CHAPTER

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. Comprehensive surveys that do exist, also known as censuses, are often undertaken periodically (for example, every ten years in India, the United States of America and the United Kingdom). This is because the process of conducting a census is an extraordinary commitment, consisting of designing census forms, managing and conducting surveys, and compiling and analyzing data. Censuses across the world have a long history, with accounts of censuses recorded by Babylonians in 4000 BC. According to historians, censuses have been conducted as far back as Ancient Rome; the Romans would track the population by making people return to their birthplace every year in order to be counted.[1] In England and other parts of Wales, a notable census was the Domesday Book, which was compiled in 1086 by William the Conqueror. The U.K. census in its current form dates back to 1801 after essays by economist Thomas Malthus (1798) inspired parliament to want to accurately know the size of the population. Over time the census has evolved from amounting to a mere headcount to the much more ambitious survey of the 2011 U.K. census costing an estimated £482 million. In India, there are accounts of censuses recorded around 300 BC, but the census in its current form has been undertaken since 1872 and every ten years since 1881. In comparison to the U.K. census of 2011, the most recent census of India, also conducted in 2011, approximately cost a mere ₹2200 crore (US$320 million)! Despite the considerable efforts made to ensure that the census records all individuals, many people 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 population of China (1.4 billion in 2018), we might survey, say, 1000 members of the population, selected at random by simple random sampling. Using statistical methods, we

[1]Source: Office for National Statistics, https://www.ons.gov.uk, accessed on August 23, 2018.

can use this sample to reach tentative conclusions—to draw statistical inferences—about characteristics of the full population.[2]

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 and 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

[2]Estimates of the 'live' population of China can be found here using the 'official' China Population Clock: http://data.stats.gov.cn/english/

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.

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 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 caret (^) notation will be used throughout this text to denote an estimator, so $\hat{\mu}_Y$ is an estimator of μ_Y.] 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.

Bias, Consistency, and Efficiency

KEY CONCEPT 3.2

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 $\tilde{\mu}_Y$ if $\text{var}(\hat{\mu}_Y) < \text{var}(\tilde{\mu}_Y)$.

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 $\widetilde{\mu}_Y$.

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.

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

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 $Y_1, \ldots, Y_n$.

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 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.

Off the Mark!

In 2009, India's general elections, also referred to as the national elections, was the largest democratic election in the world until the Indian general elections 2014 held from April 7, 2014. Shortly before the general elections, pollsters predicted a close fight between the coalition parties—the United Progressive Alliance (UPA) and the National Democratic Alliance (NDA). Psephologists envisaged that while the UPA might have had the upper hand, the NDA could not be written off. They predicted that the UPA would get between 201 and 235 seats in the 14th Lok Sabha (the lower house of India's bicameral Parliament) and the NDA between 165 and 186 seats. The actual results were surprising: UPA got 262 seats, while NDA could only manage to get 157 seats.

What could be the possible reasons for opinion polls being wide off the mark? In countries that do not have a homogenous population, such as India, caste, religion, and geographies influence electoral outcomes greatly. Vulnerable sections of the population may be afraid to disclose their actual preference. Political polls have since become much more sophisticated and adjust for sampling bias, but they still can make mistakes. If opinion polls do not randomly select samples across various locations and sections of people, they may still not hit the mark.

Source: Atul Thakur, "Why Opinion Polls Are Often Wide off the Mark," *The Times of India*, April 13, 2014.

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 "Off the Mark!" 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

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

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 earnings 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

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

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 hypothesis) 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, 0.1%), 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 $E(Y) = \mu_{Y,0}$). The *p*-value is

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

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_Y^2 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 $N(\mu_{Y,0}, \sigma_{\overline{Y}}^2)$, where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$. 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, it 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

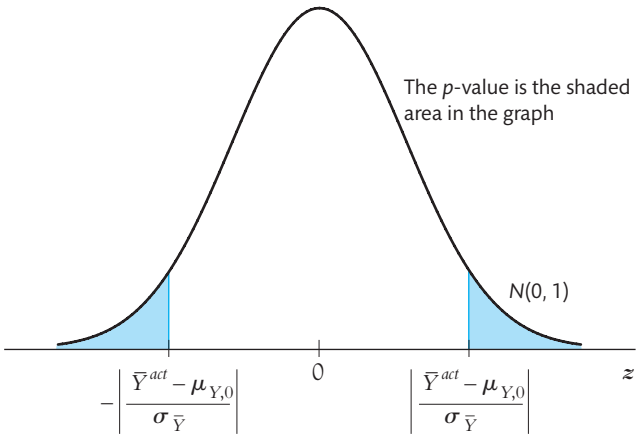
$$p\text{-value} = \Pr_{H_0}\left(\left|\frac{\overline{Y} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right| > \left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right) = 2\Phi\left(-\left|\frac{\overline{Y}^{act} - \mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right), \tag{3.6}$$

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 $\pm\left|\overline{Y}^{act} - \mu_{Y,0}\right|/\sigma_{\overline{Y}}$.

The formula for the p-value in Equation (3.6) depends on the variance of the population distribution, σ_Y^2. In practice, this variance is typically unknown. [An exception is when Y_i is binary, so that its distribution is Bernoulli, in which case the variance is determined by the null hypothesis; see Equation (2.7) and Exercise 3.2.] Because in general σ_Y^2 must be estimated before the p-value can be computed, we now turn to the problem of estimating σ_Y^2.

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 $N(\mu_{Y,0}, \sigma_{\overline{Y}}^2)$ 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 $\pm\left|(\overline{Y}^{act} - \mu_{Y,0})/\sigma_{\overline{Y}}\right|$.



The Sample Variance, Sample Standard Deviation, and Standard Error

The sample variance, s_Y^2, is an estimator of the population variance, σ_Y^2; 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_Y^2, is

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

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 $(Y_i - \mu_Y)^2$, $i = 1, \cdots, n$, with two modifications: First, μ_Y is replaced by $\overline{Y}$, and second, the average uses the divisor $n - 1$ instead of n.

The 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:

$$s_Y^2 \xrightarrow{p} \sigma_Y^2. \tag{3.8}$$

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. 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 $\overline{Y}$

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.,

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}. \tag{3.9}$$

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 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%.

This framework for testing statistical hypotheses has some specialized terminology, summarized in Key Concept 3.5. The significance level of the test in

The Terminology of Hypothesis Testing

KEY CONCEPT 3.5

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 set of 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.

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? This is a question of active debate. Historically, statisticians and econometricians have used a 5% significance level. If you were to test many statistical hypotheses at the 5% level, you would incorrectly reject the null, on average, once in 20 cases. Whether this is a small number depends on how you look at it. If only a small fraction of all null hypotheses tested are, in fact, false, then among those tests that reject, the probability of the null actually being false can be small (Exercise 3.22). This probability—the fraction of incorrect rejections among all rejections—is called the false positive rate. The false positive rate can have great practical importance. For example, for newly reported statistically

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 $|t^{act}| > 1.96$).

significant findings of effective medical treatments, it is the fraction for which the treatment is in fact ineffective. Concern that the false positive rate can be high when the 5% significance level is used has led some statisticians to recommend using instead a 0.5% significance level when reporting new results (Benjamin et al., 2017). Similar concerns can apply in a legal setting, where justice might aim to keep the fraction of false convictions low. Using a 0.5% significance level leads to two-sided rejection when the t-statistic exceeds 2.81 in absolute value. In such cases, a p-value between 0.05 and 0.005 can be viewed as suggestive, but not conclusive, evidence against the null that merits further investigation.

The choice of significance level requires judgment and depends on the application. In some economic applications, a false positive might be less of a problem than in a medical context, where the false positive could lead to patients receiving ineffective treatments. In such cases, a 5% significance level could be appropriate.

Whatever the significance level, it is important to keep in mind that p-values are designed for tests of a null hypothesis, so they, like t-statistics, are useful only when the null hypothesis itself is of interest. This section uses the example of earnings. Even though many interns are unpaid, nobody thinks that, on average, workers earn nothing at all, so the null hypothesis that earnings are zero is economically uninteresting and not worth testing. In contrast, the null hypothesis that the mean earnings of men and of women are the same is interesting and of societal importance, and that null hypothesis is examined in Section 3.4.

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

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

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, so 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. 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%.

KEY CONCEPT 3.7

Confidence Intervals for the Population Mean

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, 90%, 95%, and 99% confidence intervals for μ_Y are:

$$\text{90\% confidence interval for } \mu_Y = \{\overline{Y} \pm 1.64SE(\overline{Y})\},$$
$$\text{95\% confidence interval for } \mu_Y = \{\overline{Y} \pm 1.96SE(\overline{Y})\}, \text{ and}$$
$$\text{99\% confidence interval for } \mu_Y = \{\overline{Y} \pm 2.58SE(\overline{Y})\}.$$

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 $\overline{Y} - 1.96SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.96SE(\overline{Y})$. 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? Answering 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 earnings 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

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

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 sampling 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 ± 1.96 standard errors of $\overline{Y}_m - \overline{Y}_w$:

95% confidence interval for $d = \mu_m - \mu_w$ 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 "Social Class or Education? Childhood Circumstances and Adult Earnings Revisited" contains an empirical investigation of differences in earnings of different households in the United Kingdom.

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, 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

Social Class or Education? Childhood Circumstances and Adult Earnings Revisited

The box in Chapter 2 "The Distribution of Adulthood Earnings in the United Kingdom by Childhood Socioeconomic Circumstances" suggests that when an individual's father has a "routine" occupation, the individual, as an adult, goes on to live in a household with lower average income.

Are there any other factors that affect it? Yes, it is possible that there are relevant intermediate factors like education. It is generally hypothesized and observed that more education is associated with greater income, which will allow individuals to increase their contribution to household income.

Table 3.1 breaks down the differences in mean household income for individuals according to their father's NS-SEC occupation type, and considers these differences for selected highest level of educational qualification. These categories include those with no qualifications, those whose highest qualification level is GCSE (exams generally taken at age 16), those whose highest educational qualification is A-Level (exams generally taken at age 18), and those with an undergraduate degree or higher. For simplicity, only individuals whose father's NS-SEC occupational category was either the highest ("higher") or the lowest ("routine") are included in this analysis.

The data shows that, as expected, within both groups according to the NS-SEC of a father's occupation, those with higher qualifications are part of households with higher total income. The income gap between those with qualifications of at least one degree and those with no qualifications stands at £1467.38 where the father's NS-SEC category is higher, and at a comparable £1527.98 where the father's NS-SEC category is routine.

It is interesting to note the differences between mean income by the father's occupational categorization ($Y_h - Y_r$) for each of the educational groupings. For instance, individuals with no qualifications whose father's NS-SEC job categorization was higher are part of households with a mean income of £2223.13 while for the classification routine this value stood at £1842.98. This implies a difference in means of £380.15, with a standard error of $\sqrt{2115.12^2/1129 + 1487.29^2/6383} = £65.64$ with

TABLE 3.1 Differences in Household Income According to Childhood Socioeconomic Circumstances, Grouped by Level of Highest Qualification

	Father's NS-SEC = Higher			Father's NS-SEC = Routine			Difference, Higher vs. Routine			
Qualification	Y_h	s_h	n_h	Y_r	s_r	n_r	$Y_h - Y_r$	$SE(Y_h - Y_r)$	95% Confidence Interval for d	
None	£2,223.13	£2,115.12	1129	£1,842.98	£1,487.29	6383	£380.15	£65.64	£251.38	£508.93
GCSE/O-Level	£2,837.18	£1,819.73	1962	£2,596.93	£1,738.47	4042	£240.25	£49.35	£143.49	£337.00
A-Level	£3,045.99	£2,451.81	1216	£2,745.70	£1,912.50	1169	£300.30	£89.85	£124.11	£476.49
Undergraduate degree or more	£3,690.51	£2,743.55	4359	£3,370.96	£2,443.58	2505	£319.55	£64.11	£193.86	£445.23
All categories	£3,215.71	£2,497.73	8666	£2405.45	£1,886.86	14099	£810.25	£31.18	£749.13	£871.38

Source: Understanding Society.

a 95% confidence interval of (£251.38, £508.93). It is worth noting the difference in income, pooling these educational categories together, between those whose father's NS-SEC categorization is "higher" and those where this categorization is lower is £810.25. The results in the table suggest a difference in composition by educational attainment of these groupings according to the father's NS-SEC category. When broken down in this way, however, the estimated difference for every qualification level is substantially lower than £810.25. All of these estimated differences are significantly different from zero.

This empirical analysis suggests that levels of education do play some part in explaining the difference in household income according to the socioeconomic status of the father. However, does this analysis tell us the full story? Are individuals with higher levels of education likely to be in households with more than one earner? Does the difference in household income arise from an individual's own contribution to household income or, if the individual is cohabiting, also from her or his partner's contribution to household income? Is this relationship affected by changing patterns of educational attainment that are correlated with age? We will examine questions such as these further once we have introduced the basics of multivariate regression in later chapters.

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 Way to Increase Voter Turnout" provides an example of such a quasi-experiment that yielded some surprising conclusions.

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.

A Way to Increase Voter Turnout

Apathy among citizens toward political participation, especially in voting, has been noted in the United Kingdom and other democratic countries. This kind of behavior is generally seen in economies where people have greater mobility, maintain an intensive work culture, and work for private corporate entities. Apart from these, there could be other dominant factors that have had a negative impact on the citizens' willingness to participate in elections—politicians failing to keep their promises, inappropriately using public funds.

In 2005, during the campaign period before the general election, a study was conducted in a Manchester constituency in the United Kingdom. The constituency's voter turnout rate in the 2001 general election had been 48.6%, while the national average had been 59.4%. Thus, voter participation in this constituency was far below the national average. For the experiment, three groups (two treatment groups and one control group) were randomly selected out of the registered voters from whom landline numbers could be obtained. One of the treatment groups was exposed to strong canvassing in the form of telephone calls, and the other treatment group was exposed to strong canvassing in the form of door-to-door visits. No contacts were made with the control group. The callers and the door-to-door canvassers were given instructions to ask respondents three questions, namely, whether the respondents thought voting is important, whether the respondents intended to vote, and whether they would vote by post. The conversations were informal and the main objective of this exercise was to persuade citizens to vote, by focusing on the importance of voting. The callers and canvassers were also advised to respond to any concerns of the voters regarding the voting process.

The researchers got interesting results from the elections. The participation rate was 55.1% in the group, which was exposed to canvassing. The participation rate for the treatment group, which was treated with telephone calls, was 55%. Both these rates had a difference with the control group, which was not exposed to any experiment. Further calculations using suitable methodologies gave estimates of the effects of canvassing and telephone calls. 6.7% and 7.3% were the estimates of the two. The overall experiment was a success as the two interventions done on the two treatments groups by a non-partisan source had impacts that were statistically significant.

This exercise illustrated that citizens can be nudged to participate in elections by creating awareness through personal contacts. In yet another democracy, India, the 2014 general election saw a record voter turnout. A top Election Commission official has said that the Election Commission's efforts to increase voters' awareness and their registration has helped the process.

Sources: 1. Alice Moseley, Corinne Wales, Gerry Stoker, Graham Smith, Liz Richardson, Peter John, and Sarah Cotterill, "Nudge, Nudge, Think, Think Experimenting with Ways to Change Civic Behaviour," *Bloomsbury Academic*, March 2013. 2. "Lok Sabha Polls 2014: Country Records Highest Voter Turnout since Independence," *The Economic Times*, May 13, 2014.

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 $Y_1, \ldots, Y_n$ are i.i.d. draws from a normal distribution, 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 18.4 and 19.4.)

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 $n = 8$, so that the degrees of freedom is $n - 1 = 7$. From Appendix Table 2, the 5% two-sided critical value for the t_7 distribution is 2.36. Because the *t*-statistic is smaller in absolute value than the critical value $(2.15 < 2.36)$, the null hypothesis would not be rejected at the 5% significance level against the two-sided alternative. The 95% confidence interval for μ_Y, constructed using the t_7 distribution, would be $\overline{Y} \pm 2.36SE(\overline{Y})$. This confidence interval is 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

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

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 $N(\mu_m, \sigma^2_m)$, if the population distribution of Y in group w is $N(\mu_w, \sigma^2_w)$, *and* if the two group variances are the same (that is, $\sigma^2_m = \sigma^2_w$), 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^2_{pooled} 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 Adulthood Earnings in the United Kingdom" and "The Unpegging of the Swiss Franc"). Even if the 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.

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 text, 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.

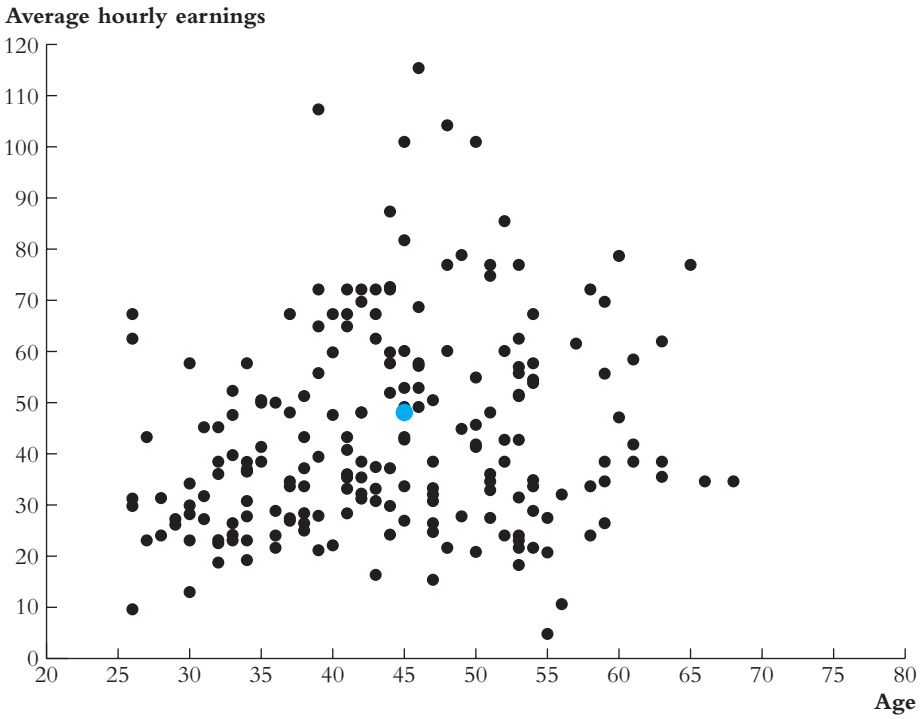
Scatterplots

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 2016 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 45 years old and earns \$49.15 per hour; this worker's age and earnings are indicated by the highlighted dot in Figure 3.2. The scatterplot shows a positive 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), $i = 1, \ldots, n$.

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 45-year-old worker who earns $49.15 per hour. The data are for computer and information systems managers from the March 2016 CPS.

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 unit free 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 the correlation is 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).

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 $s_A = 9.57$ years, and the sample standard deviation of earnings is $s_E = \$19.93$ per hour. The sample covariance between age and earnings is $s_{AE} = 91.51$ (the units are years $\times$ dollars per hour, not readily interpretable). Thus the sample correlation coefficient is $r_{AE} = 91.51/(9.57 \times 19.93) = 0.48$. The correlation of 0.48 means that there is a positive relationship between age and earnings, but as is evident in the scatterplot, this relationship is far from perfect.

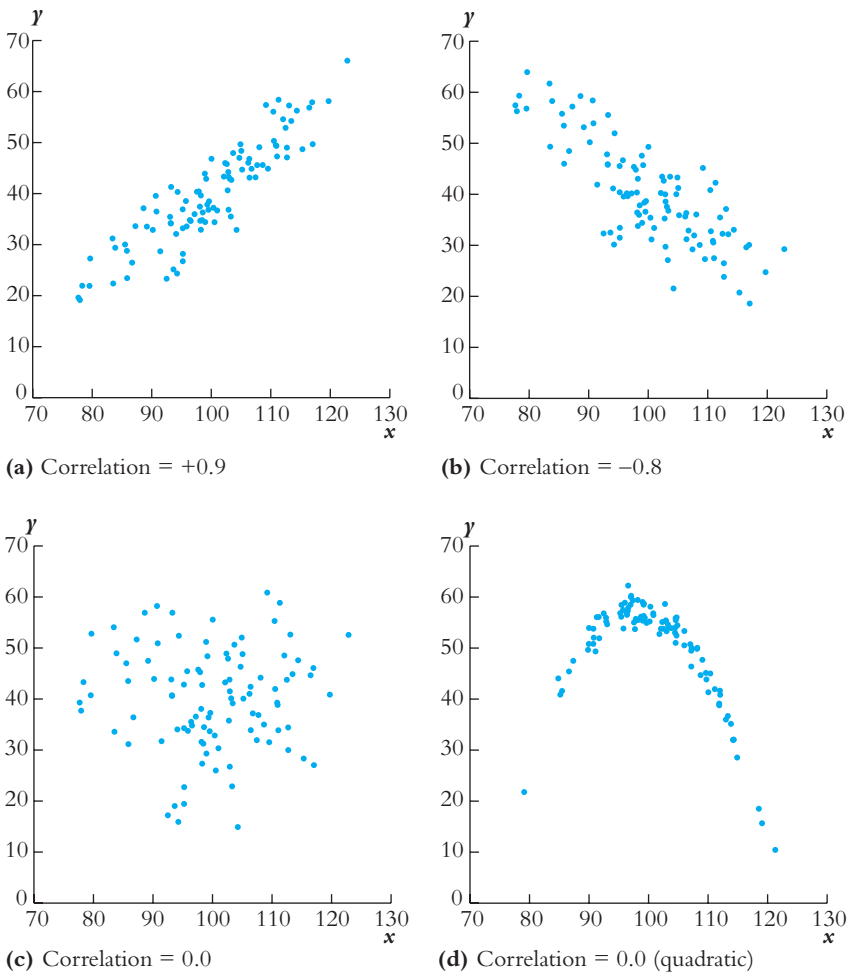
To verify that the correlation does not depend on the units of measurement, suppose that earnings had been reported in cents, in which case the sample standard deviation of earnings is 1993¢ per hour and the covariance between age and earnings is 9151 (units are years $\times$ cents per hour); then the correlation is $9151/(9.57 \times 1993) = 0.48$, or 48%.

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

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)

correlation is 0. 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 0; 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 $\sigma^2_{\overline{Y}} = \sigma^2_Y/n$;
   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. The p-value 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. 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 (105)
estimate (105)
bias (106)
consistency (106)
efficiency (106)
BLUE (Best Linear Unbiased Estimator) (107)
least squares estimator (107)
hypothesis tests (109)
null hypothesis (109)
alternative hypothesis (109)
two-sided alternative hypothesis (109)
p-value (significance probability) (110)
sample variance (112)
sample standard deviation (112)
degrees of freedom (112)
standard error of $\overline{Y}$ (113)
t-statistic (113)
t-ratio (113)
test statistic (113)
type I error (115)
type II error (115)
significance level (115)
critical value (115)

rejection region (115)
acceptance region (115)
size of a test (115)
power of a test (115)
one-sided alternative hypothesis (116)
confidence set (117)
confidence level (117)
confidence interval (117)
coverage probability (118)
test for the difference between two means (119)
causal effect (121)
treatment effect (121)
scatterplot (127)
sample covariance (128)
sample correlation coefficient (sample correlation) (128)

Review the Concepts

3.1 Explain the difference between an unbiased estimator and a consistent estimator.

3.2 What is meant by the efficiency of an estimator? Which estimator is known as BLUE?

3.3 A population distribution has a mean of 15 and a variance of 10. Determine the mean and variance of $\overline{Y}$ from an i.i.d. sample from this population for (a) $n = 5$; (b) $n = 500$; and (c) $n = 5000$. Relate your answers to the law of large numbers.

3.4 What is the difference between standard error and standard deviation? How is the standard error of the sample mean calculated?

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 What is a scatterplot? What statistical features of a dataset can be represented using a scatterplot diagram?

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; and (e) 0.0.

Exercises

3.1 In a population, $\mu_Y = 75$ and $\sigma_Y^2 = 45$. Use the central limit theorem to answer the following questions:

- **a.** In a random sample of size $n = 50$, find $\Pr(\overline{Y} < 73)$.
- **b.** In a random sample of size $n = 90$, find $\Pr(76 < \overline{Y} < 77)$.
- **c.** In a random sample of size $n = 120$, find $\Pr(\overline{Y} > 69)$.

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 poll of 500 likely voters, 270 responded that they would vote for the candidate from the democratic party, while 230 responded that they would vote for the candidate from the republican party. Let p denote the fraction of all likely voters who preferred the democratic candidate at the time of the poll, and let $\hat{p}$ be the fraction of survey respondents who preferred the democratic candidate.

- **a.** Use the poll 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 of H_0: $p = 0.5$, vs. H_1: $p \neq 0.5$?
- **d.** What is the p-value for the test of H_0: $p = 0.5$, vs. H_1: $p > 0.5$?
- **e.** Why do the results from (c) and (d) differ?
- **f.** Did the poll contain statistically significant evidence that the democratic candidate was ahead of the republican candidate at the time of the poll? 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)?
- **d.** 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 1000 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: $p = 0.4$ vs. H_1: $p \neq 0.4$. Suppose that you decide to reject H_0 if $|\hat{p} - 0.4| > 0.01$.

i. What is the size of this test?

ii. Compute the power of this test if $p = 0.45$.

b. In the survey, $\hat{p} = 0.44$.

i. Test H_0: $p = 0.4$ vs. H_1: $p \neq 0.4$ using a 10% significance level.

ii. Test H_0: $p = 0.4$ vs. H_1: $p < 0.4$ using a 10% significance level.

iii. Construct a 90% confidence interval for p.

iv. Construct a 99% confidence interval for p.

v. Construct a 60% confidence interval for p.

c. Suppose that the survey is carried out 30 times, using independently selected voters in each survey. For each of these 30 surveys, a 90% confidence interval for p is constructed.

i. What is the probability that the true value of p is contained in all 30 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 the 95% confidence interval. Suppose you want to design a survey that has a margin of error of at most 0.5%. That is, you want $\Pr(|\hat{p} - \mathrm{p}| > 0.005 \leq 0.005)$. 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: $\mu = 10$ vs. H_1: $\mu \neq 10$ using the usual t-statistic yields a p-value of 0.07.

a. Does the 90% confidence interval contain $\mu = 10$? Explain.

b. Can you determine if $\mu = 8$ is contained in the 95% confidence interval? Explain.

3.7 In a given population, 50% of the likely voters are women. A survey using a simple random sample of 1000 landline telephone numbers finds 55% women. Is there evidence that the survey is biased? Explain.

3.8 A new version of the SAT is given to 1500 randomly selected high school seniors. The sample mean test score is 1230, and the sample standard deviation is 145. Construct a 95% confidence interval for the population mean test score for high school seniors.

3.9 Suppose that a plant manufactures integrated circuits with a mean life of 1000 hours and a standard deviation of 100 hours. An inventor claims to have developed an improved process that produces integrated circuits with a longer mean life and the same standard deviation. The plant manager randomly selects 50 integrated circuits produced by the process. She says that she will believe the inventor's claim if the sample mean life of the integrated circuits

is greater than 1100 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: $\mu = 1000$ vs. H_1: $\mu > 1000$.

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 integrated circuit life of 1150 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 1%?

3.10 Suppose a new standardized test is given to 150 randomly selected third-grade students in Amsterdam. The sample average score $\overline{Y}$ on the test is 42 points, and the sample standard deviation, s_Y, is 6 points.

a. The authors plan to administer the test to all third-grade students in Amsterdam. Construct a 99% confidence interval for the mean score of all third graders in Amsterdam.

b. Suppose the same test is given to 300 randomly selected third graders from Rotterdam, producing a sample average of 48 points and sample standard deviation of 10 points. Construct a 95% confidence interval for the difference in mean scores between Rotterdam and Amsterdam.

c. Can you conclude with a high degree of confidence that the population means for Rotterdam and Amsterdam 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 British firm, a sample of 120 men and 150 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	£8200	£450	120
Women	£7900	£520	150

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 400 school districts in Brussels yield average score $\overline{Y} = 712.1$ and standard deviation $s_Y = 23.2$.

a. Construct a 90% 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 Salary ($\overline{Y}$)	Standard Deviation (s_Y)	n
Small	721.8	24.4	150
Large	710.9	20.6	250

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 200 male college students. The resulting summary statistics are $\overline{X} = 71.2$ in., $\overline{Y} = 164$ lb, $s_X = 1.9$ in., $s_Y = 16.4$ lb, $s_{XY} = 22.54$ in. $\times$ lb, and $r_{XY} = 0.8$. Convert these statistics to the metric system (meters and kilograms).

3.15 Y_a and Y_b are Bernoulli random variables from two different populations, denoted a and b. Suppose $E(Y_a) = p_a$ and $E(Y_b) = p_b$. A random sample of size n_a is chosen from population a, with a sample average denoted $\hat{p}_a$, and a random sample of size n_b is chosen from population b, with a 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 $\text{var}(\hat{p}_a - \hat{p}_b) = \dfrac{p_a(1 - p_a)}{n_a} + \dfrac{p_b(1 - p_b)}{n_b}$.
(*Hint:* Remember that the samples are independent.)

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

3.16 Assume that grades on a standardized test are known to have a mean of 500 for students in Europe. The test is administered to 600 randomly selected students in Ukraine; in this sample, the mean is 508, and the standard deviation (s) is 75.

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

b. Is there statistically significant evidence that Ukrainian students perform differently than other students in Europe?

c. Another 500 students are selected at random from Ukraine. They are given a 3-hour preparation course before the test is administered. Their average test score is 514, with a standard deviation of 65.

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 600 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 7 points, and the standard deviation of the change is 40 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 "Social Class or Education? Childhood Circumstances and Adult Earnings Revisited" in Section 3.5.

a. Construct a 95% confidence interval for the difference in the household earnings of people whose father NS-SEC classification was higher between those with no educational qualifications and those with an undergraduate degree or more.

b. Construct a 95% confidence interval for the difference in the household earnings of people whose father NS-SEC classification was routine between those with no educational qualifications and those with an undergraduate degree or more.

c. Construct a 95% confidence interval for the difference between your answers calculated in parts **a** and **b**.

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 (2.32) 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 (2.34) 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 (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 $(n_m = n_w)$.

3.22 Suppose $Y_i \sim i.i.d.N(\mu_Y, \sigma_Y^2)$ for $i = 1, \ldots, n$. With σ_Y^2 known, the t-statistic for testing H_0: $\mu_Y = 0$ vs. H_1: $\mu_Y > 0$ is $t = (\overline{Y} - 0)/SE(\overline{Y})$, where $SE(\overline{Y}) = \sigma_Y/\sqrt{n}$. Suppose $\sigma_Y = 10$ and $n = 100$, so that $SE(\overline{Y}) = 1$. Using a test with a size of 5%, the null hypothesis is rejected if $t > 1.64$.

a. Suppose $\mu_Y = 0$, so the null hypothesis is true. What is the probability that the null hypothesis is rejected?

b. Suppose $\mu_Y = 2$, so the alternative hypothesis is true. What is the probability that the null hypothesis is rejected?

c. Suppose that in 90% of cases the data are drawn from a population where the null is true ($\mu_Y = 0$) and in 10% of cases the data come from a population where the alternative is true and $\mu_Y = 2$. Your data came from either the first or the second population, but you don't know which.

i. You compute the t-statistic. What is the probability that $t > 1.64$—that is, that you reject the null hypothesis?

ii. Suppose you reject the null hypothesis; that is, $t > 1.64$. What is the probability that the sample data were drawn from the $\mu_Y = 0$ population?

d. It is hard to discover a new effective drug. Suppose 90% of new drugs are ineffective and only 10% are effective. Let Y denote the drop in the level of a specific blood toxin for a patient taking a new drug. If the drug is ineffective, $\mu_Y = 0$ and $\sigma_Y = 10$; if the drug is effective, $\mu_Y = 2$ and $\sigma_Y = 10$.

i. A new drug is tested on a random sample of $n = 100$ patients, data are collected, and the resulting t-statistic is found to be greater than 1.64. What is the probability that the drug is ineffective (i.e., what is the false positive rate for the test using $t > 1.64$)?

ii. Suppose the one-sided test uses instead the 0.5% significance level. What is the probability that the drug is ineffective (i.e., what is the false positive rate)?

Empirical Exercises

E3.1 On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **CPS96_15**, which contains an extended version of the data set used in Table 3.1 of the text for the years 1996 and 2015. It contains data on full-time workers, ages 25–34, with a high school diploma or a B.A./B.S. as their highest degree. A detailed description is given in **CPS96_15_Description**, available on the website. Use these data to complete the following.

a. i. Compute the sample mean for average hourly earnings (AHE) in 1996 and 2015.

ii. Compute the sample standard deviation for AHE in 1996 and 2015.

iii. Construct a 95% confidence interval for the population means of AHE in 1996 and 2015.

iv. Construct a 95% confidence interval for the change in the population means of AHE between 1996 and 2015.

b. In 2015, the value of the Consumer Price Index (CPI) was 237.0. In 1996, the value of the CPI was 156.9. Repeat (a), but use AHE measured in real 2015 dollars ($2015); that is, adjust the 1996 data for the price inflation that occurred between 1996 and 2015.

c. If you were interested in the change in workers' purchasing power from 1996 to 2015, would you use the results from (a) or (b)? Explain.

d. Using the data for 2015:

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 1996 data expressed in $2015.

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 1996 to 2015?

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 1996 and 2015 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.[3] 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 text website, **http://www.pearsonglobaleditions.com**, in the files **Sportscards** and **Sportscards_Description**.[4]

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 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 they were given for the other good.

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?

[3]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.

[4]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.

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 54,000 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 those 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 persons employed more than 35 hours per week for at least 48 weeks in the previous year.

More details on the data can be found in the replication materials for this chapter, available at http://www.pearsonglobaleditions.com.

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 0:

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

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 0, 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:

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

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, $s_Y^2 = (\frac{n}{n-1})(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2) \xrightarrow{p} \sigma_Y^2$ follows from Equation (3.29) and $(\frac{n}{n-1}) \to 1$.