Appendix B

Fundamentals of Probability

This appendix covers key concepts from basic probability. Appendices B and C are primarily for review; they are not intended to replace a course in probability and statistics. However, all of the probability and statistics concepts that we use in the text are covered in these appendices.

Probability is of interest in its own right for students in business, economics, and other social sciences. For example, consider the problem of an airline trying to decide how many reservations to accept for a flight that has 100 available seats. If fewer than 100 people want reservations, then these should all be accepted. But what if more than 100 people request reservations? A safe solution is to accept at most 100 reservations. However, because some people book reservations and then do not show up for the flight, there is some chance that the plane will not be full even if 100 reservations are booked. This results in lost revenue to the airline. A different strategy is to book more than 100 reservations and to hope that some people do not show up, so the final number of passengers is as close to 100 as possible. This policy runs the risk of the airline having to compensate people who are necessarily bumped from an overbooked flight.

A natural question in this context is: Can we decide on the optimal (or best) number of reservations the airline should make? This is a nontrivial problem. Nevertheless, given certain information (on airline costs and how frequently people show up for reservations), we can use basic probability to arrive at a solution.

B-1 Random Variables and Their Probability Distributions

Suppose that we flip a coin 10 times and count the number of times the coin turns up heads. This is an example of an **experiment**. Generally, an experiment is any procedure that can, at least in theory, be infinitely repeated and has a well-defined set of outcomes. We could, in principle, carry out the coin-flipping procedure again and again. Before we flip the coin, we know that the number of heads appearing is an integer from 0 to 10, so the outcomes of the experiment are well defined.

A **random variable** is one that takes on numerical values and has an outcome that is determined by an experiment. In the coin-flipping example, the number of heads appearing in 10 flips of a coin is an example of a random variable. Before we flip the coin 10 times, we do not know how many

times the coin will come up heads. Once we flip the coin 10 times and count the number of heads, we obtain the outcome of the random variable for this particular trial of the experiment. Another trial can produce a different outcome.

In the airline reservation example mentioned earlier, the number of people showing up for their flight is a random variable: before any particular flight, we do not know how many people will show up.

To analyze data collected in business and the social sciences, it is important to have a basic understanding of random variables and their properties. Following the usual conventions in probability and statistics throughout Appendices B and C, we denote random variables by uppercase letters, usually W, X, Y, and Z; particular outcomes of random variables are denoted by the corresponding lowercase letters, w, x, y, and z. For example, in the coin-flipping experiment, let X denote the number of heads appearing in 10 flips of a coin. Then, X is not associated with any particular value, but we know X will take on a value in the set $\{0, 1, 2, \ldots, 10\}$. A particular outcome is, say, $x = 6$.

We indicate large collections of random variables by using subscripts. For example, if we record last year's income of 20 randomly chosen households in the United States, we might denote these random variables by $X_1, X_2, \ldots, X_{20}$; the particular outcomes would be denoted $x_1, x_2, \ldots, x_{20}$.

As stated in the definition, random variables are always defined to take on numerical values, even when they describe qualitative events. For example, consider tossing a single coin, where the two outcomes are heads and tails. We can define a random variable as follows: $X = 1$ if the coin turns up heads, and $X = 0$ if the coin turns up tails.

A random variable that can only take on the values zero and one is called a **Bernoulli** (or **binary**) **random variable**. In basic probability, it is traditional to call the event $X = 1$ a "success" and the event $X = 0$ a "failure." For a particular application, the success-failure nomenclature might not correspond to our notion of a success or failure, but it is a useful terminology that we will adopt.

B-1a Discrete Random Variables

A **discrete random variable** is one that takes on only a finite or countably infinite number of values. The notion of "countably infinite" means that even though an infinite number of values can be taken on by a random variable, those values can be put in a one-to-one correspondence with the positive integers. Because the distinction between "countably infinite" and "uncountably infinite" is somewhat subtle, we will concentrate on discrete random variables that take on only a finite number of values. Larsen and Marx (1986, Chapter 3) provide a detailed treatment.

A Bernoulli random variable is the simplest example of a discrete random variable. The only thing we need to completely describe the behavior of a Bernoulli random variable is the probability that it takes on the value one. In the coin-flipping example, if the coin is "fair," then $P(X = 1) = 1/2$ (read as "the probability that X equals one is one-half"). Because probabilities must sum to one, $P(X = 0) = 1/2$, also.

Social scientists are interested in more than flipping coins, so we must allow for more general situations. Again, consider the example where the airline must decide how many people to book for a flight with 100 available seats. This problem can be analyzed in the context of several Bernoulli random variables as follows: for a randomly selected customer, define a Bernoulli random variable as $X = 1$ if the person shows up for the reservation, and $X = 0$ if not.

There is no reason to think that the probability of any particular customer showing up is 1/2; in principle, the probability can be any number between 0 and 1. Call this number θ, so that

$$P(X = 1) = \theta \qquad [B.1]$$

$$P(X = 0) = 1 - \theta. \qquad [B.2]$$

For example, if $\theta = .75$, then there is a 75% chance that a customer shows up after making a reservation and a 25% chance that the customer does not show up. Intuitively, the value of θ is crucial in determining the airline's strategy for booking reservations. Methods for *estimating* θ, given historical data on airline reservations, are a subject of mathematical statistics, something we turn to in Appendix C.

More generally, any discrete random variable is completely described by listing its possible values and the associated probability that it takes on each value. If X takes on the k possible values $\{x_1, \ldots, x_k\}$, then the probabilities $p_1, p_2, \ldots, p_k$ are defined by

$$p_j = \mathrm{P}(X = x_j), j = 1, 2, \ldots, k, \qquad \text{[B.3]}$$

where each p_j is between 0 and 1 and

$$p_1 + p_2 + \ldots + p_k = 1. \qquad \text{[B.4]}$$

Equation (B.3) is read as: "The probability that X takes on the value x_j is equal to p_j."

Equations (B.1) and (B.2) show that the probabilities of success and failure for a Bernoulli random variable are determined entirely by the value of θ. Because Bernoulli random variables are so prevalent, we have a special notation for them: $X \sim \text{Bernoulli}(\theta)$ is read as "X has a Bernoulli distribution with probability of success equal to θ."

The **probability density function (pdf)** of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, j = 1, 2, \ldots, k, \qquad \text{[B.5]}$$

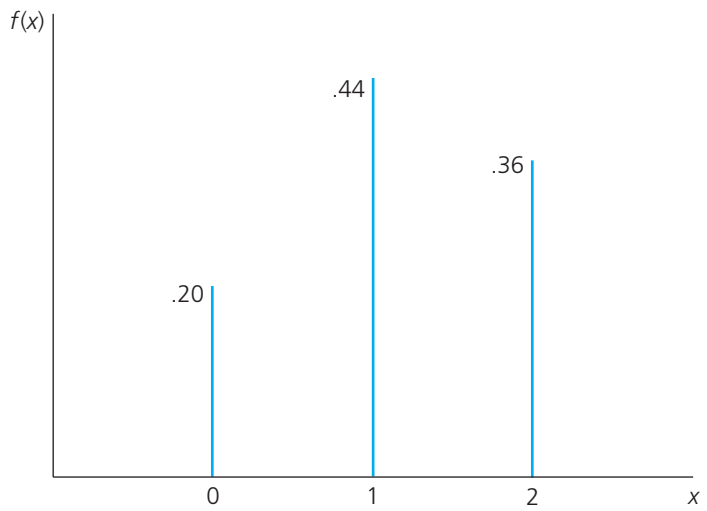
with $f(x) = 0$ for any x not equal to x_j for some j. In other words, for any real number x, $f(x)$ is the probability that the random variable X takes on the particular value x. When dealing with more than one random variable, it is sometimes useful to subscript the pdf in question: f_X is the pdf of X, f_Y is the pdf of Y, and so on.

Given the pdf of any discrete random variable, it is simple to compute the probability of any event involving that random variable. For example, suppose that X is the number of free throws made by a basketball player out of two attempts, so that X can take on the three values $\{0, 1, 2\}$. Assume that the pdf of X is given by

$$f(0) = .20, f(1) = .44, \text{ and } f(2) = .36.$$

The three probabilities sum to one, as they must. Using this pdf, we can calculate the probability that the player makes *at least* one free throw: $\mathrm{P}(X \geq 1) = \mathrm{P}(X = 1) + \mathrm{P}(X = 2) = .44 + .36 = .80$. The pdf of X is shown in Figure B.1.

FIGURE B.1 The pdf of the number of free throws made out of two attempts.



B-1b Continuous Random Variables

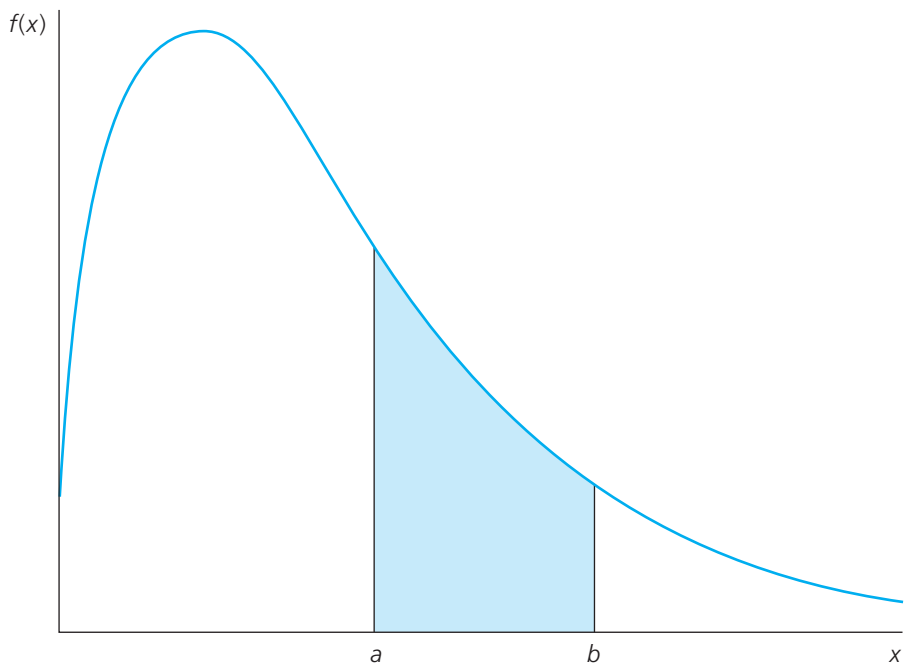
A variable X is a **continuous random variable** if it takes on any real value with *zero* probability. This definition is somewhat counterintuitive because in any application we eventually observe some outcome for a random variable. The idea is that a continuous random variable X can take on so many possible values that we cannot count them or match them up with the positive integers, so logical consistency dictates that X can take on each value with probability zero. While measurements are always discrete in practice, random variables that take on numerous values are best treated as continuous. For example, the most refined measure of the price of a good is in terms of cents. We can imagine listing all possible values of price in order (even though the list may continue indefinitely), which technically makes price a discrete random variable. However, there are so many possible values of price that using the mechanics of discrete random variables is not feasible.

We can define a probability density function for continuous random variables, and, as with discrete random variables, the pdf provides information on the likely outcomes of the random variable. However, because it makes no sense to discuss the probability that a continuous random variable takes on a particular value, we use the pdf of a continuous random variable only to compute events involving a range of values. For example, if a and b are constants where $a < b$, the probability that X lies between the numbers a and b, $P(a \leq X \leq b)$, is the *area* under the pdf between points a and b, as shown in Figure B.2. If you are familiar with calculus, you recognize this as the *integral* of the function f between the points a and b. The entire area under the pdf must always equal one.

When computing probabilities for continuous random variables, it is easiest to work with the **cumulative distribution function (cdf)**. If X is any random variable, then its cdf is defined for any real number x by

$$F(x) \equiv P(X \leq x). \qquad [B.6]$$

FIGURE B.2 The probability that *X* lies between the points *a* and *b*.



For discrete random variables, (B.6) is obtained by summing the pdf over all values x_j such that $x_j \leq x$. For a continuous random variable, $F(x)$ is the area under the pdf, f, to the left of the point x. Because $F(x)$ is simply a probability, it is always between 0 and 1. Further, if $x_1 < x_2$, then $P(X \leq x_1) \leq P(X \leq x_2)$, that is, $F(x_1) \leq F(x_2)$. This means that a cdf is an increasing (or at least a nondecreasing) function of x.

Two important properties of cdfs that are useful for computing probabilities are the following:

$$\text{For any number } c,\ P(X > c) = 1 - F(c). \qquad \text{[B.7]}$$

$$\text{For any numbers } a < b,\ P(a < X \leq b) = F(b) - F(a). \qquad \text{[B.8]}$$

In our study of econometrics, we will use cdfs to compute probabilities only for continuous random variables, in which case it does not matter whether inequalities in probability statements are strict or not. That is, for a continuous random variable X,

$$P(X \geq c) = P(X > c), \qquad \text{[B.9]}$$

and

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \qquad \text{[B.10]}$$

Combined with (B.7) and (B.8), equations (B.9) and (B.10) greatly expand the probability calculations that can be done using continuous cdfs.

Cumulative distribution functions have been tabulated for all of the important continuous distributions in probability and statistics. The most well known of these is the normal distribution, which we cover along with some related distributions in Section B-5.

B-2 Joint Distributions, Conditional Distributions, and Independence

In economics, we are usually interested in the occurrence of events involving more than one random variable. For example, in the airline reservation example referred to earlier, the airline might be interested in the probability that a person who makes a reservation shows up *and* is a business traveler; this is an example of a *joint probability*. Or, the airline might be interested in the following *conditional probability*: conditional on the person being a business traveler, what is the probability of his or her showing up? In the next two subsections, we formalize the notions of joint and conditional distributions and the important notion of *independence* of random variables.

B-2a Joint Distributions and Independence

Let X and Y be discrete random variables. Then, (X, Y) have a **joint distribution**, which is fully described by the *joint probability density function* of (X, Y):

$$f_{X,Y}(x, y) = P(X = x, Y = y), \qquad \text{[B.11]}$$

where the right-hand side is the probability that $X = x$ and $Y = y$. When X and Y are continuous, a joint pdf can also be defined, but we will not cover such details because joint pdfs for continuous random variables are not used explicitly in this text.

In one case, it is easy to obtain the joint pdf if we are given the pdfs of X and Y. In particular, random variables X and Y are said to be independent if, and only if,

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \qquad \text{[B.12]}$$

for all x and y, where f_X is the pdf of X and f_Y is the pdf of Y. In the context of more than one random variable, the pdfs f_X and f_Y are often called *marginal probability density functions* to distinguish them from the joint pdf $f_{X,Y}$. This definition of independence is valid for discrete and continuous random variables.

To understand the meaning of (B.12), it is easiest to deal with the discrete case. If X and Y are discrete, then (B.12) is the same as

$$P(X = x, Y = y) = P(X = x)P(Y = y); \qquad [B.13]$$

in other words, the probability that $X = x$ and $Y = y$ is the product of the two probabilities $P(X = x)$ and $P(Y = y)$. One implication of (B.13) is that joint probabilities are fairly easy to compute, since they only require knowledge of $P(X = x)$ and $P(Y = y)$.

If random variables are not independent, then they are said to be *dependent*.

EXAMPLE B.1 Free Throw Shooting

Consider a basketball player shooting two free throws. Let X be the Bernoulli random variable equal to one if she or he makes the first free throw, and zero otherwise. Let Y be a Bernoulli random variable equal to one if he or she makes the second free throw. Suppose that she or he is an 80% free throw shooter, so that $P(X = 1) = P(Y = 1) = .8$. What is the probability of the player making both free throws?

If X and Y are independent, we can easily answer this question: $P(X = 1, Y = 1) = P(X = 1)P(Y = 1) = (.8)(.8) = .64$. Thus, there is a 64% chance of making both free throws. If the chance of making the second free throw depends on whether the first was made—that is, X and Y are not independent—then this simple calculation is not valid.

Independence of random variables is a very important concept. In the next subsection, we will show that if X and Y are independent, then knowing the outcome of X does not change the probabilities of the possible outcomes of Y, and vice versa. One useful fact about independence is that if X and Y are independent and we define new random variables $g(X)$ and $h(Y)$ for any functions g and h, then these new random variables are also independent.

There is no need to stop at two random variables. If $X_1, X_2, \ldots, X_n$ are discrete random variables, then their joint pdf is $f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. The random variables $X_1, X_2, \ldots, X_n$ are **independent random variables** if, and only if, their joint pdf is the product of the individual pdfs for any $(x_1, x_2, \ldots, x_n)$. This definition of independence also holds for continuous random variables.

The notion of independence plays an important role in obtaining some of the classic distributions in probability and statistics. Earlier, we defined a Bernoulli random variable as a zero-one random variable indicating whether or not some event occurs. Often, we are interested in the number of successes in a sequence of *independent* Bernoulli trials. A standard example of independent Bernoulli trials is flipping a coin again and again. Because the outcome on any particular flip has nothing to do with the outcomes on other flips, independence is an appropriate assumption.

Independence is often a reasonable approximation in more complicated situations. In the airline reservation example, suppose that the airline accepts n reservations for a particular flight. For each $i = 1, 2, \ldots, n$, let Y_i denote the Bernoulli random variable indicating whether customer i shows up: $Y_i = 1$ if customer i appears, and $Y_i = 0$ otherwise. Letting θ again denote the probability of success (using reservation), each Y_i has a Bernoulli(θ) distribution. As an approximation, we might assume that the Y_i are independent of one another, although this is not exactly true in reality: some people travel in groups, which means that whether or not a person shows up is not truly independent of whether all others show up. Modeling this kind of dependence is complex, however, so we might be willing to use independence as an approximation.

The variable of primary interest is the total number of customers showing up out of the n reservations; call this variable X. Since each Y_i is unity when a person shows up, we can write

$X = Y_1 + Y_2 + \ldots + Y_n$. Now, assuming that each Y_i has probability of success θ and that the Y_i are independent, X can be shown to have a **binomial distribution**. That is, the probability density function of X is

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0, 1, 2, \ldots, n, \quad \text{[B.14]}$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$, and for any integer n, $n!$ (read "n factorial") is defined as $n! = n \cdot (n-1) \cdot (n-2) \cdots 1$. By convention, $0! = 1$. When a random variable X has the pdf given in (B.14), we write $X \sim \text{Binomial}(n, \theta)$. Equation (B.14) can be used to compute $P(X = x)$ for any value of x from 0 to n.

If the flight has 100 available seats, the airline is interested in $P(X > 100)$. Suppose, initially, that $n = 120$, so that the airline accepts 120 reservations, and the probability that each person shows up is $\theta = .85$. Then, $P(X > 100) = P(X = 101) + P(X = 102) + \ldots + P(X = 120)$, and each of the probabilities in the sum can be found from equation (B.14) with $n = 120$, $\theta = .85$, and the appropriate value of x (101 to 120). This is a difficult hand calculation, but many statistical packages have commands for computing this kind of probability. In this case, the probability that more than 100 people will show up is about .659, which is probably more risk of overbooking than the airline wants to tolerate. If, instead, the number of reservations is 110, the probability of more than 100 passengers showing up is only about .024.

B-2b Conditional Distributions

In econometrics, we are usually interested in how one random variable, call it Y, is related to one or more other variables. For now, suppose that there is only one variable whose effects we are interested in, call it X. The most we can know about how X affects Y is contained in the **conditional distribution** of Y given X. This information is summarized by the *conditional probability density function*, defined by

$$f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x) \quad \text{[B.15]}$$

for all values of x such that $f_X(x) > 0$. The interpretation of (B.15) is most easily seen when X and Y are discrete. Then,

$$f_{Y|X}(y|x) = P(Y = y|X = x), \quad \text{[B.16]}$$

where the right-hand side is read as "the probability that $Y = y$ given that $X = x$." When Y is continuous, $f_{Y|X}(y|x)$ is not interpretable directly as a probability, for the reasons discussed earlier, but conditional probabilities are found by computing areas under the conditional pdf.

An important feature of conditional distributions is that, if X and Y are independent random variables, knowledge of the value taken on by X tells us nothing about the probability that Y takes on various values (and vice versa). That is, $f_{Y|X}(y|x) = f_Y(y)$, and $f_{X|Y}(x|y) = f_X(x)$.

EXAMPLE B.2 **Free Throw Shooting**

Consider again the basketball-shooting example, where two free throws are to be attempted. Assume that the conditional density is

$$f_{Y|X}(1|1) = .85, f_{Y|X}(0|1) = .15$$
$$f_{Y|X}(1|0) = .70, f_{Y|X}(0|0) = .30.$$

This means that the probability of the player making the second free throw depends on whether the first free throw was made: if the first free throw is made, the chance of making the second is .85; if the

first free throw is missed, the chance of making the second is .70. This implies that X and Y are *not* independent; they are dependent.

We can still compute $P(X = 1, Y = 1)$ provided we know $P(X = 1)$. Assume that the probability of making the first free throw is .8, that is, $P(X = 1) = .8$. Then, from (B.15), we have

$$P(X = 1, Y = 1) = P(Y = 1|X = 1) \cdot P(X = 1) = (.85)(.8) = .68.$$

B-3 Features of Probability Distributions

For many purposes, we will be interested in only a few aspects of the distributions of random variables. The features of interest can be put into three categories: measures of central tendency, measures of variability or spread, and measures of association between two random variables. We cover the last of these in Section B-4.

B-3a A Measure of Central Tendency: The Expected Value

The expected value is one of the most important probabilistic concepts that we will encounter in our study of econometrics. If X is a random variable, the **expected value** (or expectation) of X, denoted $E(X)$ and sometimes μX or simply μ, is a weighted average of all possible values of X. The weights are determined by the probability density function. Sometimes, the expected value is called the *population mean*, especially when we want to emphasize that X represents some variable in a population.

The precise definition of expected value is simplest in the case that X is a discrete random variable taking on a finite number of values, say, $\{x_1, \ldots, x_k\}$. Let $f(x)$ denote the probability density function of X. The expected value of X is the weighted average

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j). \qquad [B.17]$$

This is easily computed given the values of the pdf at each possible outcome of X.

EXAMPLE B.3 Computing an Expected Value

Suppose that X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8, respectively. Then,

$$E(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

This example illustrates something curious about expected values: the expected value of X can be a number that is not even a possible outcome of X. We know that X takes on the values -1, 0, or 2, yet its expected value is 5/8. This makes the expected value deficient for summarizing the central tendency of certain discrete random variables, but calculations such as those just mentioned can be useful, as we will see later.

If X is a continuous random variable, then $E(X)$ is defined as an integral:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \qquad [B.18]$$

which we assume is well defined. This can still be interpreted as a weighted average. For the most common continuous distributions, $E(X)$ is a number that is a possible outcome of X. In this text, we will not need to compute expected values using integration, although we will draw on some well-known results from probability for expected values of special random variables.

Given a random variable X and a function $g(\cdot)$, we can create a new random variable $g(X)$. For example, if X is a random variable, then so is X^2 and $\log(X)$ (if $X > 0$). The expected value of $g(X)$ is, again, simply a weighted average:

$$\mathrm{E}[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j) \qquad [B.19]$$

or, for a continuous random variable,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \qquad [B.20]$$

EXAMPLE B.4 **Expected Value of X^2**

For the random variable in Example B.3, let $g(X) = X^2$. Then,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

In Example B.3, we computed $\mathrm{E}(X) = 5/8$, so that $[\mathrm{E}(X)]^2 = 25/64$. This shows that $\mathrm{E}(X^2)$ is *not* the same as $[\mathrm{E}(X)]^2$. In fact, for a nonlinear function $g(X)$, $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (except in very special cases).

If X and Y are random variables, then $g(X, Y)$ is a random variable for any function g, and so we can define its expectation. When X and Y are both discrete, taking on values $\{x_1, x_2, \ldots, x_k\}$ and $\{y_1, y_2, \ldots, y_m\}$, respectively, the expected value is

$$\mathrm{E}[g(X, Y)] = \sum_{h=1}^{k} \sum_{j=1}^{m} g(x_h, y_j) f_{X,Y}(x_h, y_j),$$

where $f_{X,Y}$ is the joint pdf of (X, Y). The definition is more complicated for continuous random variables since it involves integration; we do not need it here. The extension to more than two random variables is straightforward.

B-3b Properties of Expected Values

In econometrics, we are not so concerned with computing expected values from various distributions; the major calculations have been done many times, and we will largely take these on faith. We will need to manipulate some expected values using a few simple rules. These are so important that we give them labels:

Property E.1: For any constant c, $\mathrm{E}(c) = c$.

Property E.2: For any constants a and b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

One useful implication of E.2 is that, if $\mu = \mathrm{E}(X)$, and we define a new random variable as $Y = X - \mu$, then $\mathrm{E}(Y) = 0$; in E.2, take $a = 1$ and $b = -\mu$.

As an example of Property E.2, let X be the temperature measured in Celsius at noon on a particular day at a given location; suppose the expected temperature is $\mathrm{E}(X) = 25$. If Y is the temperature measured in Fahrenheit, then $Y = 32 + (9/5)X$. From Property E.2, the expected temperature in Fahrenheit is $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

Generally, it is easy to compute the expected value of a linear function of many random variables.

Property E.3: If $\{a_1, a_2, \ldots, a_n\}$ are constants and $\{X_1, X_2, \ldots, X_n\}$ are random variables, then

$$\mathrm{E}(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mathrm{E}(X_1) + a_2\mathrm{E}(X_2) + \ldots + a_n\mathrm{E}(X_n).$$

Or, using summation notation,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i\mathrm{E}(X_i). \qquad [B.21]$$

As a special case of this, we have (with each $a_i = 1$)

$$\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i), \qquad [B.22]$$

so that the expected value of the sum is the sum of expected values. This property is used often for derivations in mathematical statistics.

EXAMPLE B.5 Finding Expected Revenue

Let X_1, X_2, and X_3 be the numbers of small, medium, and large pizzas, respectively, sold during the day at a pizza parlor. These are random variables with expected values $\mathrm{E}(X_1) = 25$, $\mathrm{E}(X_2) = 57$, and $\mathrm{E}(X_3) = 40$. The prices of small, medium, and large pizzas are \$5.50, \$7.60, and \$9.15. Therefore, the expected revenue from pizza sales on a given day is

$$\begin{aligned}\mathrm{E}(5.50\,X_1 + 7.60\,X_2 + 9.15\,X_3) &= 5.50\,\mathrm{E}(X_1) + 7.60\,\mathrm{E}(X_2) + 9.15\,\mathrm{E}(X_3)\\ &= 5.50(25) + 7.60(57) + 9.15(40) = 936.70,\end{aligned}$$

that is, \$936.70. The actual revenue on any particular day will generally differ from this value, but this is the *expected* revenue.

We can also use Property E.3 to show that if $X \sim \text{Binomial}(n, \theta)$, then $\mathrm{E}(X) = n\theta$. That is, the expected number of successes in n Bernoulli trials is simply the number of trials times the probability of success on any particular trial. This is easily seen by writing X as $X = Y_1 + Y_2 + \ldots + Y_n$, where each $Y_i \sim \text{Bernoulli}(\theta)$. Then,

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

We can apply this to the airline reservation example, where the airline makes $n = 120$ reservations, and the probability of showing up is $\theta = .85$. The *expected* number of people showing up is $120(.85) = 102$. Therefore, if there are 100 seats available, the expected number of people showing up is too large; this has some bearing on whether it is a good idea for the airline to make 120 reservations.

Actually, what the airline should do is define a profit function that accounts for the net revenue earned per seat sold and the cost per passenger bumped from the flight. This profit function is random because the actual number of people showing up is random. Let r be the net revenue from each passenger. (You can think of this as the price of the ticket for simplicity.) Let c be the compensation owed to any passenger bumped from the flight. Neither r nor c is random; these are assumed to be known to the airline. Let Y denote profits for the flight. Then, with 100 seats available,

$$\begin{aligned}Y &= rX \text{ if } X \leq 100\\ &= 100r - c(X - 100) \text{ if } X > 100.\end{aligned}$$

The first equation gives profit if no more than 100 people show up for the flight; the second equation is profit if more than 100 people show up. (In the latter case, the net revenue from ticket sales is $100r$, since all 100 seats are sold, and then $c(X - 100)$ is the cost of making more than 100 reservations.) Using the fact that X has a Binomial(n,.85) distribution, where n is the number of reservations made, expected profits, E(Y), can be found as a function of n (and r and c). Computing E(Y) directly would be quite difficult, but it can be found quickly using a computer. Once values for r and c are given, the value of n that maximizes expected profits can be found by searching over different values of n.

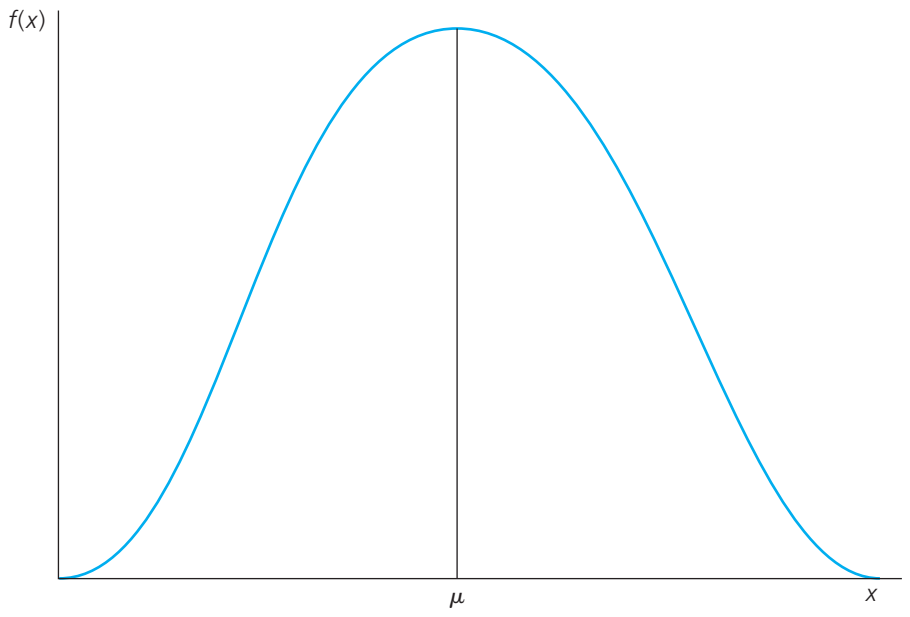
B-3c Another Measure of Central Tendency: The Median

The expected value is only one possibility for defining the central tendency of a random variable. Another measure of central tendency is the **median**. A general definition of *median* is too complicated for our purposes. If X is continuous, then the median of X, say, m, is the value such that one-half of the area under the pdf is to the left of m, and one-half of the area is to the right of m.

When X is discrete and takes on a finite number of odd values, the median is obtained by ordering the possible values of X and then selecting the value in the middle. For example, if X can take on the values $\{-4, 0, 2, 8, 10, 13, 17\}$, then the median value of X is 8. If X takes on an even number of values, there are really two median values; sometimes, these are averaged to get a unique median value. Thus, if X takes on the values $\{-5, 3, 9, 17\}$, then the median values are 3 and 9; if we average these, we get a median equal to 6.

In general, the median, sometimes denoted Med(X), and the expected value, E(X), are different. Neither is "better" than the other as a measure of central tendency; they are both valid ways to measure the center of the distribution of X. In one special case, the median and expected value (or mean) are the same. If X has a **symmetric distribution** about the value μ, then μ is both the expected value and the median. Mathematically, the condition is $f(\mu + x) = f(\mu - x)$ for all x. This case is illustrated in Figure B.3.

FIGURE B.3 A symmetric probability distribution.



B-3d Measures of Variability: Variance and Standard Deviation

Although the central tendency of a random variable is valuable, it does not tell us everything we want to know about the distribution of a random variable. Figure B.4 shows the pdfs of two random variables with the same mean. Clearly, the distribution of X is more tightly centered about its mean than is the distribution of Y. We would like to have a simple way of summarizing differences in the spreads of distributions.

B-3e Variance

For a random variable X, let $\mu = \mathrm{E}(X)$. There are various ways to measure how far X is from its expected value, but the simplest one to work with algebraically is the squared difference, $(X - \mu)^2$. (The squaring eliminates the sign from the distance measure; the resulting positive value corresponds to our intuitive notion of distance and treats values above and below μ symmetrically.) This distance is itself a random variable since it can change with every outcome of X. Just as we needed a number to summarize the central tendency of X, we need a number that tells us how far X is from μ, *on average*. One such number is the **variance**, which tells us the expected distance from X to its mean:

$$\mathrm{Var}(X) \equiv \mathrm{E}[(X - \mu)^2]. \qquad \text{[B.23]}$$

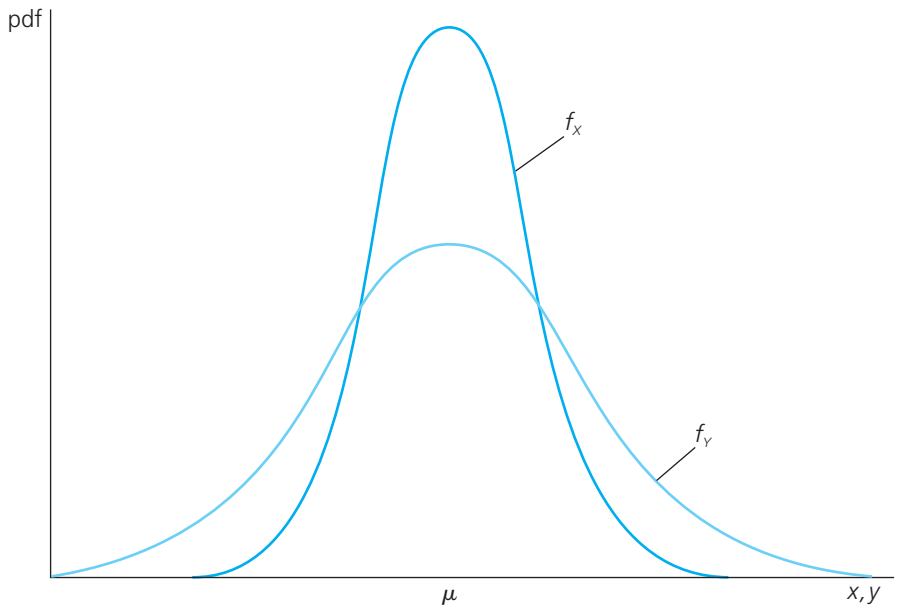
Variance is sometimes denoted σ_X^2, or simply σ^2, when the context is clear. From (B.23), it follows that the variance is always nonnegative.

As a computational device, it is useful to observe that

$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \qquad \text{[B.24]}$$

In using either (B.23) or (B.24), we need not distinguish between discrete and continuous random variables: the definition of variance is the same in either case. Most often, we first compute $\mathrm{E}(X)$, then $\mathrm{E}(X^2)$, and then we use the formula in (B.24). For example, if $X \sim \mathrm{Bernoulli}(\theta)$, then $\mathrm{E}(X) = \theta$, and, since $X^2 = X$, $\mathrm{E}(X^2) = \theta$. It follows from equation (B.24) that $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

FIGURE B.4 Random variables with the same mean but different distributions.



Two important properties of the variance follow.

Property VAR.1: $\text{Var}(X) = 0$ if, and only if, there is a constant c such that $P(X = c) = 1$, in which case $\text{E}(X) = c$.

This first property says that the variance of any constant is zero and if a random variable has zero variance, then it is essentially constant.

Property VAR.2: For any constants a and b, $\text{Var}(aX + b) = a^2\text{Var}(X)$.

This means that adding a constant to a random variable does not change the variance, but multiplying a random variable by a constant increases the variance by a factor equal to the *square* of that constant. For example, if X denotes temperature in Celsius and $Y = 32 + (9/5)X$ is temperature in Fahrenheit, then $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

B-3f Standard Deviation

The **standard deviation** of a random variable, denoted $\text{sd}(X)$, is simply the positive square root of the variance: $\text{sd}(X) \equiv +\sqrt{\text{Var}(X)}$. The standard deviation is sometimes denoted σ_X, or simply σ, when the random variable is understood. Two standard deviation properties immediately follow from Properties VAR.1 and VAR.2.

Property SD.1: For any constant c, $\text{sd}(c) = 0$.

Property SD.2: For any constants a and b,

$$\text{sd}(aX + b) = |a|\text{sd}(X).$$

In particular, if $a > 0$, then $\text{sd}(aX) = a \cdot \text{sd}(X)$.

This last property makes the standard deviation more natural to work with than the variance. For example, suppose that X is a random variable measured in thousands of dollars, say, income. If we define $Y = 1{,}000X$, then Y is income measured in dollars. Suppose that $\text{E}(X) = 20$, and $\text{sd}(X) = 6$. Then, $\text{E}(Y) = 1{,}000\text{E}(X) = 20{,}000$, and $\text{sd}(Y) = 1{,}000 \cdot \text{sd}(X) = 6{,}000$, so that the expected value and standard deviation both increase by the same factor, 1,000. If we worked with variance, we would have $\text{Var}(Y) = (1{,}000)^2\text{Var}(X)$, so that the variance of Y is one million times larger than the variance of X.

B-3g Standardizing a Random Variable

As an application of the properties of variance and standard deviation—and a topic of practical interest in its own right—suppose that given a random variable X, we define a new random variable by subtracting off its mean m and dividing by its standard deviation σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \text{[B.25]}$$

which we can write as $Z = aX + b$, where $a \equiv (1/\sigma)$ and $b \equiv -(\mu/\sigma)$. Then, from Property E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

From Property VAR.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Thus, the random variable Z has a mean of zero and a variance (and therefore a standard deviation) equal to one. This procedure is sometimes known as *standardizing* the random variable X, and Z is called a **standardized random variable**. (In introductory statistics courses, it is sometimes called the *z-transform* of X.) It is important to remember that the standard deviation, not the variance, appears in the denominator of (B.25). As we will see, this transformation is frequently used in statistical inference.

As a specific example, suppose that $E(X) = 2$, and $Var(X) = 9$. Then, $Z = (X - 2)/3$ has expected value zero and variance one.

B-3h Skewness and Kurtosis

We can use the standardized version of a random variable to define other features of the distribution of a random variable. These features are described by using what are called *higher order moments*. For example, the third moment of the random variable Z in (B.25) is used to determine whether a distribution is symmetric about its mean. We can write

$$E(Z^3) = E[(X - \mu)^3]/\sigma^3.$$

If X has a symmetric distribution about μ, then Z has a symmetric distribution about zero. (The division by σ^3 does not change whether the distribution is symmetric.) That means the density of Z at any two points z and $-z$ is the same, which means that, in computing $E(Z^3)$, positive values z^3 when $z > 0$ are exactly offset with the negative value $(-z)^3 = -z^3$. It follows that, if X is symmetric about zero, then $E(Z) = 0$. Generally, $E[(X - \mu)^3]/\sigma^3$ is viewed as a measure of **skewness** in the distribution of X. In a statistical setting, we might use data to estimate $E(Z^3)$ to determine whether an underlying population distribution appears to be symmetric. (Computer Exercise C5.4 in Chapter 5 provides an illustration.)

It also can be informative to compute the fourth moment of Z,

$$E(Z^4) = E[(X - \mu)^4]/\sigma^4.$$

Because $Z^4 \geq 0$, $E(Z^4) \geq 0$ (and, in any interesting case, strictly greater than zero). Without having a reference value, it is difficult to interpret values of $E(Z^4)$, but larger values mean that the tails in the distribution of X are thicker. The fourth moment $E(Z^4)$ is called a measure of **kurtosis** in the distribution of X. In Section B-5, we will obtain $E(Z^4)$ for the normal distribution.

B-4 Features of Joint and Conditional Distributions

B-4a Measures of Association: Covariance and Correlation

While the joint pdf of two random variables completely describes the relationship between them, it is useful to have summary measures of how, on average, two random variables vary with one another. As with the expected value and variance, this is similar to using a single number to summarize something about an entire distribution, which in this case is a joint distribution of two random variables.

B-4b Covariance

Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$ and consider the random variable $(X - \mu_X)(Y - \mu_Y)$. Now, if X is above its mean and Y is above its mean, then $(X - \mu_X)(Y - \mu_Y) > 0$. This is also true if $X < \mu_X$ and $Y < \mu_Y$. On the other hand, if $X > \mu_X$ and $Y < \mu_Y$, or vice versa, then $(X - \mu_X)(Y - \mu_Y) < 0$. How, then, can this product tell us anything about the relationship between X and Y?

The **covariance** between two random variables X and Y, sometimes called the *population covariance* to emphasize that it concerns the relationship between two variables describing a population, is defined as the expected value of the product $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X, Y) \equiv \text{E}[(X - \mu_X)(Y - \mu_Y)], \qquad \text{[B.26]}$$

which is sometimes denoted σ_{XY}. If $\sigma_{XY} > 0$, then, on average, when X is above its mean, Y is also above its mean. If $\sigma_{XY} < 0$, then, on average, when X is above its mean, Y is below its mean.

Several expressions useful for computing $\text{Cov}(X, Y)$ are as follows:

$$\begin{aligned}\text{Cov}(X, Y) &= \text{E}[(X - \mu_X)(Y - \mu_Y)] = \text{E}[(X - \mu_X)Y] \\ &= \text{E}[X(Y - \mu_Y)] = \text{E}(XY) - \mu_X\mu_Y. \qquad \text{[B.27]}\end{aligned}$$

It follows from (B.27), that if $\text{E}(X) = 0$ or $\text{E}(Y) = 0$, then $\text{Cov}(X, Y) = \text{E}(XY)$.

Covariance measures the amount of *linear* dependence between two random variables. A positive covariance indicates that two random variables move in the same direction, while a negative covariance indicates they move in opposite directions. Interpreting the *magnitude* of a covariance can be a little tricky, as we will see shortly.

Because covariance is a measure of how two random variables are related, it is natural to ask how covariance is related to the notion of independence. This is given by the following property.

Property COV.1: If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

This property follows from equation (B.27) and the fact that $\text{E}(XY) = \text{E}(X)\text{E}(Y)$ when X and Y are independent. It is important to remember that the converse of COV.1 is *not* true: zero covariance between X and Y does not imply that X and Y are independent. In fact, there are random variables X such that, if $Y = X^2$, $\text{Cov}(X, Y) = 0$. [Any random variable with $\text{E}(X) = 0$ and $\text{E}(X^3) = 0$ has this property.] If $Y = X^2$, then X and Y are clearly not independent: once we know X, we know Y. It seems rather strange that X and X^2 could have zero covariance, and this reveals a weakness of covariance as a general measure of association between random variables. The covariance is useful in contexts when relationships are at least approximately linear.

The second major property of covariance involves covariances between linear functions.

Property COV.2: For any constants a_1, b_1, a_2, and b_2,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X, Y). \qquad \text{[B.28]}$$

An important implication of COV.2 is that the covariance between two random variables can be altered simply by multiplying one or both of the random variables by a constant. This is important in economics because monetary variables, inflation rates, and so on can be defined with different units of measurement without changing their meaning.

Finally, it is useful to know that the absolute value of the covariance between any two random variables is bounded by the product of their standard deviations; this is known as the *Cauchy-Schwartz inequality*.

Property COV.3: $|\text{Cov}(X, Y)| \leq \text{sd}(X)\text{sd}(Y)$.

B-4c Correlation Coefficient

Suppose we want to know the relationship between amount of education and annual earnings in the working population. We could let X denote education and Y denote earnings and then compute their covariance. But the answer we get will depend on how we choose to measure education and earnings.

Property COV.2 implies that the covariance between education and earnings depends on whether earnings are measured in dollars or thousands of dollars, or whether education is measured in months or years. It is pretty clear that how we measure these variables has no bearing on how strongly they are related. But the covariance between them does depend on the units of measurement.

The fact that the covariance depends on units of measurement is a deficiency that is overcome by the **correlation coefficient** between X and Y:

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{sd}(X)\cdot\text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \qquad \text{[B.29]}$$

the correlation coefficient between X and Y is sometimes denoted ρ_{XY} (and is sometimes called the *population correlation*).

Because σ_X and σ_Y are positive, $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ always have the same sign, and $\text{Corr}(X, Y) = 0$ if, and only if, $\text{Cov}(X, Y) = 0$. Some of the properties of covariance carry over to correlation. If X and Y are independent, then $\text{Corr}(X, Y) = 0$, but zero correlation does not imply independence. (Like the covariance, the correlation coefficient is also a measure of linear dependence.) However, the magnitude of the correlation coefficient is easier to interpret than the size of the covariance due to the following property.

Property CORR.1: $-1 \leq \text{Corr}(X, Y) \leq 1$.

If $\text{Corr}(X, Y) = 0$, or equivalently $\text{Cov}(X, Y) = 0$, then there is no linear relationship between X and Y, and X and Y are said to be **uncorrelated random variables**; otherwise, X and Y are *correlated*. $\text{Corr}(X, Y) = 1$ implies a perfect positive linear relationship, which means that we can write $Y = a + bX$ for some constant a and some constant $b > 0$. $\text{Corr}(X, Y) = -1$ implies a perfect negative linear relationship, so that $Y = a + bX$ for some $b < 0$. The extreme cases of positive or negative 1 rarely occur. Values of ρ_{XY} closer to 1 or -1 indicate stronger linear relationships.

As mentioned earlier, the correlation between X and Y is invariant to the units of measurement of either X or Y. This is stated more generally as follows.

Property CORR.2: For constants a_1, b_1, a_2, and b_2, with $a_1a_2 > 0$,

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = \text{Corr}(X, Y).$$

If $a_1a_2 < 0$, then

$$\text{Corr}(a_1X + b_1, a_2Y + b_2) = -\text{Corr}(X, Y).$$

As an example, suppose that the correlation between earnings and education in the working population is .15. This measure does not depend on whether earnings are measured in dollars, thousands of dollars, or any other unit; it also does not depend on whether education is measured in years, quarters, months, and so on.

B-4d Variance of Sums of Random Variables

Now that we have defined covariance and correlation, we can complete our list of major properties of the variance.

Property VAR.3: For constants a and b,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y).$$

It follows immediately that, if X and Y are uncorrelated—so that $\text{Cov}(X, Y) = 0$—then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \quad \text{[B.30]}$$

and

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \quad \text{[B.31]}$$

In the latter case, note how the variance of the difference is the *sum of the variances*, not the difference in the variances.

As an example of (B.30), let X denote profits earned by a restaurant during a Friday night and let Y be profits earned on the following Saturday night. Then, $Z = X + Y$ is profits for the two nights. Suppose X and Y each have an expected value of \$300 and a standard deviation of \$15 (so that the variance is 225). Expected profits for the two nights is $\text{E}(Z) = \text{E}(X) + \text{E}(Y) = 2\cdot(300) = 600$ dollars. If X and Y are independent, and therefore uncorrelated, then the variance of total profits is the sum of the variances: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2\cdot(225) = 450$. It follows that the standard deviation of total profits is $\sqrt{450}$ or about \$21.21.

Expressions (B.30) and (B.31) extend to more than two random variables. To state this extension, we need a definition. The random variables $\{X_1, \ldots, X_n\}$ are **pairwise uncorrelated random variables** if each variable in the set is uncorrelated with every other variable in the set. That is, $\text{Cov}(X_i, X_j) = 0$, for all $i \neq j$.

Property VAR.4: If $\{X_1, \ldots, X_n\}$ are pairwise uncorrelated random variables and a_i: $i = 1, \ldots, n$ are constants, then

$$\text{Var}(a_1X_1 + \ldots + a_nX_n) = a_1^2\text{Var}(X_1) + \ldots + a_n^2\text{Var}(X_n).$$

In summation notation, we can write

$$\text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(X_i). \quad \text{[B.32]}$$

A special case of Property VAR.4 occurs when we take $a_i = 1$ for all i. Then, for pairwise uncorrelated random variables, the variance of the sum is the sum of the variances:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \quad \text{[B.33]}$$

Because independent random variables are uncorrelated (see Property COV.1), the variance of a sum of independent random variables is the sum of the variances.

If the X_i are not pairwise uncorrelated, then the expression for $\text{Var}(\sum_{i=1}^{n} a_iX_i)$ is much more complicated; we must add to the right-hand side of (B.32) the terms $2a_ia_j\text{Cov}(x_i, x_j)$ for all $i > j$.

We can use (B.33) to derive the variance for a binomial random variable. Let $X \sim \text{Binomial}(n, \theta)$ and write $X = Y_1 + \ldots + Y_n$, where the Y_i are independent Bernoulli (θ) random variables. Then, by (B.33), $\text{Var}(X) = \text{Var}(Y_1) + \ldots + \text{Var}(Y_n) = n\theta(1 - \theta)$.

In the airline reservation example with $n = 120$ and $\theta = .85$, the variance of the number of passengers arriving for their reservations is $120(.85)(.15) = 15.3$, so the standard deviation is about 3.9.

B-4e Conditional Expectation

Covariance and correlation measure the linear relationship between two random variables and treat them symmetrically. More often in the social sciences, we would like to explain one variable, called Y, in terms of another variable, say, X. Further, if Y is related to X in a nonlinear fashion, we would like

to know this. Call Y the explained variable and X the explanatory variable. For example, Y might be hourly wage, and X might be years of formal education.

We have already introduced the notion of the conditional probability density function of Y given X. Thus, we might want to see how the distribution of wages changes with education level. However, we usually want to have a simple way of summarizing this distribution. A single number will no longer suffice, since the distribution of Y given $X = x$ generally depends on the value of x. Nevertheless, we can summarize the relationship between Y and X by looking at the **conditional expectation** of Y given X, sometimes called the *conditional mean.* The idea is this. Suppose we know that X has taken on a particular value, say, x. Then, we can compute the expected value of Y, given that we know this outcome of X. We denote this expected value by $\mathrm{E}(Y|X = x)$, or sometimes $\mathrm{E}(Y|x)$ for shorthand. Generally, as x changes, so does $\mathrm{E}(Y|x)$.

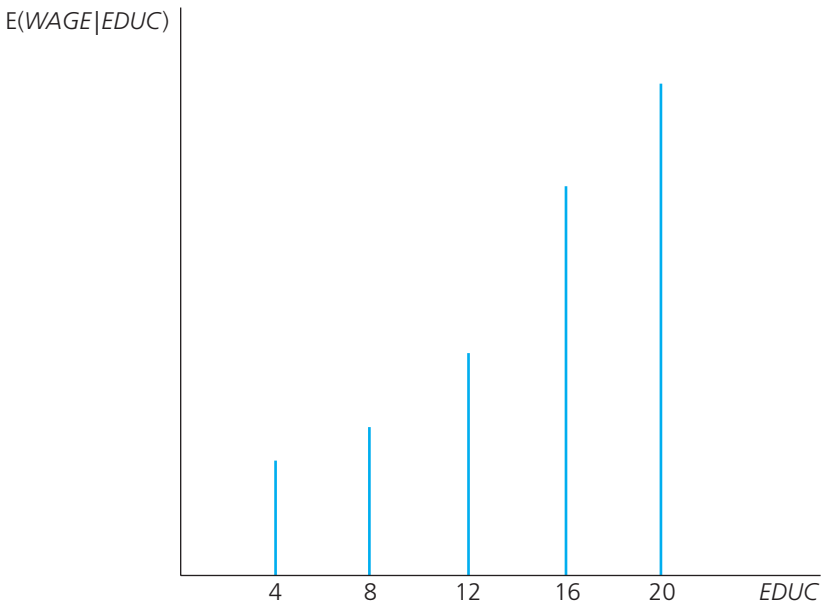
When Y is a discrete random variable taking on values $\{y_1, \ldots, y_m\}$, then

$$\mathrm{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

When Y is continuous, $\mathrm{E}(Y|x)$ is defined by integrating $y f_{Y|X}(y|x)$ over all possible values of y. As with unconditional expectations, the conditional expectation is a weighted average of possible values of Y, but now the weights reflect the fact that X has taken on a specific value. Thus, $E(Y|x)$ is just some function of x, which tells us how the expected value of Y varies with x.

As an example, let (X, Y) represent the population of all working individuals, where X is years of education and Y is hourly wage. Then, $\mathrm{E}(Y|X = 12)$ is the average hourly wage for all people in the population with 12 years of education (roughly a high school education). $\mathrm{E}(Y|X = 16)$ is the average hourly wage for all people with 16 years of education. Tracing out the expected value for various levels of education provides important information on how wages and education are related. See Figure B.5 for an illustration.

FIGURE B.5 The expected value of hourly wage given various levels of education.



In principle, the expected value of hourly wage can be found at each level of education, and these expectations can be summarized in a table. Because education can vary widely—and can even be measured in fractions of a year—this is a cumbersome way to show the relationship between average wage and amount of education. In econometrics, we typically specify simple functions that capture this relationship. As an example, suppose that the expected value of *WAGE* given *EDUC* is the linear function

$$\mathrm{E}(WAGE|EDUC) = 1.05 + .45\,EDUC.$$

If this relationship holds in the population of working people, the average wage for people with eight years of education is $1.05 + .45(8) = 4.65$, or \$4.65. The average wage for people with 16 years of education is 8.25, or \$8.25. The coefficient on *EDUC* implies that each year of education increases the expected hourly wage by .45, or 45¢.

Conditional expectations can also be nonlinear functions. For example, suppose that $\mathrm{E}(Y|x) = 10/x$, where X is a random variable that is always greater than zero. This function is graphed in Figure B.6. This could represent a demand function, where Y is quantity demanded and X is price. If Y and X are related in this way, an analysis of linear association, such as correlation analysis, would be incomplete.

B-4f Properties of Conditional Expectation

Several basic properties of conditional expectations are useful for derivations in econometric analysis.

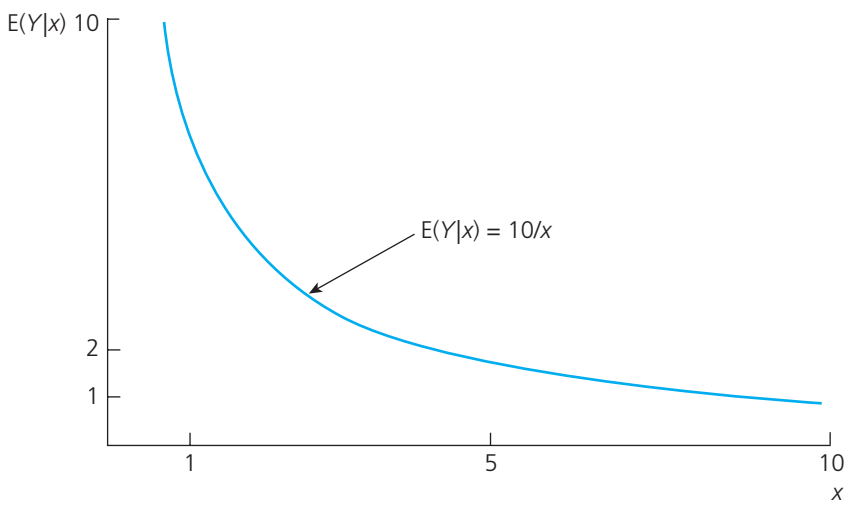
Property CE.1: $\mathrm{E}[c(X)|X] = c(X)$, for any function $c(X)$.

This first property means that functions of X behave as constants when we compute expectations conditional on X. For example, $\mathrm{E}(X^2|X) = X^2$. Intuitively, this simply means that if we know X, then we also know X^2.

Property CE.2: For functions $a(X)$ and $b(X)$,

$$\mathrm{E}[a(X)Y + b(X)|X] = a(X)\mathrm{E}(Y|X) + b(X).$$

FIGURE B.6 Graph of $\mathrm{E}(Y|X) = 10/x$.



For example, we can easily compute the conditional expectation of a function such as $XY + 2X^2$: $\mathrm{E}(XY + 2X^2|X) = X\mathrm{E}(Y|X) + 2X^2$.

The next property ties together the notions of independence and conditional expectations.

Property CE.3: If X and Y are independent, then $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

This property means that, if X and Y are independent, then the expected value of Y given X does not depend on X, in which case, $\mathrm{E}(Y|X)$ always equals the (unconditional) expected-value of Y. In the wage and education example, if wages were independent of education, then the average wages of high school and college graduates would be the same. Since this is almost certainly false, we cannot assume that wage and education are independent.

A special case of Property CE.3 is the following: if U and X are independent and $\mathrm{E}(U) = 0$, then $\mathrm{E}(U|X) = 0$.

There are also properties of the conditional expectation that have to do with the fact that $\mathrm{E}(Y|X)$ is a function of X, say, $\mathrm{E}(Y|X) = \mu(X)$. Because X is a random variable, $\mu(X)$ is also a random variable. Furthermore, $\mu(X)$ has a probability distribution and therefore an expected value. Generally, the expected value of $\mu(X)$ could be very difficult to compute directly. The **law of iterated expectations** says that the expected value of $\mu(X)$ is simply equal to the expected value of Y. We write this as follows.

Property CE.4: $\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$.

This property is a little hard to grasp at first. It means that, if we first obtain $\mathrm{E}(Y|X)$ as a function of X and take the expected value of this (with respect to the distribution of X, of course), then we end up with $\mathrm{E}(Y)$. This is hardly obvious, but it can be derived using the definition of expected values.

As an example of how to use Property CE.4, let $Y = WAGE$ and $X = EDUC$, where $WAGE$ is measured in hours and $EDUC$ is measured in years. Suppose the expected value of $WAGE$ given $EDUC$ is $\mathrm{E}(WAGE|EDUC) = 4 + .60\,EDUC$. Further, $\mathrm{E}(EDUC) = 11.5$. Then, the law of iterated expectations implies that $\mathrm{E}(WAGE) = \mathrm{E}(4 + .60\,EDUC) = 4 + .60\,\mathrm{E}(EDUC) = 4 + .60(11.5) = 10.90$, or \$10.90 an hour.

The next property states a more general version of the law of iterated expectations.

Property CE.4': $\mathrm{E}(Y|X) = \mathrm{E}[\mathrm{E}(Y|X, Z)|X]$.

In other words, we can find $\mathrm{E}(Y|X)$ in two steps. First, find $\mathrm{E}(Y|X, Z)$ for any other random variable Z. Then, find the expected value of $\mathrm{E}(Y|X, Z)$, conditional on X.

Property CE.5: If $\mathrm{E}(Y|X) = \mathrm{E}(Y)$, then $\mathrm{Cov}(X, Y) = 0$ [and so $\mathrm{Corr}(X, Y) = 0$]. In fact, *every* function of X is uncorrelated with Y.

This property means that, if knowledge of X does not change the expected value of Y, then X and Y *must* be uncorrelated, which implies that if X and Y are correlated, then $\mathrm{E}(Y|X)$ must depend on X. The converse of Property CE.5 is not true: if X and Y are uncorrelated, $\mathrm{E}(Y|X)$ *could* still depend on X. For example, suppose $Y = X^2$. Then, $\mathrm{E}(Y|X) = X^2$, which is clearly a function of X. However, as we mentioned in our discussion of covariance and correlation, it is possible that X and X^2 are uncorrelated. The conditional expectation captures the nonlinear relationship between X and Y that correlation analysis would miss entirely.

Properties CE.4 and CE.5 have two important implications: if U and X are random variables such that $\mathrm{E}(U|X) = 0$, then $\mathrm{E}(U) = 0$, and U and X are uncorrelated.

Property CE.6: If $\mathrm{E}(Y^2) < \infty$ and $\mathrm{E}[g(X)^2] < \infty$ for some function g, then $\mathrm{E}\{[Y - \mu(X)]^2|X\} \leq \mathrm{E}\{[Y - g(X)]^2|X\}$ and $\mathrm{E}\{[Y - \mu(X)]^2\} \leq \mathrm{E}\{[Y - g(X)]^2\}$.

Property CE.6 is very useful in predicting or forecasting contexts. The first inequality says that, if we measure prediction inaccuracy as the *expected* squared prediction error, conditional on X, then the conditional mean is better than any other function of X for predicting Y. The conditional mean also minimizes the unconditional expected squared prediction error.

B-4g Conditional Variance

Given random variables X and Y, the variance of Y, conditional on $X = x$, is simply the variance associated with the conditional distribution of Y, given $X = x$: $\mathrm{E}\{[Y - \mathrm{E}(Y|x)]^2|x\}$. The formula

$$\mathrm{Var}(Y|X = x) = \mathrm{E}(Y^2|x) - [\mathrm{E}(Y|x)]^2$$

is often useful for calculations. Only occasionally will we have to compute a conditional variance. But we will have to make assumptions about and manipulate conditional variances for certain topics in regression analysis.

As an example, let $Y = SAVING$ and $X = INCOME$ (both of these measured annually for the population of all families). Suppose that $\mathrm{Var}(SAVING|INCOME) = 400 + .25\,INCOME$. This says that, as income increases, the variance in saving levels also increases. It is important to see that the relationship between the variance of *SAVING* and *INCOME* is totally separate from that between the *expected value* of *SAVING* and *INCOME*.

We state one useful property about the conditional variance.

Property CV.1: If X and Y are independent, then $\mathrm{Var}(Y|X) = \mathrm{Var}(Y)$.

This property is pretty clear, since the distribution of Y given X does not depend on X, and $\mathrm{Var}(Y|X)$ is just one feature of this distribution.

B-5 The Normal and Related Distributions

B-5a The Normal Distribution

The normal distribution and those derived from it are the most widely used distributions in statistics and econometrics. Assuming that random variables defined over populations are normally distributed simplifies probability calculations. In addition, we will rely heavily on the normal and related distributions to conduct inference in statistics and econometrics—even when the underlying population is not necessarily normal. We must postpone the details, but be assured that these distributions will arise many times throughout this text.

A normal random variable is a continuous random variable that can take on any value. Its probability density function has the familiar bell shape graphed in Figure B.7.

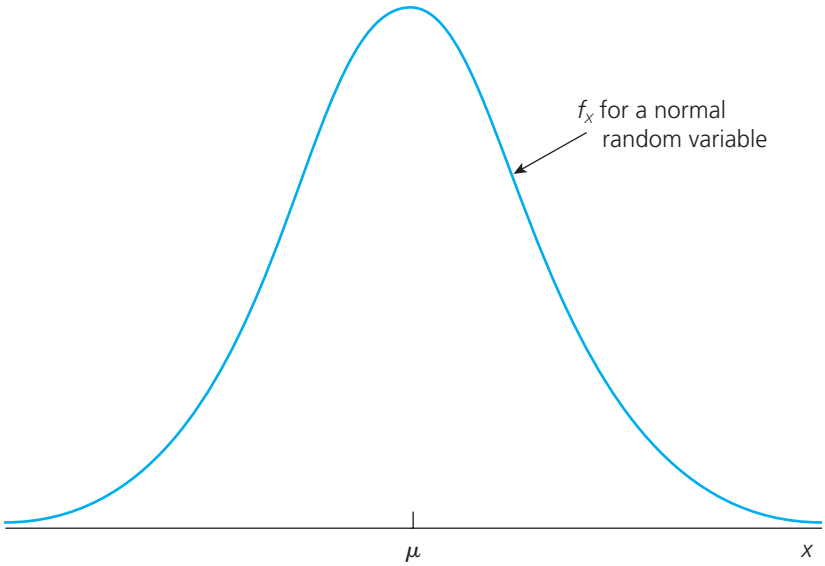
Mathematically, the pdf of X can be written as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-(x - \mu)^2/2\sigma^2], \quad -\infty < x < \infty, \qquad \text{[B.34]}$$

where $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. We say that X has a **normal distribution** with expected value μ and variance σ^2, written as $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Because the normal distribution is symmetric about μ, μ is also the median of X. The normal distribution is sometimes called the *Gaussian distribution* after the famous mathematician C. F. Gauss.

Certain random variables appear to roughly follow a normal distribution. Human heights and weights, test scores, and county unemployment rates have pdfs roughly the shape in Figure B.7. Other distributions, such as income distributions, do not appear to follow the normal probability function. In most countries, income is not symmetrically distributed about any value; the distribution is skewed toward the upper tail. In some cases, a variable can be transformed to achieve normality. A popular transformation is

FIGURE B.7 The general shape of the normal probability density function.



the natural log, which makes sense for positive random variables. If X is a positive random variable, such as income, and $Y = \log(X)$ has a normal distribution, then we say that X has a *lognormal distribution*. It turns out that the lognormal distribution fits income distribution pretty well in many countries. Other variables, such as prices of goods, appear to be well described as lognormally distributed.

B-5b The Standard Normal Distribution

One special case of the normal distribution occurs when the mean is zero and the variance (and, therefore, the standard deviation) is unity. If a random variable Z has a Normal(0,1) distribution, then we say it has a **standard normal distribution**. The pdf of a standard normal random variable is denoted $\phi(z)$; from (B.34), with $\mu = 0$ and $\sigma^2 = 1$, it is given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2), \quad -\infty < z < \infty. \qquad \text{[B.35]}$$

The standard normal cumulative distribution function is denoted $\Phi(z)$ and is obtained as the area under ϕ, to the left of z; see Figure B.8. Recall that $\Phi(z) = \mathrm{P}(Z \leq z)$; because Z is continuous, $\Phi(z) = \mathrm{P}(Z < z)$ as well.

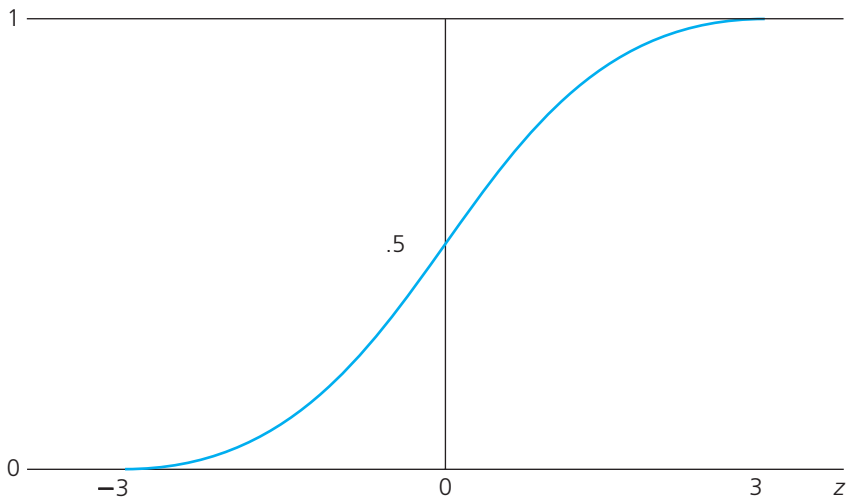
No simple formula can be used to obtain the values of $\Phi(z)$ [because $\Phi(z)$ is the integral of the function in (B.35), and this integral has no closed form]. Nevertheless, the values for $\Phi(z)$ are easily tabulated; they are given for z between −3.1 and 3.1 in Table G.1 in Appendix G. For $z \leq -3.1$, $\Phi(z)$ is less than .001, and for $z \geq 3.1$, $\Phi(z)$ is greater than .999. Most statistics and econometrics software packages include simple commands for computing values of the standard normal cdf, so we can often avoid printed tables entirely and obtain the probabilities for any value of z.

Using basic facts from probability—and, in particular, properties (B.7) and (B.8) concerning cdfs—we can use the standard normal cdf for computing the probability of any event involving a standard normal random variable. The most important formulas are

$$\mathrm{P}(Z > z) = 1 - \Phi(z), \qquad \text{[B.36]}$$

$$\mathrm{P}(Z < -z) = \mathrm{P}(Z > z), \qquad \text{[B.37]}$$

FIGURE B.8 The standard normal cumulative distribution function.



and

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \qquad \text{[B.38]}$$

Because Z is a continuous random variable, all three formulas hold whether or not the inequalities are strict. Some examples include $P(Z > .44) = 1 - .67 = .33$, $P(Z < -.92) = P(Z > .92) = 1 - .821 = .179$, and $P(-1 < Z \leq .5) = .692 - .159 = .533$.

Another useful expression is that, for any $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \qquad \text{[B.39]} \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned}$$

Thus, the probability that the absolute value of Z is bigger than some positive constant c is simply twice the probability $P(Z > c)$; this reflects the symmetry of the standard normal distribution.

In most applications, we start with a normally distributed random variable, $X \sim \text{Normal}(\mu, \sigma^2)$, where μ is different from zero and $\sigma^2 \neq 1$. Any normal random variable can be turned into a standard normal using the following property.

Property Normal.1: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim \text{Normal}(0, 1)$.

Property Normal.1 shows how to turn any normal random variable into a standard normal. Thus, suppose $X \sim \text{Normal}(3, 4)$, and we would like to compute $P(X \leq 1)$. The steps always involve the normalization of X to a standard normal:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = .159. \end{aligned}$$

EXAMPLE B.6 Probabilities for a Normal Random Variable

First, let us compute $P(2 < X \leq 6)$ when $X \sim$ Normal(4,9) (whether we use $<$ or $\leq$ is irrelevant because X is a continuous random variable). Now,

$$P(2 < X \leq 6) = P\left(\frac{2-4}{3} < \frac{X-4}{3} \leq \frac{6-4}{3}\right) = P(-2/3 < Z \leq 2/3)$$
$$= \Phi(.67) - \Phi(-.67) = .749 - .251 = .498.$$

Now, let us compute $P(|X| > 2)$:

$$P(|X| > 2) = P(X > 2) + P(X < -2)$$
$$= P[(X-4)/3 > (2-4)/3] + P[(X-4)/3 < (-2-4)/3]$$
$$= 1 - \Phi(-2/3) + \Phi(-2)$$
$$= 1 - .251 + .023 = .772.$$

B-5c Additional Properties of the Normal Distribution

We end this subsection by collecting several other facts about normal distributions that we will later use.

Property Normal.2: If $X \sim \text{Normal}(\mu, \sigma^2)$, then $aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

Thus, if $X \sim$ Normal(1,9), then $Y = 2X + 3$ is distributed as normal with mean $2E(X) + 3 = 5$ and variance $2^2 \cdot 9 = 36$; $sd(Y) = 2sd(X) = 2 \cdot 3 = 6$.

Earlier, we discussed how, in general, zero correlation and independence are not the same. In the case of normally distributed random variables, it turns out that zero correlation suffices for independence.

Property Normal.3: If X and Y are jointly normally distributed, then they are independent if, and only if, $\text{Cov}(X, Y) = 0$.

Property Normal.4: Any linear combination of independent, identically distributed normal random variables has a normal distribution.

For example, let X_i, for $i = 1, 2$, and 3, be independent random variables distributed as Normal(μ, σ^2). Define $W = X_1 + 2X_2 - 3X_3$. Then, W is normally distributed; we must simply find its mean and variance. Now,

$$E(W) = E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0.$$

Also,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

Property Normal.4 also implies that the average of independent, normally distributed random variables has a normal distribution. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Normal(μ, σ^2), then

$$\overline{Y} \sim \text{Normal}(\mu, \sigma^2/n). \qquad \text{[B.40]}$$

This result is critical for statistical inference about the mean in a normal population.

Other features of the normal distribution are worth knowing, although they do not play a central role in the text. Because a normal random variable is symmetric about its mean, it has zero skewness, that is, $E[(X - \mu)^3] = 0$. Further, it can be shown that

$$E[(X - \mu)^4]/\sigma^4 = 3,$$

or $E(Z^4) = 3$, where Z has a standard normal distribution. Because the normal distribution is so prevalent in probability and statistics, the measure of kurtosis for any given random variable X (whose fourth moment exists) is often defined to be $E[(X - \mu)^4]/\sigma^4 - 3$, that is, relative to the value for the standard normal distribution. If $E[(X - \mu)^4]/\sigma^4 > 3$, then the distribution of X has fatter tails than the normal distribution (a somewhat common occurrence, such as with the t distribution to be introduced shortly); if $E[(X - \mu)^4]/\sigma^4 < 3$, then the distribution has thinner tails than the normal (a rarer situation).

B-5d The Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \qquad [B.41]$$

Then, X has what is known as a **chi-square distribution** with n **degrees of freedom** (or *df* for short). We write this as $X \sim \chi_n^2$. The *df* in a chi-square distribution corresponds to the number of terms in the sum in (B.41). The concept of degrees of freedom will play an important role in our statistical and econometric analyses.

The pdf for chi-square distributions with varying degrees of freedom is given in Figure B.9; we will not need the formula for this pdf, and so we do not reproduce it here. From equation (B.41), it is clear that a chi-square random variable is always nonnegative, and that, unlike the normal distribution, the chi-square distribution is not symmetric about any point. It can be shown that if $X \sim \chi_n^2$, then the expected value of X is n [the number of terms in (B.41)], and the variance of X is $2n$.

B-5e The *t* Distribution

The t distribution is the workhorse in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal and a chi-square random variable.

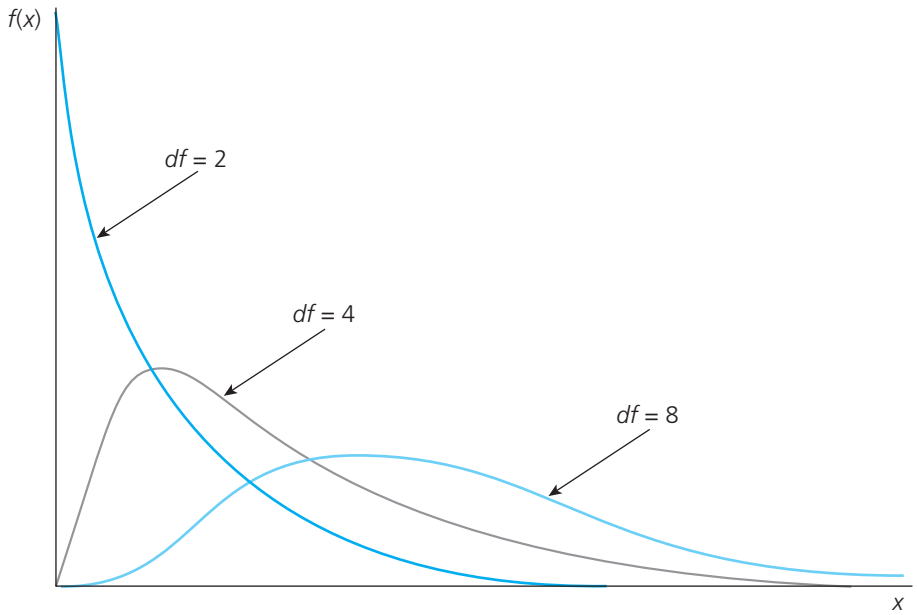
Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Further, assume that Z and X are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{X/n}} \qquad [B.42]$$

has a ***t* distribution** with n degrees of freedom. We will denote this by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi-square random variable in the denominator of (B.42).

The pdf of the t distribution has a shape similar to that of the standard normal distribution, except that it is more spread out and therefore has more area in the tails. The expected value of a t distributed random variable is zero (strictly speaking, the expected value exists only for $n > 1$),

FIGURE B.9 The chi-square distribution with various degrees of freedom.



and the variance is $n/(n-2)$ for $n > 2$. (The variance does not exist for $n \leq 2$ because the distribution is so spread out.) The pdf of the t distribution is plotted in Figure B.10 for various degrees of freedom. As the degrees of freedom gets large, the t distribution approaches the standard normal distribution.

B-5f The *F* Distribution

Another important distribution for statistics and econometrics is the F distribution. In particular, the F distribution will be used for testing hypotheses in the context of multiple regression analysis.

To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \qquad \text{[B.43]}$$

has an ***F* distribution** with (k_1, k_2) degrees of freedom. We denote this as $F \sim F_{k_1, k_2}$. The pdf of the F distribution with different degrees of freedom is given in Figure B.11.

The order of the degrees of freedom in F_{k_1, k_2} is critical. The integer k_1 is called the *numerator degrees of freedom* because it is associated with the chi-square variable in the numerator. Likewise, the integer k_2 is called the *denominator degrees of freedom* because it is associated with the chi-square variable in the denominator. This can be a little tricky because (B.43) can also be written as $(X_1k_2)/(X_2k_1)$, so that k_1 appears in the denominator. Just remember that the numerator *df* is the integer associated with the chi-square variable in the numerator of (B.43), and similarly for the denominator *df*.

FIGURE B.10 The *t* distribution with various degrees of freedom.

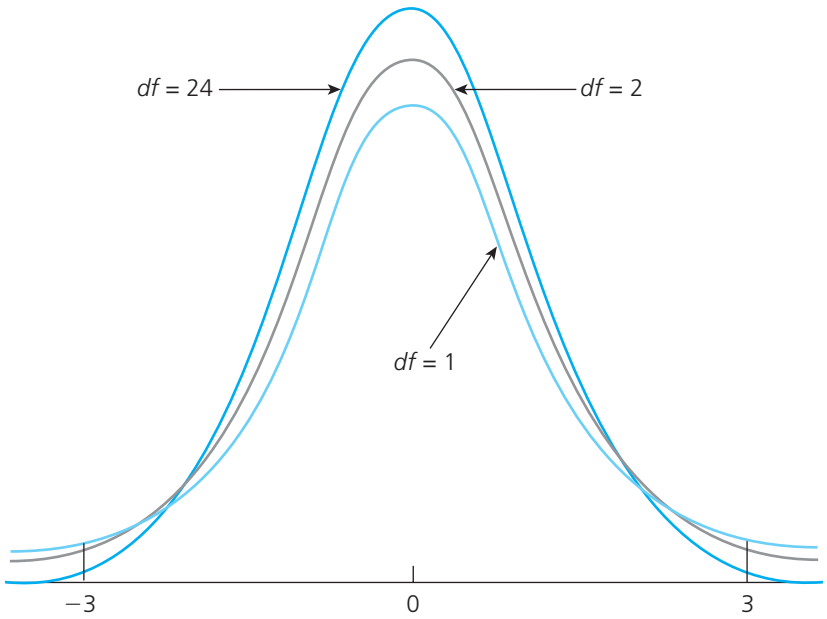
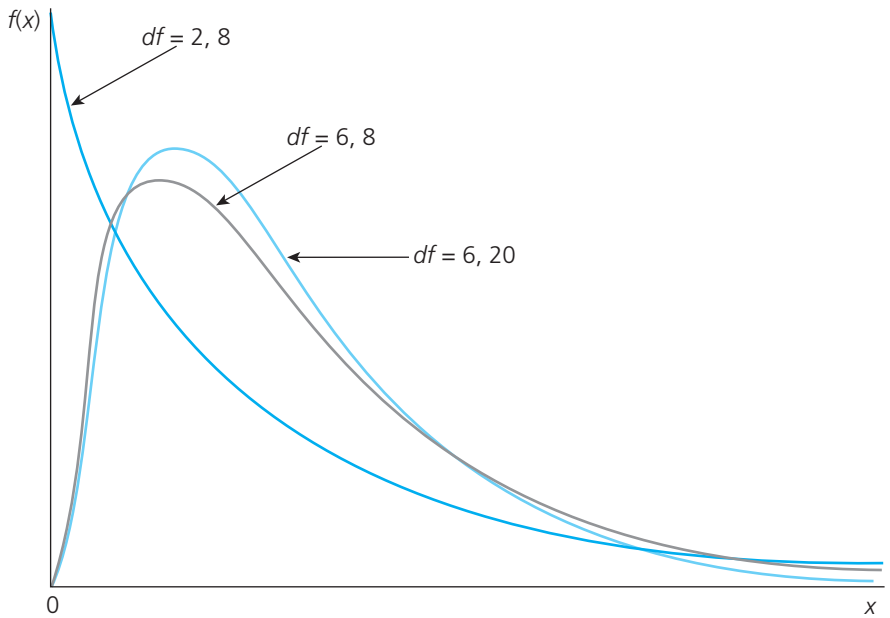


FIGURE B.11 The F_{k_1, k_2} distribution for various degrees of freedom, k_1 and k_2.



Summary

In this appendix, we have reviewed the probability concepts that are needed in econometrics. Most of the concepts should be familiar from your introductory course in probability and statistics. Some of the more advanced topics, such as features of conditional expectations, do not need to be mastered now—there is time for that when these concepts arise in the context of regression analysis in Part 1.

In an introductory statistics course, the focus is on calculating means, variances, covariances, and so on for particular distributions. In Part 1, we will not need such calculations: we mostly rely on the *properties* of expectations, variances, and so on that have been stated in this appendix.

Key Terms

Bernoulli (or Binary) Random Variable
Binomial Distribution
Chi-Square Distribution
Conditional Distribution
Conditional Expectation
Continuous Random Variable
Correlation Coefficient
Covariance
Cumulative Distribution Function (cdf)
Degrees of Freedom
Discrete Random Variable
Expected Value
Experiment
F Distribution
Independent Random Variables
Joint Distribution
Kurtosis
Law of Iterated Expectations
Median
Normal Distribution
Pairwise Uncorrelated Random Variables
Probability Density Function (pdf)
Random Variable
Skewness
Standard Deviation
Standard Normal Distribution
Standardized Random Variable
Symmetric Distribution
t Distribution
Uncorrelated Random Variables
Variance

Problems

1 Suppose that a high school student is preparing to take the SAT exam. Explain why his or her eventual SAT score is properly viewed as a random variable.

2 Let X be a random variable distributed as Normal(5,4). Find the probabilities of the following events:
(i) $P(X \leq 6)$.
(ii) $P(X > 4)$.
(iii) $P(|X - 5| > 1)$.

3 Much is made of the fact that certain mutual funds outperform the market year after year (that is, the return from holding shares in the mutual fund is higher than the return from holding a portfolio such as the S&P 500). For concreteness, consider a 10-year period and let the population be the 4,170 mutual funds reported in *The Wall Street Journal* on January 1, 1995. By saying that performance relative to the market is random, we mean that each fund has a 50–50 chance of outperforming the market in any year and that performance is independent from year to year.
(i) If performance relative to the market is truly random, what is the probability that any particular fund outperforms the market in all 10 years?
(ii) Of the 4,170 mutual funds, what is the expected number of funds that will outperform the market in all 10 years?
(iii) Find the probability that *at least* one fund out of 4,170 funds outperforms the market in all 10 years. What do you make of your answer?
(iv) If you have a statistical package that computes binomial probabilities, find the probability that at least five funds outperform the market in all 10 years.

4 For a randomly selected county in the United States, let X represent the proportion of adults over age 65 who are employed, or the elderly employment rate. Then, X is restricted to a value between zero and one. Suppose that the cumulative distribution function for X is given by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least .6 (60%).

5 Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that about 20% of the adult population believed Simpson was innocent (after much of the physical evidence in the case had been revealed to the public). Ignore the fact that this 20% is an estimate based on a subsample from the population; for illustration, take it as the true percentage of people who thought Simpson was innocent prior to jury selection. Assume that the 12 jurors were selected randomly and independently from the population (although this turned out not to be true).

(i) Find the probability that the jury had at least one member who believed in Simpson's innocence prior to jury selection. [*Hint*: Define the Binomial(12,.20) random variable X to be the number of jurors believing in Simpson's innocence.]

(ii) Find the probability that the jury had at least two members who believed in Simpson's innocence. [*Hint*: $P(X \geq 2) = 1 - P(X \leq 1)$ and $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

6 (Requires calculus) Let X denote the prison sentence, in years, for people convicted of auto theft in a particular state in the United States. Suppose that the pdf of X is given by

$$f(x) = (1/9)x^2, 0 < x < 3.$$

Use integration to find the expected prison sentence.

7 If a basketball player is a 74% free throw shooter, then, on average, how many free throws will he or she make in a game with eight free throw attempts?

8 Suppose that a college student is taking three courses: a two-credit course, a three-credit course, and a four-credit course. The expected grade in the two-credit course is 3.5, while the expected grade in the three- and four-credit courses is 3.0. What is the expected overall grade point average for the semester? (Remember that each course grade is weighted by its share of the total number of units.)

9 Let X denote the annual salary of university professors in the United States, measured in thousands of dollars. Suppose that the average salary is 52.3, with a standard deviation of 14.6. Find the mean and standard deviation when salary is measured in dollars.

10 Suppose that at a large university, college grade point average, *GPA*, and SAT score, *SAT*, are related by the conditional expectation $E(GPA|SAT) = .70 + .002\,SAT$.

(i) Find the expected *GPA* when $SAT = 800$. Find $E(GPA|SAT = 1{,}400)$. Comment on the difference.

(ii) If the average *SAT* in the university is 1,100, what is the average *GPA*? (*Hint*: Use Property CE.4.)

(iii) If a student's SAT score is 1,100, does this mean he or she will have the GPA found in part (ii)? Explain.

11 (i) Let X be a random variable taking on the values -1 and 1, each with probability 1/2. Find $E(X)$ and $E(X^2)$.

(ii) Now let X be a random variable taking on the values 1 and 2, each with probability 1/2. Find $E(X)$ and $E(1/X)$.

(iii) Conclude from parts (i) and (ii) that, in general,

$$E[g(X)] \neq g[E(X)]$$

for a nonlinear function $g(\cdot)$.

(iv) Given the definition of the F random variable in equation (B.43), show that

$$E(F) = E\left[\frac{1}{(X_2/k_2)}\right].$$

Can you conclude that $E(F) = 1$?