdot 0.99.$$

On the other hand, the events H_1 and H_2 are not independent. Intuitively, if we are told that the first toss resulted in heads, this leads us to suspect that the blue coin was selected, in which case, we expect the second toss to also result in heads. Mathematically, we use the total probability theorem to obtain

$$\mathbf{P}(H_1) = \mathbf{P}(B)\mathbf{P}(H_1 \mid B) + \mathbf{P}(B^c)\mathbf{P}(H_1 \mid B^c) = \frac{1}{2} \cdot 0.99 + \frac{1}{2} \cdot 0.01 = \frac{1}{2},$$

as should be expected from symmetry considerations. Similarly, we have $\mathbf{P}(H_2) = 1/2$. Now notice that

$$\begin{aligned}\mathbf{P}(H_1 \cap H_2) &= \mathbf{P}(B)\mathbf{P}(H_1 \cap H_2 \,|\, B) + \mathbf{P}(B^c)\mathbf{P}(H_1 \cap H_2 \,|\, B^c) \\ &= \frac{1}{2} \cdot 0.99 \cdot 0.99 + \frac{1}{2} \cdot 0.01 \cdot 0.01 \approx \frac{1}{2}.\end{aligned}$$

Thus, $\mathbf{P}(H_1 \cap H_2) \neq \mathbf{P}(H_1)\mathbf{P}(H_2)$, and the events H_1 and H_2 are dependent, even though they are conditionally independent given B.

As mentioned earlier, if A and B are independent, the occurrence of B does not provide any new information on the probability of A occurring. It is then intuitive that the non-occurrence of B should also provide no information on the probability of A. Indeed, it can be verified that if A and B are independent, the same holds true for A and B^c (see the theoretical problems).

We now summarize.

Independence

- Two events A and B are said to independent if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

  If in addition, $\mathbf{P}(B) > 0$, independence is equivalent to the condition

$$\mathbf{P}(A \,|\, B) = \mathbf{P}(A).$$

- If A and B are independent, so are A and B^c.
- Two events A and B are said to be conditionally independent, given another event C with $\mathbf{P}(C) > 0$, if

$$\mathbf{P}(A \cap B \,|\, C) = \mathbf{P}(A \,|\, C)\mathbf{P}(B \,|\, C).$$

  If in addition, $\mathbf{P}(B \cap C) > 0$, conditional independence is equivalent to the condition

$$\mathbf{P}(A \,|\, B \cap C) = \mathbf{P}(A \,|\, C).$$

- Independence does not imply conditional independence, and vice versa.

Independence of a Collection of Events

The definition of independence can be extended to multiple events.

Definition of Independence of Several Events

We say that the events $A_1, A_2, \ldots, A_n$ are **independent** if

$$\mathbf{P}\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} \mathbf{P}(A_i), \qquad \text{for every subset } S \text{ of } \{1, 2, \ldots, n\}.$$

If we have a collection of three events, A_1, A_2, and A_3, independence amounts to satisfying the four conditions

$$\begin{aligned}
\mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(A_1)\,\mathbf{P}(A_2),\\
\mathbf{P}(A_1 \cap A_3) &= \mathbf{P}(A_1)\,\mathbf{P}(A_3),\\
\mathbf{P}(A_2 \cap A_3) &= \mathbf{P}(A_2)\,\mathbf{P}(A_3),\\
\mathbf{P}(A_1 \cap A_2 \cap A_3) &= \mathbf{P}(A_1)\,\mathbf{P}(A_2)\,\mathbf{P}(A_3).
\end{aligned}$$

The first three conditions simply assert that any two events are independent, a property known as **pairwise independence**. But the fourth condition is also important and does not follow from the first three. Conversely, the fourth condition does not imply the first three; see the two examples that follow.

Example 1.20. Pairwise independence does not imply independence. Consider two independent fair coin tosses, and the following events:

$$\begin{aligned}
H_1 &= \{\text{1st toss is a head}\},\\
H_2 &= \{\text{2nd toss is a head}\},\\
D &= \{\text{the two tosses have different results}\}.
\end{aligned}$$

The events H_1 and H_2 are independent, by definition. To see that H_1 and D are independent, we note that

$$\mathbf{P}(D \,|\, H_1) = \frac{\mathbf{P}(H_1 \cap D)}{\mathbf{P}(H_1)} = \frac{1/4}{1/2} = \frac{1}{2} = \mathbf{P}(D).$$

Similarly, H_2 and D are independent. On the other hand, we have

$$\mathbf{P}(H_1 \cap H_2 \cap D) = 0 \neq \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \mathbf{P}(H_1)\mathbf{P}(H_2)\mathbf{P}(D),$$

and these three events are not independent.

Example 1.21. The equality $\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\,\mathbf{P}(A_2)\,\mathbf{P}(A_3)$ is not enough for independence. Consider two independent rolls of a fair die, and

the following events:

$$A = \{\text{1st roll is 1, 2, or 3}\},$$
$$B = \{\text{1st roll is 3, 4, or 5}\},$$
$$C = \{\text{the sum of the two rolls is 9}\}.$$

We have

$$\mathbf{P}(A \cap B) = \frac{1}{6} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbf{P}(A)\mathbf{P}(B),$$
$$\mathbf{P}(A \cap C) = \frac{1}{36} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(A)\mathbf{P}(C),$$
$$\mathbf{P}(B \cap C) = \frac{1}{12} \neq \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(B)\mathbf{P}(C).$$

Thus the three events A, B, and C are not independent, and indeed no two of these events are independent. On the other hand, we have

$$\mathbf{P}(A \cap B \cap C) = \frac{1}{36} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{4}{36} = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).$$

The intuition behind the independence of a collection of events is analogous to the case of two events. Independence means that the occurrence or non-occurrence of **any number** of the events from that collection carries no information on the remaining events or their complements. For example, if the events A_1, A_2, A_3, A_4 are independent, one obtains relations such as

$$\mathbf{P}(A_1 \cup A_2 \,|\, A_3 \cap A_4) = \mathbf{P}(A_1 \cup A_2)$$

or

$$\mathbf{P}(A_1 \cup A_2^c \,|\, A_3^c \cap A_4) = \mathbf{P}(A_1 \cup A_2^c);$$

see the theoretical problems.

Reliability

In probabilistic models of complex systems involving several components, it is often convenient to assume that the components behave "independently" of each other. This typically simplifies the calculations and the analysis, as illustrated in the following example.

Example 1.22. Network connectivity. A computer network connects two nodes A and B through intermediate nodes C, D, E, F, as shown in Fig. 1.14(a). For every pair of directly connected nodes, say i and j, there is a given probability p_{ij} that the link from i to j is up. We assume that link failures are independent

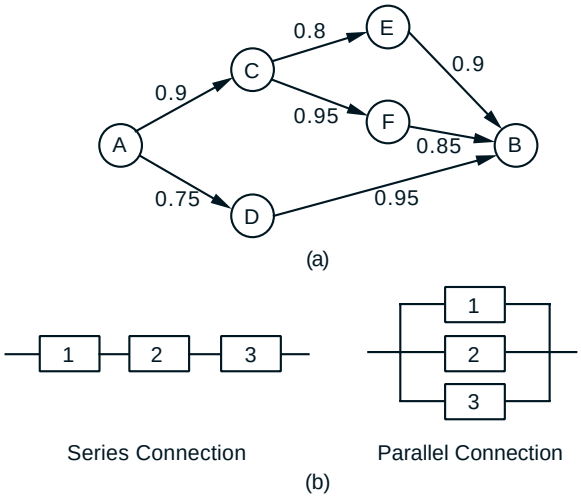

Figure 1.14: (a) Network for Example 1.22. The number next to each link (i, j) indicates the probability that the link is up. (b) Series and parallel connections of three components in a reliability problem.

of each other. What is the probability that there is a path connecting A and B in which all links are up?

This is a typical problem of assessing the reliability of a system consisting of components that can fail independently. Such a system can often be divided into subsystems, where each subsystem consists in turn of several components that are connected either in **series** or in **parallel**; see Fig. 1.14(b).

Let a subsystem consist of components $1, 2, \ldots, m$, and let p_i be the probability that component i is up ("succeeds"). Then, a series subsystem succeeds if **all** of its components are up, so its probability of success is the product of the probabilities of success of the corresponding components, i.e.,

$$\mathbf{P}(\text{series subsystem succeeds}) = p_1 p_2 \cdots p_m.$$

A parallel subsystem succeeds if **any one** of its components succeeds, so its probability of failure is the product of the probabilities of failure of the corresponding components, i.e.,

$$\begin{aligned}\mathbf{P}(\text{parallel subsystem succeeds}) &= 1 - \mathbf{P}(\text{parallel subsystem fails}) \\ &= 1 - (1 - p_1)(1 - p_2) \cdots (1 - p_m).\end{aligned}$$

Returning now to the network of Fig. 1.14(a), we can calculate the probability of success (a path from A to B is available) sequentially, using the preceding formulas, and starting from the end. Let us use the notation $X \to Y$ to denote the

event that there is a (possibly indirect) connection from node X to node Y. Then,

$$\begin{aligned}\mathbf{P}(C \to B) &= 1 - \big(1 - \mathbf{P}(C \to E \text{ and } E \to B)\big)\big(1 - \mathbf{P}(C \to F \text{ and } F \to B)\big) \\ &= 1 - (1 - p_{CE}p_{EB})(1 - p_{CF}p_{FB}) \\ &= 1 - (1 - 0.8 \cdot 0.9)(1 - 0.85 \cdot 0.95) \\ &= 0.946,\end{aligned}$$

$$\mathbf{P}(A \to C \text{ and } C \to B) = \mathbf{P}(A \to C)\mathbf{P}(C \to B) = 0.9 \cdot 0.946 = 0.851,$$

$$\mathbf{P}(A \to D \text{ and } D \to B) = \mathbf{P}(A \to D)\mathbf{P}(D \to B) = 0.75 \cdot 0.95 = 0.712,$$

and finally we obtain the desired probability

$$\begin{aligned}\mathbf{P}(A \to B) &= 1 - \big(1 - \mathbf{P}(A \to C \text{ and } C \to B)\big)\big(1 - \mathbf{P}(A \to D \text{ and } D \to B)\big) \\ &= 1 - (1 - 0.851)(1 - 0.712) \\ &= 0.957.\end{aligned}$$

Independent Trials and the Binomial Probabilities

If an experiment involves a sequence of independent but identical stages, we say that we have a sequence of **independent trials**. In the special case where there are only two possible results at each stage, we say that we have a sequence of independent **Bernoulli trials**. The two possible results can be anything, e.g., "it rains" or "it doesn't rain," but we will often think in terms of coin tosses and refer to the two results as "heads" (H) and "tails" (T).

Consider an experiment that consists of n independent tosses of a biased coin, in which the probability of "heads" is p, where p is some number between 0 and 1. In this context, independence means that the events $A_1, A_2, \ldots, A_n$ are independent, where $A_i = \{i\text{th toss is a head}\}$.

We can visualize independent Bernoulli trials by means of a sequential description, as shown in Fig. 1.15 for the case where $n = 3$. The conditional probability of any toss being a head, conditioned on the results of any preceding tosses is p, because of independence. Thus, by multiplying the conditional probabilities along the corresponding path of the tree, we see that any particular outcome (3-long sequence of heads and tails) that involves k heads and $3 - k$ tails has probability $p^k(1-p)^{3-k}$. This formula extends to the case of a general number n of tosses. We obtain that the probability of any particular n-long sequence that contains k heads and $n - k$ tails is $p^k(1-p)^{n-k}$, for all k from 0 to n.

Let us now consider the probability

$$p(k) = \mathbf{P}(k \text{ heads come up in an } n\text{-toss sequence}),$$

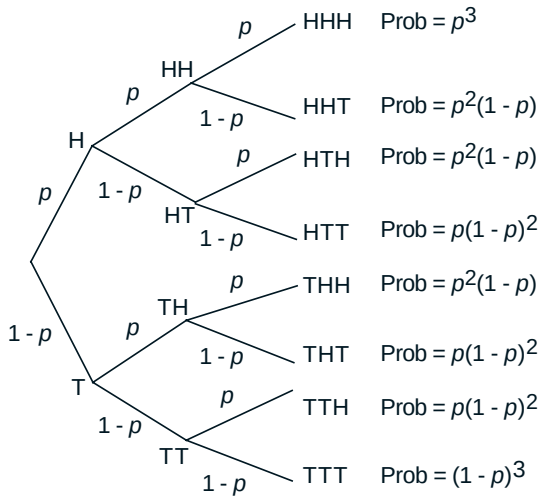

Figure 1.15: Sequential description of the sample space of an experiment involving three independent tosses of a biased coin. Along the branches of the tree, we record the corresponding conditional probabilities, and by the multiplication rule, the probability of obtaining a particular 3-toss sequence is calculated by multiplying the probabilities recorded along the corresponding path of the tree.

which will play an important role later. We showed above that the probability of any given sequence that contains k heads is $p^k(1-p)^{n-k}$, so we have

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \text{number of distinct } n\text{-toss sequences that contain } k \text{ heads.}$$

The numbers $\binom{n}{k}$ (called "n choose k") are known as the **binomial coefficients**, while the probabilities $p(k)$ are known as the **binomial probabilities**. Using a counting argument, to be given in Section 1.6, one finds that

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}, \qquad k = 0, 1, \ldots, n,$$

where for any positive integer i we have

$$i! = 1 \cdot 2 \cdots (i-1) \cdot i,$$

and, by convention, $0! = 1$. An alternative verification is sketched in the theoretical problems. Note that the binomial probabilities $p(k)$ must add to 1, thus showing the **binomial formula**

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1.$$

Example 1.23. Grade of service. An internet service provider has installed c modems to serve the needs of a population of n customers. It is estimated that at a given time, each customer will need a connection with probability p, independently of the others. What is the probability that there are more customers needing a connection than there are modems?

Here we are interested in the probability that more than c customers simultaneously need a connection. It is equal to

$$\sum_{k=c+1}^{n} p(k),$$

where

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

are the binomial probabilities.

This example is typical of problems of sizing the capacity of a facility to serve the needs of a homogeneous population, consisting of independently acting customers. The problem is to select the size c to achieve a certain threshold probability (sometimes called *grade of service*) that no user is left unserved.

1.6 COUNTING*

The calculation of probabilities often involves counting of the number of outcomes in various events. We have already seen two contexts where such counting arises.

(a) When the sample space Ω has a finite number of equally likely outcomes, so that the discrete uniform probability law applies. Then, the probability of any event A is given by

$$\mathbf{P}(A) = \frac{\text{Number of elements of } A}{\text{Number of elements of } \Omega},$$

and involves counting the elements of A and of Ω.

(b) When we want to calculate the probability of an event A with a finite number of equally likely outcomes, each of which has an already known probability p. Then the probability of A is given by

$$\mathbf{P}(A) = p \cdot (\text{Number of elements of } A),$$

and involves counting of the number of elements of A. An example of this type is the calculation of the probability of k heads in n coin tosses (the binomial probabilities). We saw there that the probability of each distinct sequence involving k heads is easily obtained, but the calculation of the number of all such sequences is somewhat intricate, as will be seen shortly.

While counting is in principle straightforward, it is frequently challenging; the art of counting constitutes a large portion of a field known as **combinatorics**. In this section, we present the basic principle of counting and apply it to a number of situations that are often encountered in probabilistic models.

The Counting Principle

The counting principle is based on a divide-and-conquer approach, whereby the counting is broken down into stages through the use of a tree. For example, consider an experiment that consists of two consecutive stages. The possible results of the first stage are $a_1, a_2, \ldots, a_m$; the possible results of the second stage are $b_1, b_2, \ldots, b_n$. Then, the possible results of the two-stage experiment are all possible **ordered** pairs (a_i, b_j), $i = 1, \ldots, m$, $j = 1, \ldots, n$. Note that the number of such ordered pairs is equal to mn. This observation can be generalized as follows (see also Fig. 1.16).

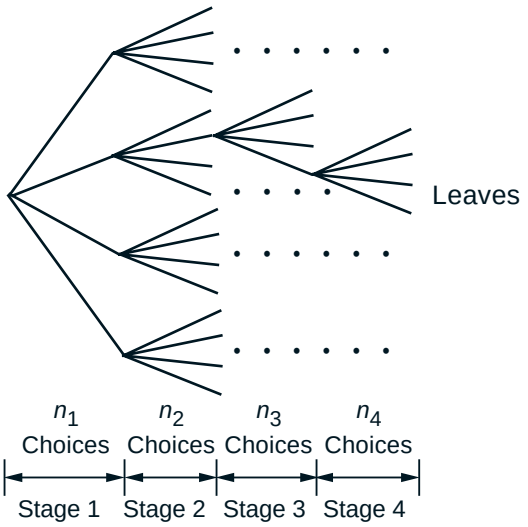

Figure 1.16: Illustration of the basic counting principle. The counting is carried out in r stages ($r = 4$ in the figure). The first stage has n_1 possible results. For every possible result of the first $i - 1$ stages, there are n_i possible results at the ith stage. The number of leaves is $n_1 n_2 \cdots n_r$. This is the desired count.

The Counting Principle

Consider a process that consists of r stages. Suppose that:

(a) There are n_1 possible results for the first stage.

(b) For every possible result of the first stage, there are n_2 possible results at the second stage.

(c) More generally, for all possible results of the first $i-1$ stages, there are n_i possible results at the ith stage.

Then, the total number of possible results of the r-stage process is

$$n_1 \cdot n_2 \cdots n_r.$$

Example 1.24. The number of telephone numbers. A telephone number is a 7-digit sequence, but the first digit has to be different from 0 or 1. How many distinct telephone numbers are there? We can visualize the choice of a sequence as a sequential process, where we select one digit at a time. We have a total of 7 stages, and a choice of one out of 10 elements at each stage, except for the first stage where we only have 8 choices. Therefore, the answer is

$$8 \cdot \underbrace{10 \cdot 10 \cdots 10}_{6 \text{ times}} = 8 \cdot 10^6.$$

Example 1.25. The number of subsets of an n-element set. Consider an n-element set $\{s_1, s_2, \ldots, s_n\}$. How many subsets does it have (including itself and the empty set)? We can visualize the choice of a subset as a sequential process where we examine one element at a time and decide whether to include it in the set or not. We have a total of n stages, and a binary choice at each stage. Therefore the number of subsets is

$$\underbrace{2 \cdot 2 \cdots 2}_{n \text{ times}} = 2^n.$$

It should be noted that the Counting Principle remains valid even if each first-stage result leads to a different set of potential second-stage results, etc. The only requirement is that the number of possible second-stage results is constant, regardless of the first-stage result. This observation is used in the sequel.

In what follows, we will focus primarily on two types of counting arguments that involve the selection of k objects out of a collection of n objects. If the order of selection matters, the selection is called a **permutation**, and otherwise, it is

called a **combination**. We will then discuss a more general type of counting, involving a **partition** of a collection of n objects into multiple subsets.

k-permutations

We start with n distinct objects, and let k be some positive integer, with $k \leq n$. We wish to count the number of different ways that we can pick k out of these n objects and arrange them in a sequence, i.e., the number of distinct k-object sequences. We can choose any of the n objects to be the first one. Having chosen the first, there are only $n-1$ possible choices for the second; given the choice of the first two, there only remain $n-2$ available objects for the third stage, etc. When we are ready to select the last (the kth) object, we have already chosen $k-1$ objects, which leaves us with $n-(k-1)$ choices for the last one. By the Counting Principle, the number of possible sequences, called **k-permutations**, is

$$\begin{aligned} n(n-1)\cdots(n-k+1) &= \frac{n(n-1)\cdots(n-k+1)(n-k)\cdots 2\cdot 1}{(n-k)\cdots 2\cdot 1} \\ &= \frac{n!}{(n-k)!}. \end{aligned}$$

In the special case where $k=n$, the number of possible sequences, simply called **permutations**, is

$$n\cdot(n-1)\cdot(n-2)\cdots 2\cdot 1 = n!.$$

(Let $k=n$ in the formula for the number of k-permutations, and recall the convention $0!=1$.)

Example 1.26. Let us count the number of words that consist of four distinct letters. This is the problem of counting the number of 4-permutations of the 26 letters in the alphabet. The desired number is

$$\frac{n!}{(n-k)!} = \frac{26!}{22!} = 26\cdot 25\cdot 24\cdot 23 = 358,800.$$

The count for permutations can be combined with the Counting Principle to solve more complicated counting problems.

Example 1.27. You have n_1 classical music CDs, n_2 rock music CDs, and n_3 country music CDs. In how many different ways can you arrange them so that the CDs of the same type are contiguous?

We break down the problem in two stages, where we first select the order of the CD types, and then the order of the CDs of each type. There are 3! ordered sequences of the types of CDs (such as classical/rock/country, rock/country/classical, etc), and there are $n_1!$ (or $n_2!$, or $n_3!$) permutations of the classical (or rock, or

country, respectively) CDs. Thus for each of the 3! CD type sequences, there are $n_1!\,n_2!\,n_3!$ arrangements of CDs, and the desired total number is $3!\,n_1!\,n_2!\,n_3!$.

Combinations

There are n people and we are interested in forming a committee of k. How many different committees are there? More abstractly, this is the same as the problem of counting the number of k-element subsets of a given n-element set. Notice that forming a combination is different than forming a k-permutation, because **in a combination there is no ordering of the selected elements**. Thus for example, whereas the 2-permutations of the letters A, B, C, and D are

AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC,

the combinations of two out four of these letters are

AB, AC, AD, BC, BD, CD.

There is a close connection between the number of combinations and the binomial coefficient that was introduced in Section 1.5. To see this note that specifying an n-toss sequence with k heads is the same as picking k elements (those that correspond to heads) out of the n-element set of tosses. Thus, the number of combinations is the same as the binomial coefficient $\binom{n}{k}$ introduced in Section 1.5.

To count the number of combinations, note that selecting a k-permutation is the same as first selecting a combination of k items and then ordering them. Since there are $k!$ ways of ordering the k selected items, we see that the number of k-permutations is equal to the number of combinations times $k!$. Hence, the number of possible combinations, is given by

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Example 1.28. The number of combinations of two out of the four letters A, B, C, and D is found by letting $n = 4$ and $k = 2$. It is

$$\binom{4}{2} = \frac{4!}{2!\,2!} = 6,$$

consistently with the listing given earlier.

It is worth observing that counting arguments sometimes lead to formulas that are rather difficult to derive algebraically. One example is the **binomial formula**

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$$

discussed in Section 1.5. Here is another example. Since $\binom{n}{k}$ is the number of k-element subsets of a given n-element subset, the sum over k of $\binom{n}{k}$ counts the number of subsets of all possible cardinalities. It is therefore equal to the number of all subsets of an n-element set, which is 2^n, and we obtain

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

Partitions

Recall that a combination is a choice of k elements out of an n-element set without regard to order. This is the same as partitioning the set in two: one part contains k elements and the other contains the remaining $n-k$. We now generalize by considering partitions in more than two subsets.

We have n distinct objects and we are given nonnegative integers $n_1, n_2, \ldots, n_r$, whose sum is equal to n. The n items are to be divided into r disjoint groups, with the ith group containing exactly n_i items. Let us count in how many ways this can be done.

We form the groups one at a time. We have $\binom{n}{n_1}$ ways of forming the first group. Having formed the first group, we are left with $n - n_1$ objects. We need to choose n_2 of them in order to form the second group, and we have $\binom{n-n_1}{n_2}$ choices, etc. Using the Counting Principle for this r-stage process, the total number of choices is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r},$$

which is equal to

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

We note that several terms cancel and we are left with

$$\frac{n!}{n_1!\,n_2!\cdots n_r!}.$$

This is called the **multinomial coefficient** and is usually denoted by

$$\binom{n}{n_1, n_2, \ldots, n_r}.$$

Example 1.29. Anagrams. How many different letter sequences can be obtained by rearranging the letters in the word TATTOO? There are six positions to be filled

by the available letters. Each rearrangement corresponds to a partition of the set of the six positions into a group of size 3 (the positions that get the letter T), a group of size 1 (the position that gets the letter A), and a group of size 2 (the positions that get the letter O). Thus, the desired number is

$$\frac{6!}{1!\,2!\,3!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 1 \cdot 2 \cdot 1 \cdot 2 \cdot 3} = 60.$$

It is instructive to rederive this answer using an alternative argument. (This argument can also be used to rederive the multinomial coefficient formula; see the theoretical problems.) Let us rewrite TATTOO in the form $T_1AT_2T_3O_1O_2$ pretending for a moment that we are dealing with 6 distinguishable objects. These 6 objects can be rearranged in 6! different ways. However, any of the 3! possible permutations of T_1, T_1, and T_3, as well as any of the 2! possible permutations of O_1 and O_2, lead to the same word. Thus, when the subscripts are removed, there are only $6!/(3!\,2!)$ different words.

Example 1.30. A class consisting of 4 graduate and 12 undergraduate students is randomly divided into four groups of 4. What is the probability that each group includes a graduate student? This is the same as Example 1.11 in Section 1.3, but we will now obtain the answer using a counting argument.

We first determine the nature of the sample space. A typical outcome is a particular way of partitioning the 16 students into four groups of 4. We take the term "randomly" to mean that every possible partition is equally likely, so that the probability question can be reduced to one of counting.

According to our earlier discussion, there are

$$\binom{16}{4,4,4,4} = \frac{16!}{4!\,4!\,4!\,4!}$$

different partitions, and this is the size of the sample space.

Let us now focus on the event that each group contains a graduate student. Generating an outcome with this property can be accomplished in two stages:

(a) Take the four graduate students and distribute them to the four groups; there are four choices for the group of the first graduate student, three choices for the second, two for the third. Thus, there is a total of 4! choices for this stage.

(b) Take the remaining 12 undergraduate students and distribute them to the four groups (3 students in each). This can be done in

$$\binom{12}{3,3,3,3} = \frac{12!}{3!\,3!\,3!\,3!}$$

different ways.

By the Counting Principle, the event of interest can materialize in

$$\frac{4!\,12!}{3!\,3!\,3!\,3!}$$

different ways. The probability of this event is

$$\frac{\dfrac{4!\,12!}{3!\,3!\,3!\,3!}}{\dfrac{16!}{4!\,4!\,4!\,4!}}.$$

After some cancellations, we can see that this is the same as the answer $12 \cdot 8 \cdot 4/(15 \cdot 14 \cdot 13)$ obtained in Example 1.11.

Here is a summary of all the counting results we have developed.

Summary of Counting Results

- Permutations of n objects: $n!$
- k-permutations of n objects: $n!/(n-k)!$
- Combinations of k out of n objects: $\dbinom{n}{k} = \dfrac{n!}{k!(n-k)!}$
- Partitions of n objects into r groups with the ith group having n_i objects:

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

1.7 SUMMARY AND DISCUSSION

A probability problem can usually be broken down into a few basic steps:

1. The description of the sample space, that is, the set of possible outcomes of a given experiment.
2. The (possibly indirect) specification of the probability law (the probability of each event).
3. The calculation of probabilities and conditional probabilities of various events of interest.

The probabilities of events must satisfy the nonnegativity, additivity, and normalization axioms. In the important special case where the set of possible outcomes is finite, one can just specify the probability of each outcome and obtain the probability of any event by adding the probabilities of the elements of the event.

Conditional probabilities can be viewed as probability laws on the same sample space. We can also view the conditioning event as a new universe, be-

cause only outcomes contained in the conditioning event can have positive conditional probability. Conditional probabilities are derived from the (unconditional) probability law using the definition $\mathbf{P}(A \mid B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$. However, the reverse process is often convenient, that is, first specify some conditional probabilities that are natural for the real situation that we wish to model, and then use them to derive the (unconditional) probability law. Two important tools in this context are the multiplication rule and the total probability theorem.

We have illustrated through examples three methods of specifying probability laws in probabilistic models:

(1) The **counting method**. This method applies to the case where the number of possible outcomes is finite, and all outcomes are equally likely. To calculate the probability of an event, we count the number of elements in the event and divide by the number of elements of the sample space.

(2) The **sequential method**. This method applies when the experiment has a sequential character, and suitable conditional probabilities are specified or calculated along the branches of the corresponding tree (perhaps using the counting method). The probabilities of various events are then obtained by multiplying conditional probabilities along the corresponding paths of the tree, using the multiplication rule.

(3) The **divide-and-conquer method**. Here, the probabilities $\mathbf{P}(B)$ of various events B are obtained from conditional probabilities $\mathbf{P}(B \mid A_i)$, where the A_i are suitable events that form a partition of the sample space and have known probabilities $\mathbf{P}(A_i)$. The probabilities $\mathbf{P}(B)$ are then obtained by using the total probability theorem.

Finally, we have focused on a few side topics that reinforce our main themes. We have discussed the use of Bayes' rule in inference, which is an important application context. We have also discussed some basic principles of counting and combinatorics, which are helpful in applying the counting method.