CHAPTER

2 Review of Probability

This chapter reviews the core ideas of the theory of probability that are needed to understand regression analysis and econometrics. We assume that you have taken an introductory course in probability and statistics. If your knowledge of probability is stale, you should refresh it by reading this chapter. If you feel confident with the material, you still should skim the chapter and the terms and concepts at the end to make sure you are familiar with the ideas and notation.

Most aspects of the world around us have an element of randomness. The theory of probability provides mathematical tools for quantifying and describing this randomness. Section 2.1 reviews probability distributions for a single random variable, and Section 2.2 covers the mathematical expectation, mean, and variance of a single random variable. Most of the interesting problems in economics involve more than one variable, and Section 2.3 introduces the basic elements of probability theory for two random variables. Section 2.4 discusses three special probability distributions that play a central role in statistics and econometrics: the normal, chi-squared, and F distributions.

The final two sections of this chapter focus on a specific source of randomness of central importance in econometrics: the randomness that arises by randomly drawing a sample of data from a larger population. For example, suppose you survey ten recent college graduates selected at random, record (or "observe") their earnings, and compute the average earnings using these ten data points (or "observations"). Because you chose the sample at random, you could have chosen ten different graduates by pure random chance; had you done so, you would have observed ten different earnings, and you would have computed a different sample average. Because the average earnings vary from one randomly chosen sample to the next, the sample average is itself a random variable. Therefore, the sample average has a probability distribution, which is referred to as its sampling distribution because this distribution describes the different possible values of the sample average that would have occurred had a different sample been drawn.

Section 2.5 discusses random sampling and the sampling distribution of the sample average. This sampling distribution is, in general, complicated. When the sample size is sufficiently large, however, the sampling distribution of the sample average is approximately normal, a result known as the central limit theorem, which is discussed in Section 2.6.

2.1 Random Variables and Probability Distributions

Probabilities, the Sample Space, and Random Variables

Probabilities and outcomes. The sex of the next new person you meet, your grade on an exam, and the number of times your wireless network connection fails while you are writing a term paper all have an element of chance or randomness. In each of these examples, there is something not yet known that is eventually revealed.

The mutually exclusive potential results of a random process are called the **outcomes**. For example, while writing your term paper, the wireless connection might never fail, it might fail once, it might fail twice, and so on. Only one of these outcomes will actually occur (the outcomes are mutually exclusive), and the outcomes need not be equally likely.

The **probability** of an outcome is the proportion of the time that the outcome occurs in the long run. If the probability of your wireless connection not failing while you are writing a term paper is 80%, then over the course of writing many term papers, you will complete 80% without a wireless connection failure.

The sample space and events. The set of all possible outcomes is called the **sample space**. An **event** is a subset of the sample space; that is, an event is a set of one or more outcomes. The event "my wireless connection will fail no more than once" is the set consisting of two outcomes: "no failures" and "one failure."

Random variables. A random variable is a numerical summary of a random outcome. The number of times your wireless connection fails while you are writing a term paper is random and takes on a numerical value, so it is a random variable.

Some random variables are discrete and some are continuous. As their names suggest, a **discrete random variable** takes on only a discrete set of values, like $0, 1, 2, \ldots$, whereas a **continuous random variable** takes on a continuum of possible values.

Probability Distribution of a Discrete Random Variable

Probability distribution. The **probability distribution** of a discrete random variable is the list of all possible values of the variable and the probability that each value will occur. These probabilities sum to 1.

For example, let M be the number of times your wireless network connection fails while you are writing a term paper. The probability distribution of the random variable M is the list of probabilities of all possible outcomes: The probability that $M = 0$, denoted $\Pr(M = 0)$, is the probability of no wireless connection failures; $\Pr(M = 1)$ is the probability of a single connection failure; and so forth. An example of a probability distribution for M is given in the first row of Table 2.1. According to this distribution, the probability of no connection failures is 80%; the probability of one failure is 10%; and the probabilities of two, three, and four failures are,

TABLE 2.1 Probability of Your Wireless Network Connection Failing M Times

	Outcome (number of failures)				
	0	1	2	3	4
Probability distribution	0.80	0.10	0.06	0.03	0.01
Cumulative probability distribution	0.80	0.90	0.96	0.99	1.00

respectively, 6%, 3%, and 1%. These probabilities sum to 100%. This probability distribution is plotted in Figure 2.1.

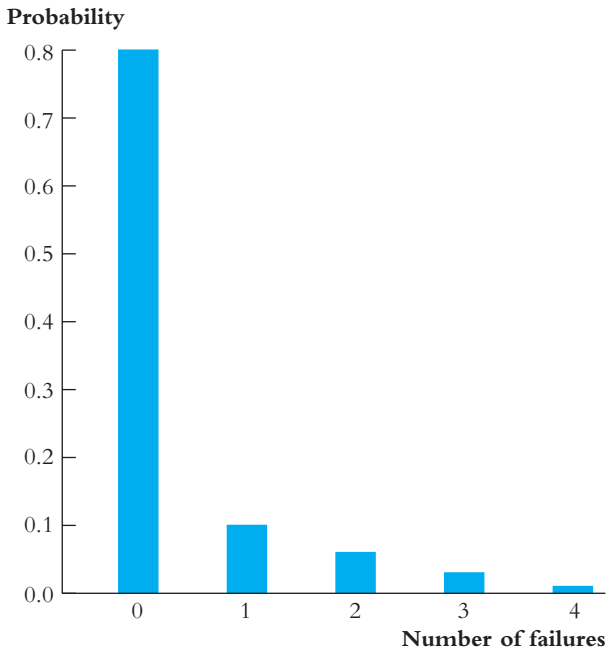
Probabilities of events. The probability of an event can be computed from the probability distribution. For example, the probability of the event of one or two failures is the sum of the probabilities of the constituent outcomes. That is, $\Pr(M = 1 \text{ or } M = 2) = \Pr(M = 1) + \Pr(M = 2) = 0.10 + 0.06 = 0.16$, or 16%.

Cumulative probability distribution. The **cumulative probability distribution** is the probability that the random variable is less than or equal to a particular value. The final row of Table 2.1 gives the cumulative probability distribution of the random variable M. For example, the probability of at most one connection failure, $\Pr(M \leq 1)$, is 90%, which is the sum of the probabilities of no failures (80%) and of one failure (10%).

A cumulative probability distribution is also referred to as a **cumulative distribution function**, a **c.d.f.**, or a **cumulative distribution**.

FIGURE 2.1 Probability Distribution of the Number of Wireless Network Connection Failures

The height of each bar is the probability that the wireless connection fails the indicated number of times. The height of the first bar is 0.8, so the probability of 0 connection failures is 80%. The height of the second bar is 0.1, so the probability of 1 failure is 10%, and so forth for the other bars.



The Bernoulli distribution. An important special case of a discrete random variable is when the random variable is binary; that is, the outcome is 0 or 1. A binary random variable is called a **Bernoulli random variable** (in honor of the 17th-century Swiss mathematician and scientist Jacob Bernoulli), and its probability distribution is called the **Bernoulli distribution**.

For example, let G be the sex of the next new person you meet, where $G = 0$ indicates that the person is male and $G = 1$ indicates that the person is female. The outcomes of G and their probabilities thus are

$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p, \end{cases} \tag{2.1}$$

where p is the probability of the next new person you meet being a woman. The probability distribution in Equation (2.1) is the Bernoulli distribution.

Probability Distribution of a Continuous Random Variable

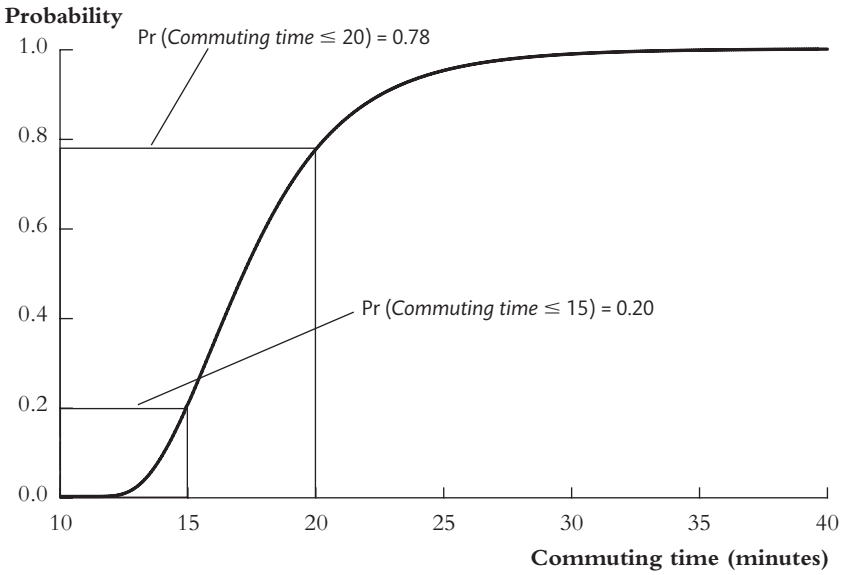
Cumulative probability distribution. The cumulative probability distribution for a continuous variable is defined just as it is for a discrete random variable. That is, the cumulative probability distribution of a continuous random variable is the probability that the random variable is less than or equal to a particular value.

For example, consider a student who drives from home to school. This student's commuting time can take on a continuum of values, and because it depends on random factors such as the weather and traffic conditions, it is natural to treat it as a continuous random variable. Figure 2.2a plots a hypothetical cumulative distribution of commuting times. For example, the probability that the commute takes less than 15 minutes is 20%, and the probability that it takes less than 20 minutes is 78%.

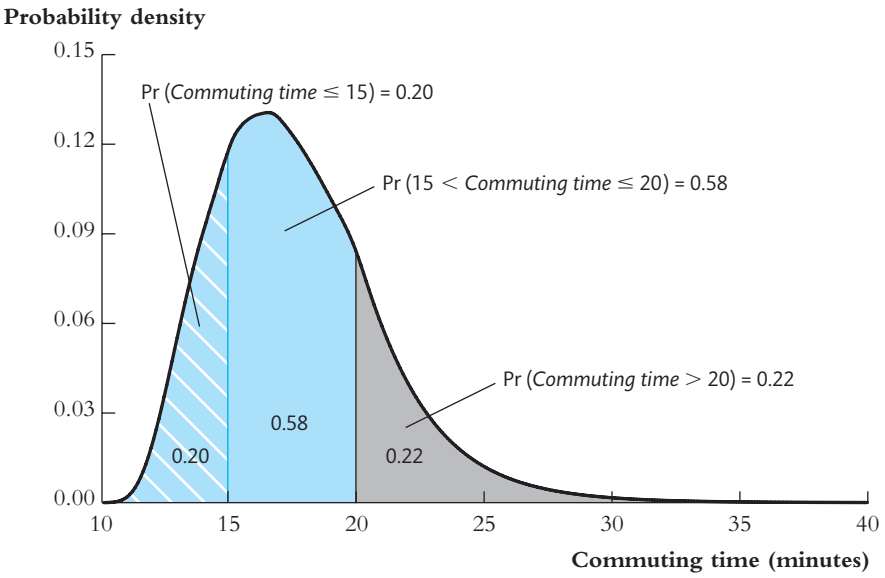
Probability density function. Because a continuous random variable can take on a continuum of possible values, the probability distribution used for discrete variables, which lists the probability of each possible value of the random variable, is not suitable for continuous variables. Instead, the probability is summarized by the **probability density function**. The area under the probability density function between any two points is the probability that the random variable falls between those two points. A probability density function is also called a **p.d.f.**, a **density function**, or simply a **density**.

Figure 2.2b plots the probability density function of commuting times corresponding to the cumulative distribution in Figure 2.2a. The probability that the commute takes between 15 and 20 minutes is given by the area under the p.d.f. between 15 minutes and 20 minutes, which is 0.58, or 58%. Equivalently, this probability can be seen on the cumulative distribution in Figure 2.2a as the difference between the probability that the commute is less than 20 minutes (78%) and the probability that it is less than 15 minutes (20%). Thus the probability density function and the cumulative probability distribution show the same information in different formats.

FIGURE 2.2 Cumulative Probability Distribution and Probability Density Functions of Commuting Time



(a) Cumulative probability distribution function of commuting times



(b) Probability density function of commuting times

Figure 2.2a shows the cumulative probability distribution function (c.d.f.) of commuting times. The probability that a commuting time is less than 15 minutes is 0.20 (or 20%), and the probability that it is less than 20 minutes is 0.78 (78%). Figure 2.2b shows the probability density function (or p.d.f.) of commuting times. Probabilities are given by areas under the p.d.f. The probability that a commuting time is between 15 and 20 minutes is 0.58 (58%) and is given by the area under the curve between 15 and 20 minutes.

2.2 Expected Values, Mean, and Variance

The Expected Value of a Random Variable

Expected value. The **expected value** of a random variable Y, denoted $E(Y)$, is the long-run average value of the random variable over many repeated trials or occurrences. The expected value of a discrete random variable is computed as a weighted average of the possible outcomes of that random variable, where the weights are the probabilities of that outcome. The expected value of Y is also called the **expectation** of Y or the **mean** of Y and is denoted μ_Y.

For example, suppose you loan a friend \$100 at 10% interest. If the loan is repaid, you get \$110 (the principal of \$100 plus interest of \$10), but there is a risk of 1% that your friend will default and you will get nothing at all. Thus the amount you are repaid is a random variable that equals \$110 with probability 0.99 and equals \$0 with probability 0.01. Over many such loans, 99% of the time you would be paid back \$110, but 1% of the time you would get nothing, so on average you would be repaid $\$110 \times 0.99 + \$0 \times 0.01 = \$108.90$. Thus the expected value of your repayment is \$108.90.

As a second example, consider the number of wireless network connection failures M with the probability distribution given in Table 2.1. The expected value of M—that is, the mean of M—is the average number of failures over many term papers, weighted by the frequency with which a given number of failures occurs. Accordingly,

$$E(M) = 0 \times 0.80 + 1 \times 0.10 + 2 \times 0.06 + 3 \times 0.03 + 4 \times 0.01 = 0.35. \quad (2.2)$$

That is, the expected number of connection failures while writing a term paper is 0.35. Of course, the actual number of failures must always be an integer; it makes no sense to say that the wireless connection failed 0.35 times while writing a particular term paper! Rather, the calculation in Equation (2.2) means that the average number of failures over many such term papers is 0.35.

The formula for the expected value of a discrete random variable Y that can take on k different values is given in Key Concept 2.1. (Key Concept 2.1 uses summation notation, which is reviewed in Exercise 2.25.)

KEY CONCEPT 2.1

Expected Value and the Mean

Suppose that the random variable Y takes on k possible values, $y_1, \ldots, y_k$, where y_1 denotes the first value, y_2 denotes the second value, and so forth, and that the probability that Y takes on y_1 is p_1, the probability that Y takes on y_2 is p_2, and so forth. The expected value of Y, denoted $E(Y)$, is

$$E(Y) = y_1p_1 + y_2p_2 + \cdots + y_kp_k = \sum_{i=1}^{k} y_ip_i, \quad (2.3)$$

where the notation $\sum_{i=1}^{k} y_ip_i$ means "the sum of y_ip_i for i running from 1 to k." The expected value of Y is also called the mean of Y or the expectation of Y and is denoted μ_Y.

Expected value of a Bernoulli random variable. An important special case of the general formula in Key Concept 2.1 is the mean of a Bernoulli random variable. Let G be the Bernoulli random variable with the probability distribution in Equation (2.1). The expected value of G is

$$E(G) = 0 \times (1 - p) + 1 \times p = p. \tag{2.4}$$

Thus the expected value of a Bernoulli random variable is p, the probability that it takes on the value 1.

Expected value of a continuous random variable. The expected value of a continuous random variable is also the probability-weighted average of the possible outcomes of the random variable. Because a continuous random variable can take on a continuum of possible values, the formal mathematical definition of its expectation involves calculus and its definition is given in Appendix 18.1.

The Standard Deviation and Variance

The variance and standard deviation measure the dispersion or the "spread" of a probability distribution. The **variance** of a random variable Y, denoted $\text{var}(Y)$, is the expected value of the square of the deviation of Y from its mean: $\text{var}(Y) = E[(Y - \mu_Y)^2]$.

Because the variance involves the square of Y, the units of the variance are the units of the square of Y, which makes the variance awkward to interpret. It is therefore common to measure the spread by the **standard deviation**, which is the square root of the variance and is denoted σ_Y. The standard deviation has the same units as Y. These definitions are summarized in Key Concept 2.2.

For example, the variance of the number of connection failures M is the probability-weighted average of the squared difference between M and its mean, 0.35:

$$\begin{aligned}\text{var}(M) = {} & (0 - 0.35)^2 \times 0.80 + (1 - 0.35)^2 \times 0.10 + (2 - 0.35)^2 \times 0.06 \\ & + (3 - 0.35)^2 \times 0.03 + (4 - 0.35)^2 \times 0.01 = 0.6475.\end{aligned} \tag{2.5}$$

The standard deviation of M is the square root of the variance, so $\sigma_M = \sqrt{0.64750} \cong 0.80$.

Variance and Standard Deviation

KEY CONCEPT 2.2

The variance of the discrete random variable Y, denoted σ_Y^2, is

$$\sigma_Y^2 = \text{var}(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \tag{2.6}$$

The standard deviation of Y is σ_Y, the square root of the variance. The units of the standard deviation are the same as the units of Y.

Variance of a Bernoulli random variable. The mean of the Bernoulli random variable G with the probability distribution in Equation (2.1) is $\mu_G = p$ [Equation (2.4)], so its variance is

$$\text{var}(G) = \sigma_G^2 = (0-p)^2 \times (1-p) + (1-p)^2 \times p = p(1-p). \quad (2.7)$$

Thus the standard deviation of a Bernoulli random variable is $\sigma_G = \sqrt{p(1-p)}$.

Mean and Variance of a Linear Function of a Random Variable

This section discusses random variables (say, X and Y) that are related by a linear function. For example, consider an income tax scheme under which a worker is taxed at a rate of 20% on his or her earnings and then given a (tax-free) grant of $2000. Under this tax scheme, after-tax earnings Y are related to pre-tax earnings X by the equation

$$Y = 2000 + 0.8X. \quad (2.8)$$

That is, after-tax earnings Y is 80% of pre-tax earnings X, plus $2000.

Suppose an individual's pre-tax earnings next year are a random variable with mean μ_X and variance σ_X^2. Because pre-tax earnings are random, so are after-tax earnings. What are the mean and standard deviations of her after-tax earnings under this tax? After taxes, her earnings are 80% of the original pre-tax earnings, plus $2000. Thus the expected value of her after-tax earnings is

$$E(Y) = \mu_Y = 2000 + 0.8\mu_X. \quad (2.9)$$

The variance of after-tax earnings is the expected value of $(Y-\mu_Y)^2$. Because $Y = 2000 + 0.8X$, $Y - \mu_Y = 2000 + 0.8X - (2000 + 0.8\mu_X) = 0.8(X-\mu_X)$. Thus $E[(Y-\mu_Y)^2] = E\{[0.8(X-\mu_X)]^2\} = 0.64E[(X-\mu_X)^2]$. It follows that $\text{var}(Y) = 0.64\text{var}(X)$, so, taking the square root of the variance, the standard deviation of Y is

$$\sigma_Y = 0.8\sigma_X. \quad (2.10)$$

That is, the standard deviation of the distribution of her after-tax earnings is 80% of the standard deviation of the distribution of her pre-tax earnings.

This analysis can be generalized so that Y depends on X with an intercept a (instead of $2000) and a slope b (instead of 0.8) so that

$$Y = a + bX. \quad (2.11)$$

Then the mean and variance of Y are

$$\mu_Y = a + b\mu_X \quad \text{and} \quad (2.12)$$

$$\sigma_Y^2 = b^2\sigma_X^2, \quad (2.13)$$

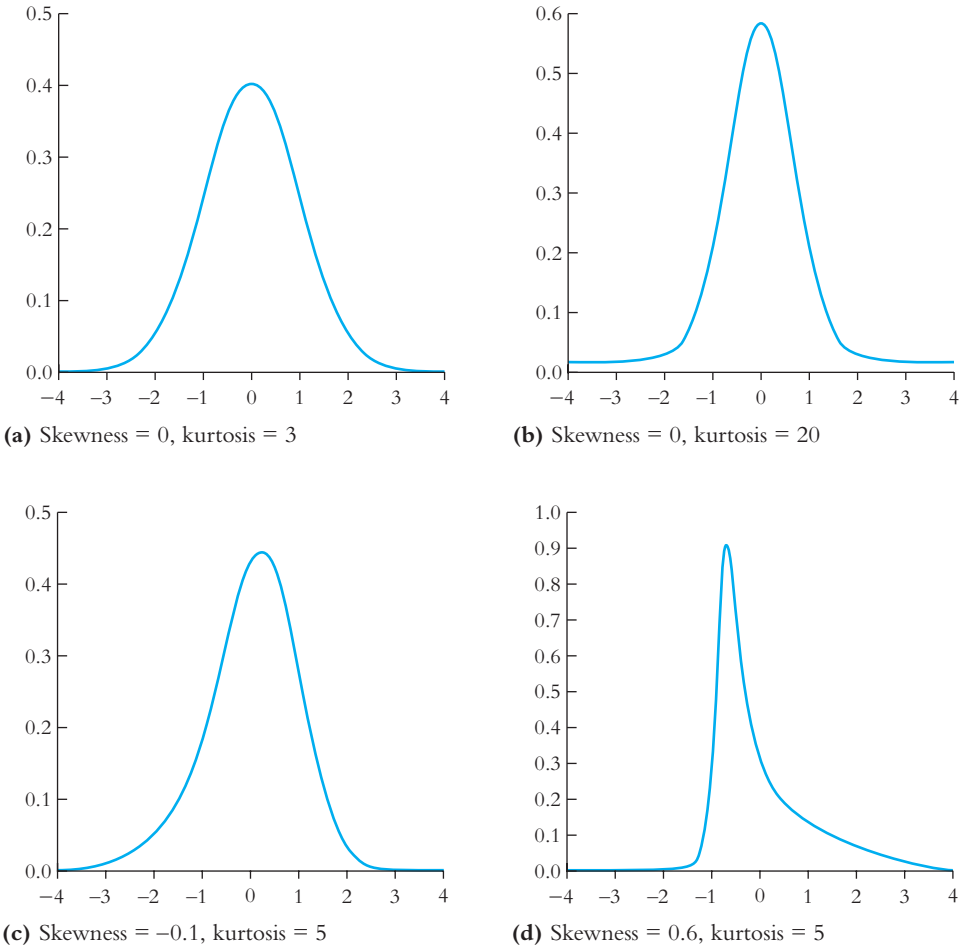
and the standard deviation of Y is $\sigma_Y = b\sigma_X$. The expressions in Equations (2.9) and (2.10) are applications of the more general formulas in Equations (2.12) and (2.13) with $a = 2000$ and $b = 0.8$.

Other Measures of the Shape of a Distribution

The mean and standard deviation measure two important features of a distribution: its center (the mean) and its spread (the standard deviation). This section discusses measures of two other features of a distribution: the skewness, which measures the lack of symmetry of a distribution, and the kurtosis, which measures how thick, or "heavy," are its tails. The mean, variance, skewness, and kurtosis are all based on what are called the **moments of a distribution**.

Skewness. Figure 2.3 plots four distributions, two that are symmetric (Figures 2.3a and 2.3b) and two that are not (Figures 2.3c and 2.3d). Visually, the distribution in Figure 2.3d appears to deviate more from symmetry than does the distribution in

FIGURE 2.3 Four Distributions with Different Skewness and Kurtosis



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = −0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

All of these distributions have a mean of 0 and a variance of 1. The distributions with skewness of 0 (a and b) are symmetric; the distributions with nonzero skewness (c and d) are not symmetric. The distributions with kurtosis exceeding 3 (b, c, and d) have heavy tails.

Figure 2.3c. The skewness of a distribution provides a mathematical way to describe how much a distribution deviates from symmetry.

The **skewness** of the distribution of a random variable Y is

$$\text{Skewness} = \frac{E[(Y - \mu_Y)^3]}{\sigma_Y^3}, \tag{2.14}$$

where σ_Y is the standard deviation of Y. For a symmetric distribution, a value of Y a given amount above its mean is just as likely as a value of Y the same amount below its mean. If so, then positive values of $(Y - \mu_Y)^3$ will be offset on average (in expectation) by equally likely negative values. Thus, for a symmetric distribution, $E(Y - \mu_Y)^3 = 0$: The skewness of a symmetric distribution is 0. If a distribution is not symmetric, then a positive value of $(Y - \mu_Y)^3$ generally is not offset on average by an equally likely negative value, so the skewness is nonzero for a distribution that is not symmetric. Dividing by σ_Y^3 in the denominator of Equation (2.14) cancels the units of Y^3 in the numerator, so the skewness is unit free; in other words, changing the units of Y does not change its skewness.

Below each of the four distributions in Figure 2.3 is its skewness. If a distribution has a long right tail, positive values of $(Y - \mu_Y)^3$ are not fully offset by negative values, and the skewness is positive. If a distribution has a long left tail, its skewness is negative.

Kurtosis. The **kurtosis** of a distribution is a measure of how much mass is in its tails and therefore is a measure of how much of the variance of Y arises from extreme values. An extreme value of Y is called an **outlier**. The greater the kurtosis of a distribution, the more likely are outliers.

The kurtosis of the distribution of Y is

$$\text{Kurtosis} = \frac{E[(Y - \mu_Y)^4]}{\sigma_Y^4}. \tag{2.15}$$

If a distribution has a large amount of mass in its tails, then some extreme departures of Y from its mean are likely, and these departures will lead to large values, on average (in expectation), of $(Y - \mu_Y)^4$. Thus, for a distribution with a large amount of mass in its tails, the kurtosis will be large. Because $(Y - \mu_Y)^4$ cannot be negative, the kurtosis cannot be negative.

The kurtosis of a normally distributed random variable is 3, so a random variable with kurtosis exceeding 3 has more mass in its tails than a normal random variable. A distribution with kurtosis exceeding 3 is called **leptokurtic** or, more simply, heavy-tailed. Like skewness, the kurtosis is unit free, so changing the units of Y does not change its kurtosis.

Below each of the four distributions in Figure 2.3 is its kurtosis. The distributions in Figures 2.3b–d are heavy-tailed.

Moments. The mean of Y, $E(Y)$, is also called the first moment of Y, and the expected value of the square of Y, $E(Y^2)$, is called the second moment of Y. In general, the

expected value of Y^r is called the ***r*th moment** of the random variable Y. That is, the r^{th} moment of Y is $E(Y^r)$. The skewness is a function of the first, second, and third moments of Y, and the kurtosis is a function of the first through fourth moments of Y.

Standardized Random Variables

A random variable can be transformed into a random variable with mean 0 and variance 1 by subtracting its mean and then dividing by its standard deviation, a process called standardization. Specifically, let Y have mean μ_Y and variance σ_Y^2. Then the **standardized random variable** computed from Y is $(Y - \mu_Y)/\sigma_Y$. The mean of the standardized random variable is $E(Y - \mu_Y)/\sigma_Y = (EY - \mu_Y)/\sigma_Y = 0$, and its variance is $\text{var}[(Y - \mu_Y)/\sigma_Y] = \text{var}(Y)/\sigma_Y^2 = 1$. Standardized random variables do not have any units, such as dollars or meters, because the units of Y are canceled by dividing through by σ_Y, which also has the units of Y.

2.3 Two Random Variables

Most of the interesting questions in economics involve two or more variables. Are college graduates more likely to have a job than nongraduates? How does the distribution of income for women compare to that for men? These questions concern the distribution of two random variables, considered together (education and employment status in the first example, income and sex in the second). Answering such questions requires an understanding of the concepts of joint, marginal, and conditional probability distributions.

Joint and Marginal Distributions

Joint distribution. The **joint probability distribution** of two discrete random variables, say X and Y, is the probability that the random variables simultaneously take on certain values, say x and y. The probabilities of all possible (x, y) combinations sum to 1. The joint probability distribution can be written as the function $\Pr(X = x, Y = y)$.

For example, weather conditions—whether or not it is raining—affect the commuting time of the student commuter in Section 2.1. Let Y be a binary random variable that equals 1 if the commute is short (less than 20 minutes) and that equals 0 otherwise, and let X be a binary random variable that equals 0 if it is raining and 1 if not. Between these two random variables, there are four possible outcomes: it rains and the commute is long $(X = 0, Y = 0)$; rain and short commute $(X = 0, Y = 1)$; no rain and long commute $(X = 1, Y = 0)$; and no rain and short commute $(X = 1, Y = 1)$. The joint probability distribution is the frequency with which each of these four outcomes occurs over many repeated commutes.

An example of a joint distribution of these two variables is given in Table 2.2. According to this distribution, over many commutes, 15% of the days have rain and a long commute $(X = 0, Y = 0)$; that is, the probability of a long rainy commute is

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

15%, or $\Pr(X = 0, Y = 0) = 0.15$. Also, $\Pr(X = 0, Y = 1) = 0.15$, $\Pr(X = 1, Y = 0) = 0.07$, and $\Pr(X = 1, Y = 1) = 0.63$. These four possible outcomes are mutually exclusive and constitute the sample space, so the four probabilities sum to 1.

Marginal probability distribution. The **marginal probability distribution** of a random variable Y is just another name for its probability distribution. This term is used to distinguish the distribution of Y alone (the marginal distribution) from the joint distribution of Y and another random variable.

The marginal distribution of Y can be computed from the joint distribution of X and Y by adding up the probabilities of all possible outcomes for which Y takes on a specified value. If X can take on l different values $x_1, \ldots, x_l$, then the marginal probability that Y takes on the value y is

$$\Pr(Y = y) = \sum_{i=1}^{l} \Pr(X = x_i, Y = y). \tag{2.16}$$

For example, in Table 2.2, the probability of a long rainy commute is 15%, and the probability of a long commute with no rain is 7%, so the probability of a long commute (rainy or not) is 22%. The marginal distribution of commuting times is given in the final column of Table 2.2. Similarly, the marginal probability that it will rain is 30%, as shown in the final row of Table 2.2.

Conditional Distributions

Conditional distribution. The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the **conditional distribution** of Y given X. The conditional probability that Y takes on the value y when X takes on the value x is written $\Pr(Y = y | X = x)$.

For example, what is the probability of a long commute ($Y = 0$) if you know it is raining ($X = 0$)? From Table 2.2, the joint probability of a rainy short commute is 15%, and the joint probability of a rainy long commute is 15%, so if it is raining, a long commute and a short commute are equally likely. Thus the probability of a long commute ($Y = 0$) conditional on it being rainy ($X = 0$) is 50%, or $\Pr(Y = 0 | X = 0) = 0.50$. Equivalently, the marginal probability of rain is 30%; that is, over many commutes, it rains 30% of the time. Of this 30% of commutes, 50% of the time the commute is long ($0.15/0.30$).

TABLE 2.3 Joint and Conditional Distributions of Number of Wireless Connection Failures (M) and Network Age (A)

A. Joint Distribution

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old network ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.50
New network ($A = 1$)	0.45	0.035	0.01	0.005	0.00	0.50
Total	0.80	0.10	0.06	0.03	0.01	1.00

B. Conditional Distributions of *M* given *A*

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
$\Pr(M \mid A = 0)$	0.70	0.13	0.10	0.05	0.02	1.00
$\Pr(M \mid A = 1)$	0.90	0.07	0.02	0.01	0.00	1.00

In general, the conditional distribution of Y given $X = x$ is

$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x,\ Y = y)}{\Pr(X = x)}. \tag{2.17}$$

For example, the conditional probability of a long commute given that it is rainy is $\Pr(Y = 0 \mid X = 0) = \Pr(X = 0,\ Y = 0) / \Pr(X = 0) = 0.15 / 0.30 = 0.50$.

As a second example, consider a modification of the network connection failure example. Suppose that half the time you write your term paper in the school library, which has a new wireless network; otherwise, you write it in your room, which has an old wireless network. If we treat the location where you write the term paper as random, then the network age A ($= 1$ if the network is new, $= 0$ if it is old) is a random variable. Suppose the joint distribution of the random variables M and A is given in Part A of Table 2.3. Then the conditional distributions of connection failures given the age of the network are shown in Part B of the table. For example, the joint probability of $M = 0$ and $A = 0$ is 0.35; because half the time you use the old network, the conditional probability of no failures given that you use the old network is $\Pr(M = 0 \mid A = 0) = \Pr(M = 0,\ A = 0) / \Pr(A = 0) = 0.35 / 0.50 = 0.70$, or 70%. In contrast, the conditional probability of no failures given that you use the new network is 90%. According to the conditional distributions in Part B of Table 2.3, the new network is less likely to fail than the old one; for example, the probability of three failures is 5% using the old network but 1% using the new network.

Conditional expectation. The **conditional expectation** of Y given X, also called the **conditional mean** of Y given X, is the mean of the conditional distribution of Y given X. That is, the conditional expectation is the expected value of Y, computed using the conditional distribution of Y given X. If Y takes on k values $y_1, \ldots, y_k$, then the conditional mean of Y given $X = x$ is

$$E(Y \mid X = x) = \sum_{i=1}^{k} y_i \Pr(Y = y_i \mid X = x). \tag{2.18}$$

For example, based on the conditional distributions in Table 2.3, the expected number of connection failures, given that the network is old, is $E(M|A = 0) = 0 \times 0.70 + 1 \times 0.13 + 2 \times 0.10 + 3 \times 0.05 + 4 \times 0.02 = 0.56$. The expected number of failures, given that the network is new, is $E(M|A = 1) = 0.14$, less than for the old network.

The conditional expectation of Y given $X = x$ is just the mean value of Y when $X = x$. In the example of Table 2.3, the mean number of failures is 0.56 for the old network, so the conditional expectation of Y given that the network is old is 0.56. Similarly, for the new network, the mean number of failures is 0.14; that is, the conditional expectation of Y given that the network is new is 0.14.

The law of iterated expectations. The mean of Y is the weighted average of the conditional expectation of Y given X, weighted by the probability distribution of X. For example, the mean height of adults is the weighted average of the mean height of men and the mean height of women, weighted by the proportions of men and women. Stated mathematically, if X takes on the l values $x_1, \ldots, x_l$, then

$$E(Y) = \sum_{i=1}^{l} E(Y|X = x_i)\Pr(X = x_i). \tag{2.19}$$

Equation (2.19) follows from Equations (2.18) and (2.17) (see Exercise 2.19).

Stated differently, the expectation of Y is the expectation of the conditional expectation of Y given X,

$$E(Y) = E[E(Y|X)], \tag{2.20}$$

where the inner expectation on the right-hand side of Equation (2.20) is computed using the conditional distribution of Y given X and the outer expectation is computed using the marginal distribution of X. Equation (2.20) is known as the **law of iterated expectations**.

For example, the mean number of connection failures M is the weighted average of the conditional expectation of M given that it is old and the conditional expectation of M given that it is new, so $E(M) = E(M|A = 0) \times \Pr(A = 0) + E(M|A = 1) \times \Pr(A = 1) = 0.56 \times 0.50 + 0.14 \times 0.50 = 0.35$. This is the mean of the marginal distribution of M, as calculated in Equation (2.2).

The law of iterated expectations implies that if the conditional mean of Y given X is 0, then the mean of Y is 0. This is an immediate consequence of Equation (2.20): if $E(Y|X) = 0$, then $E(Y) = E[E(Y|X)] = E[0] = 0$. Said differently, if the mean of Y given X is 0, then it must be that the probability-weighted average of these conditional means is 0; that is, the mean of Y must be 0.

The law of iterated expectations also applies to expectations that are conditional on multiple random variables. For example, let X, Y, and Z be random variables that are jointly distributed. Then the law of iterated expectations says that $E(Y) = E[E(Y|X, Z)]$, where $E(Y|X, Z)$ is the conditional expectation of Y

given both X and Z. For example, in the network connection illustration of Table 2.3, let P denote the number of people using the network; then $E(M|A, P)$ is the expected number of failures for a network with age A that has P users. The expected number of failures overall, $E(M)$, is the weighted average of the expected number of failures for a network with age A and number of users P, weighted by the proportion of occurrences of both A and P.

Exercise 2.20 provides some additional properties of conditional expectations with multiple variables.

Conditional variance. The variance of Y conditional on X is the variance of the conditional distribution of Y given X. Stated mathematically, the **conditional variance** of Y given X is

$$\text{var}(Y|X = x) = \sum_{i=1}^{k} [y_i - E(Y|X = x)]^2 \Pr(Y = y_i|X = x). \quad (2.21)$$

For example, the conditional variance of the number of failures given that the network is old is $\text{var}(M|A = 0) = (0 - 0.56)^2 \times 0.70 + (1 - 0.56)^2 \times 0.13 + (2 - 0.56)^2 \times 0.10 + (3 - 0.56)^2 \times 0.05 + (4 - 0.56)^2 \times 0.02 \cong 0.99$. The standard deviation of the conditional distribution of M given that $A = 0$ is thus $\sqrt{0.99} = 0.99$. The conditional variance of M given that $A = 1$ is the variance of the distribution in the second row of Part B of Table 2.3, which is 0.22, so the standard deviation of M for the new network is $\sqrt{0.22} = 0.47$. For the conditional distributions in Table 2.3, the expected number of failures for the new network (0.14) is less than that for the old network (0.56), and the spread of the distribution of the number of failures, as measured by the conditional standard deviation, is smaller for the new network (0.47) than for the old (0.99).

Bayes' rule. **Bayes' rule** says that the conditional probability of Y given X is the conditional probability of X given Y times the relative marginal probabilities of Y and X:

$$\Pr(Y = y|X = x) = \frac{\Pr(X = x|Y = y)\Pr(Y = y)}{\Pr(X = x)} \text{ (Bayes' rule).} \quad (2.22)$$

Equation (2.22) obtains from the definition of the conditional distribution in Equation (2.17), which implies that $\Pr(X = x, Y = y) = \Pr(Y = y|X = x)\Pr(X = x)$ and that $\Pr(X = x, Y = y) = \Pr(X = x|Y = y)\Pr(Y = y)$; equating the second parts of these two equalities and rearranging gives Bayes' rule.

Bayes' rule can be used to deduce conditional probabilities from the reverse conditional probability, with the help of marginal probabilities. For example, suppose you told your friend that you were dropped by the network three times last night while working on your term paper and your friend knows that half the time you work in the library and half the time you work in your room. Then your friend could deduce from Table 2.3 that the probability you worked in your room last night given three network failures is 83% (Exercise 2.28).

The conditional mean is the minimum mean squared error prediction. The conditional mean plays a central role in prediction; in fact it is, in a precise sense, the optimal prediction of Y given $X = x$.

A common formulation of the statistical prediction problem is to posit that the cost of making a prediction error increases with the square of that error. The motivation for this squared-error prediction loss is that small errors in prediction might not matter much, but large errors can be very costly in real-world applications. Stated mathematically, the prediction problem thus is: what is the function $g(X)$ that minimizes the mean squared prediction error, $E\{[Y - g(X)]^2\}$? The answer is the conditional mean $E(Y|X)$: Of all possible ways to use the information X, the conditional mean minimizes the mean squared prediction error. This result is proven in Appendix 2.2.

Independence

Two random variables X and Y are **independently distributed**, or **independent**, if knowing the value of one of the variables provides no information about the other. Specifically, X and Y are independent if the conditional distribution of Y given X equals the marginal distribution of Y. That is, X and Y are independently distributed if, for all values of x and y,

$$\Pr(Y = y|X = x) = \Pr(Y = y) \text{ (independence of } X \text{ and } Y). \tag{2.23}$$

Substituting Equation (2.23) into Equation (2.17) gives an alternative expression for independent random variables in terms of their joint distribution. If X and Y are independent, then

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y). \tag{2.24}$$

That is, the joint distribution of two independent random variables is the product of their marginal distributions.

Covariance and Correlation

Covariance. One measure of the extent to which two random variables move together is their covariance. The **covariance** between X and Y is the expected value $E[(X - \mu_X)(Y - \mu_Y)]$, where μ_X is the mean of X and μ_Y is the mean of Y. The covariance is denoted $\text{cov}(X, Y)$ or σ_{XY}. If X can take on l values and Y can take on k values, then the covariance is given by the formula

$$\begin{aligned}\text{cov}(X, Y) &= \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X)(y_i - \mu_Y)\Pr(X = x_j, Y = y_i).\end{aligned} \tag{2.25}$$

To interpret this formula, suppose that when X is greater than its mean (so that $X - \mu_X$ is positive), then Y tends be greater than its mean (so that $Y - \mu_Y$ is

positive) and that when X is less than its mean (so that $X - \mu_X < 0$), then Y tends to be less than its mean (so that $Y - \mu_Y < 0$). In both cases, the product $(X - \mu_X) \times (Y - \mu_Y)$ tends to be positive, so the covariance is positive. In contrast, if X and Y tend to move in opposite directions (so that X is large when Y is small, and vice versa), then the covariance is negative. Finally, if X and Y are independent, then the covariance is 0 (see Exercise 2.19).

Correlation. Because the covariance is the product of X and Y, deviated from their means, its units are, awkwardly, the units of X multiplied by the units of Y. This "units" problem can make numerical values of the covariance difficult to interpret.

The correlation is an alternative measure of dependence between X and Y that solves the "units" problem of the covariance. Specifically, the **correlation** between X and Y is the covariance between X and Y divided by their standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\ \text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.26}$$

Because the units of the numerator in Equation (2.26) are the same as those of the denominator, the units cancel, and the correlation is unit free. The random variables X and Y are said to be **uncorrelated** if $\text{corr}(X, Y) = 0$.

The correlation always is between –1 and 1; that is, as proven in Appendix 2.1,

$$-1 \le \text{corr}(X, Y) \le 1 \quad (\text{correlation inequality}). \tag{2.27}$$

Correlation and conditional mean. If the conditional mean of Y does not depend on X, then Y and X are uncorrelated. That is,

$$\text{if } E(Y|X) = \mu_Y, \text{ then } \text{cov}(Y, X) = 0 \text{ and } \text{corr}(Y, X) = 0. \tag{2.28}$$

We now show this result. First, suppose Y and X have mean 0, so that $\text{cov}(Y, X) = E[(Y - \mu_Y)(X - \mu_X)] = E(YX)$. By the law of iterated expectations [Equation (2.20)], $E(YX) = E[E(YX|X)] = E[E(Y|X)X] = 0$ because $E(Y|X) = 0$, so $\text{cov}(Y, X) = 0$. Equation (2.28) follows by substituting $\text{cov}(Y, X) = 0$ into the definition of correlation in Equation (2.26). If Y and X do not have mean 0, subtract off their means, and then the preceding proof applies.

It is *not* necessarily true, however, that if X and Y are uncorrelated, then the conditional mean of Y given X does not depend on X. Said differently, it is possible for the conditional mean of Y to be a function of X but for Y and X nonetheless to be uncorrelated. An example is given in Exercise 2.23.

The Mean and Variance of Sums of Random Variables

The mean of the sum of two random variables, X and Y, is the sum of their means:

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y. \tag{2.29}$$

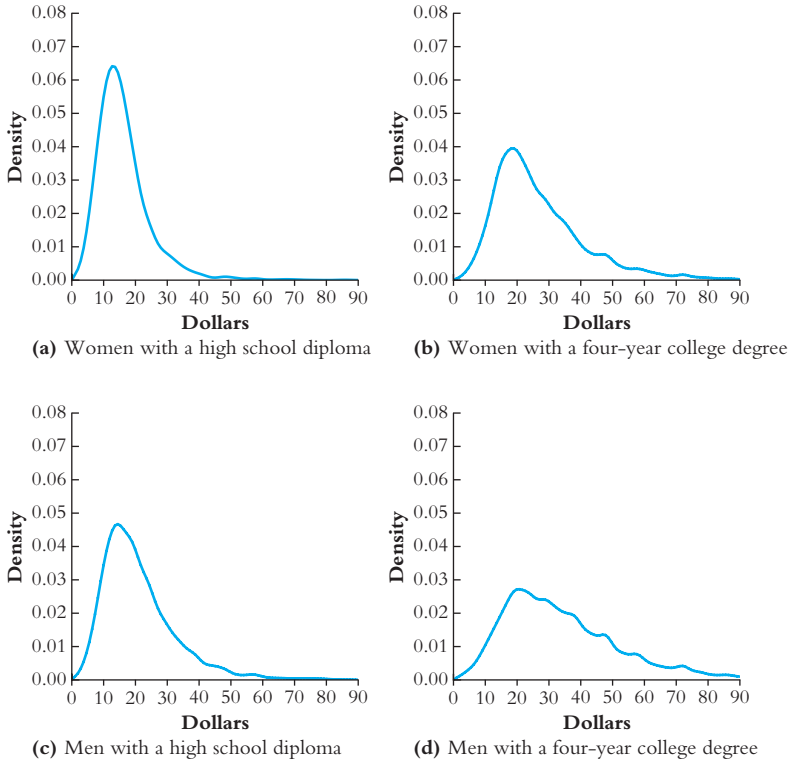
The Distribution of Earnings in the United States in 2015

Some parents tell their children that they will be able to get a better, higher-paying job if they get a college degree than if they skip higher education. Are these parents right? Does the distribution of earnings differ between workers who are college graduates and workers who have only a high school diploma and, if so, how? Among workers with a similar education, does the distribution of earnings for men and women differ? For example, do the best-paid college-educated women earn as much as the best-paid college-educated men?

One way to answer these questions is to examine the distribution of earnings of full-time workers, conditional on the highest educational degree achieved (high school diploma or bachelor's degree) and on sex. These four conditional distributions are shown in Figure 2.4, and the mean, standard deviation, and

FIGURE 2.4 Conditional Distributions of Average Hourly Earnings of U.S. Full-Time Workers in 2015, Given Education Level and Sex

The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



(a) Women with a high school diploma

(b) Women with a four-year college degree

(c) Men with a high school diploma

(d) Men with a four-year college degree

TABLE 2.4 Summary of the Conditional Distributions of Average Hourly Earnings of U.S. Full-Time Workers in 2015 Given Education Level and Sex

			Percentile			
	Mean	Standard Deviation	25%	50% (median)	75%	90%
(a) Women with a high school diploma	$16.28	$8.91	$10.99	$14.42	$19.23	$25.64
(b) Women with a four-year college degree	27.23	16.18	16.83	23.56	33.65	47.60
(c) Men with a high school diploma	21.22	11.96	13.22	19.12	26.10	36.06
(d) Men with a four-year college degree	35.10	20.36	20.67	30.92	44.71	60.90

Average hourly earnings are the sum of annual pre-tax wages, salaries, tips, and bonuses divided by the number of hours worked annually.

some percentiles of the conditional distributions are presented in Table 2.4.[1] For example, the conditional mean of earnings for women whose highest degree is a high school diploma—that is, $E(Earnings|Highest\ degree = high\ school\ diploma, Sex = female)$—is $16.28 per hour.

The distribution of average hourly earnings for female college graduates (Figure 2.4b) is shifted to the right of the distribution for women with only a high school diploma (Figure 2.4a); the same shift can be seen for the two groups of men (Figure 2.4d and Figure 2.4c). For both men and women, mean earnings are higher for those with a college degree (Table 2.4, first numeric column). Interestingly, the spread of the distribution of earnings, as measured by the standard deviation, is greater for those with a college degree than for those with a high school diploma. In addition, for both men and women, the 90th percentile of earnings is much higher for workers with a college degree than for workers with only a high school diploma. This final comparison is consistent with the parental admonition that a college degree opens doors that remain closed to individuals with only a high school diploma.

Another feature of these distributions is that the distribution of earnings for men is shifted to the right of the distribution of earnings for women for a given level of education. This "gender gap" in earnings is an important—and, to many, troubling—aspect of the distribution of earnings. We return to this topic in later chapters.

[1]The distributions were estimated using data from the March 2016 Current Population Survey, which is discussed in more detail in Appendix 3.1.

KEY CONCEPT 2.3

Means, Variances, and Covariances of Sums of Random Variables

Let X, Y, and V be random variables; let μ_X and σ_X^2 be the mean and variance of X and let σ_{XY} be the covariance between X and Y (and so forth for the other variables); and let a, b, and c be constants. Equations (2.30) through (2.36) follow from the definitions of the mean, variance, and covariance:

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y, \tag{2.30}$$

$$\text{var}(a + bY) = b^2\sigma_Y^2, \tag{2.31}$$

$$\text{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \tag{2.32}$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \tag{2.33}$$

$$\text{cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \tag{2.34}$$

$$E(XY) = \sigma_{XY} + \mu_X\mu_Y, \tag{2.35}$$

$$|\text{corr}(X, Y)| \leq 1 \text{ and } |\sigma_{XY}| \leq \sqrt{\sigma_X^2\sigma_Y^2} \text{ (correlation inequality)}. \tag{2.36}$$

The variance of the sum of X and Y is the sum of their variances plus two times their covariance:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \tag{2.37}$$

If X and Y are independent, then the covariance is 0, and the variance of their sum is the sum of their variances:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) = \sigma_X^2 + \sigma_Y^2 \quad (\text{if } X \text{ and } Y \text{ are independent}). \tag{2.38}$$

Useful expressions for means, variances, and covariances involving weighted sums of random variables are collected in Key Concept 2.3. The results in Key Concept 2.3 are derived in Appendix 2.1.

2.4 The Normal, Chi-Squared, Student *t*, and *F* Distributions

The probability distributions most often encountered in econometrics are the normal, chi-squared, Student t, and F distributions.

The Normal Distribution

A continuous random variable with a **normal distribution** has the familiar bell-shaped probability density shown in Figure 2.5. The function defining the normal probability density is given in Appendix 18.1. As Figure 2.5 shows, the normal density with mean μ and variance σ^2 is symmetric around its mean and has 95% of its probability between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$.

Some special notation and terminology have been developed for the normal distribution. The normal distribution with mean μ and variance σ^2 is expressed concisely as $N(\mu, \sigma^2)$. The **standard normal distribution** is the normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ and is denoted $N(0, 1)$. Random variables that have a $N(0, 1)$ distribution are often denoted Z, and the standard normal cumulative distribution function is denoted by the Greek letter Φ; accordingly, $\Pr(Z \leq c) = \Phi(c)$, where c is a constant. Values of the standard normal cumulative distribution function are tabulated in Appendix Table 1.

To look up probabilities for a normal variable with a general mean and variance, we must first standardize the variable. For example, suppose Y is distributed $N(1, 4)$—that is, Y is normally distributed with a mean of 1 and a variance of 4. What is the probability that $Y \leq 2$—that is, what is the shaded area in Figure 2.6a? The standardized version of Y is Y minus its mean, divided by its standard deviation; that is, $(Y - 1)/\sqrt{4} = \frac{1}{2}(Y - 1)$. Accordingly, the random variable $\frac{1}{2}(Y - 1)$ is normally distributed with mean 0 and variance 1 (see Exercise 2.8); it has the standard normal

FIGURE 2.5 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.

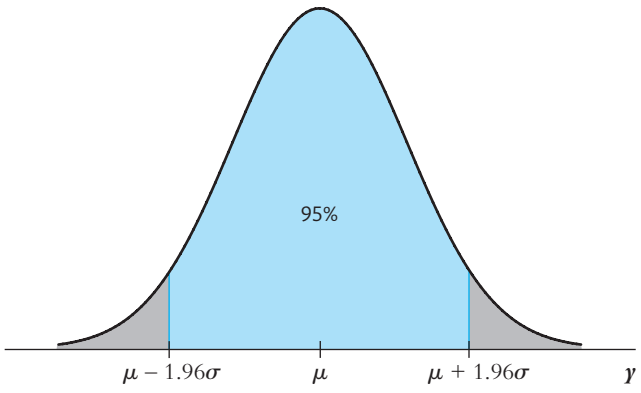
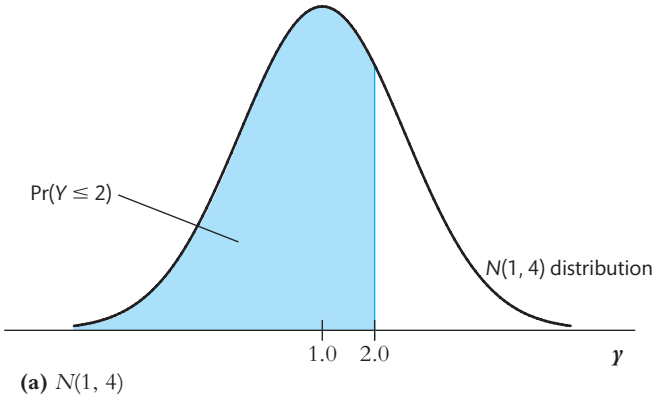
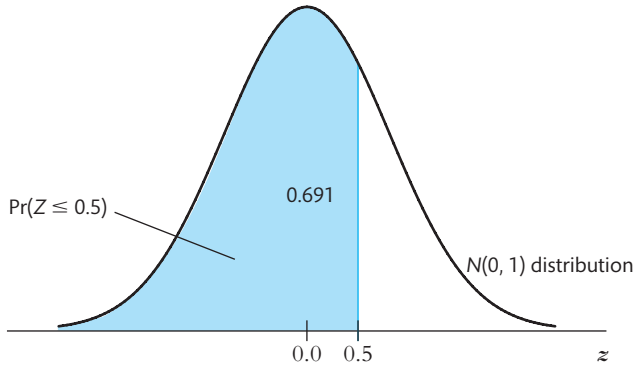


FIGURE 2.6 Calculating the Probability That $Y \leq 2$ When Y Is Distributed $N(1, 4)$

To calculate $\Pr(Y \leq 2)$, standardize Y, then use the standard normal distribution table. Y is standardized by subtracting its mean ($\mu = 1$) and dividing by its standard deviation ($\sigma = 2$). The probability that $Y \leq 2$ is shown in Figure 2.6a, and the corresponding probability after standardizing Y is shown in Figure 2.6b. Because the standardized random variable, $(Y - 1)/2$, is a standard normal (Z) random variable, $\Pr(Y \leq 2) = \Pr(\frac{Y-1}{2} \leq \frac{2-1}{2}) = \Pr(Z \leq 0.5)$. From Appendix Table 1, $\Pr(Z \leq 0.5) = \Phi(0.5) = 0.691$.



(a) $N(1, 4)$



(b) $N(0, 1)$

KEY CONCEPT 2.4

Computing Probabilities and Involving Normal Random Variables

Suppose Y is normally distributed with mean μ and variance σ^2; in other words, Y is distributed $N(\mu, \sigma^2)$. Then Y is standardized by subtracting its mean and dividing by its standard deviation, that is, by computing $Z = (Y - \mu)/\sigma$.

Let c_1 and c_2 denote two numbers with $c_1 < c_2$, and let $d_1 = (c_1 - \mu)/\sigma$ and $d_2 = (c_2 - \mu)/\sigma$. Then

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2), \tag{2.39}$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1), \tag{2.40}$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \tag{2.41}$$

The normal cumulative distribution function Φ is tabulated in Appendix Table 1.

distribution shown in Figure 2.6b. Now $Y \leq 2$ is equivalent to $\frac{1}{2}(Y - 1) \leq \frac{1}{2}(2 - 1)$; that is, $\frac{1}{2}(Y - 1) \leq \frac{1}{2}$. Thus

$$\Pr(Y \leq 2) = \Pr\left[\tfrac{1}{2}(Y - 1) \leq \tfrac{1}{2}\right] = \Pr(Z \leq \tfrac{1}{2}) = \Phi(0.5) = 0.691, \quad (2.42)$$

where the value 0.691 is taken from Appendix Table 1.

The same approach can be used to compute the probability that a normally distributed random variable exceeds some value or that it falls in a certain range. These steps are summarized in Key Concept 2.4. The box "A Bad Day on Wall Street" presents an unusual application of the cumulative normal distribution.

The normal distribution is symmetric, so its skewness is 0. The kurtosis of the normal distribution is 3.

The multivariate normal distribution. The normal distribution can be generalized to describe the joint distribution of a set of random variables. In this case, the distribution is called the **multivariate normal distribution** or, if only two variables are being considered, the **bivariate normal distribution**. The formula for the bivariate normal p.d.f. is given in Appendix 18.1, and the formula for the general multivariate normal p.d.f. is given in Appendix 19.2.

The multivariate normal distribution has four important properties. If X and Y have a bivariate normal distribution with covariance σ_{XY} and if a and b are two constants, then $aX + bY$ has the normal distribution:

$$aX + bY \text{ is distributed } N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}) \quad (X, Y \text{ bivariate normal}). \quad (2.43)$$

A Bad Day on Wall Street

On a typical day, the overall value of stocks traded on the U.S. stock market can rise or fall by 1% or even more. This is a lot—but nothing compared to what happened on Monday, October 19, 1987. On "Black Monday," the Dow Jones Industrial Average (an average of 30 large industrial stocks) fell by 22.6%! From January 1, 1980, to September 29, 2017, the standard deviation of daily percentage price changes on the Dow was 1.08%, so the drop of 22.6% was a negative return of 21 ($= 22.6/1.08$) standard deviations. The enormity of this drop can be seen in Figure 2.7, a plot of the daily returns on the Dow during the 1980s.

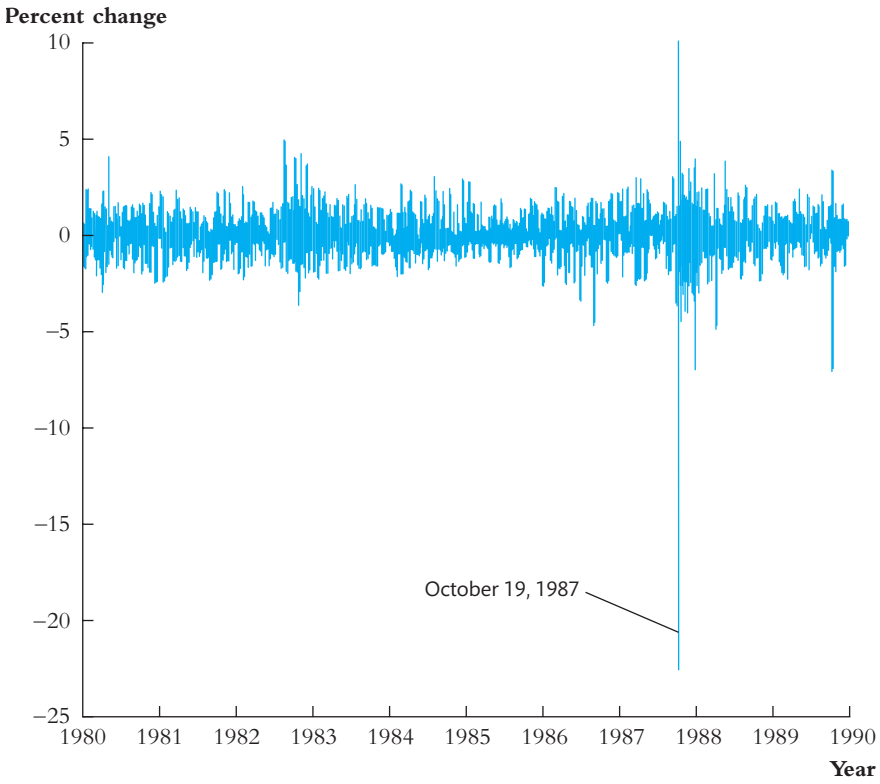
If daily percentage price changes are normally distributed, then the probability of a change of at least 21 standard deviations is $\Pr(|Z| \geq 21) = 2 \times \Phi(-21)$. You will not find this value in Appendix Table 1, but you can calculate it using a computer (try it!). This probability is 6.6×10^{-98}—that is, 0.000 . . . 00066, where there are a total of 97 zeros!

How small is 6.6×10^{-98}? Consider the following:

- The world population is about 7.6 billion, so the probability of winning a random lottery among all living people is about 1 in 7.6 billion, or 1.3×10^{-10}.

continued on next page

FIGURE 2.7 Daily Percentage Changes in the Dow Jones Industrial Average in the 1980s



From January 1980 through September 2017, the average percentage daily change of "the Dow" index was 0.04% and its standard deviation was 1.08%. On October 19, 1987—"Black Monday"—the Dow fell 22.6%, or 21 standard deviations.

- The universe is believed to have existed for 14 billion years, or about 5×10^{17} seconds, so the probability of choosing a particular second at random from all the seconds since the beginning of time is 2×10^{-18}.
- There are approximately 10^{43} molecules of gas in the first kilometer above the earth's surface. The probability of choosing a particular molecule at random is 10^{-43}.

Although Wall Street *did* have a bad day, the fact that it happened at all suggests its probability was more than 6.6×10^{-98}. In fact, there have been many days—good and bad—with stock price changes too large to be consistent with a normal distribution with a constant variance. Table 2.5 lists the ten largest daily percentage price changes in the Dow Jones Industrial Average in the 9521 trading days between January 1, 1980 and September 29, 2017, along with the standardized change using the mean and variance over this period. All ten changes exceed 6.6 standard deviations, an extremely rare event if stock prices are normally distributed.

Clearly, stock price percentage changes have a distribution with heavier tails than the normal distribution. For this reason, finance professionals use other models of stock price changes. One such model treats stock price changes as normally distributed with a variance that evolves over time, so periods

TABLE 2.5 The Ten Largest Daily Percentage Changes in the Dow Jones Industrial Average, January 1980–September 2017, and the Normal Probability of a Change at Least as Large

Date	Percentage Change (x)	Standardized Change $z = (x - \mu)/\sigma$	Normal Probability of a Change at Least This Large $\Pr(\lvert Z\rvert \geq \lvert z\rvert) = 2\Phi(-\lvert z\rvert)$
October 19, 1987	–22.6	–21.0	6.6×10^{-98}
October 13, 2008	11.1	10.2	1.5×10^{-24}
October 28, 2008	10.9	10.0	1.0×10^{-23}
October 21, 1987	10.1	9.4	7.7×10^{-21}
October 26, 1987	–8.0	–7.5	7.2×10^{-14}
October 15, 2008	–7.9	–7.3	2.3×10^{-13}
December 01, 2008	–7.7	–7.2	7.4×10^{-13}
October 09, 2008	–7.3	–6.8	8.5×10^{-12}
October 27, 1997	–7.2	–6.7	2.2×10^{-11}
September 17, 2001	–7.1	–6.6	3.1×10^{-11}

like October 1987 and the financial crisis in the fall of 2008 have higher volatility than others (models with time-varying variances are discussed in Chapter 17). Other models abandon the normal distribution in favor of distributions with heavier tails, an idea popularized in Nassim Taleb's 2007 book, *The Black Swan*. These models are more consistent with the very bad—and very good—days we actually see on Wall Street.

More generally, if n random variables have a multivariate normal distribution, then any linear combination of these variables (such as their sum) is normally distributed.

Second, if a set of variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal [this follows from Equation (2.43) by setting $a = 1$ and $b = 0$].

Third, if variables with a multivariate normal distribution have covariances that equal 0, then the variables are independent. Thus, if X and Y have a bivariate normal distribution and $\sigma_{XY} = 0$, then X and Y are independent (this is shown in Appendix 18.1). In Section 2.3, it was shown that if X and Y are independent, then, regardless of their joint distribution, $\sigma_{XY} = 0$. If X and Y are jointly normally distributed, then the converse is also true. This result—that 0 covariance implies independence—is a special property of the multivariate normal distribution that is not true in general.

Fourth, if X and Y have a bivariate normal distribution, then the conditional expectation of Y given X is linear in X; that is, $E(Y|X = x) = a + bx$, where a and b are constants (Exercise 18.11). Joint normality implies linearity of conditional expectations, but linearity of conditional expectations does not imply joint normality.

The Chi-Squared Distribution

The chi-squared distribution is used when testing certain types of hypotheses in statistics and econometrics.

The **chi-squared distribution** is the distribution of the sum of m squared independent standard normal random variables. This distribution depends on m, which is called the degrees of freedom of the chi-squared distribution. For example, let Z_1, Z_2, and Z_3 be independent standard normal random variables. Then $Z_1^2 + Z_2^2 + Z_3^2$ has a chi-squared distribution with 3 degrees of freedom. The name for this distribution derives from the Greek letter used to denote it: A chi-squared distribution with m degrees of freedom is denoted χ_m^2.

Selected percentiles of the χ_m^2 distribution are given in Appendix Table 3. For example, Appendix Table 3 shows that the 95th percentile of the χ_3^2 distribution is 7.81, so $\Pr(Z_1^2 + Z_2^2 + Z_3^3 \leq 7.81) = 0.95$.

The Student *t* Distribution

The **Student *t* distribution** with m degrees of freedom is defined to be the distribution of the ratio of a standard normal random variable to the square root of an independently distributed chi-squared random variable with m degrees of freedom divided by m. That is, let Z be a standard normal random variable, let W be a random variable with a chi-squared distribution with m degrees of freedom, and let Z and W be independently distributed. Then the random variable $Z/\sqrt{W/m}$ has a Student t distribution (also called the ***t* distribution**) with m degrees of freedom. This distribution is denoted t_m. Selected percentiles of the Student t distribution are given in Appendix Table 2.

The Student t distribution depends on the degrees of freedom m. Thus the 95th percentile of the t_m distribution depends on the degrees of freedom m. The Student t distribution has a bell shape similar to that of the normal distribution, but it has more mass in the tails; that is, it is a "fatter" bell shape than the normal. When m is 30 or more, the Student t distribution is well approximated by the standard normal distribution, and the t_∞ distribution equals the standard normal distribution.

The *F* Distribution

The ***F* distribution** with m and n degrees of freedom, denoted $F_{m,n}$, is defined to be the distribution of the ratio of a chi-squared random variable with degrees of freedom m, divided by m, to an independently distributed chi-squared random variable with degrees of freedom n, divided by n. To state this mathematically, let W be a chi-squared random variable with m degrees of freedom and let V be a chi-squared random variable with n degrees of freedom, where W and V are independently distributed. Then $\frac{W/m}{V/n}$ has an $F_{m,n}$ distribution—that is, an F distribution with numerator degrees of freedom m and denominator degrees of freedom n.

In statistics and econometrics, an important special case of the F distribution arises when the denominator degrees of freedom is large enough that the $F_{m,n}$

distribution can be approximated by the $F_{m,\infty}$ distribution. In this limiting case, the denominator random variable V/n is the mean of infinitely many squared standard normal random variables, and that mean is 1 because the mean of a squared standard normal random variable is 1 (see Exercise 2.24). Thus the $F_{m,\infty}$ distribution is the distribution of a chi-squared random variable with m degrees of freedom divided by m: W/m is distributed $F_{m,\infty}$. For example, from Appendix Table 4, the 95th percentile of the $F_{3,\infty}$ distribution is 2.60, which is the same as the 95th percentile of the χ^2_3 distribution, 7.81 (from Appendix Table 2), divided by the degrees of freedom, which is $3(7.81/3 = 2.60)$.

The 90th, 95th, and 99th percentiles of the $F_{m,n}$ distribution are given in Appendix Table 5 for selected values of m and n. For example, the 95th percentile of the $F_{3,30}$ distribution is 2.92, and the 95th percentile of the $F_{3,90}$ distribution is 2.71. As the denominator degrees of freedom n increases, the 95th percentile of the $F_{3,n}$ distribution tends to the $F_{3,\infty}$ limit of 2.60.

2.5 Random Sampling and the Distribution of the Sample Average

Almost all the statistical and econometric procedures used in this text involve averages or weighted averages of a sample of data. Characterizing the distributions of sample averages therefore is an essential step toward understanding the performance of econometric procedures.

This section introduces some basic concepts about random sampling and the distributions of averages that are used throughout the book. We begin by discussing random sampling. The act of random sampling—that is, randomly drawing a sample from a larger population—has the effect of making the sample average itself a random variable. Because the sample average is a random variable, it has a probability distribution, which is called its sampling distribution. This section concludes with some properties of the sampling distribution of the sample average.

Random Sampling

Simple random sampling. Suppose our commuting student from Section 2.1 aspires to be a statistician and decides to record her commuting times on various days. She selects these days at random from the school year, and her daily commuting time has the cumulative distribution function in Figure 2.2a. Because these days were selected at random, knowing the value of the commuting time on one of these randomly selected days provides no information about the commuting time on another of the days; that is, because the days were selected at random, the values of the commuting time on the different days are independently distributed random variables.

The situation described in the previous paragraph is an example of the simplest sampling scheme used in statistics, called **simple random sampling**, in which n objects are

KEY CONCEPT 2.5

Simple Random Sampling and i.i.d. Random Variables

In a simple random sample, n objects are drawn at random from a population, and each object is equally likely to be drawn. The value of the random variable Y for the i^{th} randomly drawn object is denoted Y_i. Because each object is equally likely to be drawn and the distribution of Y_i is the same for all i, the random variables $Y_1, \ldots, Y_n$ are independently and identically distributed (i.i.d.); that is, the distribution of Y_i is the same for all $i = 1, \ldots, n$, and Y_1 is distributed independently of $Y_2, \ldots, Y_n$ and so forth.

selected at random from a **population** (the population of commuting days) and each member of the population (each day) is equally likely to be included in the sample.

The n observations in the sample are denoted $Y_1, \ldots, Y_n$, where Y_1 is the first observation, Y_2 is the second observation, and so forth. In the commuting example, Y_1 is the commuting time on the first of the n randomly selected days, and Y_i is the commuting time on the i^{th} of the randomly selected days.

Because the members of the population included in the sample are selected at random, the values of the observations $Y_1, \ldots, Y_n$ are themselves random. If different members of the population are chosen, their values of Y will differ. Thus the act of random sampling means that $Y_1, \ldots, Y_n$ can be treated as random variables. Before they are sampled, $Y_1, \ldots, Y_n$ can take on many possible values; after they are sampled, a specific value is recorded for each observation.

i.i.d. draws. Because $Y_1, \ldots, Y_n$ are randomly drawn from the same population, the marginal distribution of Y_i is the same for each $i = 1, \ldots, n$; this marginal distribution is the distribution of Y in the population being sampled. When Y_i has the same marginal distribution for $i = 1, \ldots, n$, then $Y_1, \ldots, Y_n$ are said to be **identically distributed**.

Under simple random sampling, knowing the value of Y_1 provides no information about Y_2, so the conditional distribution of Y_2 given Y_1 is the same as the marginal distribution of Y_2. In other words, under simple random sampling, Y_1 is distributed independently of $Y_2, \ldots, Y_n$.

When $Y_1, \ldots, Y_n$ are drawn from the same distribution and are independently distributed, they are said to be **independently and identically distributed (i.i.d.)**.

Simple random sampling and i.i.d. draws are summarized in Key Concept 2.5.

The Sampling Distribution of the Sample Average

The **sample average** or **sample mean**, $\overline{Y}$, of the n observations $Y_1, \ldots, Y_n$ is

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.44}$$

An essential concept is that the act of drawing a random sample has the effect of making the sample average $\overline{Y}$ a random variable. Because the sample was drawn at random, the value of each Y_i is random. Because $Y_1, \ldots, Y_n$ are random, their average is random. Had a different sample been drawn, then the observations and their sample average would have been different: The value of $\overline{Y}$ differs from one randomly drawn sample to the next.

For example, suppose our student commuter selected five days at random to record her commute times, then computed the average of those five times. Had she chosen five different days, she would have recorded five different times—and thus would have computed a different value of the sample average.

Because $\overline{Y}$ is random, it has a probability distribution. The distribution of $\overline{Y}$ is called the **sampling distribution** of $\overline{Y}$ because it is the probability distribution associated with possible values of $\overline{Y}$ that could be computed for different possible samples $Y_1, \ldots, Y_n$.

The sampling distribution of averages and weighted averages plays a central role in statistics and econometrics. We start our discussion of the sampling distribution of $\overline{Y}$ by computing its mean and variance under general conditions on the population distribution of Y.

Mean and variance of $\overline{Y}$. Suppose that the observations $Y_1, \ldots, Y_n$ are i.i.d., and let μ_Y and σ_Y^2 denote the mean and variance of Y_i (because the observations are i.i.d., the mean is the same for all $i = 1, \ldots, n$, and so is the variance). When $n = 2$, the mean of the sum $Y_1 + Y_2$ is given by applying Equation (2.29): $E(Y_1 + Y_2) = \mu_Y + \mu_Y = 2\mu_Y$. Thus the mean of the sample average is $E[\frac{1}{2}(Y_1 + Y_2)] = \frac{1}{2} \times 2\mu_Y = \mu_Y$. In general,

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y. \tag{2.45}$$

The variance of $\overline{Y}$ is found by applying Equation (2.38). For example, for $n = 2$, $\text{var}(Y_1 + Y_2) = 2\sigma_Y^2$, so [by applying Equation (2.32) with $a = b = \frac{1}{2}$ and $\text{cov}(Y_1, Y_2) = 0$], $\text{var}(\overline{Y}) = \frac{1}{2}\sigma_Y^2$. For general n, because $Y_1, \ldots, Y_n$ are i.i.d., Y_i and Y_j are independently distributed for $i \neq j$, so $\text{cov}(Y_i, Y_j) = 0$. Thus

$$\begin{aligned} \text{var}(\overline{Y}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \text{cov}(Y_i, Y_j) \\ &= \frac{\sigma_Y^2}{n}. \end{aligned} \tag{2.46}$$

The standard deviation of $\overline{Y}$ is the square root of the variance, $\sigma_Y/\sqrt{n}$.

Financial Diversification and Portfolios

The principle of diversification says that you can reduce your risk by holding small investments in multiple assets, compared to putting all your money into one asset. That is, you shouldn't put all your eggs in one basket.

The math of diversification follows from Equation (2.46). Suppose you divide \$1 equally among n assets. Let Y_i represent the payout in one year of \$1 invested in the i^{th} asset. Because you invested $1/n$ dollars in each asset, the actual payoff of your portfolio after one year is $(Y_1 + Y_2 + \cdots + Y_n)/n = \overline{Y}$. To keep things simple, suppose that each asset has the same expected payout, μ_Y, the same variance, σ^2, and the same positive correlation, ρ, across assets [so that $\text{cov}(Y_i, Y_j) = \rho\sigma^2$]. Then the expected payout is $E(\overline{Y}) = \mu_Y$, and for large n, the variance of the portfolio payout is $\text{var}(\overline{Y}) = \rho\sigma^2$ (Exercise 2.26). Putting all your money into one asset or spreading it equally across all n assets has the same expected payout, but diversifying reduces the variance from σ^2 to $\rho\sigma^2$.

The math of diversification has led to financial products such as stock mutual funds, in which the fund holds many stocks and an individual owns a share of the fund, thereby owning a small amount of many stocks. But diversification has its limits: For many assets, payouts are positively correlated, so $\text{var}(\overline{Y})$ remains positive even if n is large. In the case of stocks, risk is reduced by holding a portfolio, but that portfolio remains subject to the unpredictable fluctuations of the overall stock market.

In summary, if $Y_1, \ldots, Y_n$ are i.i.d., the mean, the variance, and the standard deviation of $\overline{Y}$ are

$$E(\overline{Y}) = \mu_Y, \tag{2.47}$$

$$\text{var}(\overline{Y}) = \sigma_{\overline{Y}}^2 = \frac{\sigma_Y^2}{n}, \text{ and} \tag{2.48}$$

$$\text{std.dev}(\overline{Y}) = \sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}}. \tag{2.49}$$

These results hold whatever the distribution of Y is; that is, the distribution of Y does not need to take on a specific form, such as the normal distribution, for Equations (2.47) through (2.49) to hold.

The notation $\sigma_{\overline{Y}}^2$ denotes the variance of the sampling distribution of the sample average $\overline{Y}$. In contrast, σ_Y^2 is the variance of each individual Y_i, that is, the variance of the population distribution from which the observation is drawn. Similarly, $\sigma_{\overline{Y}}$ denotes the standard deviation of the sampling distribution of $\overline{Y}$.

Sampling distribution of $\overline{Y}$ when Y is normally distributed. Suppose that $Y_1, \ldots, Y_n$ are i.i.d. draws from the $N(\mu_Y, \sigma_Y^2)$ distribution. As stated following Equation (2.43), the sum of n normally distributed random variables is itself normally distributed. Because the mean of $\overline{Y}$ is μ_Y and the variance of $\overline{Y}$ is σ_Y^2/n, this means that, if $Y_1, \ldots, Y_n$ are i.i.d. draws from the $N(\mu_Y, \sigma_Y^2)$ distribution, then $\overline{Y}$ is distributed $N(\mu_Y, \sigma_Y^2/n)$.

2.6 Large-Sample Approximations to Sampling Distributions

Sampling distributions play a central role in the development of statistical and econometric procedures, so it is important to know, in a mathematical sense, what the sampling distribution of $\overline{Y}$ is. There are two approaches to characterizing sampling distributions: an "exact" approach and an "approximate" approach.

The exact approach entails deriving a formula for the sampling distribution that holds exactly for any value of n. The sampling distribution that exactly describes the distribution of $\overline{Y}$ for any n is called the **exact distribution** or **finite-sample distribution** of $\overline{Y}$. For example, if Y is normally distributed and $Y_1, \ldots, Y_n$ are i.i.d., then (as discussed in Section 2.5) the exact distribution of $\overline{Y}$ is normal with mean μ_Y and variance σ_Y^2/n. Unfortunately, if the distribution of Y is not normal, then in general the exact sampling distribution of $\overline{Y}$ is very complicated and depends on the distribution of Y.

The approximate approach uses approximations to the sampling distribution that rely on the sample size being large. The large-sample approximation to the sampling distribution is often called the **asymptotic distribution**—"asymptotic" because the approximations become exact in the limit that $n \to \infty$. As we see in this section, these approximations can be very accurate even if the sample size is only $n = 30$ observations. Because sample sizes used in practice in econometrics typically number in the hundreds or thousands, these asymptotic distributions can be counted on to provide very good approximations to the exact sampling distribution.

This section presents the two key tools used to approximate sampling distributions when the sample size is large: the law of large numbers and the central limit theorem. The law of large numbers says that when the sample size is large, $\overline{Y}$ will be close to μ_Y with very high probability. The central limit theorem says that when the sample size is large, the sampling distribution of the standardized sample average, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, is approximately normal.

Although exact sampling distributions are complicated and depend on the distribution of Y, the asymptotic distributions are simple. Moreover—remarkably—the asymptotic normal distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ does *not* depend on the distribution of Y. This normal approximate distribution provides enormous simplifications and underlies the theory of regression used throughout this text.

The Law of Large Numbers and Consistency

The **law of large numbers** states that, under general conditions, $\overline{Y}$ will be near μ_Y with very high probability when n is large. This is sometimes called the "law of averages." When a large number of random variables with the same mean are averaged together, the large values tend to balance the small values, and their sample average is close to their common mean.

For example, consider a simplified version of our student commuter's experiment in which she simply records whether her commute was short (less than

KEY CONCEPT 2.6

Convergence in Probability, Consistency, and the Law of Large Numbers

The sample average $\overline{Y}$ converges in probability to μ_Y (or, equivalently, $\overline{Y}$ is consistent for μ_Y) if the probability that $\overline{Y}$ is in the range $(\mu_Y - c)$ to $(\mu_Y + c)$ becomes arbitrarily close to 1 as n increases for any constant $c > 0$. The convergence of $\overline{Y}$ to μ_Y in probability is written $\overline{Y} \xrightarrow{p} \mu_Y$.

The law of large numbers says that if $Y_1, \ldots, Y_n$ are independently and identically distributed with $E(Y_i) = \mu_Y$ and if large outliers are unlikely (technically if $\text{var}(Y_i) = \sigma_Y^2 < \infty$), then $\overline{Y} \xrightarrow{p} \mu_Y$.

20 minutes) or long. Let $Y_i = 1$ if her commute was short on the i^{th} randomly selected day and $Y_i = 0$ if it was long. Because she used simple random sampling, $Y_1, \ldots, Y_n$ are i.i.d. Thus $Y_1, \ldots, Y_n$ are i.i.d. draws of a Bernoulli random variable, where (from Table 2.2) the probability that $Y_i = 1$ is 0.78. Because the expectation of a Bernoulli random variable is its success probability, $E(Y_i) = \mu_Y = 0.78$. The sample average $\overline{Y}$ is the fraction of days in her sample in which her commute was short.

Figure 2.8 shows the sampling distribution of $\overline{Y}$ for various sample sizes n. When $n = 2$ (Figure 2.8a), $\overline{Y}$ can take on only three values: 0, $\frac{1}{2}$, and 1 (neither commute was short, one was short, and both were short), none of which is particularly close to the true proportion in the population, 0.78. As n increases, however (Figures 2.8b–d), $\overline{Y}$ takes on more values, and the sampling distribution becomes tightly centered on μ_Y.

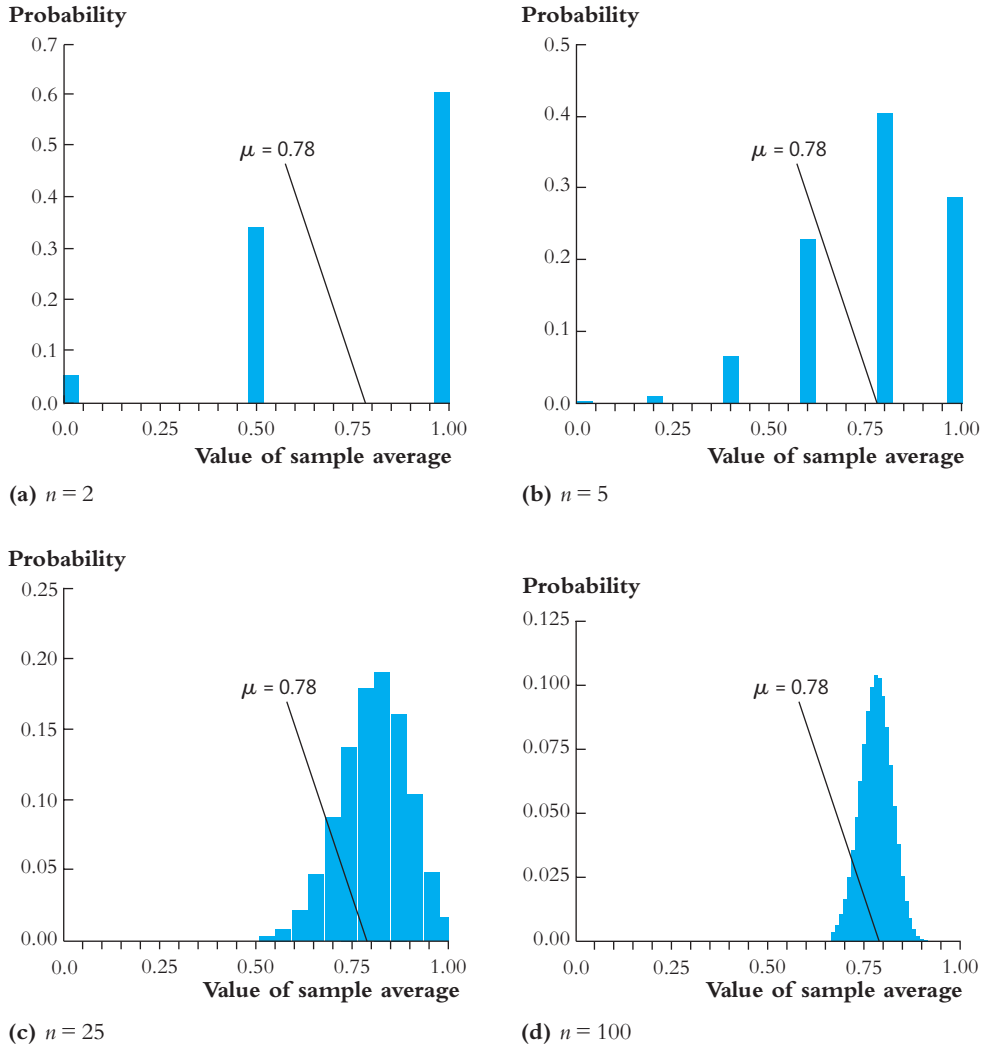
The property that $\overline{Y}$ is near μ_Y with probability increasing to 1 as n increases is called **convergence in probability** or, more concisely, **consistency** (see Key Concept 2.6). The law of large numbers states that under certain conditions $\overline{Y}$ converges in probability to μ_Y or, equivalently, that $\overline{Y}$ is consistent for μ_Y.

The conditions for the law of large numbers that we will use in this text are that $Y_1, \ldots, Y_n$ are i.i.d. and that the variance of Y_i, σ_Y^2, is finite. The mathematical role of these conditions is made clear in Section 18.2, where the law of large numbers is proven. If the data are collected by simple random sampling, then the i.i.d. assumption holds. The assumption that the variance is finite says that extremely large values of Y_i—that is, outliers—are unlikely and are observed infrequently; otherwise, these large values could dominate $\overline{Y}$, and the sample average would be unreliable. This assumption is plausible for the applications in this text. For example, because there is an upper limit to our student's commuting time (she could park and walk if the traffic is dreadful), the variance of the distribution of commuting times is finite.

The Central Limit Theorem

The **central limit theorem** says that, under general conditions, the distribution of $\overline{Y}$ is well approximated by a normal distribution when n is large. Recall that the mean of $\overline{Y}$ is μ_Y and its variance is $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. According to the central limit theorem, when

FIGURE 2.8 Sampling Distribution of the Sample Average of n Bernoulli Random Variables

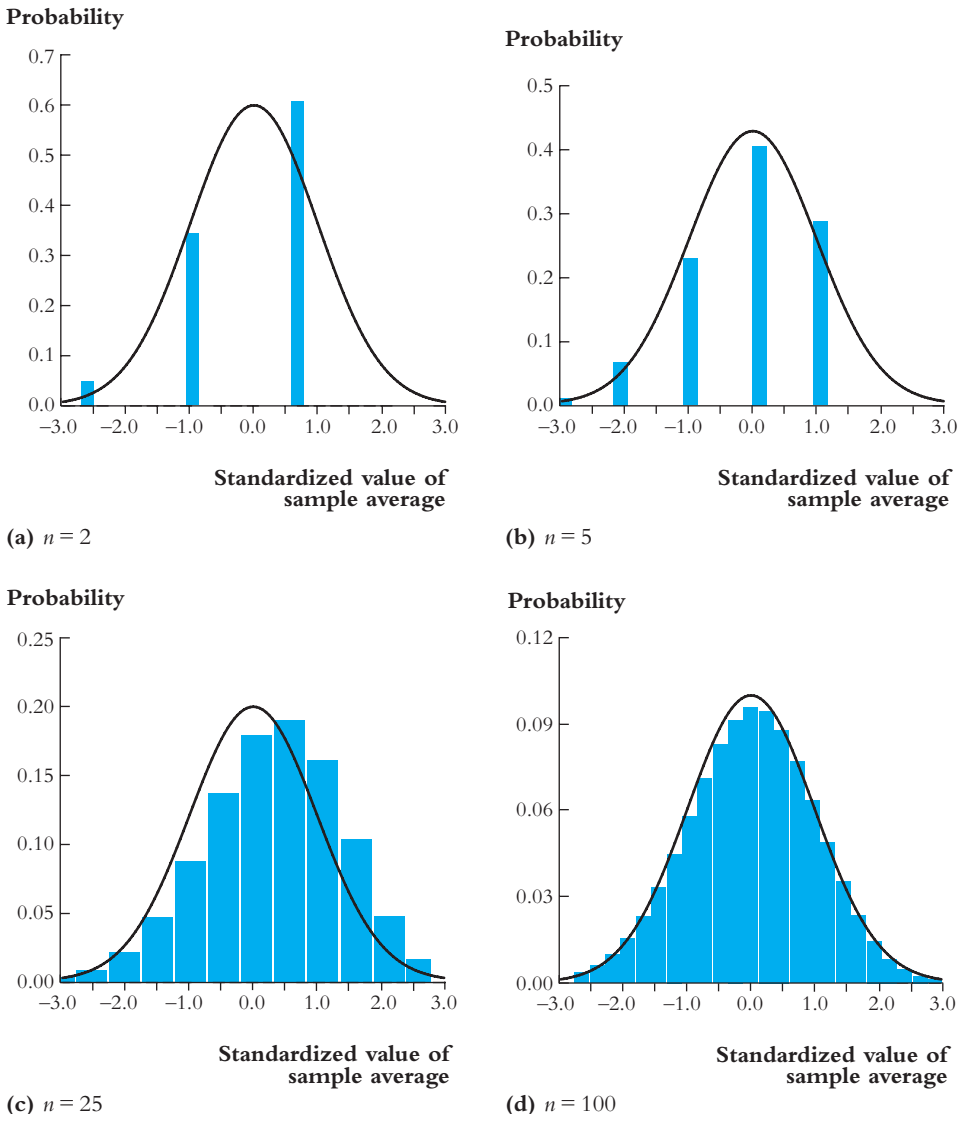


The distributions are the sampling distributions of $\overline{Y}$, the sample average of n independent Bernoulli random variables with $p = \Pr(Y_i = 1) = 0.78$ (the probability of a short commute is 78%). The variance of the sampling distribution of $\overline{Y}$ decreases as n gets larger, so the sampling distribution becomes more tightly concentrated around its mean, $\mu = 0.78$, as the sample size n increases.

n is large, the distribution of $\overline{Y}$ is approximately $N(\mu_Y, \sigma^2_{\overline{Y}})$. As discussed at the end of Section 2.5, the distribution of $\overline{Y}$ is *exactly* $N(\mu_Y, \sigma^2_{\overline{Y}})$ when the sample is drawn from a population with the normal distribution $N(\mu_Y, \sigma^2_{\overline{Y}})$. The central limit theorem says that this same result is *approximately* true when n is large even if $Y_1, \ldots, Y_n$ are not themselves normally distributed.

The convergence of the distribution of $\overline{Y}$ to the bell-shaped, normal approximation can be seen (a bit) in Figure 2.8. However, because the distribution gets quite tight for large n, this requires some squinting. It would be easier to see the shape of

FIGURE 2.9 Distribution of the Standardized Sample Average of n Bernoulli Random Variables with $p = 0.78$



The sampling distributions of $\overline{Y}$ in Figure 2.8 are plotted here after standardizing $\overline{Y}$. Standardization centers the distributions in Figure 2.8 and magnifies the scale on the horizontal axis by a factor of $\sqrt{n}$. When the sample size is large, the sampling distributions are increasingly well approximated by the normal distribution (the solid line), as predicted by the central limit theorem. The normal distribution is scaled so that the height of the distribution is approximately the same in all figures.

the distribution of $\overline{Y}$ if you used a magnifying glass or had some other way to zoom in or to expand the horizontal axis of the figure.

One way to do this is to standardize $\overline{Y}$ so that it has a mean of 0 and a variance of 1. This process leads to examining the distribution of the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$. According to the central limit theorem, this distribution should be well approximated by a $N(0, 1)$ distribution when n is large.

The distribution of the standardized average $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ is plotted in Figure 2.9 for the distributions in Figure 2.8; the distributions in Figure 2.9 are exactly the same as in Figure 2.8, except that the scale of the horizontal axis is changed so that the standardized variable has a mean of 0 and a variance of 1. After this change of scale, it is easy to see that, if n is large enough, the distribution of $\overline{Y}$ is well approximated by a normal distribution.

One might ask, how large is "large enough"? That is, how large must n be for the distribution of $\overline{Y}$ to be approximately normal? The answer is, "It depends." The quality of the normal approximation depends on the distribution of the underlying Y_i that make up the average. At one extreme, if the Y_i are themselves normally distributed, then $\overline{Y}$ is exactly normally distributed for all n. In contrast, when the underlying Y_i themselves have a distribution that is far from normal, then this approximation can require $n = 30$ or even more.

This point is illustrated in Figure 2.10 for a population distribution, shown in Figure 2.10a, that is quite different from the Bernoulli distribution. This distribution has a long right tail (it is skewed to the right). The sampling distribution of $\overline{Y}$, after centering and scaling, is shown in Figures 2.10b–d for $n = 5$, 25, and 100, respectively. Although the sampling distribution is approaching the bell shape for $n = 25$, the normal approximation still has noticeable imperfections. By $n = 100$, however, the normal approximation is quite good. In fact, for $n \geq 100$, the normal approximation to the distribution of $\overline{Y}$ typically is very good for a wide variety of population distributions.

The central limit theorem is a remarkable result. While the "small n" distributions of $\overline{Y}$ in parts b and c of Figures 2.9 and 2.10 are complicated and quite different from each other, the "large n" distributions in Figures 2.9d and 2.10d are simple and, amazingly, have a similar shape. Because the distribution of $\overline{Y}$ approaches the normal as n grows large, $\overline{Y}$ is said to have an **asymptotic normal distribution**.

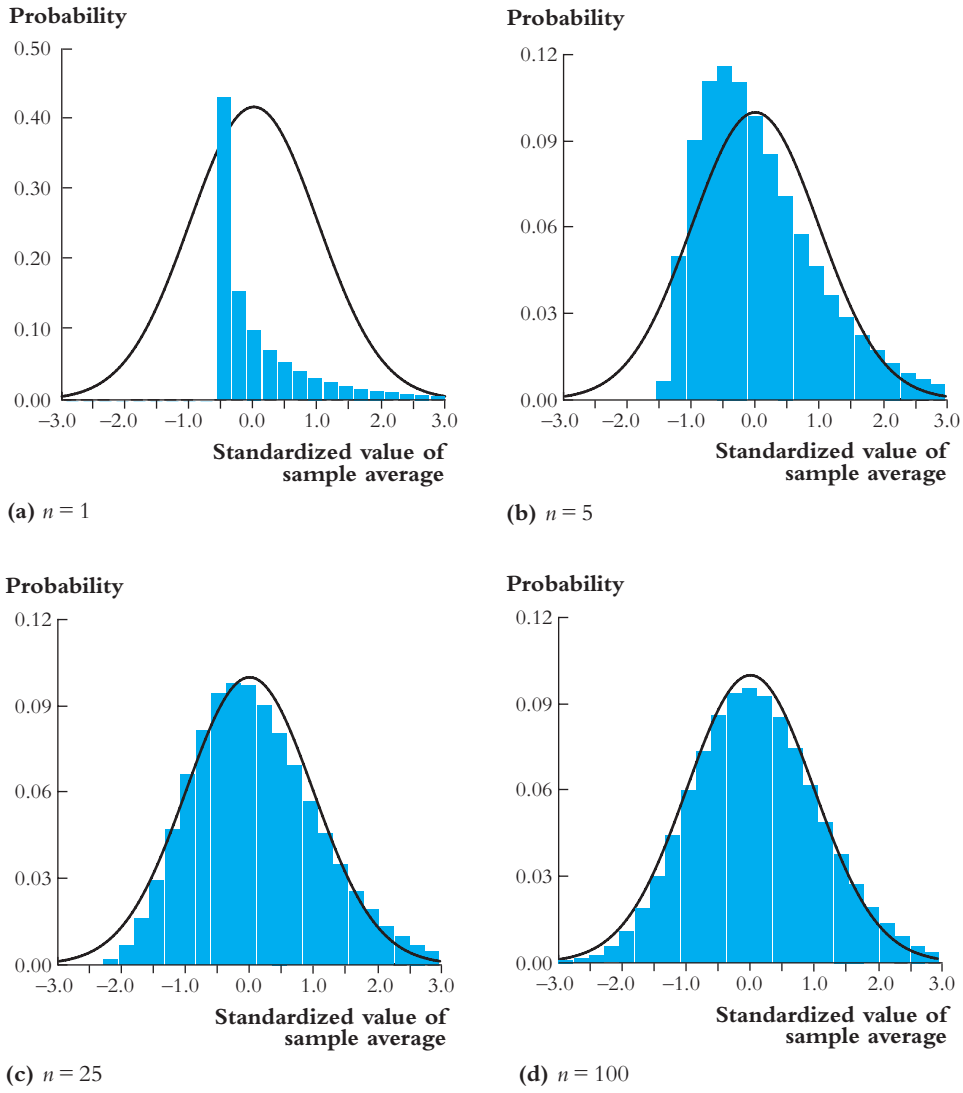
The convenience of the normal approximation, combined with its wide applicability because of the central limit theorem, makes it a key underpinning of applied econometrics. The central limit theorem is summarized in Key Concept 2.7.

The Central Limit Theorem

KEY CONCEPT 2.7

Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with $E(Y_i) = \mu_Y$ and $\text{var}(Y_i) = \sigma_Y^2$, where $0 < \sigma_Y^2 < \infty$. As $n \to \infty$, the distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ (where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$) becomes arbitrarily well approximated by the standard normal distribution.

FIGURE 2.10 Distribution of the Standardized Sample Average of n Draws from a Skewed Population Distribution



The figures show sampling distributions of the standardized sample average of n draws from the skewed (asymmetric) population distribution shown in Figure 2.10a. When n is small ($n = 5$), the sampling distribution, like the population distribution, is skewed. But when n is large ($n = 100$), the sampling distribution is well approximated by a standard normal distribution (solid line), as predicted by the central limit theorem. The normal distribution is scaled so that the height of the distribution is approximately the same in all figures.

Summary

1. The probabilities with which a random variable takes on different values are summarized by the cumulative distribution function, the probability distribution function (for discrete random variables), and the probability density function (for continuous random variables).
2. The expected value of a random variable Y (also called its mean, μ_Y), denoted $E(Y)$, is its probability-weighted average value. The variance of Y is $\sigma_Y^2 = E[(Y - \mu_Y)^2]$, and the standard deviation of Y is the square root of its variance.
3. The joint probabilities for two random variables, X and Y, are summarized by their joint probability distribution. The conditional probability distribution of Y given $X = x$ is the probability distribution of Y, conditional on X taking on the value x.
4. A normally distributed random variable has the bell-shaped probability density in Figure 2.5. To calculate a probability associated with a normal random variable, first standardize the variable, and then use the standard normal cumulative distribution tabulated in Appendix Table 1.
5. Simple random sampling produces n random observations, $Y_1, \ldots, Y_n$, that are independently and identically distributed (i.i.d.).
6. The sample average, $\overline{Y}$, varies from one randomly chosen sample to the next and thus is a random variable with a sampling distribution. If $Y_1, \ldots, Y_n$ are i.i.d., then
   a. the sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$;
   b. the law of large numbers says that $\overline{Y}$ converges in probability to μ_Y; and
   c. the central limit theorem says that the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$, has a standard normal distribution [$N(0, 1)$ distribution] when n is large.

Key Terms

outcomes (14)
probability (14)
sample space (14)
event (14)
discrete random variable (14)
continuous random variable (14)
probability distribution (14)
cumulative probability distribution (15)
cumulative distribution function (c.d.f.) (15)
cumulative distribution (15)
Bernoulli random variable (16)
Bernoulli distribution (16)
probability density function (p.d.f.) (16)
density function (16)
density (16)
expected value (18)
expectation (18)
mean (18)
variance (19)
standard deviation (19)
moments of a distribution (21)
skewness (22)
kurtosis (22)
outlier (22)
leptokurtic (22)

r^{th} moment (23)
standardized random variable (23)
joint probability distribution (23)
marginal probability distribution (24)
conditional distribution (24)
conditional expectation (25)
conditional mean (25)
law of iterated expectations (26)
conditional variance (27)
Bayes' rule (27)
independently distributed (28)
independent (28)
covariance (28)
correlation (29)
uncorrelated (29)
normal distribution (33)
standard normal distribution (33)
multivariate normal distribution (35)
bivariate normal distribution (35)
chi-squared distribution (38)
Student t distribution (38)
t distribution (38)
F distribution (38)
simple random sampling (39)
population (40)
identically distributed (40)
independently and identically distributed (i.i.d.) (40)
sample average (40)
sample mean (40)
sampling distribution (41)
exact (finite-sample) distribution (43)
asymptotic distribution (43)
law of large numbers (43)
convergence in probability (44)
consistency (44)
central limit theorem (44)
asymptotic normal distribution (47)

Review the Concepts

2.1 Examples of random variables used in this chapter included (a) the sex of the next person you meet, (b) the number of times a wireless network fails, (c) the time it takes to commute to school, and (d) whether it is raining or not. Explain why each can be thought of as random.

2.2 Suppose that the random variables X and Y are independent and you know their distributions. Explain why knowing the value of X tells you nothing about the value of Y.

2.3 Suppose that X denotes the amount of rainfall in your hometown during a randomly selected month and Y denotes the number of children born in Los Angeles during the same month. Are X and Y independent? Explain.

2.4 An econometrics class has 80 students, and the mean student weight is 145 lb. A random sample of 4 students is selected from the class, and their average weight is calculated. Will the average weight of the students in the sample equal 145 lb? Why or why not? Use this example to explain why the sample average, $\overline{Y}$, is a random variable.

2.5 Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with a $N(1, 4)$ distribution. Sketch the probability density of $\overline{Y}$ when $n = 2$. Repeat this for $n = 10$ and $n = 100$. In words, describe how the densities differ. What is the relationship between your answer and the law of large numbers?

2.6 Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with the probability distribution given in Figure 2.10a. You want to calculate $\Pr(\overline{Y} \leq 0.1)$. Would it be reasonable to use the normal approximation if $n = 5$? What about $n = 25$ or $n = 100$? Explain.

2.7 Y is a random variable with $\mu_Y = 0$, $\sigma_Y = 1$, skewness $= 0$, and kurtosis $= 100$. Sketch a hypothetical probability distribution of Y. Explain why n random variables drawn from this distribution might have some large outliers.

Exercises

2.1 Let Y denote the number of "heads" that occur when two coins are tossed.

- **a.** Derive the probability distribution of Y.
- **b.** Derive the cumulative probability distribution of Y.
- **c.** Derive the mean and variance of Y.

2.2 Use the probability distribution given in Table 2.2 to compute (a) $E(Y)$ and $E(X)$; (b) σ_X^2 and σ_Y^2; and (c) σ_{XY} and $\text{corr}(X, Y)$.

2.3 Using the random variables X and Y from Table 2.2, consider two new random variables, $W = 3 + 6X$ and $V = 20 - 7Y$. Compute (a) $E(W)$ and $E(V)$; (b) σ_W^2 and σ_V^2; and (c) σ_{WV} and $\text{corr}(W, V)$.

2.4 Suppose X is a Bernoulli random variable with $\Pr(X = 1) = p$.

- **a.** Show $E(X^3) = p$.
- **b.** Show $E(X^k) = p$ for $k > 0$.
- **c.** Suppose that $p = 0.3$. Compute the mean, variance, skewness, and kurtosis of X. (*Hint:* You might find it helpful to use the formulas given in Exercise 2.21.)

2.5 In September, Seattle's daily high temperature has a mean of 70° F and a standard deviation of 7° F. What are the mean, standard deviation, and variance in degrees Celsius?

2.6 The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working-age U.S. population for September 2017.

Joint Distribution of Employment Status and College Graduation in the U.S. Population Aged 25 and Older, September 2017

	Unemployed ($Y = 0$)	Employed ($Y = 1$)	Total
Non–college grads ($X = 0$)	0.026	0.576	0.602
College grads ($X = 1$)	0.009	0.389	0.398
Total	0.035	0.965	1.000

a. Compute $E(Y)$.

b. The unemployment rate is the fraction of the labor force that is unemployed. Show that the unemployment rate is given by $1 - E(Y)$.

c. Calculate $E(Y|X = 1)$ and $E(Y|X = 0)$.

d. Calculate the unemployment rate for (i) college graduates and (ii) non–college graduates.

e. A randomly selected member of this population reports being unemployed. What is the probability that this worker is a college graduate? A non–college graduate?

f. Are educational achievement and employment status independent? Explain.

2.7 In a given population of two-earner male-female couples, male earnings have a mean of \$40,000 per year and a standard deviation of \$12,000. Female earnings have a mean of \$45,000 per year and a standard deviation of \$18,000. The correlation between male and female earnings for a couple is 0.80. Let C denote the combined earnings for a randomly selected couple.

a. What is the mean of C?

b. What is the covariance between male and female earnings?

c. What is the standard deviation of C?

d. Convert the answers to (a) through (c) from U.S. dollars (\$) to euros (€).

2.8 The random variable Y has a mean of 1 and a variance of 4. Let $Z = \frac{1}{2}(Y - 1)$. Show that $\mu_Z = 0$ and $\sigma_Z^2 = 1$.

2.9 X and Y are discrete random variables with the following joint distribution:

		Value of Y				
		14	**22**	**30**	**40**	**65**
Value of X	**1**	0.02	0.05	0.10	0.03	0.01
	5	0.17	0.15	0.05	0.02	0.01
	8	0.02	0.03	0.15	0.10	0.09

That is, $\Pr(X = 1,\ Y = 14) = 0.02$, and so forth.

a. Calculate the probability distribution, mean, and variance of Y.

b. Calculate the probability distribution, mean, and variance of Y given $X = 8$.

c. Calculate the covariance and correlation between X and Y.

2.10 Compute the following probabilities:

a. If Y is distributed $N(1, 4)$, find $\Pr(Y \leq 3)$.

b. If Y is distributed $N(3, 9)$, find $\Pr(Y > 0)$.

c. If Y is distributed $N(50, 25)$, find $\Pr(40 \leq Y \leq 52)$.

d. If Y is distributed $N(5, 2)$, find $\Pr(6 \leq Y \leq 8)$.

2.11 Compute the following probabilities:

a. If Y is distributed χ^2_4, find $\Pr(Y \leq 7.78)$.

b. If Y is distributed χ^2_{10}, find $\Pr(Y > 18.31)$.

c. If Y is distributed $F_{10,\infty}$, find $\Pr(Y > 1.83)$.

d. Why are the answers to (b) and (c) the same?

e. If Y is distributed χ^2_1, find $\Pr(Y \leq 1.0)$. (*Hint:* Use the definition of the χ^2_1 distribution.)

2.12 Compute the following probabilities:

a. If Y is distributed t_{15}, find $\Pr(Y > 1.75)$.

b. If Y is distributed t_{90}, find $\Pr(-1.99 \leq Y \leq 1.99)$.

c. If Y is distributed $N(0, 1)$, find $\Pr(-1.99 \leq Y \leq 1.99)$.

d. Why are the answers to (b) and (c) approximately the same?

e. If Y is distributed $F_{7,4}$, find $\Pr(Y > 4.12)$.

f. If Y is distributed $F_{7,120}$, find $\Pr(Y > 2.79)$.

2.13 X is a Bernoulli random variable with $\Pr(X = 1) = 0.99$; Y is distributed $N(0, 1)$; W is distributed $N(0, 100)$; and X, Y, and W are independent. Let $S = XY + (1 - X)W$. (That is, $S = Y$ when $X = 1$, and $S = W$ when $X = 0$.)

a. Show that $E(Y^2) = 1$ and $E(W^2) = 100$.

b. Show that $E(Y^3) = 0$ and $E(W^3) = 0$. (*Hint:* What is the skewness for a symmetric distribution?)

c. Show that $E(Y^4) = 3$ and $E(W^4) = 3 \times 100^2$. (*Hint:* Use the fact that the kurtosis is 3 for a normal distribution.)

d. Derive $E(S)$, $E(S^2)$, $E(S^3)$, and $E(S^4)$. (*Hint:* Use the law of iterated expectations conditioning on $X = 0$ and $X = 1$.)

e. Derive the skewness and kurtosis for S.

2.14 In a population, $\mu_Y = 100$ and $\sigma_Y^2 = 43$. Use the central limit theorem to answer the following questions:

a. In a random sample of size $n = 100$, find $\Pr(\overline{Y} \leq 101)$.

b. In a random sample of size $n = 165$, find $\Pr(\overline{Y} > 98)$.

c. In a random sample of size $n = 64$, find $\Pr(101 \leq \overline{Y} \leq 103)$.

2.15 Suppose $Y_1, \ldots, Y_n$ are i.i.d. random variables, each distributed $N(10, 4)$.

a. Compute $\Pr(9.6 \leq \overline{Y} \leq 10.4)$ when (i) $n = 20$, (ii) $n = 100$, and (iii) $n = 1000$.

b. Suppose c is a positive number. Show that $\Pr(10 - c \leq \overline{Y} \leq 10 + c)$ becomes close to 1.0 as n grows large.

c. Use your answer in (b) to argue that $\overline{Y}$ converges in probability to 10.

2.16 Y is distributed $N(5, 100)$, and you want to calculate $\Pr(Y < 3.6)$. Unfortunately, you do not have your textbook, and do not have access to a normal probability table like Appendix Table 1. However, you do have your computer and a computer program that can generate i.i.d. draws from the $N(5, 100)$ distribution. Explain how you can use your computer to compute an accurate approximation for $\Pr(Y < 3.6)$.

2.17 $Y_1, \ldots, Y_n$ are i.i.d. Bernoulli random variables with $p = 0.4$. Let $\overline{Y}$ denote the sample mean.

a. Use the central limit to compute approximations for

i. $\Pr(\overline{Y} \geq 0.43)$ when $n = 100$.
ii. $\Pr(\overline{Y} \leq 0.37)$ when $n = 400$.

b. How large would n need to be to ensure that $\Pr(0.39 \leq \overline{Y} \leq 0.41) \geq 0.95$? (Use the central limit theorem to compute an approximate answer.)

2.18 In any year, the weather can inflict storm damage to a home. From year to year, the damage is random. Let Y denote the dollar value of damage in any given year. Suppose that in 95% of the years $Y = 0$ but in 5% of the years $Y = \$20{,}000$.

a. What are the mean and standard deviation of the damage in any year?

b. Consider an "insurance pool" of 100 people whose homes are sufficiently dispersed so that, in any year, the damage to different homes can be viewed as independently distributed random variables. Let $\overline{Y}$ denote the average damage to these 100 homes in a year. (i) What is the expected value of the average damage, $\overline{Y}$? (ii) What is the probability that $\overline{Y}$ exceeds \$2000?

2.19 Consider two random variables, X and Y. Suppose that Y takes on k values $y_1, \ldots, y_k$ and that X takes on l values $x_1, \ldots, x_l$.

a. Show that $\Pr(Y = y_j) = \sum_{i=1}^{l} \Pr(Y = y_j | X = x_i) \Pr(X = x_i)$. [*Hint:* Use the definition of $\Pr(Y = y_j | X = x_i)$.]

b. Use your answer to (a) to verify Equation (2.19).

c. Suppose that X and Y are independent. Show that $\sigma_{XY} = 0$ and $\text{corr}(X, Y) = 0$.

2.20 Consider three random variables, X, Y, and Z. Suppose that Y takes on k values $y_1, \ldots, y_k$; that X takes on l values $x_1, \ldots, x_l$; and that Z takes on m values $z_1, \ldots, z_m$. The joint probability distribution of X, Y, Z is $\Pr(X = x, Y = y, Z = z)$, and the conditional probability distribution of Y given X and Z is $\Pr(Y = y | X = x, Z = z) = \frac{\Pr(Y = y, X = x, Z = z)}{\Pr(X = x, Z = z)}$.

a. Explain how the marginal probability that $Y = y$ can be calculated from the joint probability distribution. [*Hint:* This is a generalization of Equation (2.16).]

b. Show that $E(Y) = E[E(Y|X, Z)]$. [*Hint:* This is a generalization of Equations (2.19) and (2.20).]

2.21 X is a random variable with moments $E(X)$, $E(X^2)$, $E(X^3)$, and so forth.

a. Show $E(X - \mu)^3 = E(X^3) - 3[E(X^2)][E(X)] + 2[E(X)]^3$.

b. Show $E(X-\mu)^4 = E(X^4) - 4[E(X)][E(X^3)] + 6[E(X)]^2[E(X^2)] - 3[E(X)]^4$.

2.22 Suppose you have some money to invest—for simplicity, \$1—and you are planning to put a fraction w into a stock market mutual fund and the rest, $1 - w$, into a bond mutual fund. Suppose that \$1 invested in a stock fund yields R_s after one year and that \$1 invested in a bond fund yields R_b, suppose that R_s is random with mean 0.08 (8%) and standard deviation 0.07, and suppose that R_b is random with mean 0.05 (5%) and standard deviation 0.04. The correlation between R_s and R_b is 0.25. If you place a fraction w of your money in the stock fund and the rest, $1 - w$, in the bond fund, then the return on your investment is $R = wR_s + (1 - w)R_b$.

a. Suppose that $w = 0.5$. Compute the mean and standard deviation of R.

b. Suppose that $w = 0.75$. Compute the mean and standard deviation of R.

c. What value of w makes the mean of R as large as possible? What is the standard deviation of R for this value of w?

d. (Harder) What is the value of w that minimizes the standard deviation of R? (Show using a graph, algebra, or calculus.)

2.23 This exercise provides an example of a pair of random variables, X and Y, for which the conditional mean of Y given X depends on X but $\text{corr}(X, Y) = 0$. Let X and Z be two independently distributed standard normal random variables, and let $Y = X^2 + Z$.

a. Show that $E(Y|X) = X^2$.

b. Show that $\mu_Y = 1$.

c. Show that $E(XY) = 0$. (*Hint:* Use the fact that the odd moments of a standard normal random variable are all 0.)

d. Show that $\text{cov}(X, Y) = 0$ and thus $\text{corr}(X, Y) = 0$.

2.24 Suppose Y_i is distributed i.i.d. $N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

a. Show that $E(Y_i^2/\sigma^2) = 1$.

b. Show that $W = (1/\sigma^2)\sum_{i=1}^{n} Y_i^2$ is distributed χ^2_n.

c. Show that $E(W) = n$. [*Hint:* Use your answer to (a).]

d. Show that $V = Y_1 \Big/ \sqrt{\dfrac{\sum_{i=2}^{n} Y_i^2}{n-1}}$ is distributed t_{n-1}.

2.25 (Review of summation notation) Let $x_1, \ldots, x_n$ denote a sequence of numbers; $y_1, \ldots, y_n$ denote another sequence of numbers; and a, b, and c denote three constants. Show that

a. $\displaystyle\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$,

b. $\displaystyle\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$,

c. $\displaystyle\sum_{i=1}^{n} a = n \times a$, and

d. $\displaystyle\sum_{i=1}^{n} (a + bx_i + cy_i)^2 = na^2 + b^2\sum_{i=1}^{n} x_i^2 + c^2\sum_{i=1}^{n} y_i^2 + 2ab\sum_{i=1}^{n} x_i + 2ac\sum_{i=1}^{n} y_i + 2bc\sum_{i=1}^{n} x_i y_i$.

2.26 Suppose that $Y_1, Y_2, \ldots, Y_n$ are random variables with a common mean μ_Y; a common variance σ_Y^2; and the same correlation ρ (so that the correlation between Y_i and Y_j is equal to ρ for all pairs i and j, where $i \neq j$).

a. Show that $\text{cov}(Y_i, Y_j) = \rho\sigma_Y^2$ for $i \neq j$.

b. Suppose that $n = 2$. Show that $E(\overline{Y}) = \mu_Y$ and $\text{var}(\overline{Y}) = \frac{1}{2}\sigma_Y^2 + \frac{1}{2}\rho\sigma_Y^2$.

c. For $n \geq 2$, show that $E(\overline{Y}) = \mu_Y$ and $\text{var}(\overline{Y}) = \sigma_Y^2/n + [(n-1)/n]\rho\sigma_Y^2$.

d. When n is very large, show that $\text{var}(\overline{Y}) \simeq \rho\sigma_Y^2$.

2.27 Consider the problem of predicting Y using another variable, X, so that the prediction of Y is some function of X, say $g(X)$. Suppose that the quality of the prediction is measured by the squared prediction error made on average over all predictions, that is, by $E\{[Y - g(X)]^2\}$. This exercise provides a non-calculus proof that of all possible prediction functions g, the best prediction is made by the conditional expectation, $E(Y|X)$.

a. Let $\hat{Y} = E(Y|X)$, and let $u = Y - \hat{Y}$ denote its prediction error. Show that $E(u) = 0$. (*Hint:* Use the law of iterated expectations.)

b. Show that $E(uX) = 0$.

c. Let $\widetilde{Y} = g(X)$ denote a different prediction of Y using X, and let $v = Y - \widetilde{Y}$ denote its error. Show that $E[(Y - \widetilde{Y})^2] > E[(Y - \hat{Y})^2]$. [*Hint:* Let $h(X) = g(X) - E(Y|X)$, so that $v = [Y - E(Y|X)] - h(X)$. Derive $E(v^2)$.]

2.28 Refer to Part B of Table 2.3 for the conditional distribution of the number of network failures M given network age A. Let $\Pr(A = 0) = 0.5$; that is, you work in your room 50% of the time.

a. Compute the probability of three network failures, $\Pr(M = 3)$.

b. Use Bayes' rule to compute $\Pr(A = 0|M = 3)$.

c. Now suppose you work in your room one-fourth of the time, so $\Pr(A = 0) = 0.25$. Use Bayes' rule to compute $\Pr(A = 0|M = 3)$.

Empirical Exercise

E2.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the spreadsheet **Age_HourlyEarnings**, which contains the joint distribution of age (*Age*) and average hourly earnings (*AHE*) for 25- to 34-year-old full-time workers in 2015 with an education level that exceeds a high school diploma. Use this joint distribution to carry out the following exercises. (*Note:* For these exercises, you need to be able to carry out calculations and construct charts using a spreadsheet.)

a. Compute the marginal distribution of *Age*.

b. Compute the mean of *AHE* for each value of *Age*; that is, compute, $E(AHE|Age = 25)$, and so forth.

c. Compute and plot the mean of *AHE* versus *Age*. Are average hourly earnings and age related? Explain.

d. Use the law of iterated expectations to compute the mean of AHE; that is, compute $E(AHE)$.

e. Compute the variance of AHE.

f. Compute the covariance between AHE and Age.

g. Compute the correlation between AHE and Age.

h. Relate your answers in (f) and (g) to the plot you constructed in (c).

APPENDIX

2.1 Derivation of Results in Key Concept 2.3

This appendix derives the equations in Key Concept 2.3.

Equation (2.30) follows from the definition of the expectation.

To derive Equation (2.31), use the definition of the variance to write $\text{var}(a + bY) = E\{[a + bY - E(a + bY)]^2\} = E\{[b(Y - \mu_Y)]^2\} = b^2E[(Y - \mu_Y)^2] = b^2\sigma_Y^2$.

To derive Equation (2.32), use the definition of the variance to write

$$\begin{aligned}
\text{var}(aX + bY) &= E\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} \\
&= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\
&= E[a^2(X - \mu_X)^2] + 2E[ab(X - \mu_X)(Y - \mu_Y)] \\
&\quad + E[b^2(Y - \mu_Y)^2] \\
&= a^2\text{var}(X) + 2ab\ \text{cov}(X, Y) + b^2\ \text{var}(Y) \\
&= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \qquad (2.50)
\end{aligned}$$

where the second equality follows by collecting terms, the third equality follows by expanding the quadratic, and the fourth equality follows by the definition of the variance and covariance.

To derive Equation (2.33), write

$$E(Y^2) = E\{(Y - \mu_Y) + \mu_Y]^2\} = E[(Y - \mu_Y)^2] + 2\ \mu_Y E(Y - \mu_Y) + \mu_Y^2 = \sigma_Y^2 + \mu_Y^2$$

because $E(Y - \mu_Y) = 0$.

To derive Equation (2.34), use the definition of the covariance to write

$$\begin{aligned}
\text{cov}(a + bX + cV, Y) &= E\{[a + bX + cV - E(a + bX + cV)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X) + c(V - \mu_V)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X)][Y - \mu_Y]\} + E\{[c(V - \mu_V)][Y - \mu_Y]\} \\
&= b\sigma_{XY} + c\sigma_{VY}, \qquad (2.51)
\end{aligned}$$

which is Equation (2.34).

To derive Equation (2.35), write

$$\begin{aligned} E(XY) &= E\{[(X-\mu_X)+\mu_X][(Y-\mu_Y)+\mu_Y]\} \\ &= E[(X-\mu_X)(Y-\mu_Y)] + \mu_X E(Y-\mu_Y) + \mu_Y E(X-\mu_X) + \mu_X\mu_Y \\ &= \sigma_{XY} + \mu_X\mu_Y. \end{aligned}$$

We now prove the correlation inequality in Equation (2.36); that is, $|\text{corr}(X, Y) \leq 1.|$ Let $a = -\sigma_{XY}/\sigma_X^2$ and $b = 1$. Applying Equation (2.32), we have,

$$\begin{aligned} \text{var}(aX + Y) &= a^2\sigma_X^2 + \sigma_Y^2 + 2a\sigma_{XY} \\ &= (-\sigma_{XY}/\sigma_X^2)^2\sigma_X^2 + \sigma_Y^2 + 2(-\sigma_{XY}/\sigma_X^2)\sigma_{XY} \\ &= \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2. \end{aligned} \tag{2.52}$$

Because $\text{var}(aX + Y)$ is a variance, it cannot be negative, so from the final line of Equation (2.52), it must be that $\sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 \geq 0$. Rearranging this inequality yields

$$\sigma_{XY}^2 \leq \sigma_X^2\sigma_Y^2 \text{ (covariance inequality)}. \tag{2.53}$$

The covariance inequality implies that $\sigma_{XY}^2/(\sigma_X^2\sigma_Y^2) \leq 1$ or, equivalently, $|\sigma_{XY}/(\sigma_X\sigma_Y)| \leq 1$, which (using the definition of the correlation) proves the correlation inequality, $|\text{corr}(X\ Y)| \leq 1$.

APPENDIX

2.2 The Conditional Mean as the Minimum Mean Squared Error Predictor

At a general level, the statistical prediction problem is, how does one best use the information in a random variable X to predict the value of another random variable Y?

To answer to this question, we must first make precise mathematically what it means for one prediction to be better than another. A common way to do so is to consider the cost of making a prediction error. This cost, which is called the prediction loss, depends on the magnitude of the prediction error. For example, if your job is to predict sales so that a production supervisor can develop a production schedule, being off by a small amount is unlikely to inconvenience customers or to disrupt the production process. But if you are off by a large amount and production is set far too low, your company might lose customers who need to wait a long time to receive a product they order, or if production is far too high, the company will have costly excess inventory on its hands. In either case, a large prediction error can be disproportionately more costly than a small one.

One way to make this logic precise is to let the cost of a prediction error depend on the square of that error, so an error twice as large is four times as costly. Specifically, suppose that your prediction of Y, given the random variable X, is $g(X)$. The prediction error is $Y - g(X)$, and the quadratic loss associated with this prediction is,

$$Loss = E\{[Y - g(X)]^2\}. \tag{2.54}$$

We now show that, of all possible functions $g(X)$, the loss in Equation (2.54) is minimized by $g(X) = E(Y|X)$. We show this result using discrete random variables, however this result extends to continuous random variables. The proof here uses calculus; Exercise 2.27 works through a non-calculus proof of this result.

First consider the simpler problem of finding a number, m, that minimizies $E[(Y - m)^2]$. From the definition of the expectation, $E[(Y - m)^2] = \sum_{i=1}^{k}(Y_i - m)^2 p_i$. To find the value of m that minimizes $E[(Y - m)^2]$, take the derivative of $\sum_{i=1}^{k}(Y_i - m)^2 p_i$ with respect to m and set it to zero:

$$\begin{aligned}\frac{d}{dm}\sum_{i=1}^{k}(Y_i - m)^2 p_i &= -2\sum_{i=1}^{k}(Y_i - m)p_i = -2\left(\sum_{i=1}^{k}Y_i p_i - m\sum_{i=1}^{k}p_i\right) \\ &= -2\left(\sum_{i=1}^{k}Y_i p_i - m\right) = 0,\end{aligned} \tag{2.55}$$

where the final equality uses the fact that probabilities sum to 1. It follow from the final equality in Equation (2.55) that the squared error prediction loss is minimized by $m = \sum_{i=1}^{k}Y_i p_i = E(Y)$, that is, by setting m equal to the mean of Y.

To find the predictor $g(X)$ that minimizes the loss in Equation (2.54), use the law of iterated expectations to write that loss as, $Loss = E\{[Y - g(X)]^2\} = E(E\{[Y - g(X)]^2|X\})$. Thus, if the function $g(X)$ minimizes $E\{[Y - g(X)]^2|X = x\}$ for each value of x, it minimizes the loss in Equation (2.54). But for a fixed value $X = x$, $g(X) = g(x)$ is a fixed number, so this problem is the same as the one just solved, and the loss is minimized by choosing $g(x)$ to be the mean of Y, given $X = x$. This is true for every value of x. Thus the squared error loss in Equation (2.54) is minimzed by $g(X) = E(Y|X)$.