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 might 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 gender of the next new person you meet, your grade on an exam, and the number of times your computer will crash 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, your computer might never crash, it might crash once, it might crash 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 computer not crashing while you are writing a term paper is 80%, then over the course of writing many term papers you will complete 80% without a crash.

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 computer will crash no more than once" is the set consisting of two outcomes: "no crashes" and "one crash."

Random variables. A random variable is a numerical summary of a random outcome. The number of times your computer crashes 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, . . . , 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 computer crashes while you are writing a term paper. The probability distribution of the random variable M is the list of probabilities of each possible outcome: The probability that $M = 0$, denoted $\Pr(M = 0)$, is the probability of no computer crashes; $\Pr(M = 1)$ is the probability of a single computer crash; and so forth. An example of a probability distribution for M is given in the second row of Table 2.1; in this distribution, if your computer crashes four times, you will quit and write the paper by hand. According to this distribution, the probability of no crashes is 80%; the probability of one crash is 10%; and the probability of two, three, or four crashes is, 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 crashes 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%.

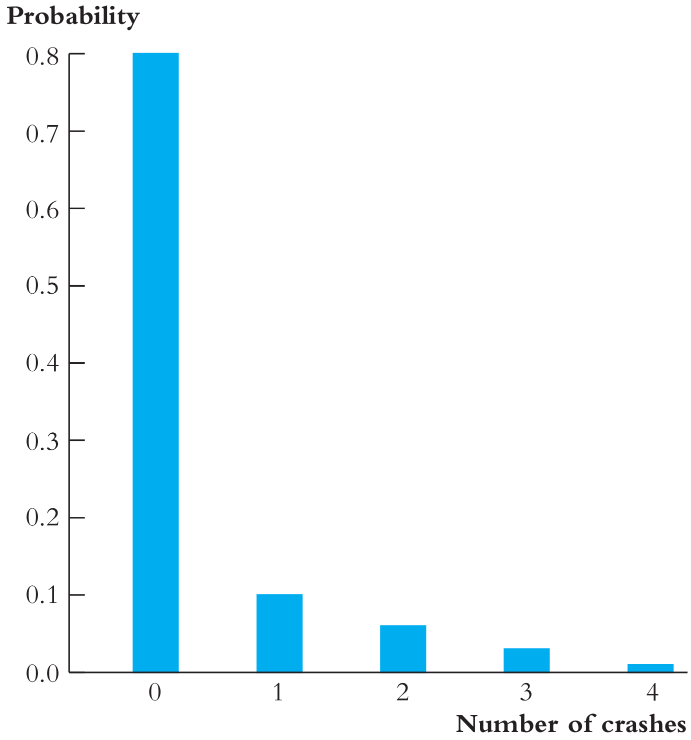
Cumulative probability distribution. The **cumulative probability distribution** is the probability that the random variable is less than or equal to a particular value. The last row of Table 2.1 gives the cumulative probability distribution of the random variable M. For example, the probability of at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

TABLE 2.1 Probability of Your Computer Crashing *M* Times

	Outcome (number of crashes)				
	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

FIGURE 2.1 Probability Distribution of the Number of Computer Crashes

The height of each bar is the probability that the computer crashes the indicated number of times. The height of the first bar is 0.8, so the probability of 0 computer crashes is 80%. The height of the second bar is 0.1, so the probability of 1 computer crash is 10%, and so forth for the other bars.



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

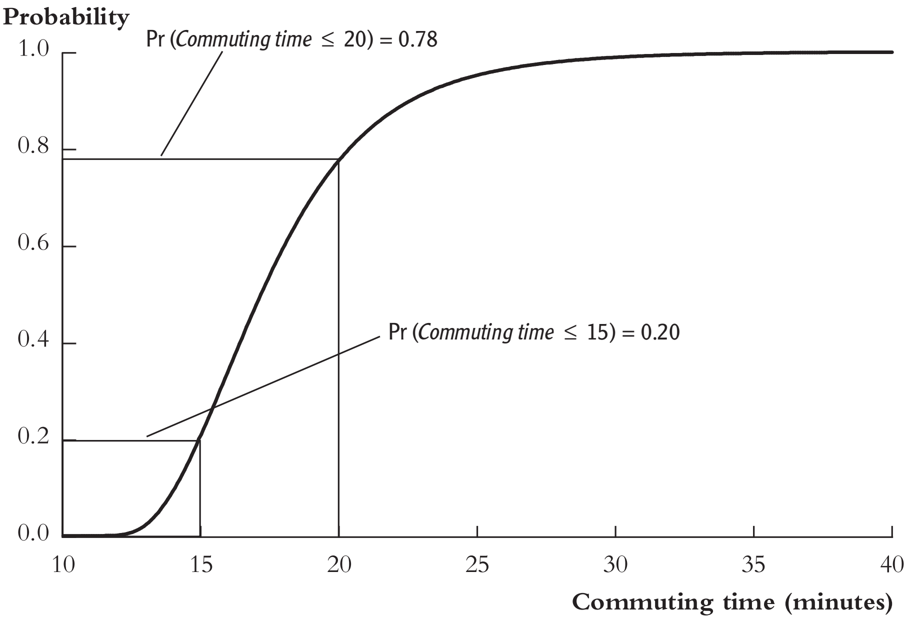
The Bernoulli distribution. An important special case of a discrete random variable is when the random variable is binary, that is, the outcomes are 0 or 1. A binary random variable is called a **Bernoulli random variable** (in honor of the seventeenth-century Swiss mathematician and scientist Jacob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

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

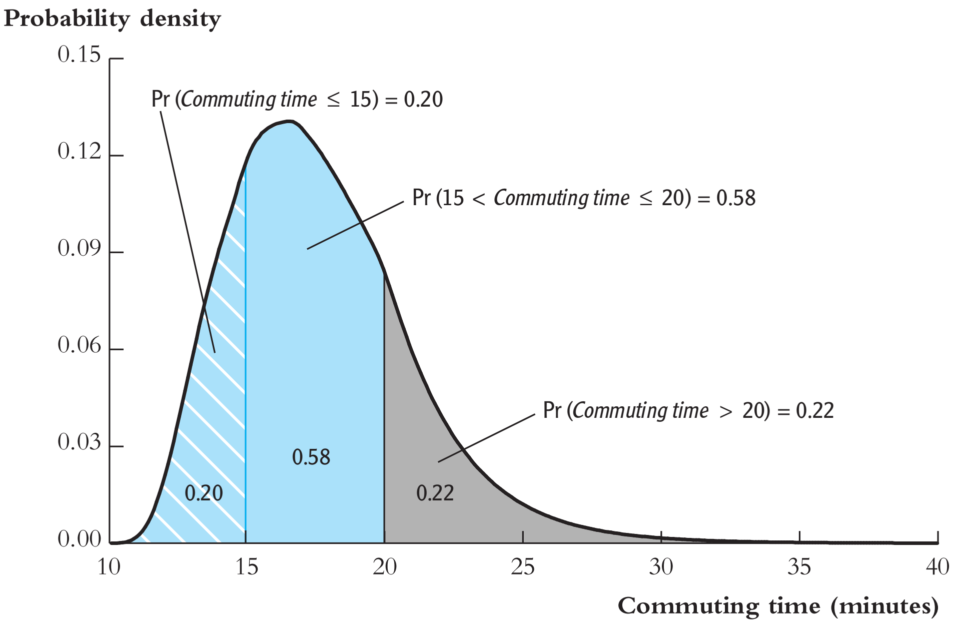
$$G = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p, \end{cases} \tag{2.1}$$

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.

FIGURE 2.2 Cumulative Distribution and Probability Density Functions of Commuting Time



(a) Cumulative distribution function of commuting time



(b) Probability density function of commuting time

Figure 2.2a shows the cumulative probability distribution (or 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.

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.

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 (or the "mean repayment") is \$108.90.

As a second example, consider the number of computer crashes M with the probability distribution given in Table 2.1. The expected value of M is the average number of crashes over many term papers, weighted by the frequency with which a crash of a given size 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 computer crashes while writing a term paper is 0.35. Of course, the actual number of crashes must always be an integer; it makes no sense to say that the computer crashed 0.35 times while writing a particular term paper! Rather, the calculation in Equation (2.2) means that the average number of crashes 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 as 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 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

$$E(Y) = y_1 p_1 + y_2 p_2 + \cdots + y_k p_k = \sum_{i=1}^{k} y_i p_i, \quad (2.3)$$

where the notation $\sum_{i=1}^{k} y_i p_i$ means "the sum of $y_i p_i$ 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

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

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 17.1.

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.

Variance and Standard Deviation

KEY CONCEPT 2.2

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

$$\sigma_Y^2 = \text{var}(Y) = E[(Y - \mu_Y)^2] = \sum_{i=1}^{k} (y_i - \mu_Y)^2 p_i. \tag{2.5}$$

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.

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

$$\begin{aligned}\text{var}(M) = {} & (0 - 0.35)^2 \times 0.80 + (1 - 0.35)^2 \times 0.10 + (2 - 0.35)^2 \times 0.06 \\ & + (3 - 0.35)^2 \times 0.03 + (4 - 0.35)^2 \times 0.01 = 0.6475. \end{aligned} \tag{2.6}$$

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

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

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

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

$$Y = 2000 + 0.8X. \tag{2.8}$$

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

$$E(Y) = \mu_Y = 2000 + 0.8\mu_X. \tag{2.9}$$

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

$$\sigma_Y = 0.8\sigma_X. \tag{2.10}$$

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

$$Y = a + bX. \tag{2.11}$$

Then the mean and variance of Y are

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

$$\sigma_Y^2 = b^2\sigma_X^2, \tag{2.13}$$

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

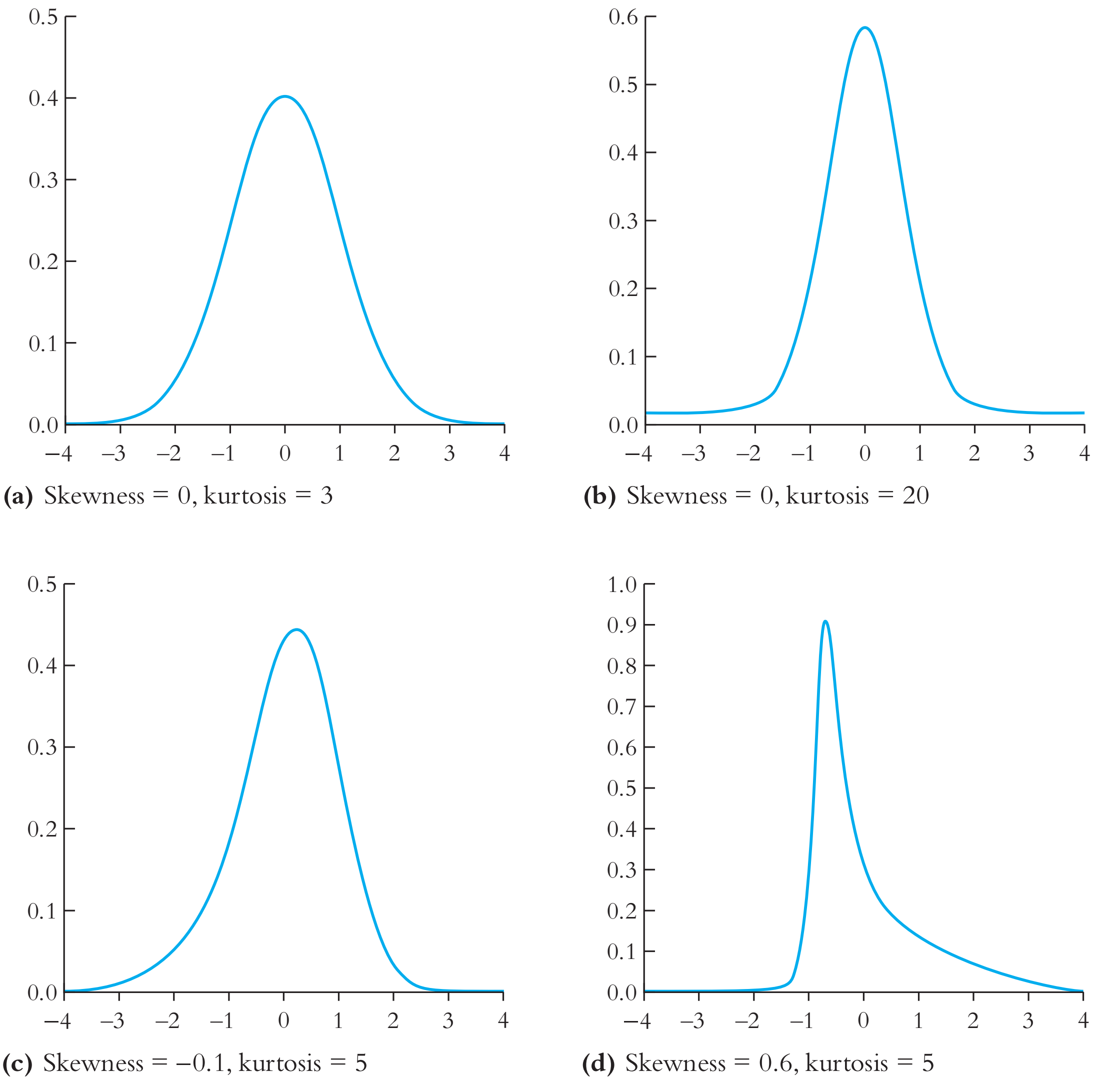
The mean and standard deviation measure two important features of a distribution: its center (the mean) and its spread (the standard deviation). This section discusses measures of two other features of a distribution: the skewness, which measures the lack of symmetry of a distribution, and the kurtosis, 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 which are symmetric (Figures 2.3a and 2.3b) and two which 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 2.3c. 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

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

FIGURE 2.3 Four Distributions with Different Skewness and Kurtosis



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–d) have heavy tails.

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 zero. 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 $\boldsymbol{r^{\text{th}}}$ **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.

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 gender 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 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 15%, or $\Pr(X = 0, Y = 0) = 0.15$. Also, $\Pr(X = 0, Y = 1) = 0.15$,

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

$\Pr(X = 1, Y = 0) = 0.07$, 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 $\Pr(Y = y \mid X = x)$.

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 $\Pr(Y = 0 \mid X = 0) = 0.50$. 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).

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

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

TABLE 2.3 Joint and Conditional Distributions of Computer Crashes (M) and Computer Age (A)

A. Joint Distribution

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	Total
Old computer ($A = 0$)	0.35	0.065	0.05	0.025	0.01	0.50
New computer ($A = 1$)	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

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 crashing computer example. Suppose you use a computer in the library to type your term paper and the librarian randomly assigns you a computer from those available, half of which are new and half of which are old. Because you are randomly assigned to a computer, the age of the computer you use, A ($= 1$ if the computer 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 distribution of computer crashes, given the age of the computer, is given in Part B of the table. For example, the joint probability $M = 0$ and $A = 0$ is 0.35; because half the computers are old, the conditional probability of no crashes, given that you are using an old computer, 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 crashes given that you are assigned a new computer is 90%. According to the conditional distributions in Part B of Table 2.3, the newer computers are less likely to crash than the old ones; for example, the probability of three crashes is 5% with an old computer but 1% with a new computer.

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 computer crashes, given that the computer is old, is $E(M \mid A = 0) = 0 \times 0.70 + 1 \times 0.13 + 2 \times 0.10 + 3 \times 0.05 + 4 \times 0.02 = 0.56$. The expected number of computer crashes, given that the computer is new, is $E(M \mid A = 1) = 0.14$, less than for the old computers.

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 crashes is 0.56 for old computers, so the conditional expectation of Y given that the computer is old is 0.56. Similarly, among new computers, the mean number of crashes is 0.14, that is, the conditional expectation of Y given that the computer 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

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

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,

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

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 crashes 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 $E(M) = E(M \mid A = 0) \times \Pr(A = 0) + E(M \mid A = 1) \times \Pr(A = 1) = 0.56 \times 0.50 + 0.14 \times 0.50 = 0.35$. 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 zero, then the mean of Y is zero. This is an immediate consequence of Equation (2.20): if $E(Y \mid X) = 0$, then $E(Y) = E[E(Y \mid X)] = E[0] = 0$. Said differently, if the mean of Y given X is zero, then it must be that the probability-weighted average of these conditional means is zero, that is, the mean of Y must be zero.

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 $E(Y) = E[E(Y \mid X, Z)]$, where $E(Y \mid X, Z)$ is the conditional expectation of Y given both X and Z. For example, in the computer crash illustration of Table 2.3, let P denote the number of programs installed on the computer; then $E(M \mid A, P)$ is the expected number of crashes for a computer with age A that has P programs installed. The expected number of crashes overall, $E(M)$, is the weighted average of the expected number of crashes for a computer with age A and number of programs P, weighted by the proportion of computers with that value 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

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

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

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,

$$\Pr(Y = y \mid X = x) = \Pr(Y = y) \quad \text{(independence of } X \text{ and } Y\text{)}. \tag{2.22}$$

Substituting Equation (2.22) 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

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y). \tag{2.23}$$

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, 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

$$\begin{aligned}\text{cov}(X, Y) = \sigma_{XY} &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}(x_j - \mu_X)(y_i - \mu_Y)\Pr(X = x_j, Y = y_i).\end{aligned} \tag{2.24}$$

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 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 zero (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:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.25}$$

Because the units of the numerator in Equation (2.25) are the same as those of the denominator, the units cancel and the correlation is unitless. 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,

$$-1 \leq \text{corr}(X, Y) \leq 1 \quad \text{(correlation inequality)}. \tag{2.26}$$

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

$$\text{if } E(Y \mid X) = \mu_Y, \text{ then cov}(Y, X) = 0 \text{ and corr}(Y, X) = 0. \tag{2.27}$$

We now show this result. First suppose that Y and X have mean zero so that $\text{cov}(Y, X) = E[(Y - \mu_Y)(X - \mu_X)] = E(YX)$. By the law of iterated expectations [Equation (2.20)], $E(YX) = E[E(YX \mid X)] = E[E(Y \mid X)X] = 0$ because $E(Y \mid X) = 0$, so $\text{cov}(Y, X) = 0$. Equation (2.27) follows by substituting $\text{cov}(Y, X) = 0$ into the definition of correlation in Equation (2.25). If Y and X do not have mean zero, first subtract off their means, 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:

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y. \tag{2.28}$$

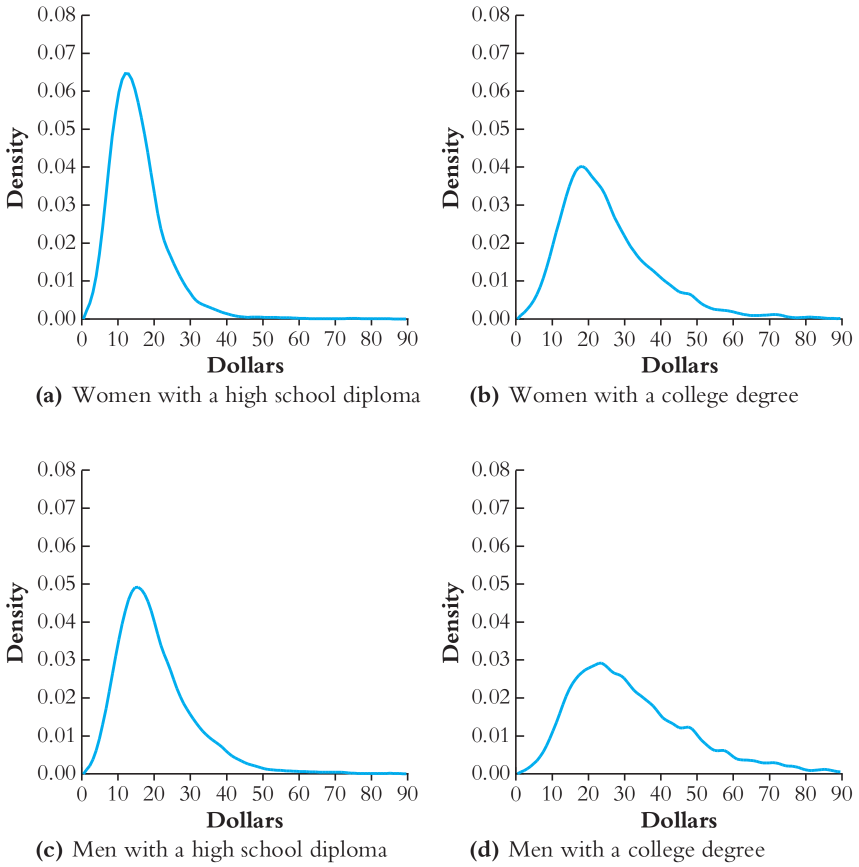
The Distribution of Earnings in the United States in 2012

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 gender. These four conditional distributions are shown in Figure 2.4, and the mean, standard deviation, and

FIGURE 2.4 Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2012, Given Education Level and Gender

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 college degree

(c) Men with a high school diploma

(d) Men with a college degree

continued on next page

TABLE 2.4 Summaries of the Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2012 Given Education Level and Gender

			Percentile			
	Mean	Standard Deviation	25%	50% (median)	75%	90%
(a) Women with high school diploma	$15.49	$8.42	$10.10	$14.03	$18.75	$24.52
(b) Women with four-year college degree	25.42	13.81	16.15	22.44	31.34	43.27
(c) Men with high school diploma	20.25	11.00	12.92	17.86	24.83	33.78
(d) Men with four-year college degree	32.73	18.11	19.61	28.85	41.68	57.30

Average hourly earnings are the sum of annual pretax 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, $E(Earnings|Highest\ degree = high\ school\ diploma,\ Gender = female)$—is $15.49 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 degree (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. 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 2013 Current Population Survey, which is discussed in more detail in Appendix 3.1.

Means, Variances, and Covariances of Sums of Random Variables

KEY CONCEPT 2.3

Let X, Y, and V be random variables, let μ_X and σ_X^2 be the mean and variance of X, 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 (2.29) through (2.35) follow from the definitions of the mean, variance, and covariance:

$$E(a + bX + cY) = a + b\mu_X + c\mu_Y, \quad (2.29)$$

$$\mathrm{var}(a + bY) = b^2\sigma_Y^2, \quad (2.30)$$

$$\mathrm{var}(aX + bY) = a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \quad (2.31)$$

$$E(Y^2) = \sigma_Y^2 + \mu_Y^2, \quad (2.32)$$

$$\mathrm{cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}, \quad (2.33)$$

$$E(XY) = \sigma_{XY} + \mu_X\mu_Y, \quad (2.34)$$

$$|\mathrm{corr}(X, Y)| \leq 1 \text{ and } |\sigma_{XY}| \leq \sqrt{\sigma_X^2\sigma_Y^2} \text{ (correlation inequality)}. \quad (2.35)$$

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

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\,\mathrm{cov}(X, Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \quad (2.36)$$

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

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) = \sigma_X^2 + \sigma_Y^2$$
$$\text{(if } X \text{ and } Y \text{ are independent)}. \quad (2.37)$$

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

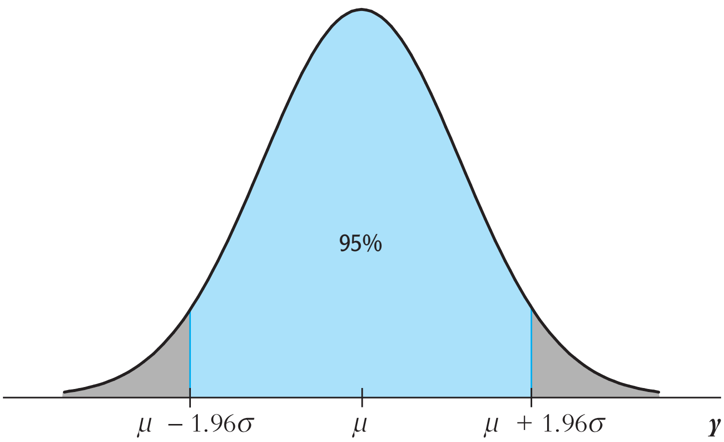
A continuous random variable with a **normal distribution** has the familiar bell-shaped probability density shown in Figure 2.5. The function defining the normal probability density is given in Appendix 17.1. As Figure 2.5 shows, the normal density with mean μ and variance σ^2 is symmetric around its 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, $\Pr(Z \le c) = \Phi(c)$, 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 **standardize the variable** by first subtracting the mean, then by dividing

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 $N(\mu, \sigma^2)$.



Computing Probabilities Involving Normal Random Variables

KEY CONCEPT 2.4

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

$$\Pr(Y \leq c_2) = \Pr(Z \leq d_2) = \Phi(d_2), \tag{2.38}$$

$$\Pr(Y \geq c_1) = \Pr(Z \geq d_1) = 1 - \Phi(d_1), \tag{2.39}$$

$$\Pr(c_1 \leq Y \leq c_2) = \Pr(d_1 \leq Z \leq d_2) = \Phi(d_2) - \Phi(d_1). \tag{2.40}$$

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

the result by the standard deviation. 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 zero and variance one (see Exercise 2.8); it has the standard normal 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,

$$\Pr(Y \leq 2) = \Pr[\tfrac{1}{2}(Y - 1) \leq \tfrac{1}{2}] = \Pr(Z \leq \tfrac{1}{2}) = \Phi(0.5) = 0.691, \tag{2.41}$$

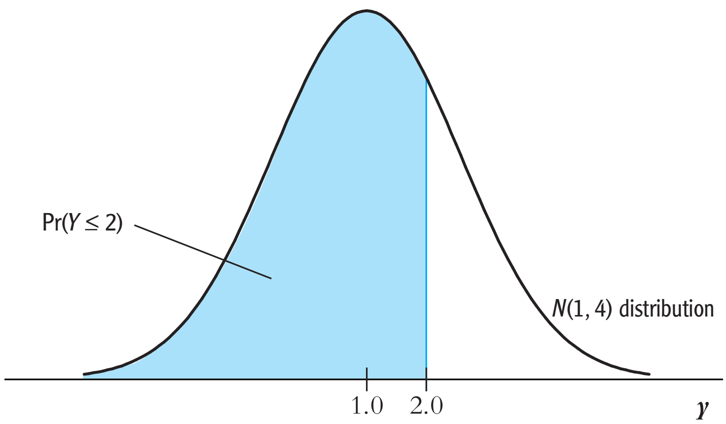
where the value 0.691 is taken from Appendix Table 1.

The same approach can be applied 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.

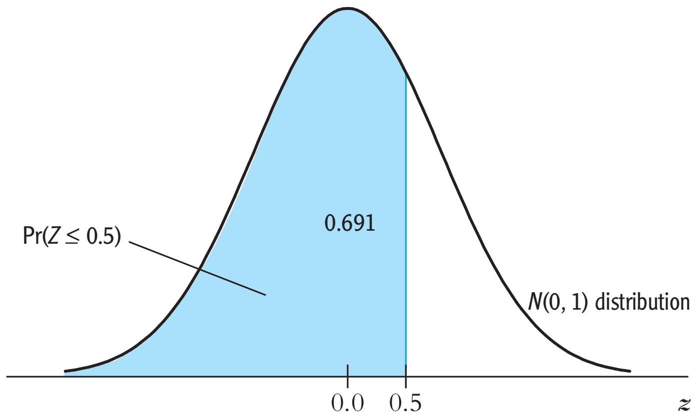
The normal distribution is symmetric, so its skewness is zero. The kurtosis of the normal distribution is 3.

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, $\Pr(Y \leq 2) = \Pr\left(\frac{Y-1}{2} \leq \frac{2-1}{2}\right) = \Pr(Z \leq 0.5)$. From Appendix Table 1, $\Pr(Z \leq 0.5) = \Phi(0.5) = 0.691$.



(a) $N(1, 4)$



(b) $N(0, 1)$

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 17.1, and the formula for the general multivariate normal p.d.f. is given in Appendix 18.1.

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:

$$aX + bY \text{ is distributed } N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}) \quad (X, Y \text{ bivariate normal}). \tag{2.42}$$

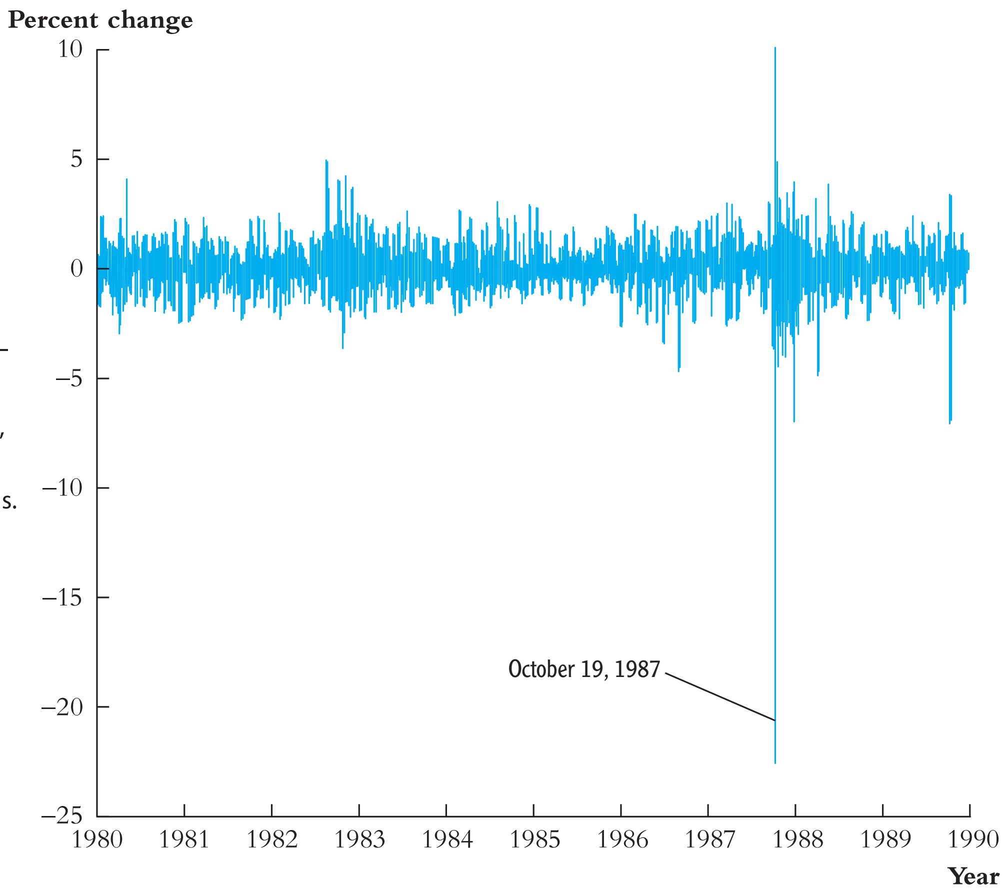
A Bad Day on Wall Street

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 December 31, 2012, the standard deviation of daily percentage price changes on the Dow was 1.12%, so the drop of 22.6% was a negative return of 20($= 22.6/1.12$) 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 20 standard deviations is $\Pr(|Z| \geq 20) = 2 \times \Phi(-20)$. You will not find this value in Appendix Table 1, but you can calculate it using a computer (try it!). This probability is 5.5×10^{-89}, that is, 0.000 . . . 00055, where there are a total of 88 zeros!

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

From 1980 through 2012, the average percentage daily change of "the Dow" index was 0.04% and its standard deviation was 1.12%. On October 19, 1987—"Black Monday"—the Dow fell 22.6%, or more than 20 standard deviations.



continued on next page

How small is 5.5×10^{-89}? Consider the following:

- The world population is about 7 billion, so the probability of winning a random lottery among all living people is about one in 7 billion, or 1.4×10^{-10}.
- 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 one 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 5.5×10^{-89}. 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 8325 trading days between January 1, 1980, and December 31, 2012, along with the standardized change using the mean and variance over this period. All ten changes exceed 6.4 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 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 16). 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.

TABLE 2.5 The Ten Largest Daily Percentage Changes in the Dow Jones Industrial Index, 1980–2012, 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 $\Pr(\lvert Z \rvert \geq z) = 2\Phi(-z)$
October 19, 1987	–22.6	–20.2	5.5×10^{-89}
October 13, 2008	11.1	9.9	6.4×10^{-23}
October 28, 2008	10.9	9.7	3.8×10^{-22}
October 21, 1987	10.1	9.0	1.8×10^{-19}
October 26, 1987	–8.0	–7.2	5.6×10^{-13}
October 15, 2008	–7.9	–7.1	1.6×10^{-12}
December 01, 2008	–7.7	–6.9	4.9×10^{-12}
October 09, 2008	–7.3	–6.6	4.7×10^{-11}
October 27, 1997	–7.2	–6.4	1.2×10^{-10}
September 17, 2001	–7.1	–6.4	1.6×10^{-10}

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 (2.42) by setting $a = 1$ and $b = 0$].

Third, if variables with a multivariate normal distribution have covariances that equal zero, 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. 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 zero 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 17.11). 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_m^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, divided by 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 when m is small (20 or less), 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 book 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 each of 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 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 her n randomly selected days and Y_i is the commuting time on the i^{th} of her 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

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.

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** (or **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

$$\overline{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.43}$$

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 and variance is the same for all $i = 1, \ldots, n$). When $n = 2$, the mean of the sum $Y_1 + Y_2$ is given by applying Equation (2.28): $E(Y_1 + Y_2) = \mu_Y + \mu_Y = 2\mu_Y$. 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,

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \mu_Y. \tag{2.44}$$

The variance of $\overline{Y}$ is found by applying Equation (2.37). For example, for $n = 2$, $\text{var}(Y_1 + Y_2) = 2\sigma_Y^2$, so [by applying Equation (2.31) 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,

$$\begin{aligned}
\text{var}(\overline{Y}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j \neq i}^{n}\text{cov}(Y_i, Y_j) \\
&= \frac{\sigma_Y^2}{n}.
\end{aligned} \tag{2.45}$$

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 (2.45). Suppose you divide \$1 equally among n assets. Let Y_i represent the payout in 1 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 1 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, the mean, the variance, and the standard deviation of $\overline{Y}$ are

$$E(\overline{Y}) = \mu_Y. \tag{2.46}$$

$$\text{var}(\overline{Y}) = \sigma^2_{\overline{Y}} = \frac{\sigma^2_Y}{n}, \text{ and} \tag{2.47}$$

$$\text{std.dev}(\overline{Y}) = \sigma_{\overline{Y}} = \frac{\sigma_Y}{\sqrt{n}}. \tag{2.48}$$

These results hold whatever the distribution of Y_i is; that is, the distribution of Y_i does not need to take on a specific form, such as the normal distribution, for Equations (2.46) through (2.48) to hold.

The notation $\sigma^2_{\overline{Y}}$ denotes the variance of the sampling distribution of the sample average $\overline{Y}$. In contrast, σ^2_Y 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 $N(\mu_Y, \sigma^2_Y)$ distribution. As stated following Equation (2.42), 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 $N(\mu_Y, \sigma_Y^2)$, 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 $n \rightarrow \infty$. 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 book.

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 $Y_i, i = 1, \ldots, n$ 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 $\overline{Y} \xrightarrow{p} \mu_Y$.

The Law of Large Numbers and Consistency

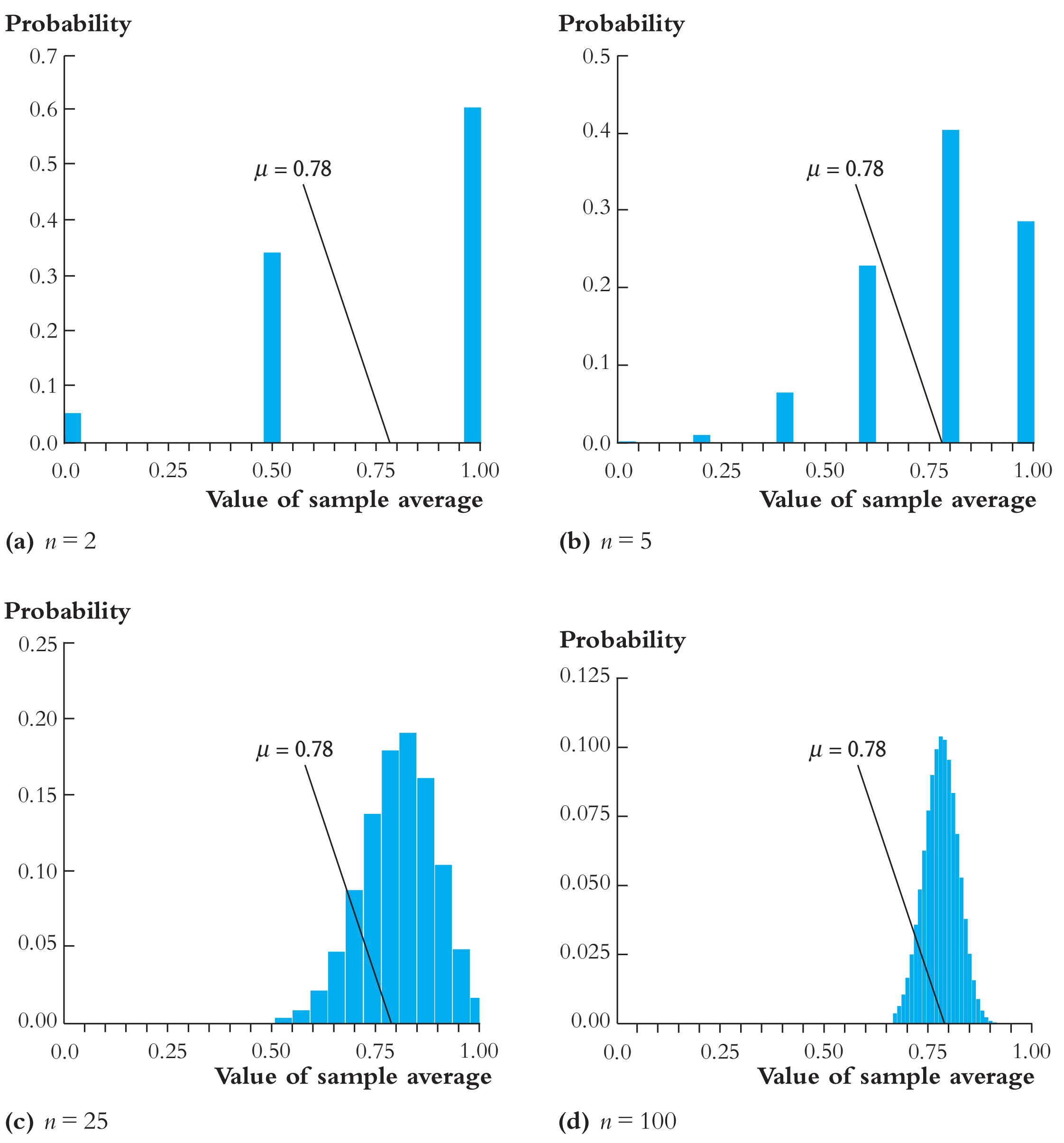
The **law of large numbers** states that, under general conditions, $\overline{Y}$ will be near μ_Y with very high probability when n is large. This is sometimes called the "law of averages." When a large number of random variables with the same mean are averaged together, the large values balance the 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 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_i, $i = 1, \ldots, n$ 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 increasing probability 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.

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.

The conditions for the law of large numbers that we will use in this book are that $Y_i, i = 1, \ldots, n$ 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 17.2, 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 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 book. 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 n is large, the distribution of $\overline{Y}$ is approximately $N(\mu_Y, \sigma_{\overline{Y}}^2)$. As discussed at the end of Section 2.5, the distribution of $\overline{Y}$ is *exactly* $N(\mu_Y, \sigma_{\overline{Y}}^2)$ when the sample is drawn from a population with the normal distribution $N(\mu_Y, \sigma_Y^2)$. 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.

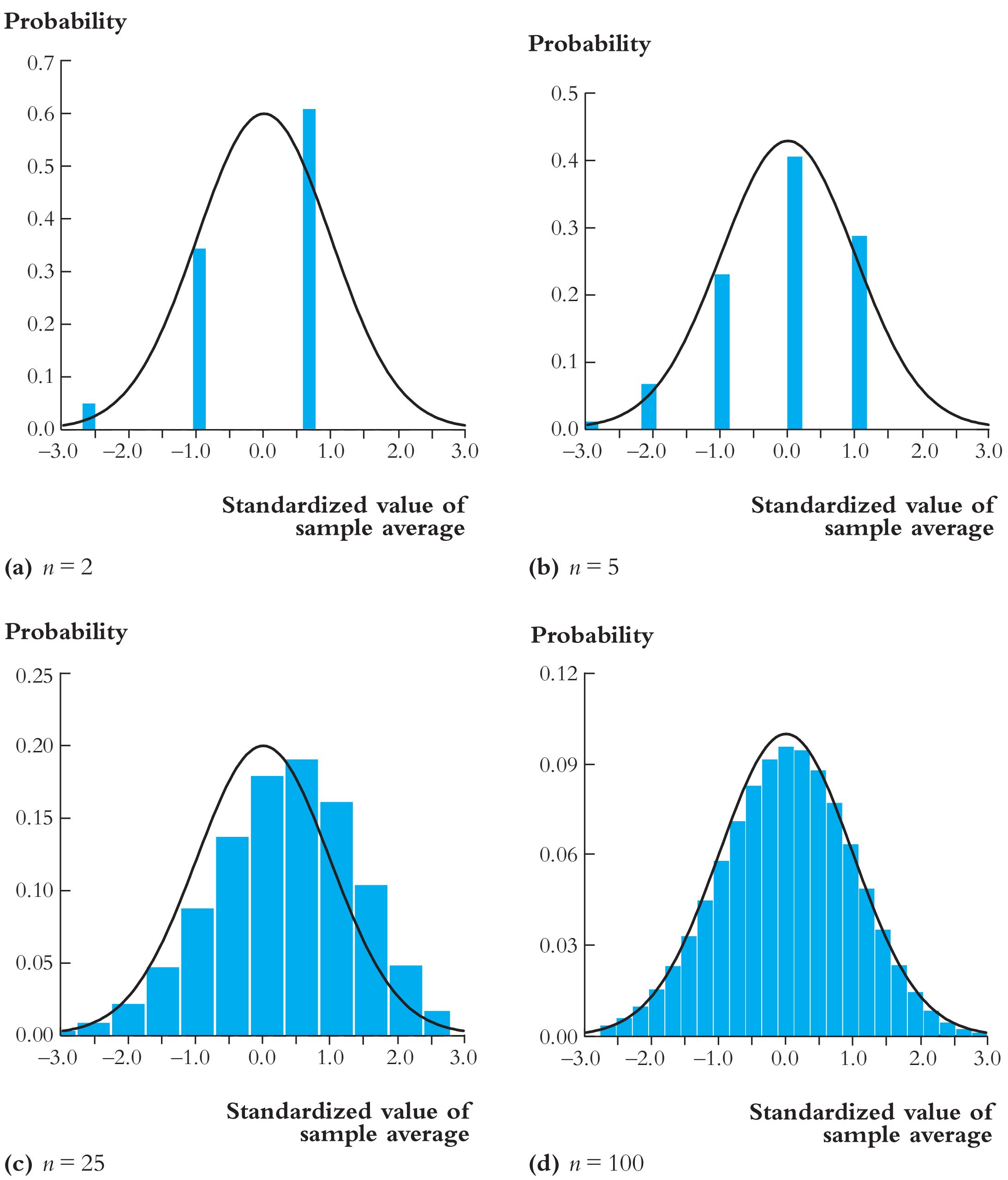
The convergence of the distribution of $\overline{Y}$ to the bell-shaped, normal approximation can be seen (a bit) in Figure 2.8. However, because the distribution gets quite tight for large n, this requires some squinting. It would be easier to see the shape of the distribution of $\overline{Y}$ if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

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



The sampling distribution of $\overline{Y}$ in Figure 2.8 is plotted here after standardizing $\overline{Y}$. This plot 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 distributions is approximately the same in all figures.

KEY CONCEPT 2.7

The Central Limit Theorem

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.

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.

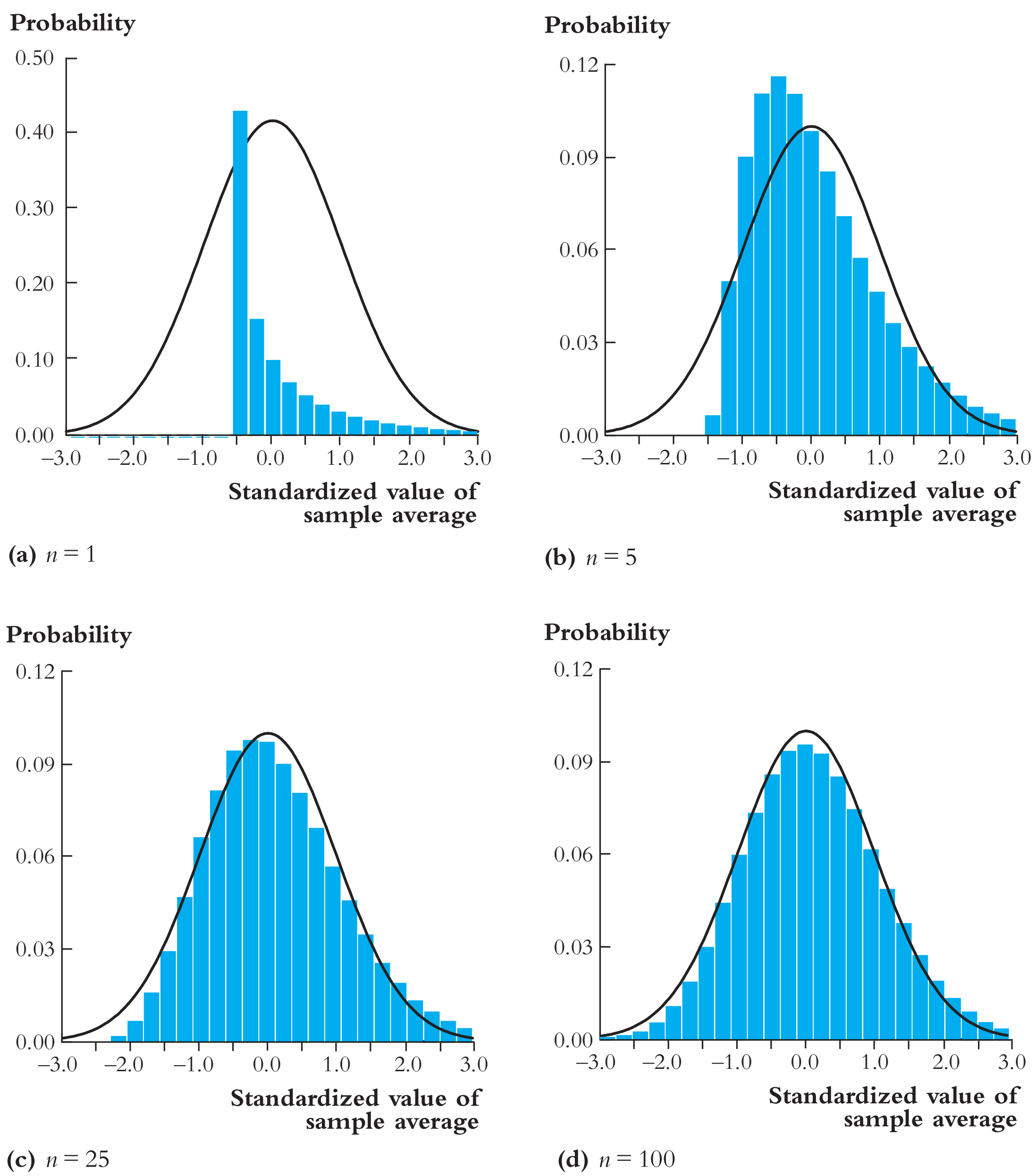
The central limit theorem is a remarkable result. While the "small n" distributions of $\overline{Y}$ in parts b and c of Figures 2.9 and 2.10 are complicated and quite different from each other, the "large n" distributions in Figures 2.9d and 2.10d are simple and, amazingly, have a similar shape. Because 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 modern applied econometrics. The central limit theorem is summarized in Key Concept 2.7.

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).

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



The figures show the sampling distribution 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 distributions is approximately the same in all figures.

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 (15)
probability (15)
sample space (15)
event (15)
discrete random variable (15)
continuous random variable (15)
probability distribution (16)
cumulative probability distribution (16)
cumulative distribution function (c.d.f.) (17)
Bernoulli random variable (17)
Bernoulli distribution (17)
probability density function (p.d.f.) (19)
density function (19)
density (19)
expected value (19)
expectation (19)
mean (19)
variance (21)
standard deviation (21)
moments of a distribution (23)
skewness (23)
kurtosis (25)
outlier (25)

leptokurtic (25)
r^{th} moment (25)
joint probability distribution (26)
marginal probability distribution (27)
conditional distribution (27)
conditional expectation (28)
conditional mean (28)
law of iterated expectations (29)
conditional variance (30)
independently distributed (31)
independent (31)
covariance (31)
correlation (32)
uncorrelated (32)
normal distribution (36)
standard normal distribution (36)
standardize a variable (36)
multivariate normal distribution (38)
bivariate normal distribution (38)
chi-squared distribution (41)
Student t distribution (41)
t distribution (42)
F distribution (42)
simple random sampling (43)
population (43)
identically distributed (44)
independently and identically distributed (i.i.d.) (44)
sample average (44)
sample mean (44)
sampling distribution (45)
exact (finite-sample) distribution (47)
asymptotic distribution (47)
law of large numbers (48)
convergence in probability (48)
consistency (48)
central limit theorem (50)
asymptotic normal distribution (52)

Review the Concepts

2.1. Examples of random variables used in this chapter included (a) the gender of the next person you meet, (b) the number of times a computer crashes, (c) the time it takes to commute to school, (d) whether the computer you are assigned in the library is new or old, and (e) 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 $P(X = 1) = p$.

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 °C?

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 2012.

Joint Distribution of Employment Status and College Graduation in the U.S. Population Aged 25 and Older, 2012

	Unemployed ($Y = 0$)	Employed ($Y = 1$)	Total
Non–college grads ($X = 0$)	0.053	0.586	0.639
College grads ($X = 1$)	0.015	0.346	0.361
Total	0.068	0.932	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_4^2, find $\Pr(Y \leq 7.78)$.

b. If Y is distributed χ_{10}^2, 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_1^2, find $\Pr(Y \leq 1.0)$. (*Hint:* Use the definition of the χ_1^2 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_i, $i = 1, 2, \ldots, n$, 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 $Y_i, i = 1, \ldots, n$, 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 $\Pr(Y = y_j) = \sum_{i=1}^{l} \Pr(Y = y_j \mid X = x_i) \Pr(X = x_i)$. [*Hint:* Use the definition of $\Pr(Y = y_j \mid X = x_i)$.]

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 $\operatorname{corr}(X, Y) = 0$.

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

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 1 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 zero.)

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_n^2.

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

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

c. $\sum_{i=1}^{n} a = na$

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

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 $\text{var}(\overline{Y}) \approx \rho\sigma_Y^2$.

2.27 X and Z are two jointly distributed random variables. Suppose you know the value of Z, but not the value of X. Let $\widetilde{X} = E(X \mid Z)$ denote a guess of the value of X using the information on Z, and let $W = X - \widetilde{X}$ denote the error associated with this guess.

a. Show that $E(W) = 0$. (*Hint:* Use the law of iterated expectations.)

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

c. Let $\hat{X} = g(Z)$ denote another guess of X using Z, and $V = X - \hat{X}$ denote its error. Show that $E(V^2) \geq E(W^2)$. [*Hint:* Let $h(Z) = g(Z) - E(X \mid Z)$, so that $V = [X - E(X \mid Z)] - h(Z)$. Derive $E(V^2)$.]

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 2012 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 parts (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 (2.29) follows from the definition of the expectation.

To derive Equation (2.30), use the definition of the variance to write $\mathrm{var}(a + bY) = E\{[a + bY - E(a + bY)]^2\} = E\{[b(Y - \mu_Y)]^2\} = b^2E[(Y - \mu_Y)^2] = b^2\sigma_Y^2$.

To derive Equation (2.31), use the definition of the variance to write

$$
\begin{aligned}
\operatorname{var}(aX + bY) &= E\{[(aX + bY) - (a\mu_X + b\mu_Y)]^2\} \\
&= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\
&= E[a^2(X - \mu_X)^2] + 2E[ab(X - \mu_X)(Y - \mu_Y)] \\
&\quad + E[b^2(Y - \mu_Y)^2] \\
&= a^2\operatorname{var}(X) + 2ab\operatorname{cov}(X, Y) + b^2\operatorname{var}(Y) \\
&= a^2\sigma_X^2 + 2ab\sigma_{XY} + b^2\sigma_Y^2, \qquad (2.49)
\end{aligned}
$$

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 (2.32), 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 (2.33), use the definition of the covariance to write

$$
\begin{aligned}
\operatorname{cov}(a + bX + cV, Y) &= E\{[a + bX + cV - E(a + bX + cV)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X) + c(V - \mu_V)][Y - \mu_Y]\} \\
&= E\{[b(X - \mu_X)][Y - \mu_Y]\} + E\{[c(V - \mu_V)][Y - \mu_Y]\} \\
&= b\sigma_{XY} + c\sigma_{VY}, \qquad (2.50)
\end{aligned}
$$

which is Equation (2.33).

To derive Equation (2.34), write $E(XY) = E\{[(X - \mu_X) + \mu_X][(Y - \mu_Y) + \mu_Y]\} = E[(X - \mu_X)(Y - \mu_Y)] + \mu_X E(Y - \mu_Y) + \mu_Y E(X - \mu_X) + \mu_X\mu_Y = \sigma_{XY} + \mu_X\mu_Y$.

We now prove the correlation inequality in Equation (2.35); that is, $|\operatorname{corr}(X, Y)| \le 1$. Let $a = -\sigma_{XY}/\sigma_X^2$ and $b = 1$. Applying Equation (2.31), we have that

$$
\begin{aligned}
\operatorname{var}(aX + Y) &= a^2\sigma_X^2 + \sigma_Y^2 + 2a\sigma_{XY} \\
&= (-\sigma_{XY}/\sigma_X^2)^2\,\sigma_X^2 + \sigma_Y^2 + 2(-\sigma_{XY}/\sigma_X^2)\sigma_{XY} \\
&= \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2. \qquad (2.51)
\end{aligned}
$$

Because $\operatorname{var}(aX + Y)$ is a variance, it cannot be negative, so from the final line of Equation (2.51), it must be that $\sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2 \ge 0$. Rearranging this inequality yields

$$\sigma_{XY}^2 \le \sigma_X^2\sigma_Y^2 \quad \text{(covariance inequality).} \qquad (2.52)$$

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

3 Review of Statistics

Statistics is the science of using data to learn about the world around us. Statistical tools help us answer questions about unknown characteristics of distributions in populations of interest. For example, what is the mean of the distribution of earnings of recent college graduates? Do mean earnings differ for men and women, and, if so, by how much?

These questions relate to the distribution of earnings in the population of workers. One way to answer these questions would be to perform an exhaustive survey of the population of workers, measuring the earnings of each worker and thus finding the population distribution of earnings. In practice, however, such a comprehensive survey would be extremely expensive. The only comprehensive survey of the U.S. population is the decennial census, which cost $13 billion to carry out in 2010. The process of designing the census forms, managing and conducting the surveys, and compiling and analyzing the data takes ten years. Despite this extraordinary commitment, many members of the population slip through the cracks and are not surveyed. Thus a different, more practical approach is needed.

The key insight of statistics is that one can learn about a population distribution by selecting a random sample from that population. Rather than survey the entire U.S. population, we might survey, say, 1000 members of the population, selected at random by simple random sampling. Using statistical methods, we can use this sample to reach tentative conclusions—to draw statistical inferences—about characteristics of the full population.

Three types of statistical methods are used throughout econometrics: estimation, hypothesis testing, and confidence intervals. Estimation entails computing a "best guess" numerical value for an unknown characteristic of a population distribution, such as its mean, from a sample of data. Hypothesis testing entails formulating a specific hypothesis about the population, then using sample evidence to decide whether it is true. Confidence intervals use a set of data to estimate an interval or range for an unknown population characteristic. Sections 3.1, 3.2, and 3.3 review estimation, hypothesis testing, and confidence intervals in the context of statistical inference about an unknown population mean.

Most of the interesting questions in economics involve relationships between two or more variables or comparisons between different populations. For example, is there a gap between the mean earnings for male and female recent college graduates? In Section 3.4, the methods for learning about the mean of a single population in Sections 3.1 through 3.3 are extended to compare means in two different populations. Section 3.5 discusses how the methods for comparing the means of two populations can be used to estimate causal effects in experiments. Sections 3.2 through 3.5 focus on the use of the normal distribution for performing hypothesis tests and for constructing confidence intervals when the sample size is large. In some special circumstances, hypothesis tests and confidence intervals can be based on the Student *t* distribution instead of the normal distribution; these special circumstances are discussed in Section 3.6. The chapter concludes with a discussion of the sample correlation and scatterplots in Section 3.7.

3.1 Estimation of the Population Mean

Suppose you want to know the mean value of Y (that is, μ_Y) in a population, such as the mean earnings of women recently graduated from college. A natural way to estimate this mean is to compute the sample average $\overline{Y}$ from a sample of n independently and identically distributed (i.i.d.) observations, $Y_1, \ldots, Y_n$ (recall that $Y_1, \ldots, Y_n$ are i.i.d. if they are collected by simple random sampling). This section discusses estimation of μ_Y and the properties of $\overline{Y}$ as an estimator of μ_Y.

Estimators and Their Properties

Estimators. The sample average $\overline{Y}$ is a natural way to estimate μ_Y, but it is not the only way. For example, another way to estimate μ_Y is simply to use the first observation, Y_1. Both $\overline{Y}$ and Y_1 are functions of the data that are designed to estimate μ_Y; using the terminology in Key Concept 3.1, both are estimators of μ_Y. When evaluated in repeated samples, $\overline{Y}$ and Y_1 take on different values (they produce different estimates) from one sample to the next. Thus the estimators $\overline{Y}$ and Y_1 both have sampling distributions. There are, in fact, many estimators of μ_Y, of which $\overline{Y}$ and Y_1 are two examples.

There are many possible estimators, so what makes one estimator "better" than another? Because estimators are random variables, this question can be phrased more precisely: What are desirable characteristics of the sampling distribution of an estimator? In general, we would like an estimator that gets as close as possible to the unknown true value, at least in some average sense; in other words, we would like the sampling distribution of an estimator to be as tightly

Estimators and Estimates

KEY CONCEPT 3.1

An **estimator** is a function of a sample of data to be drawn randomly from a population. An **estimate** is the numerical value of the estimator when it is actually computed using data from a specific sample. An estimator is a random variable because of randomness in selecting the sample, while an estimate is a nonrandom number.

centered on the unknown value as possible. This observation leads to three specific desirable characteristics of an estimator: unbiasedness (a lack of bias), consistency, and efficiency.

Unbiasedness. Suppose you evaluate an estimator many times over repeated randomly drawn samples. It is reasonable to hope that, on average, you would get the right answer. Thus a desirable property of an estimator is that the mean of its sampling distribution equals μ_Y; if so, the estimator is said to be unbiased.

To state this concept mathematically, let $\hat{\mu}_Y$ denote some estimator of μ_Y, such as $\overline{Y}$ or Y_1. The estimator $\hat{\mu}_Y$ is unbiased if $E(\hat{\mu}_Y) = \mu_Y$, where $E(\hat{\mu}_Y)$ is the mean of the sampling distribution of $\hat{\mu}_Y$; otherwise, $\hat{\mu}_Y$ is biased.

Consistency. Another desirable property of an estimator μ_Y is that, when the sample size is large, the uncertainty about the value of μ_Y arising from random variations in the sample is very small. Stated more precisely, a desirable property of $\hat{\mu}_Y$ is that the probability that it is within a small interval of the true value μ_Y approaches 1 as the sample size increases, that is, $\hat{\mu}_Y$ is consistent for μ_Y (Key Concept 2.6).

Variance and efficiency. Suppose you have two candidate estimators, $\hat{\mu}_Y$ and $\widetilde{\mu}_Y$, both of which are unbiased. How might you choose between them? One way to do so is to choose the estimator with the tightest sampling distribution. This suggests choosing between $\hat{\mu}_Y$ and $\widetilde{\mu}_Y$ by picking the estimator with the smallest variance. If $\hat{\mu}_Y$ has a smaller variance than $\widetilde{\mu}_Y$, then $\hat{\mu}_Y$ is said to be more efficient than $\widetilde{\mu}_Y$. The terminology "efficiency" stems from the notion that if $\hat{\mu}_Y$ has a smaller variance than $\widetilde{\mu}_Y$, then it uses the information in the data more efficiently than does $\widetilde{\mu}_Y$.

KEY CONCEPT 3.2

Bias, Consistency, and Efficiency

Let $\hat{\mu}_Y$ be an estimator of μ_Y. Then:

- The *bias* of $\hat{\mu}_Y$ is $E(\hat{\mu}_Y) - \mu_Y$.
- $\hat{\mu}_Y$ is an *unbiased estimator* of μ_Y if $E(\hat{\mu}_Y) = \mu_Y$.
- $\hat{\mu}_Y$ is a *consistent estimator* of μ_Y if $\hat{\mu}_Y \xrightarrow{p} \mu_Y$.
- Let $\tilde{\mu}_Y$ be another estimator of μ_Y and suppose that both $\hat{\mu}_Y$ and $\tilde{\mu}_Y$ are unbiased. Then $\hat{\mu}_Y$ is said to be more *efficient* than $\hat{\mu}_Y$ if $\text{var}(\hat{\mu}_Y) < \text{var}(\tilde{\mu}_Y)$.

Bias, **consistency**, and **efficiency** are summarized in Key Concept 3.2.

Properties of $\overline{Y}$

How does $\overline{Y}$ fare as an estimator of μ_Y when judged by the three criteria of bias, consistency, and efficiency?

Bias and consistency. The sampling distribution of $\overline{Y}$ has already been examined in Sections 2.5 and 2.6. As shown in Section 2.5, $E(\overline{Y}) = \mu_Y$, so $\overline{Y}$ is an unbiased estimator of μ_Y. Similarly, the law of large numbers (Key Concept 2.6) states that $\overline{Y} \xrightarrow{p} \mu_Y$; that is, $\overline{Y}$ is consistent.

Efficiency. What can be said about the efficiency of $\overline{Y}$? Because efficiency entails a comparison of estimators, we need to specify the estimator or estimators to which $\overline{Y}$ is to be compared.

We start by comparing the efficiency of $\overline{Y}$ to the estimator Y_1. Because $Y_1, \ldots, Y_n$ are i.i.d., the mean of the sampling distribution of Y_1 is $E(Y_1) = \mu_Y$; thus Y_1 is an unbiased estimator of μ_Y. Its variance is $\text{var}(Y_1) = \sigma_Y^2$. From Section 2.5, the variance of $\overline{Y}$ is σ_Y^2/n. Thus, for $n \geq 2$, the variance of $\overline{Y}$ is less than the variance of Y_1; that is, $\overline{Y}$ is a more efficient estimator than Y_1, so, according to the criterion of efficiency, $\overline{Y}$ should be used instead of Y_1. The estimator Y_1 might strike you as an obviously poor estimator—why would you go to the trouble of collecting a sample of n observations only to throw away all but the first?—and the concept of efficiency provides a formal way to show that $\overline{Y}$ is a more desirable estimator than Y_1.

Efficiency of $\overline{Y}$: $\overline{Y}$ Is BLUE

KEY CONCEPT 3.3

Let $\hat{\mu}_Y$ be an estimator of μ_Y that is a weighted average of $Y_1, \ldots, Y_n$, that is, $\hat{\mu}_Y = (1/n)\sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are nonrandom constants. If $\hat{\mu}_Y$ is unbiased, then $\text{var}(\overline{Y}) < \text{var}(\hat{\mu}_Y)$ unless $\hat{\mu}_Y = \overline{Y}$. Thus $\overline{Y}$ is the Best Linear Unbiased Estimator (BLUE); that is, $\overline{Y}$ is the most efficient estimator of μ_Y among all unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$.

What about a less obviously poor estimator? Consider the weighted average in which the observations are alternately weighted by $\frac{1}{2}$ and $\frac{3}{2}$:

$$\widetilde{Y} = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \frac{1}{2}Y_3 + \frac{3}{2}Y_4 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right), \tag{3.1}$$

where the number of observations n is assumed to be even for convenience. The mean of $\widetilde{Y}$ is μ_Y and its variance is $\text{var}(\widetilde{Y}) = 1.25\sigma_Y^2/n$ (Exercise 3.11). Thus $\widetilde{Y}$ is unbiased and, because $\text{var}(\widetilde{Y}) \to 0$ as $n \to \infty$, $\widetilde{Y}$ is consistent. However, $\widetilde{Y}$ has a larger variance than $\overline{Y}$. Thus $\overline{Y}$ is more efficient than $\widetilde{Y}$.

The estimators $\overline{Y}$, Y_1, and $\widetilde{Y}$ have a common mathematical structure: They are weighted averages of $Y_1, \ldots, Y_n$. The comparisons in the previous two paragraphs show that the weighted averages Y_1 and $\widetilde{Y}$ have larger variances than $\overline{Y}$. In fact, these conclusions reflect a more general result: $\overline{Y}$ is the most efficient estimator of *all* unbiased estimators that are weighted averages of $Y_1, \ldots, Y_n$. Said differently, $\overline{Y}$ is the **B**est **L**inear **U**nbiased **E**stimator (**BLUE**); that is, it is the most efficient (best) estimator among all estimators that are unbiased and are linear functions of $Y_1, \ldots, Y_n$. This result is stated in Key Concept 3.3 and is proved in Chapter 5.

$\overline{Y}$ is the least squares estimator of μ_Y. The sample average $\overline{Y}$ provides the best fit to the data in the sense that the average squared differences between the observations and $\overline{Y}$ are the smallest of all possible estimators.

Consider the problem of finding the estimator m that minimizes

$$\sum_{i=1}^{n}(Y_i - m)^2, \tag{3.2}$$

which is a measure of the total squared gap or distance between the estimator m and the sample points. Because m is an estimator of $E(Y)$, you can think of it as a

Landon Wins!

Shortly before the 1936 U.S. presidential election, the *Literary Gazette* published a poll indicating that Alf M. Landon would defeat the incumbent, Franklin D. Roosevelt, by a landslide—57% to 43%. The *Gazette* was right that the election was a landslide, but it was wrong about the winner: Roosevelt won by 59% to 41%!

How could the *Gazette* have made such a big mistake? The *Gazette*'s sample was chosen from telephone records and automobile registration files. But in 1936 many households did not have cars or telephones, and those that did tended to be richer—and were also more likely to be Republican. Because the telephone survey did not sample randomly from the population but instead undersampled Democrats, the estimator was biased and the *Gazette* made an embarrassing mistake.

Do you think surveys conducted using social media might have a similar problem with bias?

prediction of the value of Y_i, so the gap $Y_i - m$ can be thought of as a prediction mistake. The sum of squared gaps in Expression (3.2) can be thought of as the sum of squared prediction mistakes.

The estimator m that minimizes the sum of squared gaps $Y_i - m$ in Expression (3.2) is called the **least squares estimator**. One can imagine using trial and error to solve the least squares problem: Try many values of m until you are satisfied that you have the value that makes Expression (3.2) as small as possible. Alternatively, as is done in Appendix 3.2, you can use algebra or calculus to show that choosing $m = \overline{Y}$ minimizes the sum of squared gaps in Expression (3.2) so that $\overline{Y}$ is the least squares estimator of μ_Y.

The Importance of Random Sampling

We have assumed that $Y_1, \ldots, Y_n$ are i.i.d. draws, such as those that would be obtained from simple random sampling. This assumption is important because nonrandom sampling can result in $\overline{Y}$ being biased. Suppose that, to estimate the monthly national unemployment rate, a statistical agency adopts a sampling scheme in which interviewers survey working-age adults sitting in city parks at 10 A.M. on the second Wednesday of the month. Because most employed people are at work at that hour (not sitting in the park!), the unemployed are overly represented in the sample, and an estimate of the unemployment rate based on this sampling plan would be biased. This bias arises because this sampling scheme overrepresents, or oversamples, the unemployed members of the population. This example is fictitious, but the "Landon Wins!" box gives a real-world example of biases introduced by sampling that is not entirely random.

It is important to design sample selection schemes in a way that minimizes bias. Appendix 3.1 includes a discussion of what the Bureau of Labor Statistics actually does when it conducts the U.S. Current Population Survey (CPS), the survey it uses to estimate the monthly U.S. unemployment rate.

3.2 Hypothesis Tests Concerning the Population Mean

Many hypotheses about the world around us can be phrased as yes/no questions. Do the mean hourly earnings of recent U.S. college graduates equal $20 per hour? Are mean earnings the same for male and female college graduates? Both these questions embody specific hypotheses about the population distribution of earnings. The statistical challenge is to answer these questions based on a sample of evidence. This section describes **hypothesis tests** concerning the population mean (Does the population mean of hourly earnings equal $20?). Hypothesis tests involving two populations (Are mean earnings the same for men and women?) are taken up in Section 3.4.

Null and Alternative Hypotheses

The starting point of statistical hypotheses testing is specifying the hypothesis to be tested, called the **null hypothesis**. Hypothesis testing entails using data to compare the null hypothesis to a second hypothesis, called the **alternative hypothesis**, that holds if the null does not.

The null hypothesis is that the population mean, $E(Y)$, takes on a specific value, denoted $\mu_{Y,0}$. The null hypothesis is denoted H_0 and thus is

$$H_0: E(Y) = \mu_{Y,0}. \tag{3.3}$$

For example, the conjecture that, on average in the population, college graduates earn $20 per hour constitutes a null hypothesis about the population distribution of hourly earnings. Stated mathematically, if Y is the hourly earning of a randomly selected recent college graduate, then the null hypothesis is that $E(Y) = 20$; that is, $\mu_{Y,0} = 20$ in Equation (3.3).

The alternative hypothesis specifies what is true if the null hypothesis is not. The most general alternative hypothesis is that $E(Y) \neq \mu_{Y,0}$, which is called a **two-sided alternative hypothesis** because it allows $E(Y)$ to be either less than or greater than $\mu_{Y,0}$. The two-sided alternative is written as

$$H_1: E(Y) \neq \mu_{Y,0} \quad \text{(two-sided alternative).} \tag{3.4}$$

One-sided alternatives are also possible, and these are discussed later in this section.

The problem facing the statistician is to use the evidence in a randomly selected sample of data to decide whether to accept the null hypothesis H_0 or to reject it in favor of the alternative hypothesis H_1. If the null hypothesis is "accepted," this does not mean that the statistician declares it to be true; rather, it is accepted tentatively with the recognition that it might be rejected later based on additional evidence. For this reason, statistical hypothesis testing can be posed as either rejecting the null hypothesis or failing to do so.

The *p*-Value

In any given sample, the sample average $\overline{Y}$ will rarely be exactly equal to the hypothesized value $\mu_{Y,0}$. Differences between $\overline{Y}$ and $\mu_{Y,0}$ can arise because the true mean in fact does not equal $\mu_{Y,0}$ (the null hypothesis is false) or because the true mean equals $\mu_{Y,0}$ (the null hypothesis is true) but $\overline{Y}$ differs from $\mu_{Y,0}$ because of random sampling. It is impossible to distinguish between these two possibilities with certainty. Although a sample of data cannot provide conclusive evidence about the null hypothesis, it is possible to do a probabilistic calculation that permits testing the null hypothesis in a way that accounts for sampling uncertainty. This calculation involves using the data to compute the p-value of the null hypothesis.

The ***p*-value**, also called the **significance probability**, is the probability of drawing a statistic at least as adverse to the null hypothesis as the one you actually computed in your sample, assuming the null hypothesis is correct. In the case at hand, the p-value is the probability of drawing $\overline{Y}$ at least as far in the tails of its distribution under the null hypothesis as the sample average you actually computed.

For example, suppose that, in your sample of recent college graduates, the average wage is \$22.64. The p-value is the probability of observing a value of $\overline{Y}$ at least as different from \$20 (the population mean under the null) as the observed value of \$22.64 by pure random sampling variation, assuming that the null hypothesis is true. If this p-value is small, say 0.5%, then it is very unlikely that this sample would have been drawn if the null hypothesis is true; thus it is reasonable to conclude that the null hypothesis is not true. By contrast, if this p-value is large, say 40%, then it is quite likely that the observed sample average of \$22.64 could have arisen just by random sampling variation if the null hypothesis is true; accordingly, the evidence against the null hypothesis is weak in this probabilistic sense, and it is reasonable not to reject the null hypothesis.

To state the definition of the p-value mathematically, let $\overline{Y}^{act}$ denote the value of the sample average actually computed in the data set at hand and let $\Pr_{H_0}$

denote the probability computed under the null hypothesis (that is, computed assuming that $E(Y_i) = \mu_{Y,0}$). The p-value is

$$p\text{-value} = \Pr_{H_0}\left[\,|\overline{Y} - \mu_{Y,0}| > |\overline{Y}^{act} - \mu_{Y,0}|\,\right]. \tag{3.5}$$

That is, the p-value is the area in the tails of the distribution of $\overline{Y}$ under the null hypothesis beyond $\mu_{Y,0} \pm |\overline{Y}^{act} - \mu_{Y,0}|$. If the p-value is large, then the observed value $\overline{Y}^{act}$ is consistent with the null hypothesis, but if the p-value is small, it is not.

To compute the p-value, it is necessary to know the sampling distribution of $\overline{Y}$ under the null hypothesis. As discussed in Section 2.6, when the sample size is small this distribution is complicated. However, according to the central limit theorem, when the sample size is large, the sampling distribution of $\overline{Y}$ is well approximated by a normal distribution. Under the null hypothesis the mean of this normal distribution is $\mu_{Y,0}$, so under the null hypothesis $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$, where $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$. This large-sample normal approximation makes it possible to compute the p-value without needing to know the population distribution of Y, as long as the sample size is large. The details of the calculation, however, depend on whether σ^2_Y is known.

Calculating the p-Value When σ_Y Is Known

The calculation of the p-value when σ_Y is known is summarized in Figure 3.1. If the sample size is large, then under the null hypothesis the sampling distribution of $\overline{Y}$ is $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$, where $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$. Thus, under the null hypothesis, the standardized version of $\overline{Y}$, $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, has a standard normal distribution. The p-value is the probability of obtaining a value of $\overline{Y}$ farther from $\mu_{Y,0}$ than $\overline{Y}^{act}$ under the null hypothesis or, equivalently, is the probability of obtaining $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ greater than $(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}$ in absolute value. This probability is the shaded area shown in Figure 3.1. Written mathematically, the shaded tail probability in Figure 3.1 (that is, the p-value) is

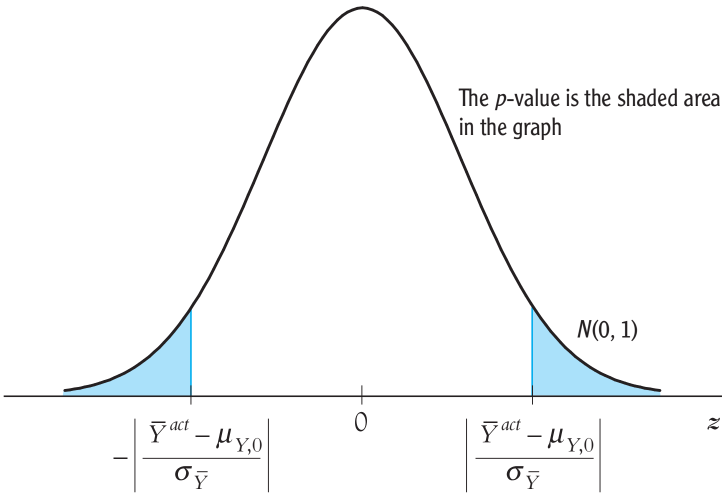
$$p\text{-value} = \Pr_{H_0}\left(\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right| > \left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right), \tag{3.6}$$

where Φ is the standard normal cumulative distribution function. That is, the p-value is the area in the tails of a standard normal distribution outside $\pm|\overline{Y}^{act} - \mu_{Y,0}|/\sigma_{\overline{Y}}$.

The formula for the p-value in Equation (3.6) depends on the variance of the population distribution, σ^2_Y. In practice, this variance is typically unknown. [An exception is when Y_i is binary so that its distribution is Bernoulli, in which case

FIGURE 3.1 Calculating a *p*-value

The p-value is the probability of drawing a value of $\overline{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\overline{Y}^{act}$. In large samples, $\overline{Y}$ is distributed $N(\mu_{Y,0}, \sigma^2_{\overline{Y}})$, under the null hypothesis, so $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$ is distributed $N(0, 1)$. Thus the p-value is the shaded standard normal tail probability outside $\pm|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}|$.



the variance is determined by the null hypothesis; see Equation (2.7) and Exercise 3.2.] Because in general σ^2_Y must be estimated before the p-value can be computed, we now turn to the problem of estimating σ^2_Y.

The Sample Variance, Sample Standard Deviation, and Standard Error

The sample variance s^2_Y is an estimator of the population variance σ^2_Y, the sample standard deviation s_Y is an estimator of the population standard deviation σ_Y, and the standard error of the sample average $\overline{Y}$ is an estimator of the standard deviation of the sampling distribution of $\overline{Y}$.

The sample variance and standard deviation. The **sample variance**, s^2_Y, is

$$s^2_Y = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{3.7}$$

The **sample standard deviation**, s_Y, is the square root of the sample variance.

The formula for the sample variance is much like the formula for the population variance. The population variance, $E(Y - \mu_Y)^2$, is the average value of $(Y - \mu_Y)^2$ in the population distribution. Similarly, the sample variance is the sample average of $(Y_i - \mu_Y)^2$, $i = 1, \ldots, n$, with two modifications: First, μ_Y is replaced by $\overline{Y}$, and second, the average uses the divisor $n - 1$ instead of n.

The Standard Error of $\overline{Y}$

KEY CONCEPT 3.4

The standard error of $\overline{Y}$ is an estimator of the standard deviation of $\overline{Y}$. The standard error of $\overline{Y}$ is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$. When $Y_1, \ldots, Y_n$ are i.i.d.,

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}. \quad (3.8)$$

The reason for the first modification—replacing μ_Y by $\overline{Y}$—is that μ_Y is unknown and thus must be estimated; the natural estimator of μ_Y is $\overline{Y}$. The reason for the second modification—dividing by $n - 1$ instead of by n—is that estimating μ_Y by $\overline{Y}$ introduces a small downward bias in $(Y_i - \overline{Y})^2$. Specifically, as is shown in Exercise 3.18, $E[(Y_i - \overline{Y})^2] = [(n-1)/n]\sigma_Y^2$. Thus $E\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = nE[(Y_i - \overline{Y})^2] = (n-1)\sigma_Y^2$. Dividing by $n - 1$ in Equation (3.7) instead of n corrects for this small downward bias, and as a result s_Y^2 is unbiased.

Dividing by $n - 1$ in Equation (3.7) instead of n is called a **degrees of freedom** correction: Estimating the mean uses up some of the information—that is, uses up 1 "degree of freedom"—in the data, so that only $n - 1$ degrees of freedom remain.

Consistency of the sample variance. The sample variance is a consistent estimator of the population variance:

$$s_Y^2 \xrightarrow{\;p\;} \sigma_Y^2. \quad (3.9)$$

In other words, the sample variance is close to the population variance with high probability when n is large.

The result in Equation (3.9) is proven in Appendix 3.3 under the assumptions that $Y_1, \ldots, Y_n$ are i.i.d. and Y_i has a finite fourth moment; that is, $E(Y_i^4) < \infty$. Intuitively, the reason that s_Y^2 is consistent is that it is a sample average, so s_Y^2 obeys the law of large numbers. But for s_Y^2 to obey the law of large numbers in Key Concept 2.6, $(Y_i - \mu_Y)^2$ must have finite variance, which in turn means that $E(Y_i^4)$ must be finite; in other words, Y_i must have a finite fourth moment.

The standard error of $\overline{Y}$. Because the standard deviation of the sampling distribution of $\overline{Y}$ is $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$, Equation (3.9) justifies using $s_Y/\sqrt{n}$ as an estimator of $\sigma_{\overline{Y}}$. The estimator of $\sigma_{\overline{Y}}$, $s_Y/\sqrt{n}$, is called the **standard error of $\overline{Y}$** and is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$ (the caret "^" over the symbol means that it is an estimator of $\sigma_{\overline{Y}}$). The standard error of $\overline{Y}$ is summarized as in Key Concept 3.4.

When $Y_1, \ldots, Y_n$ are i.i.d. draws from a Bernoulli distribution with success probability p, the formula for the variance of $\overline{Y}$ simplifies to $p(1 - p)/n$ (see Exercise 3.2). The formula for the standard error also takes on a simple form that depends only on $\overline{Y}$ and n: $SE(\overline{Y}) = \sqrt{\overline{Y}(1 - \overline{Y})/n}$.

Calculating the p-Value When σ_Y Is Unknown

Because s_Y^2 is a consistent estimator of σ_Y^2, the p-value can be computed by replacing $\sigma_{\overline{Y}}$ in Equation (3.6) by the standard error, $SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}}$. That is, when σ_Y is unknown and $Y_1, \ldots, Y_n$ are i.i.d., the p-value is calculated using the formula

$$p\text{-value} = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}\right|\right). \tag{3.10}$$

The t-Statistic

The standardized sample average $(\overline{Y} - \mu_{Y,0})/SE(\overline{Y})$ plays a central role in testing statistical hypotheses and has a special name, the ***t*-statistic** or ***t*-ratio**:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{3.11}$$

In general, a **test statistic** is a statistic used to perform a hypothesis test. The t-statistic is an important example of a test statistic.

Large-sample distribution of the* t*-statistic. When n is large, s_Y^2 is close to σ_Y^2 with high probability. Thus the distribution of the t-statistic is approximately the same as the distribution of $(\overline{Y} - \mu_{Y,0})/\sigma_{\overline{Y}}$, which in turn is well approximated by the standard normal distribution when n is large because of the central limit theorem (Key Concept 2.7). Accordingly, under the null hypothesis,

$$t \text{ is approximately distributed } N(0,1) \text{ for large } n. \tag{3.12}$$

The formula for the p-value in Equation (3.10) can be rewritten in terms of the t-statistic. Let t^{act} denote the value of the t-statistic actually computed:

$$t^{act} = \frac{\overline{Y}^{act} - \mu_{Y,0}}{SE(\overline{Y})}. \tag{3.13}$$

Accordingly, when n is large, the p-value can be calculated using

$$p\text{-value} = 2\Phi(-|t^{act}|). \tag{3.14}$$

As a hypothetical example, suppose that a sample of $n = 200$ recent college graduates is used to test the null hypothesis that the mean wage, $E(Y)$, is \$20 per hour. The sample average wage is $\overline{Y}^{act} = \$22.64$, and the sample standard deviation is $s_Y = \$18.14$. Then the standard error of $\overline{Y}$ is $s_Y/\sqrt{n} = 18.14/\sqrt{200} = 1.28$. The value of the t-statistic is $t^{act} = (22.64 - 20)/1.28 = 2.06$. From Appendix Table 1, the p-value is $2\Phi(-2.06) = 0.039$, or 3.9%. That is, assuming the null hypothesis to be true, the probability of obtaining a sample average at least as different from the null as the one actually computed is 3.9%.

Hypothesis Testing with a Prespecified Significance Level

When you undertake a statistical hypothesis test, you can make two types of mistakes: You can incorrectly reject the null hypothesis when it is true, or you can fail to reject the null hypothesis when it is false. Hypothesis tests can be performed without computing the p-value if you are willing to specify in advance the probability you are willing to tolerate of making the first kind of mistake—that is, of incorrectly rejecting the null hypothesis when it is true. If you choose a prespecified probability of rejecting the null hypothesis when it is true (for example, 5%), then you will reject the null hypothesis if and only if the p-value is less than 0.05. This approach gives preferential treatment to the null hypothesis, but in many practical situations this preferential treatment is appropriate.

Hypothesis tests using a fixed significance level. Suppose it has been decided that the hypothesis will be rejected if the p-value is less than 5%. Because the area under the tails of the standard normal distribution outside ± 1.96 is 5%, this gives a simple rule:

$$\text{Reject } H_0 \text{ if } |t^{act}| > 1.96. \tag{3.15}$$

That is, reject if the absolute value of the t-statistic computed from the sample is greater than 1.96. If n is large enough, then under the null hypothesis the t-statistic has a $N(0, 1)$ distribution. Thus the probability of erroneously rejecting the null hypothesis (rejecting the null hypothesis when it is in fact true) is 5%.

KEY CONCEPT 3.5

The Terminology of Hypothesis Testing

A statistical hypothesis test can make two types of mistakes: a **type I error**, in which the null hypothesis is rejected when in fact it is true, and a **type II error**, in which the null hypothesis is not rejected when in fact it is false. The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true—that is, the prespecified probability of a type I error—is the **significance level** of the test. The **critical value** of the test statistic is the value of the statistic for which the test just rejects the null hypothesis at the given significance level. The set of values of the test statistic for which the test rejects the null hypothesis is the **rejection region**, and the values of the test statistic for which it does not reject the null hypothesis is the **acceptance region**. The probability that the test actually incorrectly rejects the null hypothesis when it is true is the **size of the test**, and the probability that the test correctly rejects the null hypothesis when the alternative is true is the **power of the test**.

The p-value is the probability of obtaining a test statistic, by random sampling variation, at least as adverse to the null hypothesis value as is the statistic actually observed, assuming that the null hypothesis is correct. Equivalently, the p-value is the smallest significance level at which you can reject the null hypothesis.

This framework for testing statistical hypotheses has some specialized terminology, summarized in Key Concept 3.5. The significance level of the test in Equation (3.15) is 5%, the critical value of this two-sided test is 1.96, and the rejection region is the values of the t-statistic outside ±1.96. If the test rejects at the 5% significance level, the population mean μ_Y is said to be statistically significantly different from $\mu_{Y,0}$ at the 5% significance level.

Testing hypotheses using a prespecified significance level does not require computing p-values. In the previous example of testing the hypothesis that the mean earnings of recent college graduates is \$20 per hour, the t-statistic was 2.06. This value exceeds 1.96, so the hypothesis is rejected at the 5% level. Although performing the test with a 5% significance level is easy, reporting only whether the null hypothesis is rejected at a prespecified significance level conveys less information than reporting the p-value.

What significance level should you use in practice? In many cases, statisticians and econometricians use a 5% significance level. If you were to test many statistical

KEY CONCEPT 3.6

Testing the Hypothesis $E(Y) = \mu_{Y,0}$ Against the Alternative $E(Y) \neq \mu_{Y,0}$

1. Compute the standard error of $\overline{Y}$, $SE(\overline{Y})$ [Equation (3.8)].
2. Compute the t-statistic [Equation (3.13)].
3. Compute the p-value [Equation (3.14)]. Reject the hypothesis at the 5% significance level if the p-value is less than 0.05 (equivalently, if $|t^{act}| > 1.96$).

hypotheses at the 5% level, you would incorrectly reject the null on average once in 20 cases. Sometimes a more conservative significance level might be in order. For example, legal cases sometimes involve statistical evidence, and the null hypothesis could be that the defendant is not guilty; then one would want to be quite sure that a rejection of the null (conclusion of guilt) is not just a result of random sample variation. In some legal settings, the significance level used is 1%, or even 0.1%, to avoid this sort of mistake. Similarly, if a government agency is considering permitting the sale of a new drug, a very conservative standard might be in order so that consumers can be sure that the drugs available in the market actually work.

Being conservative, in the sense of using a very low significance level, has a cost: The smaller the significance level, the larger the critical value and the more difficult it becomes to reject the null when the null is false. In fact, the most conservative thing to do is never to reject the null hypothesis—but if that is your view, then you never need to look at any statistical evidence for you will never change your mind! The lower the significance level, the lower the power of the test. Many economic and policy applications can call for less conservatism than a legal case, so a 5% significance level is often considered to be a reasonable compromise.

Key Concept 3.6 summarizes hypothesis tests for the population mean against the two-sided alternative.

One-Sided Alternatives

In some circumstances, the alternative hypothesis might be that the mean exceeds $\mu_{Y,0}$. For example, one hopes that education helps in the labor market, so the relevant alternative to the null hypothesis that earnings are the same for college graduates and non–college graduates is not just that their earnings differ, but rather that graduates earn more than nongraduates. This is called a **one-sided alternative hypothesis** and can be written

$$H_1\colon E(Y) > \mu_{Y,0} \quad \text{(one-sided alternative)}. \tag{3.16}$$

The general approach to computing p-values and to hypothesis testing is the same for one-sided alternatives as it is for two-sided alternatives, with the modification that only large positive values of the t-statistic reject the null hypothesis rather than values that are large in absolute value. Specifically, to test the one-sided hypothesis in Equation (3.16), construct the t-statistic in Equation (3.13). The p-value is the area under the standard normal distribution to the right of the calculated t-statistic. That is, the p-value, based on the $N(0, 1)$ approximation to the distribution of the t-statistic, is

$$p\text{-value} = \Pr_{H_0}(Z > t^{act}) = 1 - \Phi(t^{act}). \tag{3.17}$$

The $N(0, 1)$ critical value for a one-sided test with a 5% significance level is 1.64. The rejection region for this test is all values of the t-statistic exceeding 1.64.

The one-sided hypothesis in Equation (3.16) concerns values of μ_Y exceeding $\mu_{Y,0}$. If instead the alternative hypothesis is that $E(Y) < \mu_{Y,0}$, then the discussion of the previous paragraph applies except that the signs are switched; for example, the 5% rejection region consists of values of the t-statistic less than -1.64.

3.3 Confidence Intervals for the Population Mean

Because of random sampling error, it is impossible to learn the exact value of the population mean of Y using only the information in a sample. However, it is possible to use data from a random sample to construct a set of values that contains the true population mean μ_Y with a certain prespecified probability. Such a set is called a **confidence set**, and the prespecified probability that μ_Y is contained in this set is called the **confidence level**. The confidence set for μ_Y turns out to be all the possible values of the mean between a lower and an upper limit, so that the confidence set is an interval, called a **confidence interval**.

Here is one way to construct a 95% confidence set for the population mean. Begin by picking some arbitrary value for the mean; call it $\mu_{Y,0}$. Test the null hypothesis that $\mu_Y = \mu_{Y,0}$ against the alternative that $\mu_Y \neq \mu_{Y,0}$ by computing the t-statistic; if its absolute value is less than 1.96, this hypothesized value $\mu_{Y,0}$ is not rejected at the 5% level, and write down this nonrejected value $\mu_{Y,0}$. Now pick another arbitrary value of $\mu_{Y,0}$ and test it; if you cannot reject it, write down this value on your list.

Confidence Intervals for the Population Mean

KEY CONCEPT 3.7

A 95% two-sided confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible random samples. When the sample size n is large, 95%, 90%, and 99% confidence intervals for μ_Y are

$$95\%\text{ confidence interval for }\mu_Y = \{\overline{Y} \pm 1.96SE(\overline{Y})\}.$$

$$90\%\text{ confidence interval for }\mu_Y = \{\overline{Y} \pm 1.64SE(\overline{Y})\}.$$

$$99\%\text{ confidence interval for }\mu_Y = \{\overline{Y} \pm 2.58SE(\overline{Y})\}.$$

Do this again and again; indeed, do so for all possible values of the population mean. Continuing this process yields the set of all values of the population mean that cannot be rejected at the 5% level by a two-sided hypothesis test.

This list is useful because it summarizes the set of hypotheses you can and cannot reject (at the 5% level) based on your data: If someone walks up to you with a specific number in mind, you can tell him whether his hypothesis is rejected or not simply by looking up his number on your handy list. A bit of clever reasoning shows that this set of values has a remarkable property: The probability that it contains the true value of the population mean is 95%.

The clever reasoning goes like this: Suppose the true value of μ_Y is 21.5 (although we do not know this). Then $\overline{Y}$ has a normal distribution centered on 21.5, and the t-statistic testing the null hypothesis $\mu_Y = 21.5$ has a $N(0, 1)$ distribution. Thus, if n is large, the probability of rejecting the null hypothesis $\mu_Y = 21.5$ at the 5% level is 5%. But because you tested all possible values of the population mean in constructing your set, in particular you tested the true value, $\mu_Y = 21.5$. In 95% of all samples, you will correctly accept 21.5; this means that in 95% of all samples, your list will contain the true value of μ_Y. Thus the values on your list constitute a 95% confidence set for μ_Y.

This method of constructing a confidence set is impractical, for it requires you to test all possible values of μ_Y as null hypotheses. Fortunately, there is a much easier approach. According to the formula for the t-statistic in Equation (3.13), a trial value of $\mu_{Y,0}$ is rejected at the 5% level if it is more than 1.96 standard errors away from $\overline{Y}$. Thus the set of values of μ_Y that are not rejected at the 5% level consists of those values within $\pm 1.96SE(\overline{Y})$ of $\overline{Y}$; that is, a 95% confidence interval for μ_Y is $\overline{Y} - 1.96SE(\overline{Y}) \le \mu_Y \le \overline{Y} + 1.96SE(\overline{Y})$. Key Concept 3.7 summarizes this approach.

As an example, consider the problem of constructing a 95% confidence interval for the mean hourly earnings of recent college graduates using a hypothetical random sample of 200 recent college graduates where $\overline{Y} = \$22.64$ and $SE(\overline{Y}) = 1.28$. The 95% confidence interval for mean hourly earnings is $22.64 \pm 1.96 \times 1.28 = 22.64 \pm 2.51 = [\$20.13, \$25.15]$.

This discussion so far has focused on two-sided confidence intervals. One could instead construct a one-sided confidence interval as the set of values of μ_Y that cannot be rejected by a one-sided hypothesis test. Although one-sided confidence intervals have applications in some branches of statistics, they are uncommon in applied econometric analysis.

Coverage probabilities. The **coverage probability** of a confidence interval for the population mean is the probability, computed over all possible random samples, that it contains the true population mean.

3.4 Comparing Means from Different Populations

Do recent male and female college graduates earn the same amount on average? This question involves comparing the means of two different population distributions. This section summarizes how to test hypotheses and how to construct confidence intervals for the difference in the means from two different populations.

Hypothesis Tests for the Difference Between Two Means

To illustrate a **test for the difference between two means**, let μ_w be the mean hourly earning in the population of women recently graduated from college and let μ_m be the population mean for recently graduated men. Consider the null hypothesis that mean earnings for these two populations differ by a certain amount, say d_0. Then the null hypothesis and the two-sided alternative hypothesis are

$$H_0: \mu_m - \mu_w = d_0 \text{ vs. } H_1: \mu_m - \mu_w \neq d_0. \tag{3.18}$$

The null hypothesis that men and women in these populations have the same mean earnings corresponds to H_0 in Equation (3.18) with $d_0 = 0$.

Because these population means are unknown, they must be estimated from samples of men and women. Suppose we have samples of n_m men and n_w women drawn at random from their populations. Let the sample average annual earnings be $\overline{Y}_m$ for men and $\overline{Y}_w$ for women. Then an estimator of $\mu_m - \mu_w$ is $\overline{Y}_m - \overline{Y}_w$.

To test the null hypothesis that $\mu_m - \mu_w = d_0$ using $\overline{Y}_m - \overline{Y}_w$, we need to know the distribution of $\overline{Y}_m - \overline{Y}_w$. Recall that $\overline{Y}_m$ is, according to the central limit theorem, approximately distributed $N(\mu_m, \sigma_m^2/n_m)$, where σ_m^2 is the population variance of earnings for men. Similarly, $\overline{Y}_w$ is approximately distributed $N(\mu_w, \sigma_w^2/n_w)$ where σ_w^2 is the population variance of earnings for women. Also, recall from Section 2.4 that a weighted average of two normal random variables is itself normally distributed. Because $\overline{Y}_m$ and $\overline{Y}_w$ are constructed from different randomly selected samples, they are independent random variables. Thus $\overline{Y}_m - \overline{Y}_w$ is distributed $N[\mu_m - \mu_w, (\sigma_m^2/n_m) + (\sigma_w^2/n_w)]$.

If σ_m^2 and σ_w^2 are known, then this approximate normal distribution can be used to compute p-values for the test of the null hypothesis that $\mu_m - \mu_w = d_0$. In practice, however, these population variances are typically unknown so they must be estimated. As before, they can be estimated using the sample variances, s_m^2 and s_w^2 where s_m^2 is defined as in Equation (3.7), except that the statistic is computed only for the men in the sample, and s_w^2 is defined similarly for the women. Thus the standard error of $\overline{Y}_m - \overline{Y}_w$ is

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}. \tag{3.19}$$

For a simplified version of Equation (3.19) when Y is a Bernoulli random variable, see Exercise 3.15.

The t-statistic for testing the null hypothesis is constructed analogously to the t-statistic for testing a hypothesis about a single population mean, by subtracting the null hypothesized value of $\mu_m - \mu_w$ from the estimator $\overline{Y}_m - \overline{Y}_w$ and dividing the result by the standard error of $\overline{Y}_m - \overline{Y}_w$:

$$t = \frac{(\overline{Y}_m - \overline{Y}_w) - d_0}{SE(\overline{Y}_m - \overline{Y}_w)} \quad (t\text{-statistic for comparing two means}). \tag{3.20}$$

If both n_m and n_w are large, then this t-statistic has a standard normal distribution when the null hypothesis is true.

Because the t-statistic in Equation (3.20) has a standard normal distribution under the null hypothesis when n_m and n_w are large, the p-value of the two-sided test is computed exactly as it was in the case of a single population. That is, the p-value is computed using Equation (3.14).

To conduct a test with a prespecified significance level, simply calculate the t-statistic in Equation (3.20) and compare it to the appropriate critical value. For example, the null hypothesis is rejected at the 5% significance level if the absolute value of the t-statistic exceeds 1.96.

If the alternative is one-sided rather than two-sided (that is, if the alternative is that $\mu_m - \mu_w > d_0$), then the test is modified as outlined in Section 3.2. The p-value is computed using Equation (3.17), and a test with a 5% significance level rejects when $t > 1.64$.

Confidence Intervals for the Difference Between Two Population Means

The method for constructing confidence intervals summarized in Section 3.3 extends to constructing a confidence interval for the difference between the means, $d = \mu_m - \mu_w$. Because the hypothesized value d_0 is rejected at the 5% level if $|t| > 1.96$, d_0 will be in the confidence set if $|t| \leq 1.96$. But $|t| \leq 1.96$ means that the estimated difference, $\overline{Y}_m - \overline{Y}_w$, is less than 1.96 standard errors away from d_0. Thus the 95% two-sided confidence interval for d consists of those values of d within $\pm$ 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$:

$$\text{95\% confidence interval for } d = \mu_m - \mu_w \text{ is}$$
$$(\overline{Y}_m - \overline{Y}_w) \pm 1.96SE(\overline{Y}_m - \overline{Y}_w). \tag{3.21}$$

With these formulas in hand, the box "The Gender Gap of Earnings of College Graduates in the United States" contains an empirical investigation of gender differences in earnings of U.S. college graduates.

3.5 Differences-of-Means Estimation of Causal Effects Using Experimental Data

Recall from Section 1.2 that a randomized controlled experiment randomly selects subjects (individuals or, more generally, entities) from a population of interest, then randomly assigns them either to a treatment group, which receives the experimental treatment, or to a control group, which does not receive the treatment. The difference between the sample means of the treatment and control groups is an estimator of the causal effect of the treatment.

The Causal Effect as a Difference of Conditional Expectations

The causal effect of a treatment is the expected effect on the outcome of interest of the treatment as measured in an ideal randomized controlled experiment. This effect can be expressed as the difference of two conditional expectations. Specifically, the **causal effect** on Y of treatment level x is the difference in the conditional expectations, $E(Y|X = x) - E(Y|X = 0)$, where $E(Y|X = x)$ is the expected value of Y for the treatment group (which receives treatment level $X = x$) in an ideal randomized controlled experiment and $E(Y|X = 0)$ is the expected value of Y for the control group (which receives treatment level $X = 0$). In the context of experiments, the causal effect is also called the **treatment effect**. If there are only two treatment levels (that is, if the treatment is binary), then we can let $X = 0$ denote the control group and $X = 1$ denote the treatment group. If the treatment is binary treatment, then the causal effect (that is, the treatment effect) is $E(Y|X = 1) - E(Y|X = 0)$ in an ideal randomized controlled experiment.

Estimation of the Causal Effect Using Differences of Means

If the treatment in a randomized controlled experiment is binary, then the causal effect can be estimated by the difference in the sample average outcomes between the treatment and control groups. The hypothesis that the treatment is ineffective is equivalent to the hypothesis that the two means are the same, which can be tested using the t-statistic for comparing two means, given in Equation (3.20). A 95% confidence interval for the difference in the means of the two groups is a 95% confidence interval for the causal effect, so a 95% confidence interval for the causal effect can be constructed using Equation (3.21).

A well-designed, well-run experiment can provide a compelling estimate of a causal effect. For this reason, randomized controlled experiments are commonly conducted in some fields, such as medicine. In economics, however, experiments tend to be expensive, difficult to administer, and, in some cases, ethically questionable, so they are used less often. For this reason, econometricians sometimes study "natural experiments," also called quasi-experiments, in which some event unrelated to the treatment or subject characteristics has the effect of assigning different treatments to different subjects *as if* they had been part of a randomized controlled experiment. The box "A Novel Way to Boost Retirement Savings" provides an example of such a quasi-experiment that yielded some surprising conclusions.

The Gender Gap of Earnings of College Graduates in the United States

The box in Chapter 2 "The Distribution of Earnings in the United States in 2012" shows that, on average, male college graduates earn more than female college graduates. What are the recent trends in this "gender gap" in earnings? Social norms and laws governing gender discrimination in the workplace have changed substantially in the United States. Is the gender gap in earnings of college graduates stable, or has it diminished over time?

Table 3.1 gives estimates of hourly earnings for college-educated full-time workers ages 25–34 in the United States in 1992, 1996, 2000, 2004, 2008, and 2012, using data collected by the Current Population Survey. Earnings for 1992, 1996, 2000, 2004, and 2008 were adjusted for inflation by putting them in 2012 dollars using the Consumer Price Index (CPI).[1] In 2012, the average hourly earnings of the 2004 men surveyed was \$25.30, and the standard deviation of earnings for men was \$12.09. The average hourly earnings in 2012 of the 1951 women surveyed was \$21.50, and the standard deviation of earnings was \$9.99. Thus the estimate of the gender gap in earnings for 2012 is \$3.80 (= \$25.30 − \$21.50), with a standard error of \$0.35 ($= \sqrt{12.09^2/2004 + 9.99^2/1951}$). The 95% confidence interval for the gender gap in earnings in 2012 is $3.80 \pm 1.96 \times 0.35 =$ (\$3.11, \$4.49).

The results in Table 3.1 suggest four conclusions. First, the gender gap is large. An hourly gap of \$3.80 might not sound like much, but over a year it adds up to \$7600, assuming a 40-hour workweek and 50 paid weeks per year. Second, from 1992 to 2012, the estimated gender gap increased by \$0.36 per hour in real terms, from \$3.44 per hour to \$3.80 per hour;

TABLE 3.1 Trends in Hourly Earnings in the United States of Working College Graduates, Ages 25–34, 1992 to 2012, in 2012 Dollars

	Men			Women			Difference, Men vs. Women		
Year	$\overline{Y}_m$	s_m	n_m	$\overline{Y}_w$	s_w	n_w	$\overline{Y}_m - \overline{Y}_w$	$SE(\overline{Y}_m - \overline{Y}_w)$	95% Confidence Interval for d
1992	24.83	10.85	1594	21.39	8.39	1368	3.44**	0.35	2.75–4.14
1996	23.97	10.79	1380	20.26	8.48	1230	3.71**	0.38	2.97–4.46
2000	26.55	12.38	1303	22.13	9.98	1181	4.42**	0.45	3.54–5.30
2004	26.80	12.81	1894	22.43	9.99	1735	4.37**	0.38	3.63–5.12
2008	26.63	12.57	1839	22.26	10.30	1871	4.36**	0.38	3.62–5.10
2012	25.30	12.09	2004	21.50	9.99	1951	3.80**	0.35	3.11–4.49

These estimates are computed using data on all full-time workers ages 25–34 surveyed in the Current Population Survey conducted in March of the next year (for example, the data for 2012 were collected in March 2013). The difference is significantly different from zero at the **1% significance level.

(*continued*)

however, this increase is not statistically significant at the 5% significance level (Exercise 3.17). Third, the gap is large if it is measured instead in percentage terms: According to the estimates in Table 3.1, in 2012 women earned 15% less per hour than men did ($3.80/$25.30), slightly more than the gap of 14% seen in 1992 ($3.44/$24.83). Fourth, the gender gap is smaller for young college graduates (the group analyzed in Table 3.1) than it is for all college graduates (analyzed in Table 2.4): As reported in Table 2.4, the mean earnings for all college-educated women working full-time in 2012 was $25.42, while for men this mean was $32.73, which corresponds to a gender gap of 22% $[= (32.73 - 25.42)/32.73]$ among all full-time college-educated workers.

This empirical analysis documents that the "gender gap" in hourly earnings is large and has been fairly stable (or perhaps increased slightly) over the recent past. The analysis does not, however, tell us *why* this gap exists. Does it arise from gender discrimination in the labor market? Does it reflect differences in skills, experience, or education between men and women? Does it reflect differences in choice of jobs? Or is there some other cause? We return to these questions once we have in hand the tools of multiple regression analysis, the topic of Part II.

[1]Because of inflation, a dollar in 1992 was worth more than a dollar in 2012, in the sense that a dollar in 1992 could buy more goods and services than a dollar in 2012 could. Thus earnings in 1992 cannot be directly compared to earnings in 2012 without adjusting for inflation. One way to make this adjustment is to use the CPI, a measure of the price of a "market basket" of consumer goods and services constructed by the Bureau of Labor Statistics. Over the 20 years from 1992 to 2012, the price of the CPI market basket rose by 63.6%; in other words, the CPI basket of goods and services that cost $100 in 1992 cost $163.64 in 2012. To make earnings in 1992 and 2012 comparable in Table 3.1, 1992 earnings are inflated by the amount of overall CPI price inflation, that is, by multiplying 1992 earnings by 1.636 to put them into "2012 dollars."

3.6 Using the *t*-Statistic When the Sample Size Is Small

In Sections 3.2 through 3.5, the *t*-statistic is used in conjunction with critical values from the standard normal distribution for hypothesis testing and for the construction of confidence intervals. The use of the standard normal distribution is justified by the central limit theorem, which applies when the sample size is large. When the sample size is small, the standard normal distribution can provide a poor approximation to the distribution of the *t*-statistic. If, however, the population distribution is itself normally distributed, then the exact distribution (that is, the finite-sample distribution; see Section 2.6) of the *t*-statistic testing the mean of a single population is the Student *t* distribution with $n - 1$ degrees of freedom, and critical values can be taken from the Student *t* distribution.

The *t*-Statistic and the Student *t* Distribution

The t-statistic testing the mean. Consider the *t*-statistic used to test the hypothesis that the mean of Y is $\mu_{Y,0}$, using data $Y_1, \ldots, Y_n$. The formula for this statistic is given by Equation (3.10), where the standard error of $\overline{Y}$ is given by Equation (3.8). Substitution of the latter expression into the former yields the formula for the t-statistic:

$$t = \frac{\overline{Y} - \mu_{Y,0}}{\sqrt{s_Y^2/n}}, \tag{3.22}$$

where s_Y^2 is given in Equation (3.7).

As discussed in Section 3.2, under general conditions the t-statistic has a standard normal distribution if the sample size is large and the null hypothesis is true [see Equation (3.12)]. Although the standard normal approximation to the t-statistic is reliable for a wide range of distributions of Y if n is large, it can be unreliable if n is small. The exact distribution of the t-statistic depends on the distribution of Y, and it can be very complicated. There is, however, one special case in which the exact distribution of the t-statistic is relatively simple: If Y is normally distributed, then the t-statistic in Equation (3.22) has a Student t distribution with $n - 1$ degrees of freedom. (The mathematics behind this result is provided in Sections 17.4 and 18.4.)

If the population distribution is normally distributed, then critical values from the Student t distribution can be used to perform hypothesis tests and to construct confidence intervals. As an example, consider a hypothetical problem in which $t^{act} = 2.15$ and $n = 20$ so that the degrees of freedom is $n - 1 = 19$. From Appendix Table 2, the 5% two-sided critical value for the t_{19} distribution is 2.09. Because the t-statistic is larger in absolute value than the critical value $(2.15 > 2.09)$, the null hypothesis would be rejected at the 5% significance level against the two-sided alternative. The 95% confidence interval for μ_Y, constructed using the t_{19} distribution, would be $\overline{Y} \pm 2.09\, SE(\overline{Y})$. This confidence interval is somewhat wider than the confidence interval constructed using the standard normal critical value of 1.96.

The t-statistic testing differences of means. The t-statistic testing the difference of two means, given in Equation (3.20), does not have a Student t distribution, even if the population distribution of Y is normal. (The Student t distribution does not apply here because the variance estimator used to compute the standard error in Equation (3.19) does not produce a denominator in the t-statistic with a chi-squared distribution.)

A modified version of the differences-of-means t-statistic, based on a different standard error formula—the "pooled" standard error formula—has an exact Student t distribution when Y is normally distributed; however, the pooled standard error formula applies only in the special case that the two groups have the same variance or that each group has the same number of observations (Exercise 3.21). Adopt the notation of Equation (3.19) so that the two groups are denoted as m and w. The pooled variance estimator is

$$s_{pooled}^2 = \frac{1}{n_m + n_w - 2}\left[\sum_{\substack{i=1 \\ \text{group } m}}^{n_m} (Y_i - \overline{Y}_m)^2 + \sum_{\substack{i=1 \\ \text{group } w}}^{n_w} (Y_i - \overline{Y}_m)^2\right], \quad (3.23)$$

where the first summation is for the observations in group m and the second summation is for the observations in group w. The pooled standard error of the difference in means is $SE_{pooled}(\overline{Y}_m - \overline{Y}_w) = s_{pooled} \times \sqrt{1/n_m + 1/n_w}$, and the pooled t-statistic is computed using Equation (3.20), where the standard error is the pooled standard error, $SE_{pooled}(\overline{Y}_m - \overline{Y}_w)$.

If the population distribution of Y in group m is $N(\mu_m, \sigma_m^2)$, if the population distribution of Y in group w is $N(\mu_w, \sigma_w^2)$, *and* if the two group variances are the same (that is, $\sigma_m^2 = \sigma_w^2$), then under the null hypothesis the t-statistic computed using the pooled standard error has a Student t distribution with $n_m + n_w - 2$ degrees of freedom.

The drawback of using the pooled variance estimator s_{pooled}^2 is that it applies only if the two population variances are the same (assuming $n_m \neq n_w$). If the population variances are different, the pooled variance estimator is biased and inconsistent. If the population variances are different but the pooled variance formula is used, the null distribution of the pooled t-statistic is not a Student t distribution, even if the data are normally distributed; in fact, it does not even have a standard normal distribution in large samples. Therefore, the pooled standard error and the pooled t-statistic should not be used unless you have a good reason to believe that the population variances are the same.

Use of the Student *t* Distribution in Practice

For the problem of testing the mean of Y, the Student t distribution is applicable if the underlying population distribution of Y is normal. For economic variables, however, normal distributions are the exception (for example, see the boxes in Chapter 2 "The Distribution of Earnings in the United States in 2012" and "A Bad Day on Wall Street"). Even if the underlying data are not normally distributed, the normal approximation to the distribution of the t-statistic is valid if the sample size is large. Therefore, inferences—hypothesis tests and confidence intervals—about the mean of a distribution should be based on the large-sample normal approximation.

A Novel Way to Boost Retirement Savings

Many economists think that people do not save enough for retirement. Conventional methods for encouraging retirement savings focus on financial incentives, but there also has been an upsurge in interest in unconventional ways to encourage saving for retirement.

In an important study published in 2001, Brigitte Madrian and Dennis Shea considered one such unconventional method for stimulating retirement savings. Many firms offer retirement savings plans in which the firm matches, in full or in part, savings taken out of the paycheck of participating employees. Enrollment in such plans, called 401(k) plans after the applicable section of the U.S. tax code, is always optional. However, at some firms employees are automatically enrolled in the plan, although they can opt out; at other firms, employees are enrolled only if they choose to opt in. According to conventional economic models of behavior, the method of enrollment—opt out or opt in—should not matter: The rational worker computes the optimal action, then takes it. But, Madrian and Shea wondered, could conventional economics be wrong? Does the *method of enrollment* in a savings plan directly affect its enrollment rate?

To measure the effect of the method of enrollment, Madrian and Shea studied a large firm that changed the default option for its 401(k) plan from nonparticipation to participation. They compared two groups of workers: those hired the year before the change and not automatically enrolled (but could opt in) and those hired in the year after the change and automatically enrolled (but could opt out). The financial aspects of the plan remained the same, and Madrian and Shea found no systematic differences between the workers hired before and after the change. Thus, from an econometrician's perspective, the change was like a randomly assigned treatment and the causal effect of the change could be estimated by the difference in means between the two groups.

Madrian and Shea found that the default enrollment rule made a huge difference: The enrollment rate for the "opt-in" (control) group was 37.4% ($n = 4249$), whereas the enrollment rate for the "opt-out" (treatment) group was 85.9% ($n = 5801$). The estimate of the treatment effect is 48.5% (= 85.9% − 37.4%). Because their sample is large, the 95% confidence (computed in Exercise 3.15) for the treatment effect is tight, 46.8% to 50.2%.

How could the default choice matter so much? Maybe workers found these financial choices too confusing, or maybe they just didn't want to think about growing old. Neither explanation is economically rational—but both are consistent with the predictions of the growing field of "behavioural economics," and both could lead to accepting the default enrollment option.

This research had an important practical impact. In August 2006, Congress passed the Pension Protection Act that (among other things) encouraged firms to offer 401(k) plans in which enrollment is the default. The econometric findings of Madrian and Shea and others featured prominently in testimony on this part of the legislation.

To learn more about behavioral economics and the design of retirement savings plans, see Benartzi and Thaler (2007) and Besheats, Choi, Laibson, and Madrian (2008).

When comparing two means, any economic reason for two groups having different means typically implies that the two groups also could have different variances. Accordingly, the pooled standard error formula is inappropriate, and the correct standard error formula, which allows for different group variances, is as given in Equation (3.19). Even if the population distributions are normal, the t-statistic computed using the standard error formula in Equation (3.19) does not have a Student t distribution. In practice, therefore, inferences about differences in means should be based on Equation (3.19), used in conjunction with the large-sample standard normal approximation.

Even though the Student t distribution is rarely applicable in economics, some software uses the Student t distribution to compute p-values and confidence intervals. In practice, this does not pose a problem because the difference between the Student t distribution and the standard normal distribution is negligible if the sample size is large. For $n > 15$, the difference in the p-values computed using the Student t and standard normal distributions never exceeds 0.01; for $n > 80$, the difference never exceeds 0.002. In most modern applications, and in all applications in this textbook, the sample sizes are in the hundreds or thousands, large enough for the difference between the Student t distribution and the standard normal distribution to be negligible.

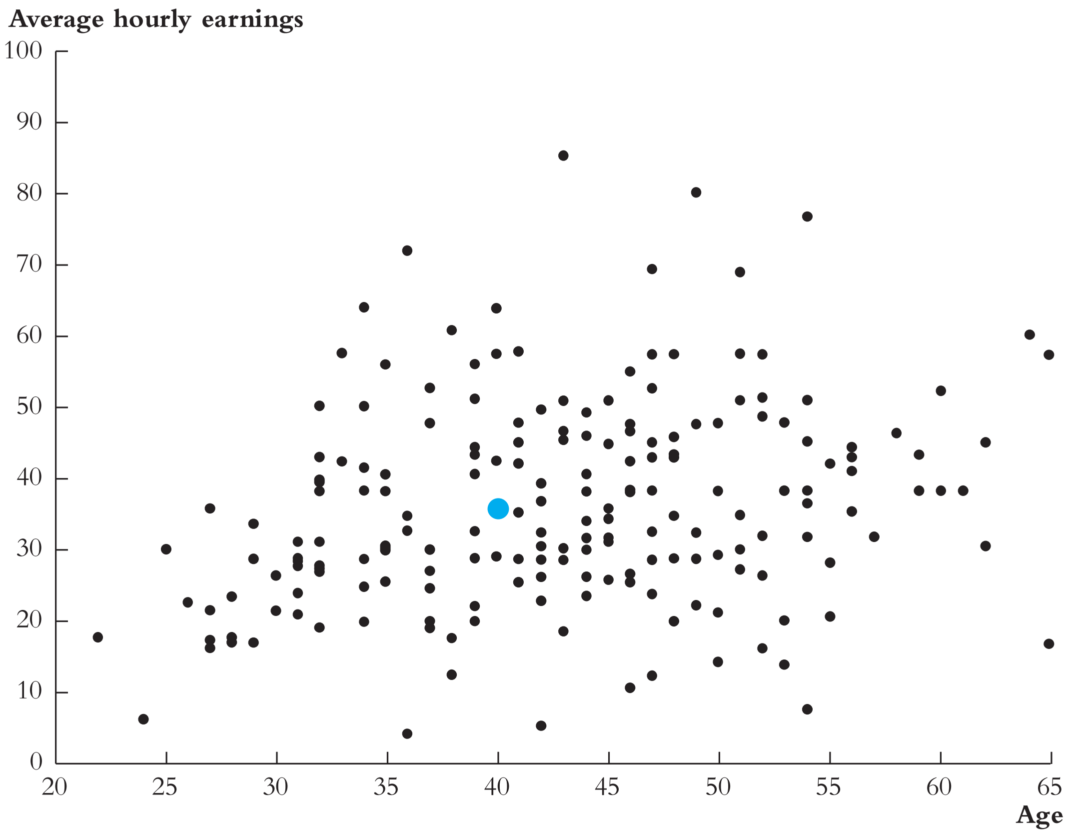
3.7 Scatterplots, the Sample Covariance, and the Sample Correlation

What is the relationship between age and earnings? This question, like many others, relates one variable, X (age), to another, Y (earnings). This section reviews three ways to summarize the relationship between variables: the scatterplot, the sample covariance, and the sample correlation coefficient.

Scatterplots

A **scatterplot** is a plot of n observations on X_i and Y_i, in which each observation is represented by the point (X_i, Y_i). For example, Figure 3.2 is a scatterplot of age (X) and hourly earnings (Y) for a sample of 200 managers in the information industry from the March 2009 CPS. Each dot in Figure 3.2 corresponds to an (X, Y) pair for one of the observations. For example, one of the workers in this sample is 40 years old and earns \$35.78 per hour; this worker's age and earnings are indicated by the highlighted dot in Figure 3.2. The scatterplot shows a positive

FIGURE 3.2 Scatterplot of Average Hourly Earnings vs. Age



Each point in the plot represents the age and average earnings of one of the 200 workers in the sample. The highlighted dot corresponds to a 40-year-old worker who earns \$35.78 per hour. The data are for computer and information systems managers from the March 2009 CPS.

relationship between age and earnings in this sample: Older workers tend to earn more than younger workers. This relationship is not exact, however, and earnings could not be predicted perfectly using only a person's age.

Sample Covariance and Correlation

The covariance and correlation were introduced in Section 2.3 as two properties of the joint probability distribution of the random variables X and Y. Because the population distribution is unknown, in practice we do not know the population covariance or correlation. The population covariance and correlation can, however, be estimated by taking a random sample of n members of the population and collecting the data $(X_i, Y_i), i = 1, \ldots, n$.

The sample covariance and correlation are estimators of the population covariance and correlation. Like the estimators discussed previously in this chapter, they are computed by replacing a population mean (the expectation) with a sample mean. The **sample covariance**, denoted s_{XY}, is

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}). \tag{3.24}$$

Like the sample variance, the average in Equation (3.24) is computed by dividing by $n - 1$ instead of n; here, too, this difference stems from using $\overline{X}$ and $\overline{Y}$ to estimate the respective population means. When n is large, it makes little difference whether division is by n or $n - 1$.

The **sample correlation coefficient**, or **sample correlation**, is denoted r_{XY} and is the ratio of the sample covariance to the sample standard deviations:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}. \tag{3.25}$$

The sample correlation measures the strength of the linear association between X and Y in a sample of n observations. Like the population correlation, the sample correlation is unitless and lies between –1 and 1: $|r_{XY}| \leq 1$.

The sample correlation equals 1 if $X_i = Y_i$ for all i and equals –1 if $X_i = -Y_i$ for all i. More generally, the correlation is ±1 if the scatterplot is a straight line. If the line slopes upward, then there is a positive relationship between X and Y and the correlation is 1. If the line slopes down, then there is a negative relationship and the correlation is –1. The closer the scatterplot is to a straight line, the closer is the correlation to ±1. A high correlation coefficient does not necessarily mean that the line has a steep slope; rather, it means that the points in the scatterplot fall very close to a straight line.

Consistency of the sample covariance and correlation. Like the sample variance, the sample covariance is consistent. That is,

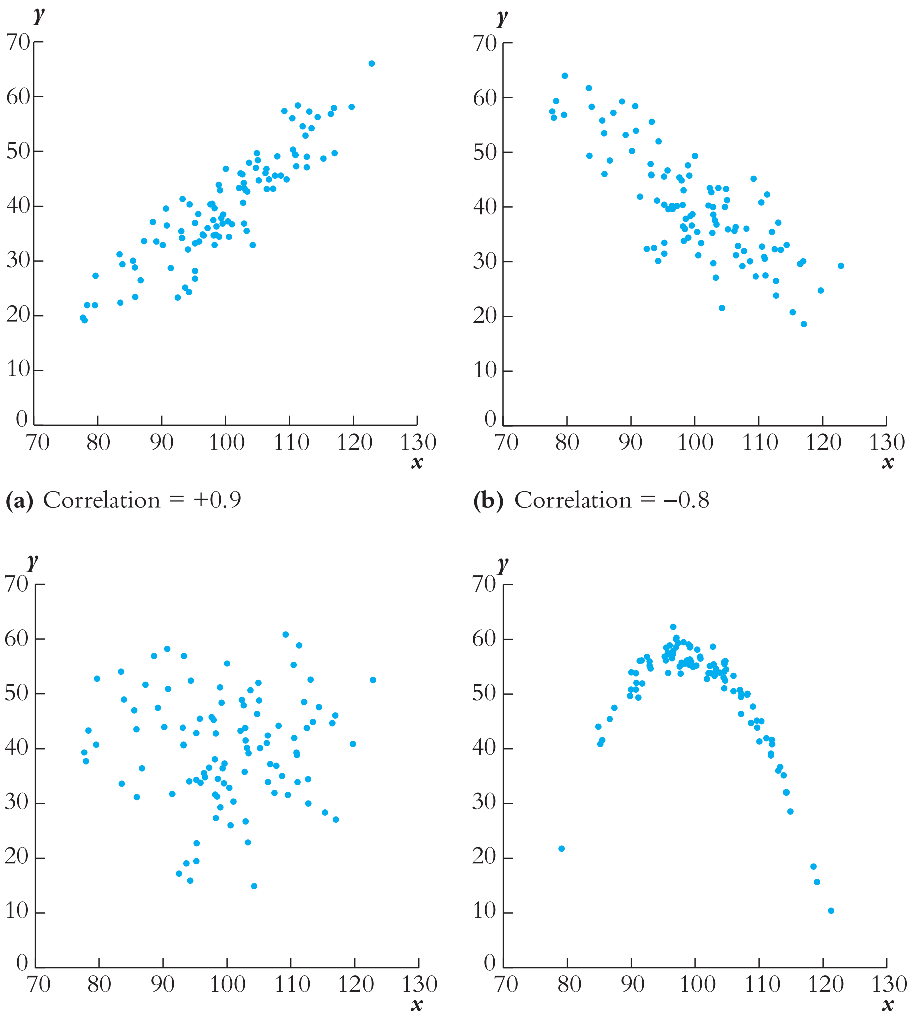
$$s_{XY} \xrightarrow{p} \sigma_{XY}. \tag{3.26}$$

In other words, in large samples the sample covariance is close to the population covariance with high probability.

The proof of the result in Equation (3.26) under the assumption that (X_i, Y_i) are i.i.d. and that X_i and Y_i have finite fourth moments is similar to the proof in Appendix 3.3 that the sample covariance is consistent and is left as an exercise (Exercise 3.20).

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y. In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.



(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

Because the sample variance and sample covariance are consistent, the sample correlation coefficient is consistent, that is, $r_{XY} \xrightarrow{p} \text{corr}(X_i, Y_i)$.

Example. As an example, consider the data on age and earnings in Figure 3.2. For these 200 workers, the sample standard deviation of age is $s_A = 9.07$ years and the sample standard deviation of earnings is $s_E = \$14.37$ per hour. The sample covariance between age and earnings is $s_{AE} = 33.16$ (the units are years $\times$ dollars per hour, not readily interpretable). Thus the sample correlation coefficient is $r_{AE} = 33.16/(9.07 \times 14.37) = 0.25$ or 25%. The correlation of 0.25 means that there is a positive relationship between age and earnings, but as is evident in the scatterplot, this relationship is far from perfect.

To verify that the correlation does not depend on the units of measurement, suppose that earnings had been reported in cents, in which case the sample standard deviations of earnings is 1437¢ per hour and the covariance between age and earnings is 3316 (units are years $\times$ cents per hour); then the correlation is $3316/(9.07 \times 1437) = 0.25$ or 25%.

Figure 3.3 gives additional examples of scatterplots and correlation. Figure 3.3a shows a strong positive linear relationship between these variables, and the sample correlation is 0.9.

Figure 3.3b shows a strong negative relationship with a sample correlation of –0.8. Figure 3.3c shows a scatterplot with no evident relationship, and the sample correlation is zero. Figure 3.3d shows a clear relationship: As X increases, Y initially increases, but then decreases. Despite this discernable relationship between X and Y, the sample correlation is zero; the reason is that, for these data, small values of Y are associated with *both* large and small values of X.

This final example emphasizes an important point: The correlation coefficient is a measure of *linear* association. There is a relationship in Figure 3.3d, but it is not linear.

Summary

1. The sample average, $\overline{Y}$, is an estimator of the population mean, μ_Y. When $Y_1, \ldots, Y_n$ are i.i.d.,
   a. the sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$;
   b. $\overline{Y}$ is unbiased;
   c. by the law of large numbers, $\overline{Y}$ is consistent; and
   d. by the central limit theorem, $\overline{Y}$ has an approximately normal sampling distribution when the sample size is large.
2. The t-statistic is used to test the null hypothesis that the population mean takes on a particular value. If n is large, the t-statistic has a standard normal sampling distribution when the null hypothesis is true.
3. The t-statistic can be used to calculate the p-value associated with the null hypothesis. A small p-value is evidence that the null hypothesis is false.
4. A 95% confidence interval for μ_Y is an interval constructed so that it contains the true value of μ_Y in 95% of all possible samples.
5. Hypothesis tests and confidence intervals for the difference in the means of two populations are conceptually similar to tests and intervals for the mean of a single population.

6. The sample correlation coefficient is an estimator of the population correlation coefficient and measures the linear relationship between two variables—that is, how well their scatterplot is approximated by a straight line.

Key Terms

estimator (67)
estimate (67)
bias, consistency, and efficiency (68)
BLUE (Best Linear Unbiased Estimator) (69)
least squares estimator (70)
hypothesis tests (71)
null hypothesis (71)
alternative hypothesis (71)
two-sided alternative hypothesis (71)
p-value (significance probability) (72)
sample variance (74)
sample standard deviation (74)
degrees of freedom (75)
standard error of $\overline{Y}$ (75)
t-statistic (*t*-ratio) (76)
test statistic (76)
type I error (78)
type II error (78)
significance level (78)
critical value (78)
rejection region (78)
acceptance region (78)
size of a test (78)
power of a test (78)
one-sided alternative hypothesis (80)
confidence set (80)
confidence level (80)
confidence interval (80)
coverage probability (82)
test for the difference between two means (82)
causal effect (85)
treatment effect (85)
scatterplot (91)
sample covariance (93)
sample correlation coefficient (sample correlation) (93)

Review the Concepts

3.1 Explain the difference between the sample average $\overline{Y}$ and the population mean.

3.2 Explain the difference between an estimator and an estimate. Provide an example of each.

3.3 A population distribution has a mean of 10 and a variance of 16. Determine the mean and variance of $\overline{Y}$ from an i.i.d. sample from this population for (a) $n = 10$; (b) $n = 100$; and (c) $n = 1000$. Relate your answers to the law of large numbers.

3.4 What role does the central limit theorem play in statistical hypothesis testing? In the construction of confidence intervals?

3.5 What is the difference between a null hypothesis and an alternative hypothesis? Among size, significance level, and power? Between a one-sided alternative hypothesis and a two-sided alternative hypothesis?

3.6 Why does a confidence interval contain more information than the result of a single hypothesis test?

3.7 Explain why the differences-of-means estimator, applied to data from a randomized controlled experiment, is an estimator of the treatment effect.

3.8 Sketch a hypothetical scatterplot for a sample of size 10 for two random variables with a population correlation of (a) 1.0; (b) –1.0; (c) 0.9; (d) –0.5; (e) 0.0.

Exercises

3.1 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 $\Pr(\overline{Y} < 101)$.

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

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

3.2 Let Y be a Bernoulli random variable with success probability $\Pr(Y = 1) = p$, and let $Y_1, \ldots, Y_n$ be i.i.d. draws from this distribution. Let $\hat{p}$ be the fraction of successes (1s) in this sample.

a. Show that $\hat{p} = \overline{Y}$.

b. Show that $\hat{p}$ is an unbiased estimator of p.

c. Show that $\text{var}(\hat{p}) = p(1 - p)/n$.

3.3 In a survey of 400 likely voters, 215 responded that they would vote for the incumbent, and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters who preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents who preferred the incumbent.

a. Use the survey results to estimate p.

b. Use the estimator of the variance of $\hat{p}$, $\hat{p}(1 - \hat{p})/n$, to calculate the standard error of your estimator.

c. What is the p-value for the test $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$?

d. What is the p-value for the test $H_0: p = 0.5$ vs. $H_1: p > 0.5$?

e. Why do the results from (c) and (d) differ?

f. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

3.4 Using the data in Exercise 3.3:

a. Construct a 95% confidence interval for p.

b. Construct a 99% confidence interval for p.

c. Why is the interval in (b) wider than the interval in (a)?

a. Without doing any additional calculations, test the hypothesis $H_0: p = 0.50$ vs. $H_1: p \neq 0.50$ at the 5% significance level.

3.5 A survey of 1055 registered voters is conducted, and the voters are asked to choose between candidate A and candidate B. Let p denote the fraction of voters in the population who prefer candidate A, and let $\hat{p}$ denote the fraction of voters in the sample who prefer Candidate A.

a. You are interested in the competing hypotheses $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$. Suppose that you decide to reject H_0 if $|\hat{p} - 0.5| > 0.02$.

i. What is the size of this test?

ii. Compute the power of this test if $p = 0.53$.

b. In the survey, $\hat{p} = 0.54$.

i. Test $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$ using a 5% significance level.

ii. Test $H_0: p = 0.5$ vs. $H_1: p > 0.5$ using a 5% significance level.

iii. Construct a 95% confidence interval for p.

iv. Construct a 99% confidence interval for p.

v. Construct a 50% confidence interval for p.

c. Suppose that the survey is carried out 20 times, using independently selected voters in each survey. For each of these 20 surveys, a 95% confidence interval for p is constructed.

i. What is the probability that the true value of p is contained in all 20 of these confidence intervals?

ii. How many of these confidence intervals do you expect to contain the true value of p?

d. In survey jargon, the "margin of error" is $1.96 \times SE(\hat{p})$; that is, it is half the length of 95% confidence interval. Suppose you want to design a survey that has a margin of error of at most 1%. That is, you want $\Pr(|\hat{p} - \mathrm{p}| > 0.01) \leq 0.05$. How large should n be if the survey uses simple random sampling?

3.6 Let $Y_1, \ldots, Y_n$ be i.i.d. draws from a distribution with mean μ. A test of $H_0: \mu = 5$ vs. $H_1: \mu \neq 5$ using the usual t-statistic yields a p-value of 0.03.

a. Does the 95% confidence interval contain $\mu = 5$? Explain.

b. Can you determine if $\mu = 6$ is contained in the 95% confidence interval? Explain.

3.7 In a given population, 11% of the likely voters are African American. A survey using a simple random sample of 600 landline telephone numbers finds 8% African Americans. Is there evidence that the survey is biased? Explain.

3.8 A new version of the SAT is given to 1000 randomly selected high school seniors. The sample mean test score is 1110, and the sample standard deviation is 123. Construct a 95% confidence interval for the population mean test score for high school seniors.

3.9 Suppose that a lightbulb manufacturing plant produces bulbs with a mean life of 2000 hours and a standard deviation of 200 hours. An inventor claims to have developed an improved process that produces bulbs with a longer mean life and the same standard deviation. The plant manager randomly selects 100 bulbs produced by the process. She says that she will believe the inventor's claim if the sample mean life of the bulbs is greater than 2100 hours; otherwise, she will conclude that the new process is no better than the old process. Let μ denote the mean of the new process. Consider the null and alternative hypotheses H_0: $\mu = 2000$ vs. H_1: $\mu > 2000$.

a. What is the size of the plant manager's testing procedure?

b. Suppose the new process is in fact better and has a mean bulb life of 2150 hours. What is the power of the plant manager's testing procedure?

c. What testing procedure should the plant manager use if she wants the size of her test to be 5%?

3.10 Suppose a new standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score $\overline{Y}$ on the test is 58 points, and the sample standard deviation, s_Y, is 8 points.

a. The authors plan to administer the test to all third-grade students in New Jersey. Construct a 95% confidence interval for the mean score of all New Jersey third graders.

b. Suppose the same test is given to 200 randomly selected third graders from Iowa, producing a sample average of 62 points and sample standard deviation of 11 points. Construct a 90% confidence interval for the difference in mean scores between Iowa and New Jersey.

c. Can you conclude with a high degree of confidence that the population means for Iowa and New Jersey students are different? (What is the standard error of the difference in the two sample means? What is the p-value of the test of no difference in means versus some difference?)

3.11 Consider the estimator $\widetilde{Y}$, defined in Equation (3.1). Show that (a) $E(\widetilde{Y}) = \mu_Y$ and (b) $\text{var}(\widetilde{Y}) = 1.25\sigma_Y^2/n$.

3.12 To investigate possible gender discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of the resulting monthly salaries follows:

	Average Salary ($\overline{Y}$)	Standard Deviation (s_Y)	n
Men	\$3100	\$200	100
Women	\$2900	\$320	64

a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that average wages of men and women are different? (To answer this question, first state the null and alternative hypotheses; second, compute the relevant t-statistic; third, compute the p-value associated with the t-statistic; and finally, use the p-value to answer the question.)

b. Do these data suggest that the firm is guilty of gender discrimination in its compensation policies? Explain.

3.13 Data on fifth-grade test scores (reading and mathematics) for 420 school districts in California yield $\overline{Y} = 646.2$ and standard deviation $s_Y = 19.5$.

a. Construct a 95% confidence interval for the mean test score in the population.

b. When the districts were divided into districts with small classes (< 20 students per teacher) and large classes (≥ 20 students per teacher), the following results were found:

Class Size	Average Score ($\overline{Y}$)	Standard Deviation (s_Y)	n
Small	657.4	19.4	238
Large	650.0	17.9	182

Is there statistically significant evidence that the districts with smaller classes have higher average test scores? Explain.

3.14 Values of height in inches (X) and weight in pounds (Y) are recorded from a sample of 300 male college students. The resulting summary statistics are $\overline{X} = 70.5\text{in.}, \overline{Y} = 158 \text{ lb.}, s_X = 1.8\text{in.}, s_Y = 14.2 \text{ lb.}, s_{XY} = 21.73 \text{ in.} \times \text{lb.}$, and $r_{XY} = 0.85$. Convert these statistics to the metric system (meters and kilograms).

3.15 Let Y_a and Y_b denote Bernoulli random variables from two different populations, denoted a and b. Suppose that $E(Y_a) = p_a$ and $E(Y_b) = p_b$. A random sample of size n_a is chosen from population a, with sample average denoted $\hat{p}_a$, and a random sample of size n_b is chosen from population b, with sample average denoted $\hat{p}_b$. Suppose the sample from population a is independent of the sample from population b.

a. Show that $E(\hat{p}_a) = p_a$ and $\text{var}(\hat{p}_a) = p_a(1 - p_a)/n_a$. Show that $E(\hat{p}_b) = p_b$ and $\text{var}(\hat{p}_b) = p_b(1 - p_b)/n_b$.

b. Show that $\text{var}(\hat{p}_a - \hat{p}_b) = \frac{p_a(1 - p_a)}{n_a} + \frac{p_b(1 - p_b)}{n_b}$. (*Hint:* Remember that the samples are independent.)

c. Suppose that n_a and n_b are large. Show that a 95% confidence interval for $p_a - p_b$ is given by $(\hat{p}_a - \hat{p}_b) \pm 1.96\sqrt{\frac{\hat{p}_a(1 - \hat{p}_a)}{n_a} + \frac{\hat{p}_b(1 - \hat{p}_b)}{n_b}}$. How would you construct a 90% confidence interval for $p_a - p_b$?

d. Read the box "A Novel Way to Boost Retirement Savings" in Section 3.6. Let population a denote the "opt-out" (treatment) group and population b denote the "opt-in" (control) group. Construct a 95% confidence interval for the treatment effect, $p_a - p_b$.

3.16 Grades on a standardized test are known to have a mean of 1000 for students in the United States. The test is administered to 453 randomly selected students in Florida; in this sample, the mean is 1013, and the standard deviation (s) is 108.

a. Construct a 95% confidence interval for the average test score for Florida students.

b. Is there statistically significant evidence that Florida students perform differently than other students in the United States?

c. Another 503 students are selected at random from Florida. They are given a 3-hour preparation course before the test is administered. Their average test score is 1019, with a standard deviation of 95.

i. Construct a 95% confidence interval for the change in average test score associated with the prep course.

ii. Is there statistically significant evidence that the prep course helped?

d. The original 453 students are given the prep course and then are asked to take the test a second time. The average change in their test scores is 9 points, and the standard deviation of the change is 60 points.

i. Construct a 95% confidence interval for the change in average test scores.

ii. Is there statistically significant evidence that students will perform better on their second attempt, after taking the prep course?

iii. Students may have performed better in their second attempt because of the prep course or because they gained test-taking experience in their first attempt. Describe an experiment that would quantify these two effects.

3.17 Read the box "The Gender Gap of Earnings of College Graduates in the United States" in Section 3.5.

a. Construct a 95% confidence interval for the change in men's average hourly earnings between 1992 and 2012.

b. Construct a 95% confidence interval for the change in women's average hourly earnings between 1992 and 2012.

c. Construct a 95% confidence interval for the change in the gender gap in average hourly earnings between 1992 and 2012. (*Hint:* $\overline{Y}_{m,1992} - \overline{Y}_{w,1992}$ is independent of $\overline{Y}_{m,2012} - \overline{Y}_{w,2012}$.)

3.18 This exercise shows that the sample variance is an unbiased estimator of the population variance when $Y_1, \ldots, Y_n$ are i.i.d. with mean μ_Y and variance σ_Y^2.

a. Use Equation (2.31) to show that $E[(Y_i - \overline{Y})^2] = \text{var}(Y_i) - 2\text{cov}(Y_i, \overline{Y}) + \text{var}(\overline{Y})$.

b. Use Equation (2.33) to show that $\text{cov}(\overline{Y}, Y_i) = \sigma_Y^2/n$.

c. Use the results in (a) and (b) to show that $E(s_Y^2) = \sigma_Y^2$.

3.19 **a.** $\overline{Y}$ is an unbiased estimator of μ_Y. Is $\overline{Y}^2$ an unbiased estimator of μ_Y^2?

b. $\overline{Y}$ is a consistent estimator of μ_Y. Is $\overline{Y}^2$ a consistent estimator of μ_Y^2?

3.20 Suppose that (X_i, Y_i) are i.i.d. with finite fourth moments. Prove that the sample covariance is a consistent estimator of the population covariance, that is, $s_{XY} \xrightarrow{p} \sigma_{XY}$, where s_{XY} is defined in Equation (3.24). (*Hint:* Use the strategy of Appendix 3.3.)

3.21 Show that the pooled standard error $[SE_{pooled}(\overline{Y}_m - \overline{Y}_w)]$ given following Equation (3.23) equals the usual standard error for the difference in means in Equation (3.19) when the two group sizes are the same $(n_m = n_w)$.

Empirical Exercises

E3.1 On the text website, **http://www.pearsonhighered.com/stock_watson/**, you will find the data file **CPS92_12**, which contains an extended version of the data set used in Table 3.1 of the text for the years 1992 and 2012. It contains data on full-time workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in

CPS92_12_Description, available on the website. Use these data to answer the following questions.

a. i. Compute the sample mean for average hourly earnings (AHE) in 1992 and 2012.

ii. Compute the sample standard deviation for AHE in 1992 and 2012.

iii. Construct a 95% confidence interval for the population means of AHE in 1992 and 2012.

iv. Construct a 95% confidence interval for the change in the population mean of AHE between 1992 and 2012.

b. In 2012, the value of the Consumer Price Index (CPI) was 229.6. In 1992, the value of the CPI was 140.3. Repeat (a) but use AHE measured in real 2012 dollars ($2012); that is, adjust the 1992 data for the price inflation that occurred between 1992 and 2012.

c. If you were interested in the change in workers' purchasing power from 1992 to 2012, would you use the results from (a) or (b)? Explain.

d. Using the data for 2012:

i. Construct a 95% confidence interval for the mean of AHE for high school graduates.

ii. Construct a 95% confidence interval for the mean of AHE for workers with a college degree.

iii. Construct a 95% confidence interval for the difference between the two means.

e. Repeat (d) using the 1992 data expressed in $2012.

f. Using appropriate estimates, confidence intervals, and test statistics, answer the following questions:

i. Did real (inflation-adjusted) wages of high school graduates increase from 1992 to 2012?

ii. Did real wages of college graduates increase?

iii. Did the gap between earnings of college and high school graduates increase? Explain.

g. Table 3.1 presents information on the gender gap for college graduates. Prepare a similar table for high school graduates, using the 1992 and 2012 data. Are there any notable differences between the results for high school and college graduates?

E3.2 A consumer is given the chance to buy a baseball card for $1, but he declines the trade. If the consumer is now given the baseball card, will he be willing to sell it for $1? Standard consumer theory suggests yes, but behavioral economists have found that "ownership" tends to increase the value of goods to consumers. That is, the consumer may hold out for some amount more than $1 (for example, $1.20) when selling the card, even though he was willing to pay only some amount less than $1 (for example, $0.88) when buying it. Behavioral economists call this phenomenon the "endowment effect." John List investigated the endowment effect in a randomized experiment involving sports memorabilia traders at a sports-card show. Traders were randomly given one of two sports collectibles, say good A or good B, that had approximately equal market value.[1] Those receiving good A were then given the option of trading good A for good B with the experimenter; those receiving good B were given the option of trading good B for good A with the experimenter. Data from the experiment and a detailed description can be found on the textbook website, **http://www.pearsonhighered.com/stock_watson/**, in the files **Sportscards** and **Sportscards_Description**.[2]

a. i. Suppose that, absent any endowment effect, all the subjects prefer good A to good B. What fraction of the experiment's subjects would you expect to trade the good that they were given for the other good? (*Hint:* Because of random assignment of the two treatments, approximately 50% of the subjects received good A and 50% received good B.)

ii. Suppose that, absent any endowment effect, 50% of the subjects prefer good A to good B, and the other 50% prefer good B to good A. What fraction of the subjects would you expect to trade the good that they were given for the other good?

iii. Suppose that, absent any endowment effect, X% of the subjects prefer good A to good B, and the other $(100 - X)$% prefer good B to good A. Show that you would expect 50% of the subjects to trade the good that they were given for the other good.

---

[1]Good A was a ticket stub from the game in which Cal Ripken, Jr., set the record for consecutive games played, and good B was a souvenir from the game in which Nolan Ryan won his 300th game.

[2]These data were provided by Professor John List of the University of Chicago and were used in his paper "Does Market Experience Eliminate Market Anomalies," *Quarterly Journal of Economics*, 2003, 118(1): 41–71.

b. Using the sports-card data, what fraction of the subjects traded the good they were given? Is the fraction significantly different from 50%? Is there evidence of an endowment effect? (*Hint:* Review Exercises 3.2 and 3.3)

c. Some have argued that the endowment effect may be present, but that it is likely to disappear as traders gain more trading experience. Half of the experimental subjects were dealers, and the other half were nondealers. Dealers have more experience than nondealers. Repeat (b) for dealers and nondealers. Is there a significant difference in their behavior? Is the evidence consistent with the hypothesis that the endowment effect disappears as traders gain more experience? (*Hint:* Review Exercise 3.15).

APPENDIX

3.1 The U.S. Current Population Survey

Each month, the U.S. Census Bureau and the U.S. Bureau of Labor Statistics conduct the Current Population Survey (CPS), which provides data on labor force characteristics of the population, including the levels of employment, unemployment, and earnings. Approximately 60,000 U.S. households are surveyed each month. The sample is chosen by randomly selecting addresses from a database of addresses from the most recent decennial census augmented with data on new housing units constructed after the last census. The exact random sampling scheme is rather complicated (first, small geographical areas are randomly selected, then housing units within these areas are randomly selected); details can be found in the *Handbook of Labor Statistics* and on the Bureau of Labor Statistics website (www.bls.gov).

The survey conducted each March is more detailed than in other months and asks questions about earnings during the previous year. The statistics in Tables 2.4 and 3.1 were computed using the March surveys. The CPS earnings data are for full-time workers, defined to be somebody employed more than 35 hours per week for at least 48 weeks in the previous year.

APPENDIX

3.2 Two Proofs That $\overline{Y}$ Is the Least Squares Estimator of μ_Y

This appendix provides two proofs, one using calculus and one not, that $\overline{Y}$ minimizes the sum of squared prediction mistakes in Equation (3.2)—that is, that $\overline{Y}$ is the least squares estimator of $E(Y)$.

Calculus Proof

To minimize the sum of squared prediction mistakes, take its derivative and set it to zero:

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = -2\sum_{i=1}^{n}(Y_i - m) = -2\sum_{i=1}^{n}Y_i + 2nm = 0. \quad (3.27)$$

Solving for the final equation for m shows that $\sum_{i=1}^{n}(Y_i - m)^2$ is minimized when $m = \overline{Y}$.

Noncalculus Proof

The strategy is to show that the difference between the least squares estimator and $\overline{Y}$ must be zero, from which it follows that $\overline{Y}$ is the least squares estimator. Let $d = \overline{Y} - m$, so that $m = \overline{Y} - d$. Then $(Y_i - m)^2 = (Y_i - [\overline{Y} - d])^2 = ([Y_i - \overline{Y}] + d)^2 = (Y_i - \overline{Y})^2 + 2d(Y_i - \overline{Y}) + d^2$. Thus the sum of squared prediction mistakes [Equation (3.2)] is

$$\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + 2d\sum_{i=1}^{n}(Y_i - \overline{Y}) + nd^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + nd^2, \quad (3.28)$$

where the second equality uses the fact that $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$. Because both terms in the final line of Equation (3.28) are nonnegative and because the first term does not depend on d, $\sum_{i=1}^{n}(Y_i - m)^2$ is minimized by choosing d to make the second term, nd^2, as small as possible. This is done by setting $d = 0$—that is, by setting $m = \overline{Y}$—so that $\overline{Y}$ is the least squares estimator of $E(Y)$.

APPENDIX

3.3 A Proof That the Sample Variance Is Consistent

This appendix uses the law of large numbers to prove that the sample variance s_Y^2 is a consistent estimator of the population variance σ_Y^2, as stated in Equation (3.9), when $Y_1, \ldots, Y_n$ are i.i.d. and $E(Y_i^4) < \infty$.

First, consider a version of the sample variance that uses n instead of $n - 1$ as a divisor:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - 2\overline{Y}\frac{1}{n}\sum_{i=1}^{n}Y_i + \overline{Y}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}Y_i^2 - \overline{Y}^2 \\ &\xrightarrow{\ p\ } (\sigma_Y^2 + \mu_Y^2) - \mu_Y^2 \\ &= \sigma_Y^2,\end{aligned} \tag{3.29}$$

where the first equality uses $(Y_i - \overline{Y})^2 = Y_i^2 - 2\overline{Y}Y_i + \overline{Y}^2$, and the second uses $\frac{1}{n}\sum_{i=1}^{n}Y_i = \overline{Y}$. The convergence in the third line follows from (i) applying the law of large numbers to $\frac{1}{n}\sum_{i=1}^{n}Y_i^2 \xrightarrow{\ p\ } E(Y^2)$ (which follows because Y_i^2 are i.i.d. and have finite variance because $E(Y_i^4)$ is finite), (ii) recognizing that $E(Y_i^2) = \sigma_Y^2 + \mu_Y^2$ (Key Concept 2.3), and (iii) noting $\overline{Y} \xrightarrow{\ p\ } \mu_Y$ so that $\overline{Y}^2 \xrightarrow{\ p\ } \mu_Y^2$. Finally, $s_Y^2 = \left(\frac{n}{n-1}\right)\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) \xrightarrow{\ p\ } \sigma_Y^2$ follows from Equation (3.29) and $\left(\frac{n}{n-1}\right) \to 1$.