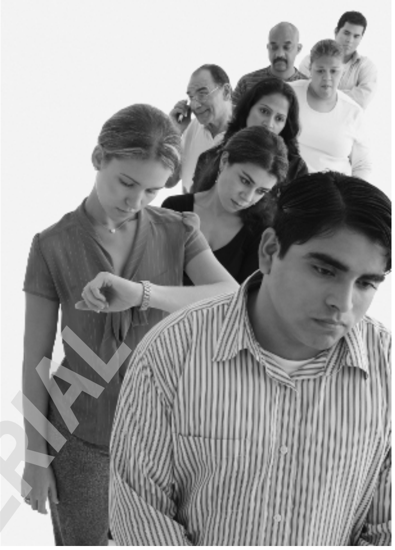
CHAPTER 1



Probability Review

CHAPTER CONTENTS

Although anyone tackling this book should have some previous exposure to probability, this chapter serves as a concise, convenient ‘‘refresher’’ summary of probability concepts that are important later. Students who are confident of their abilities in this area may elect to skim this chapter; those who feel weak on the subject should consult an introductory textbook, do all of the problems provided, and also find other problems to do. All of the subsequent material in the book depends heavily upon the basic concepts of probability.

It is worth mentioning early that we will be using a lot of probability, but not much statistics. Most introductory textbooks are so oriented toward statistics that they leave a slanted impression about probability. Students who have any confusion about the difference between the two subjects may have trouble with this book. Very concisely: Statistics is about interpreting data; probability is about representing uncertainty and/or variability. The two subjects converge when comparing data to what could have been expected from hypothetical assumptions. For example, hypothesis testing compares some numbers computed from data to numbers from a table of a specified probability function, such as the Normal, F, or Chi-squared distributions. The comparison tells whether the assumed hypothesis (corresponding to the tabulated number) is reasonably consistent with the measured data. But that is not the use of probability theory that will come into play here, so we will have little use for those tabulated distributions. Instead, we will need to understand the basic properties of random variables, distributions, and parameters so that we can manipulate them to produce predictions. We will have almost no use for the discrete uniform distribution (where permutations and combinations are used), the Normal distribution, or most of the other distributions that have names.

1.1 Interpreting and Using Probabilities

Some people learn all of the notation, rules of manipulation, and formulas, but never really "get" probabilities. Unless they are told directly to calculate a probability, they would never think of doing it on their own. So, although they may know *how* to calculate with them, they lack a full understanding of *why* probabilities are useful. Many introductory textbooks unintentionally contribute to the problem by using examples—cards, dice, colored balls, and such—that have no relevance to ordinary life. This text strives to provide more realistic examples that matter to engineers and managers. Obviously, we still have to begin with very simple situations and gradually work up to full realism.

The most immediate and obvious use of probabilities is to quantify uncertainty. Some people are very uncomfortable with uncertainty, preferring everything to be black or white, true or false, one way or the other, with no ambiguity. But most people understand that reality is not so simple—that sometimes people have to accept the fact that certainty cannot be achieved. Probabilities provide a way to quantify the "shades of gray" between impossibility and certainty. For what purpose? Generally speaking, the numbers in help to improve decision making.

When is uncertainty such an important factor that it demands quantitative treatment? Of course, if there is little at stake, you can make any choice without fear of making a big mistake. Or, if you have absolute certainty about the consequences of your actions, there is no need to assess probabilities. But let's face it—very few of the important decisions you make in life will be blessed by complete and accurate information. Almost always, you will be forced to choose with less information than you would like. On the other hand, if you have absolutely *no* information to work with, there is not much you can do with probabilities. So, we can conclude that probabilities are most useful when you need to make decisions about matters of importance and you have only partial information with which to work. That description still covers an enormous range of opportunities. In particular, engineers and managers deal with such issues routinely, because they design and control complex systems.

If you have only two choices, and want to favor the more likely event (such as betting on the winner in a two-team contest), there are ways—other than probabilities—to represent the comparative likelihoods. In sports competitions, it is common to use *point spreads*. For example, one team may be favored by three points, which means that (in someone's judgment) the first team is just as likely to score three more points than the second team's total as it is to score less than that. So, for purposes of betting, it is considered to be a fair bet when the weaker team is "given" an extra three points. Handicaps and headstarts are similar notions for equalizing chances.

Another way to express uncertainty is to use odds. If the success and failure of a certain outcome are equally likely, the odds would be 1:1 (spoken as "one to one"). If the odds against an outcome are given as 3:2 (spoken as "three to two"), it means that a bet of two units should win three units if the outcome is realized. That situation would reflect the fact or belief that the outcome is less likely to succeed than it is to fail, so the payoff should be greater than the bet to make the wager fair.

Although point spreads and odds are common in gambling situations, they do not serve very well elsewhere. In a business situation, for example, you cannot equalize the competition by handicapping the leader. Furthermore, you are commonly interested in more than winning or losing, so the two-outcome range of possibilities is far too limited. Although it is true that odds can always be translated to probabilities and vice versa (you may want to figure out how), probabilities are much easier to manipulate than odds when the situations grow complicated.

We usually speak of uncertainty as something describing the future—something that we are unsure about because it has not yet happened. But there are other sources of uncertainty that are also worth attention. Sometimes you need to answer questions about something that has already occurred, but you do not know the result. For example, a business competitor may have already taken some action that is hidden from you. Or perhaps you are in the process of conducting investigations and do not have complete answers yet.

There is another use of probability that does not involve uncertainty at all: We may have complete and accurate information about something, but that something is a set of values, rather than a single value. That is, probabilities are useful in describing a particular measurable property of individuals in the population. (Here, a population is any collection, not necessarily a biological one.) For example, if we determine the year of birth of every student in the class and then ask, ''What year was the class born in?'' we may not be able to answer with a single number even though we have all of the information. To give a full answer, we would typically have to specify a set of values—several years—and also the count of the number of students born in each year. Those counts, or frequencies, are equivalent to probabilities when they are normalized; that is, divided by the total number in the class. Hence, probabilities are useful in describing the *variation* in the population, even when that variation is fully known. In this case, the probability of any particular value corresponds to the fraction of the total population having that value. You can use ordinary probability rules to manipulate these fractions and always recover absolute numbers by multiplying the fractions by the total size of the population.

The greatest value from understanding probability comes from gaining a conceptual framework and vocabulary for dealing with uncertainty and variation. Even if you lack sufficient data to calculate anything, you can mentally weigh the factors better than people who lack that understanding. Those who have learned the concepts well use them every day.

1.2 Sample Spaces and Events

We turn now to a more formal presentation of the concepts. An *experiment* is a well-understood procedure or process whose outcome can be observed but is not known in advance with certainty. Reread that sentence; there is a lot that is said and left unsaid in those few words. For example, nothing is mentioned about being able to control anything. For this word and others that are defined here, mentally compare the formal definition and the informal use of the same word in ordinary language to be sure you understand the difference.

The set of all possible outcomes of an experiment is called the *sample space*. Whenever the sample space consists of a countable number of outcomes, it is said to be *discrete*; otherwise, it is *continuous*. An *event* is any subset of the sample space, including the empty, or null, set and the entire sample space. When the result of the experiment becomes known, we would say that a specified event *has occurred* if the observed outcome is contained in the subset which is the event. The empty set is an event that can never occur; the entire sample space is an event that is certain to occur. A set consisting of any single outcome is called an *elementary event*.

Of course, we want events to correspond to what you would ordinarily consider them to be. Often the most natural way to specify them is to describe them in words. However, the reason for defining them formally as sets is to establish a mathematical way to combine and manipulate them. The basic algebra used to manipulate events is set theory. You need to know all the rules to be able to express the real world events that we are going to model.

As a short reminder, the set theoretic union of two events produces another event. If $\mathrm{C} = \mathrm{A} \cup \mathrm{B}$, we would say in words that event C had occurred if event A *or* event B (or both) occurred. Similarly, the intersection of two events corresponds to the word *and*. The complement of any event is another event; we would say that A had occurred if A had not occurred. Two events are *mutually exclusive* if their intersection is the empty set, which can be thought of as the impossible event. In other words, two events are mutually exclusive if they could not both occur. There are several other basic rules of set theory that you should know (DeMorgan's laws, the distribution rules, and so forth); look them up if you need help in recalling them.

1.3 Probability

When the ''probability of an event'' is spoken of in everyday language, almost everyone has a rough idea of what is meant. It is fortunate that this is so, because it would be quite difficult to introduce the concept to someone who had never considered it before. There are at least three distinct ways to approach the subject, none of which is wholly satisfying.

The first to appear, historically, was the frequency concept. If an experiment were to be repeated many times, then the number of times that the event was observed to occur, divided by the number of times that the experiment was conducted, would approach a number that was defined to be the probability of the event. The ratio of the number of chances for success out of the total number of possibilities is the concept with which most elementary treatments of probability start. This definition proved to be somewhat limiting, however, because circumstances frequently prohibit the repetition of an experiment under precisely the same conditions, even conceptually. Imagine trying to determine the probability of global annihilation from a meteor collision.

To extend the notion of probability to a wider class of applications, a second approach involving the idea of ''subjective'' probabilities emerged. According to this idea, the probability of an event need not relate to the frequency with which it would occur in an infinite number of trials; it is just a measure of the degree of likelihood we *believe* the event to possess. This definition covers even hypothetical events, but seems a bit too loose for engineering applications. Different people could attach different probabilities to the same event.

Most modern texts use the third concept, which relies upon a purely axiomatic definition. According to this notion, probabilities are just elements of an abstract mathematical system obeying certain axioms. This notion is at once the most powerful and the most devoid of real-world meaning. Of course, the axioms are not purely arbitrary; they were selected to be consistent with the earlier concepts of probabilities and to provide them with all of the properties everyone would agree they should have.

We will go with the formal axiomatic system, so that we can be rigorous in the mathematics. We want to be able to calculate probabilities to assist in making good decisions. At the same time, we want to bear in mind the real-world interpretation of probabilities as measures of the likelihood of events in the world. The whole point of learning the mathematics is to be able to use it in everyday life.

A *probability* is a function, $\boldsymbol{P}(.)$, mapping events onto real numbers, and satisfying

1. $0 \leq \boldsymbol{P}(\mathrm{A}) \leq 1$, for any event A.
2. $\boldsymbol{P}(\mathrm{S}) = 1$, where S is the whole sample space, or the ''certain'' event.
3. If A_1, A_2, A_3 . . . are a set of pairwise mutually exclusive events (finite or infinite in number), then $\boldsymbol{P}(\mathrm{A}_1 \cup \mathrm{A}_2 \cup \mathrm{A}_3 \cup \ldots) = \boldsymbol{P}(\mathrm{A}_1) + \boldsymbol{P}(\mathrm{A}_2) + \boldsymbol{P}(\mathrm{A}_3) + \ldots.$

Although probabilities have a number of other properties well worth mentioning, these three axioms are sufficient to derive the others.

These three axioms are not enough to determine uniquely the probability of any event. For all but trivial sample spaces, there will exist an infinite number of ways to assign probabilities to events while satisfying the three axioms. At this point, we are merely establishing properties or rules required of *any* assignment of probabilities to events.

Some of the additional basic laws of probability (which can be proved from the above axioms) are

4. $\boldsymbol{P}(\varphi) = 0$, where φ is the empty set, or the impossible event. In words, an event that cannot occur must be assigned a probability value of zero. Usually the converse is true also; namely, if an event has a probability of zero then it cannot occur. However, that statement is not always true. When there are an infinite number of outcomes in S, there are times when possible (though extremely unlikely) events have a probability value of zero.
5. $\boldsymbol{P}(\overline{\mathrm{A}}) = 1 - \boldsymbol{P}(\mathrm{A})$. In words, the probability that an event does not occur is 1 minus the probability that it does occur. Another way to look at it is that the probability of any event plus the probability of its complement must sum to 1.
6. $\boldsymbol{P}(\mathrm{A} \cup \mathrm{B}) = \boldsymbol{P}(\mathrm{A}) + \boldsymbol{P}(\mathrm{B}) - \boldsymbol{P}(\mathrm{A} \cap \mathrm{B})$, for any two events, A and B. This is the relation that seems to give students trouble. The tendency is to want to add probabilities without considering whether the events are mutually exclusive. When they are not—that is, when there is some possibility for both A and B to occur—then you have to subtract off the probability that they both occur.
7. $\boldsymbol{P}(\mathrm{A}|\mathrm{B}) = \boldsymbol{P}(\mathrm{A} \cap \mathrm{B})/\boldsymbol{P}(\mathrm{B})$ provided $\boldsymbol{P}(\mathrm{B}) \neq 0$. This ‘‘basic law’’ is, in reality, a definition of the conditional probability of an event, A, given that another event, B, has occurred. The notation for this conditional probability is $\boldsymbol{P}(\mathrm{A}|\mathrm{B})$, (read as ‘‘the probability of A given B’’).

   Conditional probabilities are very important in modeling, and we will see a great deal more of them. The notion of conditional probability conforms to the intuitive concept of altering our estimate of the likelihood of an event as we acquire additional information. That is, $\boldsymbol{P}(\mathrm{A}|\mathrm{B})$ is the new probability of A after we know that B has occurred.

   It is common in modeling applications to know $\boldsymbol{P}(\mathrm{A}|\mathrm{B})$ directly but not to know $\boldsymbol{P}(\mathrm{A} \cap \mathrm{B})$. For that reason, rule 7 often appears in the equivalent form shown in rule 8 below.
8. $\boldsymbol{P}(\mathrm{A} \cap \mathrm{B}) = \boldsymbol{P}(\mathrm{A}|\mathrm{B})\boldsymbol{P}(\mathrm{B})$. Conditional probabilities are useful only when the events involved, A and B, have something to do with one another. If knowledge that B has occurred has no bearing upon our estimate of the likelihood of A, we would say that the two events are independent and write rule 9, shown next.
9. $\boldsymbol{P}(\mathrm{A}|\mathrm{B}) = \boldsymbol{P}(\mathrm{A})$ if and only if A and B are independent. This rule can be taken as the formal definition of independence. Combining axioms 8 and 9, rule 10 immediately follows.
10. $\boldsymbol{P}(\mathrm{A} \cap \mathrm{B}) = \boldsymbol{P}(\mathrm{A})\boldsymbol{P}(\mathrm{B})$ if, and only if, A and B are independent. Alternatively, rule 10 could be taken as the definition of independence and rule 9 would immediately follow. Rules 7, 8, 9, and 10 are all closely related; you should see them as variations of the same ‘‘fact’’ about dependent events. You should also realize that you will rarely be given numbers and asked to check the formulas to see whether dependence or independence applies. Almost always, you will have to decide for yourself whether or not the events are related, and then use the appropriate formula.

    A set of events $\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_\mathrm{n}$ constitute a *partition* of the sample space S if they are mutually exclusive and collectively exhaustive, that is,

    $$\mathrm{B}_\mathrm{i} \cap \mathrm{B}_\mathrm{j} = \varphi \text{ for every pair i and j}$$

    and

    $$\mathrm{B}_1 \cup \mathrm{B}_2 \cup \mathrm{B}_3 \cup \ldots \cup \mathrm{B}_\mathrm{n} = \mathrm{S}$$

    In simple terms, a partition is just any way of grouping and listing all possible outcomes such that no outcome appears in more than one group. When the experiment is performed, one and only one of the B_i will occur. It is easy to prove that with rule 11.

11. $\sum_i \boldsymbol{P}(B_i) = 1$ for any partition $B_1, B_2, \ldots, B_n$

12. $\boldsymbol{P}(A) = \sum_i \boldsymbol{P}(A|B_i)\boldsymbol{P}(B_i)$ for any partition $B_1, B_2, \ldots, B_n$. This is one of the most useful relationships in modeling applications. It is one expression of the so-called law of total probability, which will be discussed in detail later in the chapter.

1.4 Random Variables

Although events may be directly assigned probabilities, more commonly the events are first associated with real numbers, which are then in turn associated with probabilities. For example, if the experiment involves observing the number of heads appearing when four coins are tossed, it would be natural to associate the possible outcomes with the integers 0, 1, 2, 3, and 4. These integers are not, in themselves, events, but the event corresponding to each value is obvious. The function that assigns numbers to events is called a *random variable*. You can think of it as a coding of the events, much like identification numbers that are used for convenience but do not in any way alter the events themselves.

In most cases, the rule that provides the value in the range of the random variable to go with each real-world event is so obvious that no special attention need be given to it. It is important to realize, however, that values of random variables have probabilities associated with them only because the values correspond to events that possess the probabilities directly. Using random variables gives us an indirect way to refer to events.

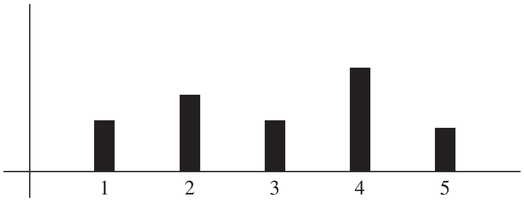
It is interesting to note that a random variable is, technically speaking, neither random nor a variable. It is conceptually convenient, however, to suppress all references to the real-world events and to regard a random variable as an ordinary variable whose value is randomly selected. In other words, once the random variable is well defined, we may speak of any value in the range of the random variable as if it were actually the event. It makes sense, thereby, to speak of the probability that a random variable, X, equals some particular number. (We really mean the probability of the event having that code value.)

If the values in the range of a random variable are integers (or, more precisely, a countable subset of the real numbers), the random variable is *discrete*. If the range consists of all values over an interval of the real numbers, the random variable is *continuous*. A discrete random variable could be either finite or infinite, depending on the number of values in the range. A continuous random variable would always have an (uncountably) infinite number of possible values, though the range could be bounded below, above, or both.

A word of caution is in order with respect to the use of the word ‘‘random.’’ Sometimes, particularly in statistical applications, the word carries the connotation of equal likelihood. For example, when we say, ‘‘Take a random sample,’’ we mean (among other things) that each member of the sampled population should have an equal chance of being selected. In general, however, the word ‘‘random’’ does not carry any such connotation.

1.5 Probability Distributions

Any rule that assigns probabilities to each of the possible values of a random variable is a probability distribution. The term is used somewhat generally, because there are several different ways to specify such a rule. More precise terms are used when a particular form is intended. However it is described, the rule essentially tells you how the total probability value of 1—that is, the amount of available probability for the entire range of possible values—is spread over those values.



■ **FIGURE 1.1** **A typical pdf**

For discrete random variables, the most obvious and commonly used method of specifying the rule is to indicate the probability for each value separately. The function p(x), defined as

$$p(x) = \boldsymbol{P}(X = x)$$

is called the *probability distribution function*, or pdf for short. (Note the different use of uppercase and lowercase letters. Uppercase is used for names of random variables; lowercase is used for values.) Although it is not essential, many people find it helpful to think of distributions in graphical terms. A discrete pdf would look something like Figure 1.1.

The values along the horizontal axis of Figure 1.1 correspond to the possible values of the random variable (which could extend to infinity in either or both directions), and the heights of the bars indicate the values of the probabilities. The overall shape of the pdf is not important; it could look like almost anything. The only necessary features are that the heights of the bars are never negative and the sum of all of them add up to 1.

An alternative, equally sufficient method to specify a probability distribution is to give the *cumulative distribution function*, or cdf for short, F(x), defined as

$$F(x) = \boldsymbol{P}(X \leq x)$$

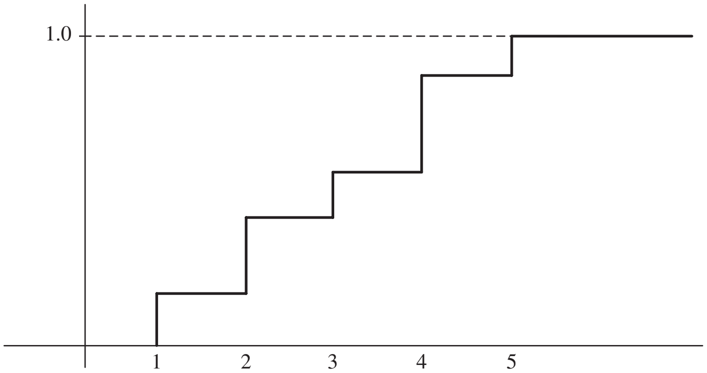
A third choice would be the *complementary cumulative distribution function*, or ccdf for short, G(x), defined as

$$G(x) = \boldsymbol{P}(X > x)$$

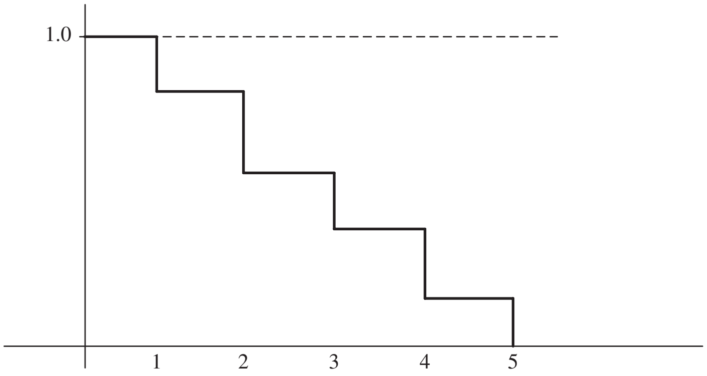
If any one of these—the pdf, cdf, or ccdf—is known, the others can be easily obtained in obvious ways. For example,

$$F(x) = 1 - G(x) \qquad \text{and} \qquad p(x) = F(x) - F(x - 1)$$

Graphs of the cdf or ccdf have characteristic forms. The cdf ''steps upward'' from 0 to 1, where the height of the step at x corresponds to the probability value at x. It can never step down because that would imply a negative probability value, which is not allowed. So the cdf is a monotonically nondecreasing function. The cdf for the pdf shown in Figure 1.1 would look like Figure 1.2.



■ **FIGURE 1.2** **A typical cdf**



■ **FIGURE 1.3 A typical ccdf**

Similarly, the ccdf in Figure 1.3 ‘‘steps down’’ from 1 to 0 and must be a monotonically nonincreasing function. Notice that, for any value of x, the cdf and ccdf sum to 1.

For continuous random variables, the situation is somewhat complicated by the fact that range of possible values is uncountably infinite. It is not consistent with the axioms of probability to allow each individual value to have positive probability. In fact, with the possible exception of a countable number of points, each individual value must be assigned the probability of zero! In contrast to the discrete case, a probability of zero does not necessarily imply that the corresponding event is impossible; it could merely mean that any one particular value is so unlikely, when considered next to the uncountably infinite set of alternatives, that the probability must be negligibly small. Consequently, it is fruitless to speak of the probabilities of particular values of random variables in the continuous case.

On the other hand, it makes perfect sense to speak of the probability that the value will fall within some interval. In particular, the cumulative distribution function, F(x), is well defined by

$$F(x) = \boldsymbol{P}(X \leq x)$$

Also, because the probability that X will exactly equal x may be zero, it can happen that

$$\boldsymbol{P}(X \leq x) = \boldsymbol{P}(X < x) + \boldsymbol{P}(X = x) = \boldsymbol{P}(X < x)$$

In other words, sometimes in the continuous case no distinction between strong and weak inequalities, or between open and closed intervals, need be made. Of course, the distinction must be scrupulously maintained in the discrete case or in continuous cases where some specific values have nonzero probability.

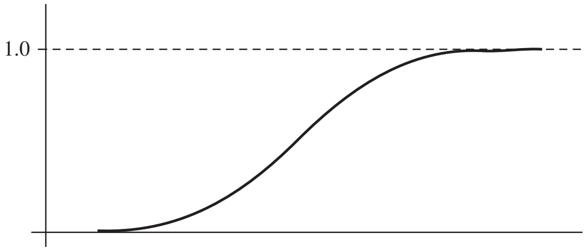
A typical continuous cdf will look something like Figure 1.4.

From the definition, it is apparent that F(x) must have the following properties:

$$0 \leq F(x) \leq 1 \qquad \text{for all } x$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$



■ **FIGURE 1.4 A typical continuous cdf** and

$$F(y) \geq F(x) \quad \text{for any } y > x$$

In words, the function F(x) must be bounded between 0 and 1, must approach zero at the left extremity of its range and one at the right extremity, and must be monotonically nondecreasing. (Actually, the last three imply the first.) Conversely, any function having these properties will qualify as a cumulative distribution function for some continuous random variable. Notice that although the figure shows a continuously rising function (which is typical), there is nothing to require that property; it could take sudden steps upward at some points.

Given the cumulative distribution function, one can easily express the probability that the random variable will assume a value within any specified region. For example,

$$\boldsymbol{P}(a \leq X \leq b) = \boldsymbol{P}(X \leq b) - \boldsymbol{P}(X \leq a) = F(b) - F(a)$$

The *complementary cumulative distribution function*, G(x), defined by

$$G(x) = \boldsymbol{P}(X > x)$$

or by

$$G(x) = 1 - F(x)$$

would also serve to describe fully the distribution. A typical continuous ccdf, as shown in Figure 1.5, would look like the cdf turned over.

The *probability density function*, f(x), is a function that, when integrated between a and b, gives the probability that the random variable will assume a value between a and b. That is,

$$\boldsymbol{P}(a \leq X \leq b) = \int_a^b f(x)dx$$

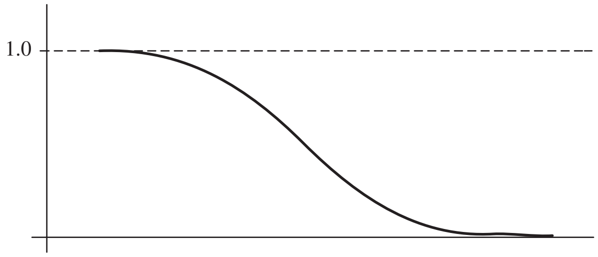
The relation between the density function and the distribution function is direct

$$F(x) = \int_{-\infty}^{x} f(y))dy$$

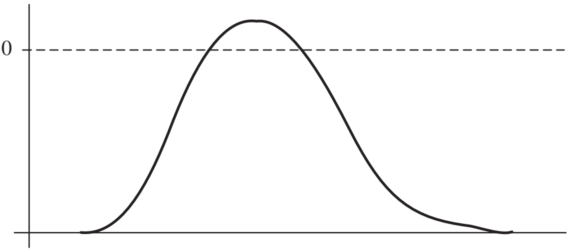
and

$$f(x) = \frac{d}{dx} F(x)$$

Although it may not seem to be the most natural way to describe a probability distribution, the density function is used more often than the cumulative or complementary cumulative distribution functions. In the case of a few distributions, only the density function



■ **FIGURE 1.5 A typical continuous ccdf**



■ **FIGURE 1.6 A typical density function**

can be expressed in closed form; the others must be expressed as integrals of the density function. It is important, therefore, that you learn to think in terms of density functions. One of the first things to get straight is that the value of the density function at some point is not a probability. The only way to get a probability from a density function is to integrate it.

The appearance of a density function is often something like Figure 1.6.

Any density function will have the properties

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

and

$$f(x) \geq 0 \qquad \text{for all x}$$

The first property is a direct consequence of the definition, but the second requires a brief argument. If f(x) were negative at any point, then there would exist two points, a and b, such that the integral of f(x) between a and b was negative. This would imply that

$$\boldsymbol{P}(a \leq X \leq b) < 0$$

which is impossible because probabilities cannot be negative. Therefore f(x) must be nonnegative everywhere.

Any function f(x) having the two properties mentioned above will qualify as a probability density function for some continuous random variable. Notice, in particular, that there is no requirement that f(x) be bounded above. The second property sometimes leads students to the mistaken presumption that f(x) cannot exceed 1. In fact, f(x) can be much greater than 1 over a narrow range, provided only that the integral over any interval does not exceed 1. Notice also that there is no requirement that f(x) be continuous. Functions that are discontinuous, or abruptly ‘‘jump’’ from one value to another, can be integrated without difficulty, provided only that the points of discontinuity are limited in number. The method, of course, is to separate the interval that you want to integrate into a sequence of intervals over each of which the density function is continuous.

Although, as already noted, it is important to keep in mind that f(x) is not a probability, it is useful in many applications to be able to substitute something involving f(x) into expressions as if it were a probability. A generally reliable device is to think of the notation f(x)dx as representing the probability that the random variable equals x. The dx part of the expression can be regarded as an interval of infinitesimal width, so the product of f(x) and dx is (roughly speaking) an area under the curve, or a probability. The presence of dx will indicate that an integration must be performed before an exact expression can be inferred.

Although the distinction between the discrete and continuous random variable cases is important, there are occasions when it is convenient to have a unified terminology to cover both cases. The letters pdf may be used to refer to either the probability distribution function, in the discrete case, or the probability density function, in the continuous case. Similarly, the letters cdf will stand for the cumulative distribution function and ccdf for the complementary cumulative distribution function, regardless of whether they are discrete or continuous.

1.6 Joint, Marginal, and Conditional Distributions

Whenever more than one random variable is involved in a single problem, there is a possibility that they are related. If so, it would not be sufficient to describe the probability distribution of each random variable in isolation; the relation between or among them must also be described. There are two methods in common use.

Suppose that two random variables, X and Y, are involved. The joint cumulative distribution function, or joint cdf, F(x, y), is defined by

$$F(x,\ y) = \boldsymbol{P}(X \le x,\ Y \le y)$$

Here, the comma represents the same as the intersection of the events implied by the terms. In words, F(x, y) is the probability that X takes on a value less than or equal to x *and* that Y takes on a value less than or equal to y. The same definition will suffice whether the random variables are both discrete, both continuous, or mixed. Conceptually, the basic idea is to extend the notion of a cumulative distribution function to two dimensions. Obviously, the same basic idea can be used to extend the notion to higher dimensions.

If both X and Y are discrete, the joint probability distribution function is defined by

$$p(x,\ y) = \boldsymbol{P}(X = x,\ Y = y)$$

If both are continuous, the joint probability density function is defined by

$$f(x,\ y) = \frac{\partial\ \ \partial}{\partial x \partial y} F(x,\ y)$$

The latter must be integrated twice in order to obtain a probability. In particular,

$$\boldsymbol{P}(r \le X \le s,\ t \le Y \le u) = \int_r^s \int_t^u f(x,\ y) dydx$$

Each of these is just a two-dimensional extension of the appropriate function for single random variables, and can be extended to higher dimensions in the obvious way. The term ''joint pdf'' will describe either function.

EXAMPLE 1.1

When X and Y are discrete and there are only a small number of possible values for each, it can be convenient to express the joint pdf in a table. For example, Table 1.1 shows a case where X can assume any of three values and Y can assume any of four. You read the cell entry directly to get the joint probability for any pair of values. For example, $\boldsymbol{P}(X = 1,\ Y = 10) = 0.2$.

■ TABLE 1.1
A Joint pdf

	Y = 5	**Y = 10**	**Y = 15**	**Y = 20**
X = 1	0.1	0.2	0	0
X = 2	0	0.25	0.25	0
X = 3	0	0	0.1	0.1

Sometimes a joint pdf for two or more random variables is given, but you want to know the pdf for just one of the random variables. That is, you might want to make a probability statement about, say, X, without regard to the value of Y. When both X and Y are discrete, the marginal probability distribution function of X is obtained from the joint pdf by

$$p(x) = \sum_{y} p(x, y)$$

When both are continuous, the marginal probability density function of X is given by

$$f(x) = \int_{y} f(x, y)dy$$

A marginal pdf is just an ordinary pdf, with all of the usual properties and interpretations. The word ‘‘marginal’’ merely conveys the information that it was obtained from a joint pdf.

If you have the joint pdf in the form of a table, as in Table 1.1, you get the marginal pdf by summing over rows or columns. For example, if you want the marginal pdf of X, you would sum across each of the three rows. In words, the probability that X takes on the value 1 is the sum of the probabilities that $X = 1$ and Y takes on any of its possible values. So, you just add across the first row to find $\boldsymbol{P}(X = 1) = 0.3$.

By symmetry, the marginal pdf of Y is obtainable from the joint pdf of X and Y by summing or integrating over all values of X. If more than two random variables are involved in a joint pdf, the marginal pdf for any one can be found by summing or integrating over all values of all random variables other than the one whose marginal pdf is sought. Although it is not often used, the marginal cdf is, if anything, even easier to obtain from a joint cdf:

$$F(x) = \lim_{y \to \infty} F(x, y)$$

Dealing with the cdf also has the advantage of permitting a single expression to cover both the discrete and continuous cases.

Independence of random variables is a property deriving from independence of the events that the random variables represent. Two random variables are independent if for all x and all y,

$$F(x, y) = F(x)F(y)$$

or, in terms of pdfs,

$$p(x, y) = p(x)p(y)$$

for discrete random variables, and

$$f(x, y) = f(x)f(y)$$

for continuous random variables. When the random variables are independent—but only then—the joint distribution can be constructed from the marginals.

Independence of random variables is an extremely important concept. Not only must you know how to manipulate the functions in the presence or absence of the property, but you also must judge whether the property can be reasonably assumed to hold in real-life situations. Because the mathematical definition may not be sufficiently revealing by itself to allow the student to grasp the concept at an intuitive level, a bit of further discussion seems warranted. When we say that the joint distribution can be obtained simply by multiplying the marginals, we are admitting that the joint distribution contains no more information than is already contained in the separate descriptions of the random variables. In other words, there is no need to account for the influence that one of the random variables might exert upon another. This would be true if, and only if, no such influence exists. Although the definition of independence of random variables is very similar in appearance to the definition of independence of events, it is actually a much stronger requirement. In order for X and Y to be independent, it is necessary that *every* event associated with X be independent of *every* event associated with Y.

The method of expressing joint pdfs or cdfs is just one of the ways to describe a relationship between two random variables. The other method is based on the idea of fixing a value for one and describing the subsequent distribution for the other. If both are discrete, the *conditional probability distribution function* of X given y (a particular value of the random variable Y) is defined by:

$$p(x|y) = \boldsymbol{P}(X = x|Y = y)$$

In $p(x|y)$, x is the argument of the function and y can be regarded as a parameter. In other words, we may insert various values of x into the function to get the probability that the random variable equals x, but this probability will be contingent upon the value of y. Through its definition as a conditional probability, the conditional pdf is easily related to the joint pdf by the expression

$$p(x|y) = \frac{p(x, y)}{p(y)} \qquad \text{provided } p(y) \neq 0$$

An analogous function exists for continuous random variables, but cannot be defined directly in terms of a conditional probability. The *conditional probability density function* of X given Y is most simply defined in terms of the joint density function

$$f(x|y) = \frac{f(x, y)}{f(y)} \qquad \text{provided } f(y) \neq 0$$

This function must be integrated with respect to x in order to yield a probability; the y simply acts as a parameter.

The conditional pdf of X given Y reduces to the marginal pdf if, and only if, X and Y are independent. In notation,

$$p(x|y) = p(x) \qquad \text{for all } x,\ y$$

or

$$f(x|y) = f(x) \qquad \text{for all } x,\ y$$

if, and only if, X and Y are independent. These expressions are entirely consistent with our earlier discussion of independence. If knowledge of the value of Y contributes nothing to a probability statement involving X, it must be that X and Y are unrelated.

Whenever a conditional distribution and one marginal distribution is given, the other marginal can be obtained. The procedure is first to obtain the joint distribution and then use that to get the desired marginal. In the discrete case, the expressions would be

$$p(x, y) = p(x|y)p(y)$$
$$p(x) = \sum_{y} p(x, y)$$

Therefore,

$$p(x) = \sum_{y} p(x|y)p(y)$$

The analogous formula in the continuous case would be

$$f(x) = \int_{y} f(x|y)f(y)dy$$

Both of these expressions are very useful in modeling applications.

1.7 Expectation

To describe a random variable completely requires a probability distribution in one of its various forms. If we were to require, however, a single number that best ‘‘summarized’’ the information contained in the distribution, we would almost certainly want to specify the ‘‘center’’ of the distribution. There are several ways to define ‘‘center,’’ but the most useful is the expectation.

The *expectation* of a random variable X, denoted $\boldsymbol{E}$(X), is defined by:

$$\boldsymbol{E}(X) = \sum_{-\infty}^{\infty} xp(x) \qquad \text{when X is discrete}$$

and

$$\boldsymbol{E}(X) = \int_{-\infty}^{\infty} xf(x)dx \qquad \text{when X is continuous.}$$

The same quantity may be called the *expected value of* X (although this term is quite misleading), the *mean* of the distribution, or the *first moment* of the distribution of X. All of these terms refer to the same thing. However, it should *not* be confused with an arithmetic average or a sample mean. The latter are statistical entities; we would compute them from data. An expectation is calculated from, and is an attribute of, a probability distribution. It can be regarded as a weighted average of the values of X, in which each possible value is weighted by the probability of its occurrence.

Although $\boldsymbol{E}$(X) is often called the expected value of X, one should be on guard against ‘‘expecting’’ $\boldsymbol{E}$(X) to occur as the value of X. Indeed, when X is discrete, $\boldsymbol{E}$(X) might not even be a *possible* value of X. It is true that if the experiment for which X is defined were to be repeated independently many times and the observed values of X were collected and averaged, then this average would be ‘‘close’’ to $\boldsymbol{E}$(X), in a certain probabilistic sense. However, this fact is a theorem of statistics (one form of the Law of Large Numbers) and has little significance for any single trial.

Typically in decision making, when one is forced to rank the options by some preference, the expectation is the value that is compared. One can easily criticize that approach, because the center of the distribution (or any other single value, for that matter) is a poor representation of the full range of possibilities. However, it is usually the most practical approach, and it can be justified at least to the extent that the expectation weights the outcomes ‘‘fairly.’’

One of the reasons that the expectation is so useful as a measure of centrality is that it has a number of very convenient properties. For any random variable X and any constants a and b,

$$\boldsymbol{E}(aX) = a\boldsymbol{E}(X)$$

and

$$\boldsymbol{E}(X + b) = \boldsymbol{E}(X) + b$$

In words, both multiplicative and additive constants can be ‘‘pulled out’’ of the expectation. For any two random variables X and Y,

$$\boldsymbol{E}(\mathrm{X}+\mathrm{Y})=\boldsymbol{E}(\mathrm{X})+\boldsymbol{E}(\mathrm{Y})$$

In words, the expected value of a sum is the sum of the expected values. The same relation can be extended to sums of more than two random variables and will hold whether or not the random variables are independent. The fact that sums ‘‘separate’’ and constants ‘‘pull out’’ of expressions in the obvious ways without any complications imply that the expectation is a linear operator. It is *always* linear, with no additional requirements on the random variables. (The same cannot be said for variances or other moments.)

Whenever X and Y are independent, the expected value of a product of random variables will decompose; that is,

$$\boldsymbol{E}(\mathrm{XY})=\boldsymbol{E}(\mathrm{X})\boldsymbol{E}(\mathrm{Y})$$

but this relation does not generally hold when the random variables are dependent. We will come to the more general case shortly.

Another convenience associated with using the expectation is the fact that the expectation of an arbitrary function of a random variable is easily expressed. Let h(X) be any function of X. Then if X is discrete,

$$\boldsymbol{E}(\mathrm{h(X)})=\sum_{\mathrm{x}}\mathrm{h(x)p(x)}$$

and if X is continuous,

$$\boldsymbol{E}(\mathrm{h(X)})=\int_{\mathrm{x}}\mathrm{h(x)f(x)dx}$$

In other words, h(x) merely replaces x in the definition of $\boldsymbol{E}$(X). These expressions are not a new definition or an obvious fact, but are derived by considering a random variable Y = h(X) and relating the distribution of Y to the distribution of X.

A concept used repeatedly in the book is that of *conditional expectation*. Formally, the conditional expectation of a random variable X given the value of a related random variable Y is defined by

$$\boldsymbol{E}(\mathrm{X|y})=\sum_{\mathrm{x}}\mathrm{xp(x|y)}\qquad\text{when X is discrete}$$

and

$$\boldsymbol{E}(\mathrm{X|y})=\int_{\mathrm{x}}\mathrm{xf(x|y)dx}\qquad\text{when X is continuous.}$$

The conditional expectation of X given y can be combined with the distribution of Y to yield the unconditional expectation of X. In notation,

$$\boldsymbol{E}(\mathrm{X})\ =\sum_{\mathrm{y}}\boldsymbol{E}(\mathrm{X|y})\mathrm{p(y)}\qquad\text{when X is discrete}$$

and

$$\boldsymbol{E}(\mathrm{X})=\int_{\mathrm{y}}\boldsymbol{E}(\mathrm{X|y})\mathrm{f(y)dy}\qquad\text{when X is continuous.}$$

A concise way to express both of these is

$$\boldsymbol{E}(X) = \boldsymbol{E}[\boldsymbol{E}(X|y)]$$

but this form does not suggest how useful the relation is as a technique for formulating an expression for $\boldsymbol{E}(X)$. The other forms suggest that the expectation of X can be thought of as a weighted average of the conditional expectations of X given y, taken over all possible conditions y, with each possible $\boldsymbol{E}(X|y)$ weighted according to the probability of occurrence. We will discuss this relation further in Section 1.9, "The Law of Total Probability," page 18.

1.8 Variance and Other Moments

After you know something about the central location of a distribution, most commonly expressed as an expectation, the next most valuable summary information would be about the spread or dispersal of the values that the random variable takes on. You could use the *range* (the interval between the highest and lowest value) if it is finite, or any of several other ways to measure the spread. But the most common measure is the *variance*, or its square root, the *standard deviation*. Computing it involves what may seem to be a nasty calculation, but the properties justify the definition.

The *n*th moment of a random variable is defined as the expectation of the *n*th power of the random variable. Since X^n is just a special case of a function of X, the *n*th moment can be expressed as

$$\boldsymbol{E}(X^n) = \sum_x x^n p(x) \qquad \text{when X is discrete}$$

and

$$\boldsymbol{E}(X^n) = \int_x x^n f(x)dx \qquad \text{when X is continuous.}$$

The first moment is, of course, the expectation. The *n*th central moment or the *n*th moment about the mean is defined as

$$\boldsymbol{E}([X - \boldsymbol{E}(X)]^n) = \sum_x [x - \boldsymbol{E}(X)]^n p(x) \qquad \text{when X is discrete}$$

and

$$\boldsymbol{E}([X - \boldsymbol{E}(X)]^n) = \int_x [x - \boldsymbol{E}(X)]^n f(x)dx \qquad \text{when X is continuous.}$$

In words, it is the expectation of the *n*th power of the random variable after it has been "shifted" by subtracting the expectation.

After the expectation, the next most important single number used to summarize distributions is the second moment about the mean, more commonly known as the *variance*. Denoting the variance of X by $\boldsymbol{V}(X)$,

$$\boldsymbol{V}(X) = \sum_x [x - \boldsymbol{E}(X)]^2 p(x) \qquad \text{when X is discrete}$$

and

$$\boldsymbol{V}(X) = \int_x [x - \boldsymbol{E}(X)]^2 f(x)dx \qquad \text{when X is continuous.}$$

In both the discrete and continuous case, the variance can be shown to equal the second moment minus the expectation squared. That is,

$$\boldsymbol{V}(\mathrm{X}) = \boldsymbol{E}[\mathrm{X}^2] - \boldsymbol{E}(\mathrm{X})^2$$

This form is often more convenient to use in algebraic manipulations. Of course, if you have the expectation, it is easy to convert between the variance and the second moment in either direction.

The variance, being defined as a weighted average of the squared deviations from the expectation, is a measure of the spread, or dispersion, of a probability distribution. One of the objections to its use for this purpose is that the units are not those of X but of X^2. The *standard deviation*, defined as the square root of the variance, overcomes this objection. We conventionally use a lowercase sigma for standard deviations, so the definition would be

$$\sigma_{\mathrm{X}} = \sqrt{\boldsymbol{V}(\mathrm{X})}$$

It is usually a little easier to work with variances than with standard deviations (just because they avoid the square root), but there are certainly times when the standard deviation is more meaningful. Either is a simple one-to-one transformation of the other, so they both convey the same information.

Sometimes we may want to express the *relative* amount of variation in a random variable, rather than an absolute measure. For example, suppose we had two random variables X and Y, which are measured on different scales (say meters and kilograms), and we wanted to say which was more variable than the other. It would make no sense to compare the variances or even the standard deviations, because the dimensional units (meters and kilograms) are not consistent.

One way to express a relative measure of variability is the *coefficient of variation*, defined as the ratio of the standard deviation to the mean,

$$\boldsymbol{C}_{\mathrm{X}} = \frac{\sigma_{\mathrm{X}}}{\boldsymbol{E}(\mathrm{X})}$$

This is a dimensionless value because the units in the numerator and denominator cancel out. A value close to zero would mean that the standard deviation is much less than the mean, and a value greater than one (or less than -1) would mean that the standard deviation is more than the mean. The standard deviation is always positive, but the mean could be negative, so the coefficient of variation could take on a negative value. However, the interpretation of relative variation would be same.

The properties of variances, standard deviations, and coefficients of variation are not so obvious as those of expectations. Whereas the behavior of expectations conforms to what intuition would suggest, considerable care must be exercised in dealing with the others. The rules for dealing with multiplicative and additive constants are

$$\boldsymbol{V}(\mathrm{aX}) = \mathrm{a}^2\boldsymbol{V}(\mathrm{X})$$

and

$$\boldsymbol{V}(\mathrm{X} + \mathrm{b}) = \boldsymbol{V}(\mathrm{X})$$

In words, a multiplicative constant can be ‘‘pulled out’’ of a variance, but must be squared; an additive constant can be ‘‘dropped out.’’ When considering a sum of random variables, the variance of the sum will be the sum of the variances, if the random variables are independent. For two independent random variables X and Y,

$$\boldsymbol{V}(\mathrm{X} + \mathrm{Y}) = \boldsymbol{V}(\mathrm{X}) + \boldsymbol{V}(\mathrm{Y})$$

On the other hand, if the random variables are dependent, this relation will not generally hold. The correct expression for the general case requires another definition.

Given two random variables X and Y, the *covariance* of X and Y is defined by

$$\boldsymbol{COV}(\mathrm{X},\ \mathrm{Y}) = \boldsymbol{E}([\mathrm{X} - \boldsymbol{E}(\mathrm{X})][\mathrm{Y} - \boldsymbol{E}(\mathrm{Y})])$$

but this expression can be shown to equal

$$\boldsymbol{COV}(\mathrm{X},\ \mathrm{Y}) = \boldsymbol{E}(XY) - \boldsymbol{E}(X)\boldsymbol{E}(Y)$$

It will be recalled that when X and Y are independent, $\boldsymbol{E}(\mathrm{XY}) = \boldsymbol{E}(\mathrm{X})\boldsymbol{E}(\mathrm{Y})$, so the covariance of independent random variables is zero. The converse does not always hold; that is, the mere knowledge that the covariance of random variables is zero would not be enough for one to conclude that they are independent. Indeed, examples can be constructed of dependent random variables for which the covariance equals zero. On the other hand, a nonzero covariance definitely implies a relationship between the random variables, so the covariance is used as a (somewhat imperfect) measure of the degree of dependence. Another related measure of dependence is the *correlation coefficient* between X and Y, usually denoted by ρ, (lowercase Greek letter rho), which is defined as

$$\rho = \frac{\boldsymbol{COV}(\mathrm{X},\ \mathrm{Y})}{\sqrt{\boldsymbol{V}(\mathrm{X})\boldsymbol{V}(\mathrm{Y})}}$$

Returning to the variance of a sum of random variables, the general equation for two random variables is

$$\boldsymbol{V}(\mathrm{X} + \mathrm{Y}) = \boldsymbol{V}(\mathrm{X}) + \boldsymbol{V}(\mathrm{Y}) + 2\boldsymbol{COV}(\mathrm{X},\ \mathrm{Y})$$

1.9 The Law of Total Probability

There is an extremely useful equation relating conditional probabilities, sometimes called the law of total probability. We will also refer to the concept as ‘‘conditioning,’’ because it is a common way to develop expressions that are helpful in computations of either probabilities or expected values. That is, when we are faced with the need to find a complicated probability or expectation, we ‘‘condition’’ on some other random variable to simplify the task. This is a very important idea—it is probably not exaggerating to call it the key idea in stochastic processes—so you should be sure that you understand what is going on. It will not be enough to remember a formula, because the notation will change with the circumstances. You must understand the idea and adapt the notation to whatever situation you are in when you need to use it.

You have already seen it in one form, back in relation 12 on page 6. It was

$$\boldsymbol{P}(\mathrm{A}) = \sum_{\mathrm{i}} \boldsymbol{P}(\mathrm{A}|\mathrm{B_i})\boldsymbol{P}(\mathrm{B_i}) \text{ for any partition } \mathrm{B_1},\ \mathrm{B_2},\ \ldots,\ \mathrm{B_n}$$

Another form, expressed in terms of discrete random variables, would be

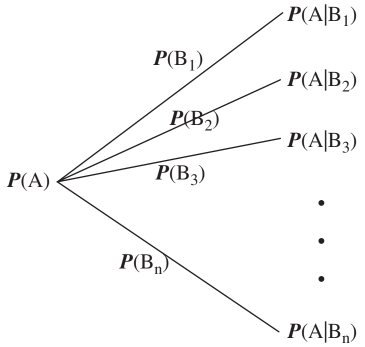
$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}) = \sum_{\mathrm{y}} \boldsymbol{P}(\mathrm{X} = \mathrm{x}|\mathrm{Y} = \mathrm{y})\boldsymbol{P}(\mathrm{Y} = \mathrm{y})$$

or, with shortened notation,

$$\mathrm{p(x)} = \sum_{\mathrm{y}} \mathrm{p(x|y)p(y)}$$

The same relation for continuous random variables would be

$$\mathrm{f(x)} = \int_{\mathrm{y}} \mathrm{f(x|y)f(y)dy}$$


$\boldsymbol{P}(\mathrm{A}|\mathrm{B}_1)$
$\boldsymbol{P}(\mathrm{B}_1)$
$\boldsymbol{P}(\mathrm{A}|\mathrm{B}_2)$
$\boldsymbol{P}(\mathrm{B}_2)$
$\boldsymbol{P}(\mathrm{A}|\mathrm{B}_3)$
$\boldsymbol{P}(\mathrm{A})$
$\boldsymbol{P}(\mathrm{B}_3)$
$\boldsymbol{P}(\mathrm{B}_n)$
$\boldsymbol{P}(\mathrm{A}|\mathrm{B}_n)$

■ **FIGURE 1.7 Decomposing an event**

Both of these expressions also appeared earlier in the chapter. There can be a lot of variations in the way the law appears, but they are all based on the same idea.

EXAMPLE 1.2

Here is the thinking you go through: Suppose, to take a concrete example, that we need to find the probability of some event A. Suppose further that the situation is too complicated to know $\boldsymbol{P}(\mathrm{A})$ directly, but we could know the probability of A if we knew which of a number of possible conditions held. That is, we know the conditional probabilities $\boldsymbol{P}(\mathrm{A}|\mathrm{B}_\mathrm{i})$ for a set of mutually exclusive, collectively exhaustive conditions, B_i. It would make some sense to "average" these various possible values for the probability of A. But if the conditions B_i, are not all equally likely, the various $\boldsymbol{P}(\mathrm{A}|\mathrm{B}_\mathrm{i})$ should not be given equal weight in the average; each should be weighted according to the probability that the condition B_i does in fact hold, or $\boldsymbol{P}(\mathrm{B}_\mathrm{i})$. This logic produces relation 12.

Another way to see the relation is to imagine breaking the event down into a set of alternative possibilities, as shown in Figure 1.7. (This is something like a decision tree.) Assuming that we have some way to get the separate conditional probabilities, all we have to do is to weight the branches by the probabilities that the separate conditions hold—the $\boldsymbol{P}(\mathrm{B}_\mathrm{i})$. Of course, the related B_i events must be such that one and only one of them will hold, which corresponds exactly to the requirement that they form a partition.

The reason that the conditioning argument is so useful in the context of stochastic processes is that the events we condition upon—the events indicated by B_i—are events that happen sometime before A in the progress of time. For example, we can often say something about the probability of A if we know what happened just before.

A very similar idea can be used to find expected values. You have seen some expressions of the idea under conditional expectations, but let's frame the issue in more intuitive terms here. Faced with the problem of expressing $\boldsymbol{E}(\mathrm{X})$, for some random variable X, one might try to find another random variable, Y, whose distribution is known or can be found, and which has the property that when the value of Y is specified, the expectation of X is easy to obtain. Usually, however, our use of the concept will be such that the conditional expectation may be known directly.

EXAMPLE 1.3

For example, suppose that we are interested in an inventory problem and X represents the number of units of some product sold during a specified period. If Y represents the number of customers who purchase some number of units during the period, and if the expectation of the number of units purchased is the same for each customer, say 3.6 units, then the conditional expectation of the number of units sold given that the number of customers is y would be 3.6y, for any y. That is, we obtain

$$\boldsymbol{E}(\mathrm{X}|\mathrm{y}) = 3.6\mathrm{y}$$

without having to use the conditional probability distribution $p(X|y)$. The details of the logic involved are probably unnecessary, but just to verify rigorously that the result is correct, we may argue as follows: The number of units sold, X, is the sum of the amounts sold to each individual customer. If the number of customers is specified to be y, then X will consist of the sum of y random variables. The expectation of a sum is the sum of the expectations; if each of these is the same, namely 3.6, the sum of y of them is 3.6y.

Once we have the conditional expectations, we put them together in the obvious way, namely, by taking a weighted sum, where the weights are the probabilities for the respective values of y. In notation, this expression is

$$\boldsymbol{E}(\mathrm{X}) = \sum_{\mathrm{y}} \boldsymbol{E}(\mathrm{X}|\mathrm{y})\mathrm{p}(\mathrm{y})$$

which is the expression shown earlier for conditional expectations.

Again, in the context of stochastic processes, the typical use of the conditioning argument is to condition on events at some prior time. You will see the idea applied numerous times over the next few chapters.

1.10 Discrete Probability Distributions

There is an infinite variety of functions that satisfy the requirements to be probability distributions. Only a few occur so commonly that they have been given names. In statistical applications, you will almost always find yourself using named distributions (for example, Normal, Student's t, F), but in real-world modeling applications you will more frequently construct distributions that are unnamed. Here we will mention just a few of the most commonly used distributions.

The Discrete Uniform Distribution

When a random variable X has only a finite number of possible values, each of which can occur with equal likelihood, the distribution is called *discrete uniform*. Without serious loss of generality, we may assume that the range of X is $x = 1, 2, \ldots, N$, in which case the probability distribution function is

$$\mathrm{p}(\mathrm{x}) = \frac{1}{\mathrm{N}} \qquad \text{for } \mathrm{x} = 1, 2, \ldots, \mathrm{N}$$

When X has this range, the mean and variance are

$$\boldsymbol{E}(X) = \frac{\mathrm{N}+1}{2}$$

and

$$\boldsymbol{V}(X) = \frac{\mathrm{N}^2 - 1}{12}$$

Of course, a shift or scaling of the range of X will have a corresponding effect upon the pdf, mean, and variance. In any case, the pdf is just 1 divided by the total number of possible values, for each value, and the expectation falls at the midpoint of the range.

Although it has many uses, the discrete uniform distribution is not so important as it is frequently thought to be by beginners in probability. Elementary textbooks often give so much emphasis to combinatorial probability—using permutations and combinations to count the number of ways that events could occur and using these counts (together with the assumption of equal likelihood) to form probabilities—that it is easy to develop a concept of probability theory that is limited to this one special case. It is important to realize that the discrete uniform distribution is just one of many useful distributions.

The Bernoulli Distribution

If a random variable must assume one of two values (usually, but not always 0 or 1), it is said to be a Bernoulli random variable. The corresponding experiment, which has only two possible outcomes, is called a Bernoulli trial. Usually the outcome that is mapped by the random variable onto the value 1 is called a *success* and the other is called a *failure*. The distribution is given by

$$p(1) = p$$

and

$$p(0) = 1 - p \text{ or } q$$

where p is the only parameter of the distribution, often referred to as the ‘‘probability of success.’’ (Note that the p on the left is a pdf, while the p on the right is a parameter.) The mean of a Bernoulli defined on 0 and 1 is p. The variance is pq.

The distribution may seem so trivial as to be undeserving of special attention. Although it is true that direct applications are limited, it turns out that a number of more important distributions can be derived from considering a sequence of independent Bernoulli trials. We will return to this subject after we see some more distributions.

The Binomial Distribution

Let X be a discrete random variable defined over the range $x = 0, 1, 2, \ldots, n$. If

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

then we say that X has a binomial distribution with parameters n and p, where n is a positive integer and $0 \leq p \leq 1$. The notation

$$\binom{n}{x}$$

refers to the so-called binomial coefficient defined by

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

A binomially distributed random variable usually can be thought of as counting the number of successes in a sequence of n independent Bernoulli trials, where the probability of success on any trial is p. Tables of binomial coefficients and the binomial distribution are readily available.

The expectation, or mean number of successes, for a binomial is np. The variance is $np(1 - p)$.

The Poisson Distribution

Let X be a discrete variable defined over the range $x = 0, 1, 2, \ldots, \infty$. If

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

then we say that X has a Poisson distribution with parameter λ, where λ must be positive. The Poisson distribution has a number of convenient properties that contribute to its usefulness in modeling. The expectation and variance are equal to one another, and are given simply by the parameter of the distribution:

$$\boldsymbol{E}(X) = \boldsymbol{V}(X) = \lambda$$

The distribution is reproductive; that is, the sum of independent Poisson distributed random variables will be another Poisson distributed random variable. The parameter of the sum random variable will be just the sum of the parameters of the constituent random variables.

One of the common usages of the Poisson distribution is as an approximation to the binomial distribution when the number of trials (n) becomes large while the probability of occurrence (p) becomes small. All that is required for the approximation is to give the two distributions the same expectation. That is, let $\lambda = np$.

Another common use of the Poisson distribution is to describe the number of events occurring within some period of time. In this context, it is the usual practice to use λt as the parameter of the distribution, where t is interpreted as the length of the period and λ is now the mean ‘‘rate’’ at which events occur. The Poisson process and its properties will be discussed in some detail in Chapter 6.

The Geometric Distribution

There are two common versions of the geometric distribution. If X is defined over the range $x = 1, 2 \ldots, \infty$ and has the pdf

$$p(x) = p(1-p)^{x-1} \qquad \text{for } x = 1, 2, 3, \ldots, \infty$$

where $0 \leq p \leq 1$, we would say that X has the geometric distribution beginning at 1. If it is defined over the range $x = 0, 1 \ldots, \infty$ (that is, starting at zero rather than 1)
and

$$p(x) = p(1-p)^{x} \qquad \text{for } x = 0, 1, 2 \ldots, \infty$$

we would say that X has the geometric distribution beginning at zero. It is apparent that one version is just a shifted version of the other, and that other shifts could be made without altering the form of the distribution. Both of these versions appear in applications and are easily confused.

The expectations and variance for the geometric distribution beginning at 1 are, respectively,

$$\boldsymbol{E}(X) = \frac{1}{p}$$

and

$$\boldsymbol{V}(X) = \frac{1-p}{p^2}$$

When the distribution begins at zero, the variance is the same, but the expectation is $(1-p)/p$.

A possible interpretation of X, when it begins at 1, is as the number of trials in a sequence of independent Bernoulli trials that will occur before the first success is observed. More precisely, it is the number of the trial on which the first success occurs. If X begins at zero, it could be thought of as counting the number of failures before the first success. In either case, X counts trials, so the geometric distribution is often regarded as a waiting-time distribution. One should not confuse this interpretation of the geometric distribution with that of the binomial distribution. The latter fixes the number of trials and counts successes.

The Negative Binomial Distribution

Let X be a discrete random variable defined over the range $x = k,\ k+1,\ \ldots,\ \infty$. We would say that X follows a negative binomial distribution if

$$p(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \qquad \text{for } x = k,\ k+1,\ \ldots,\ \infty$$

where k is an integer >1 and $0 \leq p \leq 1$ Another name for the same distribution is the Pascal distribution. When $k = 1$, the distribution reduces to the geometric. The expectation and variance are

$$\boldsymbol{E}(X) = \frac{k}{p}$$

and

$$\boldsymbol{V}(X) = \frac{k(1-p)}{p^2}$$

The explanation for this distribution just extends that of the geometric. X represents the number of the trial, in a sequence of independent Bernoulli trials, on which the *k*th success occurs. Thus the negative binomial distribution is another waiting-time distribution. Thinking of X in this way suggests that the waiting time for the *k*th success ought to be the sum of k waiting times for the one success. Because the trials are independent, this logic is valid. It is a fact that the sum of k independent geometrically distributed random variables will yield a random variable whose distribution is negative binomial with parameter k.

Sometimes the negative binomial distribution is used without any waiting-time interpretation, but simply because the parameters can be adjusted so as to fit a set of data. In this case, it may be desirable to have the range of X begin at zero, rather than k. If so, the appropriate pdf would be

$$(x) = \binom{k+x-1}{x} p^k (1-p) \qquad \text{for } x = 0,\ 1,\ 2,\ \ldots,\ \infty$$

The variance would be the same, but the expectation would be $k(1-p)/p$.

1.11 Continuous Probability Distributions

The Continuous Uniform Distribution

When a continuous random variable X is restricted to a finite range $a \leq x \leq b$ and is such that ‘‘no value is any more likely than any other,’’ then X would be appropriately described by the continuous uniform distribution. It is the obvious analog of the discrete uniform distribution, which restricted the random variable to a finite number of equally likely values. The description, “no value more likely than any other,” is somewhat loose, because, of course, the probability of any one value for a continuous random variable is zero. A better, although less intuitive, description would be, “the probability that x falls within any interval in the range of X depends only on the width of the interval and not on its location.”

In any case, the distribution is rigorously defined by its probability density function:

$$f(x) = \frac{1}{b - a} \qquad \text{for } a \leq x \leq b$$

The expectation is at the midpoint of the range,

$$\boldsymbol{E}(X) = \frac{a + b}{2}$$

and the variance is

$$\boldsymbol{V}(X) = \frac{(b - a)^2}{12}$$

The Normal Distribution

Easily the most important continuous probability distribution, the normal distribution has been useful in countless applications involving every conceivable discipline. The usefulness is due in part to the fact that the distribution has a number of properties that make it easy to deal with mathematically. More importantly, however, the distribution happens to describe quite accurately the random variables associated with a wide variety of experiments.

The range of a normally distributed random variable consists of all real numbers. The probability density function is defined by the equation

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } -\infty \leq x \leq \infty$$

where the parameter μ is unrestricted and the parameter σ is positive.

The two parameters μ and σ used to specify the distribution happen to correspond to the mean and standard deviation, respectively, of the random variable. Any linear transformation of a normally distributed random variable is also normally distributed. That is, if X is normal with mean μ and variance σ^2, and if $Y = aX + b$, then Y is normally distributed with mean

$$\boldsymbol{E}(Y) = a\mu + b$$

and with variance

$$\boldsymbol{V}(Y) = a^2\sigma^2$$

The significance of these facts is that every normal distribution, whatever the values of the parameters, can be represented in terms of the *standard* normal distribution, which has a mean of zero and variance of 1. The linear transformation required to convert a normally distributed random variable X with mean μ and variance σ^2 to the standard normal random variable Z is

$$Z = \frac{X - \mu}{\sigma}$$

The density function of the standard normal random variable is just

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$$

Unfortunately, an integral of the density function cannot be evaluated by ordinary methods of calculus, so there is no closed form expression for it, other than as an integral of the density function. However, extensive tables of the cumulative distribution function are available. Once you become familiar with the tables, virtually any desired probability can be evaluated with little trouble.

The normal distribution is reproductive; that is, the sum of two or more normally distributed random variables is itself normally distributed. The mean of the sum is, as always, the sum of the means. The variance of the sum is the sum of the variances, provided that the random variables are independent. Even if they are not, the variance of the sum can be expressed in terms of the variances and covariances of the constituents.

An even more remarkable result is established by the famous central limit theorem, which states that (under certain broad conditions) the sum of a large number of independent *arbitrarily* distributed random variables will be (approximately) normally distributed. Since quite frequently a random variable of interest may be conceptualized as being composed of a large number of independent random effects, the central limit theorem explains why the normal distribution appears so often in real-life applications. It also provides justification for *assuming* that certain random variables are normally distributed.

The Negative Exponential Distribution

Let X be a continuous random variable defined over the range 0 to ∞. If

$$f(x) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0$$

where the parameter λ is positive, we say that X has the negative exponential distribution or, sometimes, just the exponential distribution. The cumulative distribution function has, in this case, a convenient expression

$$F(x) = 1 - e^{-\lambda x}$$

The complementary cumulative distribution function is even simpler:

$$G(x) = e^{-\lambda x}$$

The expectation of a negative exponentially distributed random variable is the reciprocal of the parameter

$$\boldsymbol{E}(X) = \frac{1}{\lambda}$$

and the variance is the square of the same value

$$\boldsymbol{V}(X) = \frac{1}{\lambda^2}$$

The negative exponential distribution is used extensively to describe random variables corresponding to durations. In other words, it is a waiting time distribution. It has a number of useful properties, but since these are explored fully in Chapter 6, no further discussion need be included here.

The Erlang-k Distribution

A continuous random variable defined over the range $x \geq 0$ is Erlang-k distributed if its density function is of the form

$$f(x) = \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x} \qquad \text{for } x \geq 0$$

where the parameter λ is positive and k is an integer ≥1. When $k = 1$ the density function reduces to that of a negative exponential distribution, so the Erlang-k distribution can be thought of as a generalization of the negative exponential. In fact, if we had k independent negative exponential random variables, each with the parameter λ, then the sum of these random variables would be Erlang-k distributed with parameters λ and k. If each of the negative exponential random variables is a waiting time, the Erlang-K random variable can be thought of as the time until the *k*th event.

The expectation is most easily found as the sum of the expectations of the negative exponential random variables

$$\boldsymbol{E}(\mathrm{X}) = \boldsymbol{E}(\mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_\mathrm{k})$$
$$\boldsymbol{E}(X) = \boldsymbol{E}\left(\frac{1}{\lambda} + \frac{1}{\lambda} + \ldots + \frac{1}{\lambda}\right)$$
$$\boldsymbol{E}(\mathrm{X}) = \frac{\mathrm{k}}{\lambda}$$

and the variance is found by a similar argument

$$\boldsymbol{V}(\mathrm{X}) = \frac{\mathrm{k}}{\lambda^2}$$

In addition to its use as a waiting time for the *k*th event, the Erlang-k distribution is often considered as a candidate to fit empirical data in queueing, reliability, inventory, and replacement applications. In this case, k has no physical interpretation; it is just a parameter that may be adjusted to obtain a better fit.

There are many other distributions that are not summarized here: the hypergeometric, student's t, Chi-square, Raleigh, Pearson, Beta, and Gamma, to name a few. All of them have practical uses, but this chapter has focused on just those that will come up in later chapters of this book. You may want to make a table of them for your own reference.

1.12 Where Do Distributions Come From?

The common distributions—the ones that have names—are used often because they are relatively simple and fit certain situations. In most cases, they were derived from assumptions (rather than from statistical observations). For example, when you assume that every outcome in a finite sample space has equal likelihood, you get the uniform distribution. When the assumption of equal likelihood makes sense, you can use the uniform distribution. In other circumstances, other assumptions and therefore other distributions fit the situation better. To become a good modeler, you have to learn which assumptions go with which distributions, so that you can make logical selections. All other things being equal, you would like to pick a distribution that is easy to work with—one that has only a few parameters, that has a convenient functional form, and that has desirable properties. However, you cannot pick an easy one if the required assumptions do not fit the situation.

If you do not know very much about a particular random phenomenon, one would ordinarily attempt to acquire data representing a large number of independent samples of the random variable one has in mind. Sometimes, of course, the acquisition of adequate data may be economically infeasible or even physically impossible. In these cases, there may exist theoretical justification for believing that a certain distribution family is appropriate. For example, if the phenomenon can be thought of as the number of successes in a sequence of independent Bernoulli trials, a binomial distribution would be appropriate; if it can be thought of as consisting of the sum of a large number of independent random variables, the central limit theorem would suggest the normal distribution. On other occasions, the choice of distribution is influenced by a need for particular mathematical properties.

Preferably, however, one would like to have real-world data to provide assurance that the distribution selected really does describe the real-world phenomenon. Because it is difficult to see any pattern in a raw list of values, one would ordinarily plot a histogram as a first step in identifying an appropriate distribution. The next step, that of selecting one or more candidate distribution types, requires a familiarity with the characteristics of various distribution families. In particular, one has to know what ''shapes'' a pdf is capable of assuming, in order to decide whether there is any hope of adjusting the parameters to get a pdf that looks like the histogram. *A Guide to Probability Theory and Some of Its Applications*, by C.L. Derman et al. (referenced on page 29) provides especially good descriptions of all of the distribution types summarized only briefly here, as well as a number of others that have not even been mentioned. It also provides guidance on how to fit each distribution to particular data, and gives examples.

Once a distribution type is at least tentatively selected, the next problem is to set values for the parameters that fix the distribution within the family. Unless other external factors intervene, one would usually use the data to estimate, in the formal statistical sense, values for the parameters. In a few cases, the statistics to use are obvious. For example, the parameter λ in a Poisson distribution is estimated by the sample mean, and μ and σ^2 in a normal distribution are estimated by the sample mean and sample variance, respectively. In other cases, however, the appropriate statistic is not so obvious. The Derman et al. book also is useful in providing this kind of information.

After the parameters are adjusted so as to provide the best fit to the data that a selected distribution type can provide, one is still left with the question of whether the fit is good enough. In other words, you should validate your model by checking the goodness of fit. As a bare minimum, you could graph the precise pdf over the histogram (using vertical scales that permit comparison), and observe the discrepancies. A more formal procedure would be to perform any of several available statistical tests for goodness of fit. The chi-square and the Kolmogorov-Smirnov goodness-of-fit tests are probably the best known. Descriptions of these two tests can be found in almost all intermediate-level statistics textbooks.

One of the basic points to bear in mind about statistical goodness-of-fit testing is that the null hypothesis assumes that the candidate distribution is correct. Only if the discrepancies between the data and the candidate distribution are significantly large will the test cause you to reject the candidate. In other words, the test is, by its very nature, biased in favor of whatever distribution you have selected to test. The mere fact that the test does not reject the distribution should not be taken as strong evidence that the selected distribution is correct. Others might have selected different distributions and come up with just as much confirmation that their choices were correct. This is particularly likely to occur when the data base is small.

The word to describe the capability of a statistical test to detect that a null hypothesis is false is *power*. Other factors being equal, a greater amount of data will make for a more powerful test. To obtain a very powerful test, however, may require truly enormous quantities of data—orders of magnitude greater than would be required for good hypothesis tests about parameters. It is easy to see why this is so if you think about how many total observations are required to provide enough information about the ''tails'' of a distribution to ensure that you have obtained a proper fit.

As a final philosophical point, it is well to keep in mind that *no amount of data can confirm absolutely that you have selected the correct distribution*. Ultimately, (and this is the main point of this discussion) there is no escape from having to make assumptions. On the other hand, remember that there is no need for a model to be perfect. It only has to be adequate to be useful.

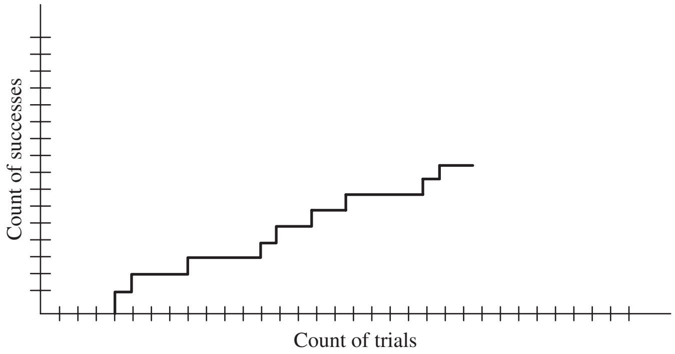
1.13 The Binomial Process

There is a very simple stochastic process that we can begin our study with. It is really too simple to do very much with, but it relates several of the discrete distributions to one another and may help to set the stage for more practical extensions.

We obtain the process by assuming we have a continuing sequence of independent Bernoulli trials. That is, on any one trial we have only two possible outcomes, called success and failure. We can get that random variable from any sample space by considering any event and its complement. For example, we could say that something happens (success) or it does not (failure). Then we imagine repeating the same experiment and keeping a running total of the number of successes. We could graph the results from any sequence on a chart like that in Figure 1.8. Each time there is a success, the count steps up by one; for each failure, the count remains level. (In Figure 1.8, we had three failures followed by two successes, and so forth.)

We call this sequence a *binomial process*, not to be confused with a binomial distribution. Of course, there is a connection. If, in advance of observing a binomial process, we specify some fixed number of trials we are going to run, n, and ask for the distribution of the total number of successes we will experience, that random variable will have a binomial distribution. (Go back and read the definition if you do not see why.) But we could also look at the binomial process in some other ways to get different random variables with different distributions. If we start the process and ask for the number of trials until the first success, we get a geometrically distributed random variable. (Again, make sure you understand why.) Or, if we want to reach a certain number of successes, say k, and ask how many trials that will take, we get a negative binomial random variable. We can even get Poisson or normal random variables if p is small and n is large. The distribution you get depends upon what question you are asking, even though the underlying process—the sequence of independent Bernoulli trials—is the same.

The binomial process is an elementary example of a stochastic process. It tracks the (uncertain) progress of a variable over time. Although it is useful for some simple things, it is limited by two constraints: the trials must be independent, and we can only increment the variable by one (or zero) unit for each time step. By the end of the next chapter, we will be able to fully escape from both of those constraints and have a much more useful class of stochastic processes.



■ FIGURE 1.8 A binomial process

1.14 Recommended Reading

If any of the topics mentioned in this chapter seems hazy, or if you would just feel more confident about proceeding if you work some problems, you should by all means devote some time to an elementary textbook on probability. There are many fine ones available. Unfortunately for the purposes of this book, the orientation of many beginning texts leans toward statistical, as opposed to modeling, applications. Also, the more recent textbooks tend to be encyclopedic in coverage, rather than concisely focused on the most important introductory topics. However, any of the older books by Clarke (1), Cramer (2), Drake (4), or Meyer (10), should serve the purpose adequately. If one does not suit your taste, feel free to select another. These older books are out of print, but can be found in the library. Feller's two volumes, (5) and (6), are classics familiar to everyone seriously interested in probability. Even beginners can find much of interest in them. The first volume deals with discrete distributions; the second, with continuous distributions. If you want to buy an inexpensive book, some of the Dover paperbacks (7), (8), and (12) are reprints of excellent older textbooks.

1. Clarke, B., and R. Disney, *Probability and Random Processes for Engineers and Scientists*. Wiley, New York, 1970.

2. Cramer, H., *The Elements of Probability Theory and Some of Its Applications*. Wiley, New York, 1955.

3. Derman, C., L. J. Gleser, and I. Olkin, *A Guide to Probability Theory and Application*. Holt, Rinehart, and Winston, New York, 1973.

4. Drake, A. W., *Fundamentals of Applied Probability Theory*. McGraw-Hill, New York, 1967.

5. Feller, W., *An Introduction to Probability Theory and Its Applications*, vol. I, 2nd ed. Wiley, New York, 1957.

6. Feller, W., *An Introduction to Probability Theory and Its Applications*, vol. II. Wiley, New York, 1966.

7. Freund, John E., *Introduction to Probability*. Dover, New York, 1973.

8. Goldberg, Samuel, *Probability: An Introduction*. Dover, New York, 1960.

9. Hsu, Hwei, *Probability, Random Variables, and Random Processes*. Schaum's Outlines, McGraw-Hill, New York, 1997.

10. Meyer, P. L., *Introductory Probability and Statistical Applications*, Addison-Wesley, Reading, MA, 1965.

11. Parzen, Emanuel, *Modern Probability Theory and Its Applications*. Wiley. New York, 1960.

12. Pfeiffer, Paul E., *Concepts of Probability Theory*. Dover, New York, 1978.

13. Ross, Sheldon M., *A First Course in Probability*. Macmillan, New York, 1976.

Chapter 1 Problems

Note: This chapter is a *review* of material you should have learned before. The following problems are designed to test your understanding of basic probability concepts and rules and to help you assess your readiness for the course. If any of them give you trouble, you should immediately begin remedial work, using some more complete introductory probability textbook.

Sets and Basic Rules of Probability

1. Imagine an experiment in which one student is selected at random from among all currently enrolled students in this university. Let A be the event that the selected student is classified as enrolled in engineering (one of the engineering schools), and let B be the event that the same selected student is currently enrolled in this class. Express in set notation the following events,

- **a.** The student is not in engineering.
- **b.** The student is in engineering and in this class.
- **c.** The student is not in engineering but is in this class.
- **d.** The student is not in engineering and is not in this class.
- **e.** The student is either in engineering or is in this class.
- **f.** The student is either in engineering or in this class, but not both.

2. A new television show has been prepared, but has not yet been broadcast. Let A be the event that, after the first appearance, it gets good reviews by critics. Let B be the event that it is popular with the public. Let C be the event that it is liked by advertisers. Express in set notation the following events.

- **a.** The show is liked by critics, the public, and advertisers.
- **b.** Critics do not like the show, but it is popular with the public and advertisers.
- **c.** Critics and advertisers like the show, but the public does not care for it.
- **d.** None of the three target audiences likes the show.

3. Suppose there are five horses in a horse race. Describe three different sample spaces for the outcomes of the race, depending upon your interest:

- **a.** You bet on a single horse and care whether you win or lose.
- **b.** You care which of the five horses wins.
- **c.** You care about which horses come in first, second, and third.

4. Suppose that an experiment has five possible outcomes, which are denoted $\{1, 2, 3, 4, 5\}$. Let A be the event $\{1, 2, 3\}$ and let B be the event $\{3, 4, 5\}$. (Notice that we did not say that the five outcomes are equally likely; the probability distributions could be anything.) For each of the following relations, tell whether it could possibly hold. If it could, give a numerical example using a probability distribution of your own choice; if it could not, explain why not (what rule is violated).

- **a.** $\boldsymbol{P}(\mathrm{A}) = \boldsymbol{P}(\mathrm{B})$
- **b.** $\boldsymbol{P}(\mathrm{A}) = 2\boldsymbol{P}(\mathrm{B})$
- **c.** $\boldsymbol{P}(\mathrm{A}) = 1 - \boldsymbol{P}(\mathrm{B})$
- **d.** $\boldsymbol{P}(\mathrm{A}) + \boldsymbol{P}(\mathrm{B}) > 1$
- **e.** $\boldsymbol{P}(\mathrm{A}) - \boldsymbol{P}(\mathrm{B}) < 0$
- **f.** $\boldsymbol{P}(\mathrm{A}) - \boldsymbol{P}(\mathrm{B}) > 1$

5. The sample space of a particular experiment is given by $\mathrm{S} = \{0, 1, 2, 3, 4, 5\}$. Let three events be defined as $\mathrm{A} = \{0, 1, 2\}$, $\mathrm{B} = \{0, 2, 4\}$, and $\mathrm{C} = \{1, 3, 5\}$. Assume that the probabilities of A, B, and C are given, but no further information is available. (Note, in particular, that we are not assuming equal likelihood for the elementary outcomes.) Express the probabilities of as many of the following events as you can.

- **a.** $\mathrm{A} \cap \mathrm{B}$
- **b.** $\mathrm{B} \cup \mathrm{C}$
- **c.** $\overline{\mathrm{A}}$
- **d.** $\mathrm{B} \cap \overline{\mathrm{C}}$
- **e.** $\overline{(\mathrm{A} \cap \mathrm{B}) \cup \mathrm{C}}$

6. Prove relation 4 on page 5 using only the axioms 1, 2, and 3, and the rules of set theory. (This is just an exercise in set theory, not a complicated proof.)

7. Prove relation 5 on page 5 using only the axioms 1, 2, and 3, and the rules of set theory. (This is just an exercise in set theory, not a complicated proof.)

Joint and Conditional Probabilities and Independence

8. For each of the following pairs of events, categorize them as independent or dependent and explain your choice.

- **a.** Rain today, rain tomorrow.
- **b.** Rain today, rain one month from today.
- **c.** Rain one year ago today, rain today.
- **d.** Receiving the grade of A in an introductory probability course; receiving grade of A in this course (same person).
- **e.** Receiving the grade of A in freshman-level physics; receiving same grade in this course.

9. If two events are known to be mutually exclusive, could they also be independent? Could they be dependent? If they are known to be independent, could they also be mutually exclusive? Could they be not mutually exclusive? If they are *not* mutually exclusive, could they be independent? Could they be dependent? If they are *not* independent, could they be mutually exclusive? Is is possible that dependent events could be not mutually exclusive? (Some of these questions are actually the same question, expressed in different words. The questions are meant to help you get the distinctions straight in your mind.)

10. If A, B, and C are events, and we know that the pair A and B are independent, and that B and C are independent, can we conclude that A and C are independent?

11. A graduating senior seeking a job has interviews with two companies. After the interviews, he estimates that his chance of getting an offer from the first company is 0.6. He thinks he has a 0.5 chance with the second company, and that the probability that at least one will reject him is 0.8. What is the probability that he gets at least one offer?

12. About 10 percent of the population is left-handed. Of those who are right-handed, about 40 percent own dogs. If you were to select a person at random, what is the joint probability that the chosen person is a right-handed and does *not* own a dog?

13. If A and B are two events, with neither being empty sets or the entire sample space, prove that if $\boldsymbol{P}(\mathrm{A}|\mathrm{B}) > \boldsymbol{P}(\mathrm{A})$ then $\boldsymbol{P}(\mathrm{B}|\overline{\mathrm{A}}) < \boldsymbol{P}(\mathrm{B})$.

Distributions

14. Here is a table giving the joint distribution of two random variables X and Y.

XY	**1**	**2**	**3**	**4**	**5**
2	0.10	0.05	0.15	0.10	0.10
4	0.04	0.02	0.06	0.04	0.04
6	0.04	0.02	0.06	0.06	0.02
8	0.02	0.01	0.03	0	0.04

What is the conditional probability $\boldsymbol{P}\{Y = 6|X = 2\}$?

15. Using the same joint distribution as shown above, are X and Y independent? (Give a yes or no answer, and explain why or why not.)

16. Using the same joint distribution as shown above, give the marginal distribution of Y.

17. Using the same joint distribution as shown above, what is the covariance of X and Y?

18. Suppose that the density function of a continuous random variable is $f(x) = 2x$, for values of x in the range [1, a] and $f(x) = 0$ elsewhere. What is the value of a?

19. Suppose that two random variables, X and Y, have a joint density function given by f(x, y) and by checking we find that $f(x, y) \neq f(x)f(y)$. That is, the two are not independent. Some of the calculated moments are: $\boldsymbol{E}(X) = 10$, $\boldsymbol{E}(Y) = 8$, $\boldsymbol{V}(X) = 9$, $\boldsymbol{V}(Y) = 4$, $\boldsymbol{E}(XY) = 84$. What is the expectation of the random variable $W = X + Y + 4$?

20. Using the same information as the previous problem, what is the variance of the random variable W?

21. Using the same information as the previous problem, what is the correlation coefficient of X and Y?

Common Distributions

22. In a sequence of 10 independent Bernoulli trials, where the probability of success is 0.4, what is the expected number of failures? (Variance?)

23. If a Bernoulli random variable X is defined so that success is given the value 10 and failure is given the value 5, and $P(\text{success}) = 0.6$, what is the expected value of X?

24. The geometric distribution describes, among other things, the waiting time until the first success in a sequence of independent Bernoulli trials. What distribution gives the corresponding waiting time to first success in continuous time?

25. If X is a Poisson distributed random variable with a mean of 3, and Y is another Poisson distributed random variable with a mean of 2, and the two are independent, what is the variance of the sum $X + Y$?

26. If X is a random variable having a binomial distribution with $n = 20$ and $p = 0.4$, and Y is a transformed version of X, where $Y = 2X + 3$, what is the expected value of Y?

Optional

27. See if you can figure out the rules that convert odds to probabilities and vice versa. For example, if you are given odds of 5:1, what should that mean in terms of probabilities? And if you are told that the probability of winning a bet is 0.25, what should the odds be? Assume that the odds reflect fair payout (rather than distorted values that leave some profit for the people handling the bet).

28. If you found problem 27 to be easy, you may want to try to express the odds equivalents of the rules of probabilities. For example, how do you combine the odds for two mutually exclusive events to get the odds of the union? (From doing this, you will learn why probabilities are so much nicer to deal with than odds.)