Appendix C

Fundamentals of Mathematical Statistics

C-1 Populations, Parameters, and Random Sampling

Statistical inference involves learning something about a population given the availability of a sample from that population. By **population**, we mean any well-defined group of subjects, which could be individuals, firms, cities, or many other possibilities. By "learning," we can mean several things, which are broadly divided into the categories of *estimation* and *hypothesis testing*.

A couple of examples may help you understand these terms. In the population of all working adults in the United States, labor economists are interested in learning about the return to education, as measured by the average percentage increase in earnings given another year of education. It would be impractical and costly to obtain information on earnings and education for the entire working population in the United States, but we can obtain data on a subset of the population. Using the data collected, a labor economist may report that his or her best estimate of the return to another year of education is 7.5%. This is an example of a *point estimate*. Or, she or he may report a range, such as "the return to education is between 5.6% and 9.4%." This is an example of an *interval estimate*.

An urban economist might want to know whether neighborhood crime watch programs are associated with lower crime rates. After comparing crime rates of neighborhoods with and without such programs in a sample from the population, he or she can draw one of two conclusions: neighborhood watch programs do affect crime, or they do not. This example falls under the rubric of hypothesis testing.

The first step in statistical inference is to identify the population of interest. This may seem obvious, but it is important to be very specific. Once we have identified the population, we can specify a model for the population relationship of interest. Such models involve probability distributions or features of probability distributions, and these depend on unknown parameters. Parameters are simply constants that determine the directions and strengths of relationships among variables. In the labor economics example just presented, the parameter of interest is the return to education in the population.

C-1a Sampling

For reviewing statistical inference, we focus on the simplest possible setting. Let Y be a random variable representing a population with a probability density function $f(y; \theta)$, which depends on the single parameter θ. The probability density function (pdf) of Y is assumed to be known except for the value of θ; different values of θ imply different population distributions, and therefore we are interested in the value of θ. If we can obtain certain kinds of samples from the population, then we can learn something about θ. The easiest sampling scheme to deal with is random sampling.

Random Sampling. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables with a common probability density function $f(y; \theta)$, then $\{Y_1, \ldots, Y_n\}$ is said to be a **random sample** from $f(y; \theta)$ [or a random sample from the population represented by $f(y; \theta)$].

When $\{Y_1, \ldots, Y_n\}$ is a random sample from the density $f(y; \theta)$, we also say that the Y_i are *independent, identically distributed* (or *i.i.d.*) random variables from $f(y; \theta)$. In some cases, we will not need to entirely specify what the common distribution is.

The random nature of $Y_1, Y_2, \ldots, Y_n$ in the definition of random sampling reflects the fact that many different outcomes are possible before the sampling is actually carried out. For example, if family income is obtained for a sample of $n = 100$ families in the United States, the incomes we observe will usually differ for each different sample of 100 families. Once a sample is obtained, we have a set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, which constitute the data that we work with. Whether or not it is appropriate to assume the sample came from a random sampling scheme requires knowledge about the actual sampling process.

Random samples from a Bernoulli distribution are often used to illustrate statistical concepts, and they also arise in empirical applications. If $Y_1, Y_2, \ldots, Y_n$ are independent random variables and each is distributed as Bernoulli(θ), so that $\mathrm{P}(Y_i = 1) = \theta$ and $\mathrm{P}(Y_i = 0) = 1 - \theta$, then $\{Y_1, Y_2, \ldots, Y_n\}$ constitutes a random sample from the Bernoulli(θ) distribution. As an illustration, consider the airline reservation example carried along in Appendix B. Each Y_i denotes whether customer i shows up for his or her reservation; $Y_i = 1$ if passenger i shows up, and $Y_i = 0$ otherwise. Here, θ is the probability that a randomly drawn person from the population of all people who make airline reservations shows up for his or her reservation.

For many other applications, random samples can be assumed to be drawn from a normal distribution. If $\{Y_1, \ldots, Y_n\}$ is a random sample from the Normal(μ, σ^2) population, then the population is characterized by two parameters, the mean μ and the variance σ^2. Primary interest usually lies in μ, but σ^2 is of interest in its own right because making inferences about μ often requires learning about σ^2.

C-2 Finite Sample Properties of Estimators

In this section, we study what are called finite sample properties of estimators. The term "finite sample" comes from the fact that the properties hold for a sample of any size, no matter how large or small. Sometimes, these are called small sample properties. In Section C-3, we cover "asymptotic properties," which have to do with the behavior of estimators as the sample size grows without bound.

C-2a Estimators and Estimates

To study properties of estimators, we must define what we mean by an estimator. Given a random sample $\{Y_1, Y_2, \ldots, Y_n\}$ drawn from a population distribution that depends on an unknown parameter θ, an **estimator** of θ is a rule that assigns each possible outcome of the sample a value of θ. The rule is specified before any sampling is carried out; in particular, the rule is the same regardless of the data actually obtained.

As an example of an estimator, let $\{Y_1, \ldots, Y_n\}$ be a random sample from a population with mean μ. A natural estimator of μ is the average of the random sample:

$$\overline{Y} = n^{-1} \sum_{i=1}^{n} Y_i. \qquad [C.1]$$

$\overline{Y}$ is called the **sample average** but, unlike in Appendix A where we defined the sample average of a set of numbers as a descriptive statistic, $\overline{Y}$ is now viewed as an estimator. Given any outcome of the random variables $Y_1, \ldots, Y_n$, we use the same rule to estimate μ: we simply average them. For actual data outcomes $\{y_1, \ldots, y_n\}$, the **estimate** is just the average in the sample: $\overline{y} = (y_1 + y_2 + \ldots + y_n)/n$.

EXAMPLE C.1 **City Unemployment Rates**

Suppose we obtain the following sample of unemployment rates for 10 cities in the United States:

City	Unemployment Rate
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6
8	3.5
9	5.8
10	7.5

Our estimate of the average city unemployment rate in the United States is $\bar{y} = 6.0$. Each sample generally results in a different estimate. But the *rule* for obtaining the estimate is the same, regardless of which cities appear in the sample, or how many.

More generally, an estimator W of a parameter θ can be expressed as an abstract mathematical formula:

$$W = h(Y_1, Y_2, \ldots Y_n), \qquad \text{[C.2]}$$

for some known function h of the random variables $Y_1, Y_2, \ldots, Y_n$. As with the special case of the sample average, W is a random variable because it depends on the random sample: as we obtain different random samples from the population, the value of W can change. When a particular set of numbers, say, $\{y_1, y_2, \ldots, y_n\}$, is plugged into the function h, we obtain an *estimate* of θ, denoted $w = h(y_1, \ldots, y_n)$. Sometimes, W is called a point estimator and w a point estimate to distinguish these from *interval* estimators and estimates, which we will come to in Section C-5.

For evaluating estimation procedures, we study various properties of the probability distribution of the random variable W. The distribution of an estimator is often called its **sampling distribution**, because this distribution describes the likelihood of various outcomes of W across different random samples. Because there are unlimited rules for combining data to estimate parameters, we need some sensible criteria for choosing among estimators, or at least for eliminating some estimators from consideration. Therefore, we must leave the realm of descriptive statistics, where we compute things such as the sample average to simply summarize a body of data. In mathematical statistics, we study the sampling distributions of estimators.

C-2b Unbiasedness

In principle, the entire sampling distribution of W can be obtained given the probability distribution of Y_i and the function h. It is usually easier to focus on a few features of the distribution of W in evaluating it as an estimator of θ. The first important property of an estimator involves its expected value.

Unbiased Estimator. An estimator, W of θ, is an **unbiased estimator** if

$$\mathrm{E}(W) = \theta, \qquad \text{[C.3]}$$

for all possible values of θ.

If an estimator is unbiased, then its probability distribution has an expected value equal to the parameter it is supposed to be estimating. Unbiasedness does *not* mean that the estimate we get with any particular sample is equal to θ, or even very close to θ. Rather, if we could *indefinitely* draw random samples on Y from the population, compute an estimate each time, and then average these estimates over all random samples, we would obtain θ. This thought experiment is abstract because, in most applications, we just have one random sample to work with.

For an estimator that is not unbiased, we define its **bias** as follows.

Bias of an Estimator. If W is a **biased estimator** of θ, its bias is defined

$$\text{Bias}(W) \equiv \text{E}(W) - \theta. \qquad \text{[C.4]}$$

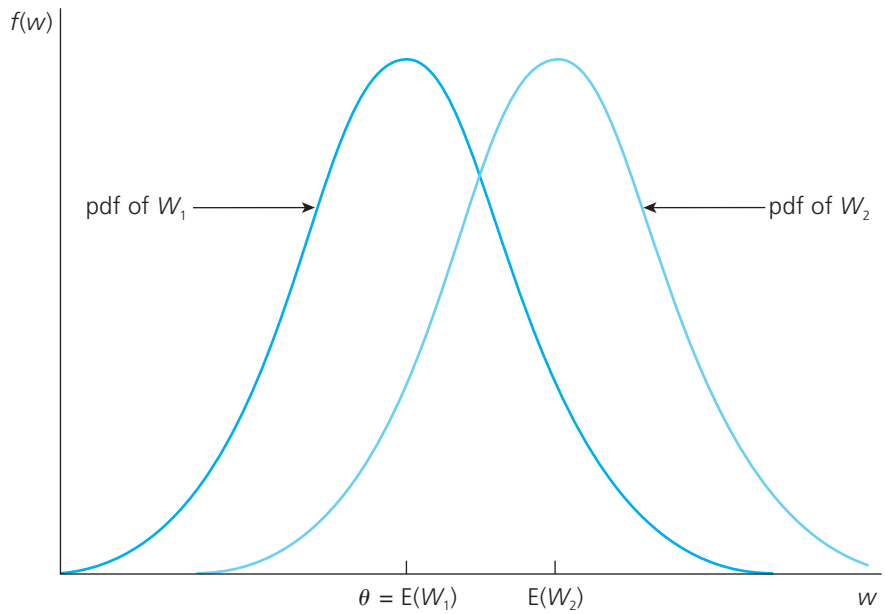
Figure C.1 shows two estimators; the first one is unbiased, and the second one has a positive bias.

The unbiasedness of an estimator and the size of any possible bias depend on the distribution of Y and on the function h. The distribution of Y is usually beyond our control (although we often choose a *model* for this distribution): it may be determined by nature or social forces. But the choice of the rule h is ours, and if we want an unbiased estimator, then we must choose h accordingly.

Some estimators can be shown to be unbiased quite generally. We now show that the sample average $\overline{Y}$ is an unbiased estimator of the population mean μ, regardless of the underlying population distribution. We use the properties of expected values (E.1 and E.2) that we covered in Section B-3:

$$\text{E}(\overline{Y}) = \text{E}\left((1/n)\sum_{i=1}^{n} Y_i\right) = (1/n)\text{E}\left(\sum_{i=1}^{n} Y_i\right) = (1/n)\left(\sum_{i=1}^{n}\text{E}(Y_i)\right)$$
$$= (1/n)\left(\sum_{i=1}^{n}\mu\right) = (1/n)(n\mu) = \mu.$$

FIGURE C.1 An unbiased estimator, W_1, and an estimator with positive bias, W_2.



For hypothesis testing, we will need to estimate the variance σ^2 from a population with mean μ. Letting $\{Y_1, \ldots, Y_n\}$ denote the random sample from the population with $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \sigma^2$, define the estimator as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \text{[C.5]}$$

which is usually called the **sample variance**. It can be shown that S^2 is unbiased for σ^2: $\mathrm{E}(S^2) = \sigma^2$. The division by $n - 1$, rather than n, accounts for the fact that the mean μ is estimated rather than known. If μ were known, an unbiased estimator of σ^2 would be $n^{-1}\sum_{i=1}^{n}(Y_i - \mu)^2$, but μ is rarely known in practice.

Although unbiasedness has a certain appeal as a property for an estimator—indeed, its antonym, "biased," has decidedly negative connotations—it is not without its problems. One weakness of unbiasedness is that some reasonable, and even some very good, estimators are not unbiased. We will see an example shortly.

Another important weakness of unbiasedness is that unbiased estimators exist that are actually quite poor estimators. Consider estimating the mean μ from a population. Rather than using the sample average $\bar{Y}$ to estimate μ, suppose that, after collecting a sample of size n, we discard all of the observations except the first. That is, our estimator of μ is simply $W \equiv Y_1$. This estimator is unbiased because $\mathrm{E}(Y_1) = \mu$. Hopefully, you sense that ignoring all but the first observation is not a prudent approach to estimation: it throws out most of the information in the sample. For example, with $n = 100$, we obtain 100 outcomes of the random variable Y, but then we use only the first of these to estimate $\mathrm{E}(Y)$.

C-2d The Sampling Variance of Estimators

The example at the end of the previous subsection shows that we need additional criteria to evaluate estimators. Unbiasedness only ensures that the sampling distribution of an estimator has a mean value equal to the parameter it is supposed to be estimating. This is fine, but we also need to know how spread out the distribution of an estimator is. An estimator can be equal to θ, on average, but it can also be very far away with large probability. In Figure C.2, W_1 and W_2 are both unbiased estimators of θ. But the distribution of W_1 is more tightly centered about θ: the probability that W_1 is greater than any given distance from θ is less than the probability that W_2 is greater than that same distance from θ. Using W_1 as our estimator means that it is less likely that we will obtain a random sample that yields an estimate very far from θ.

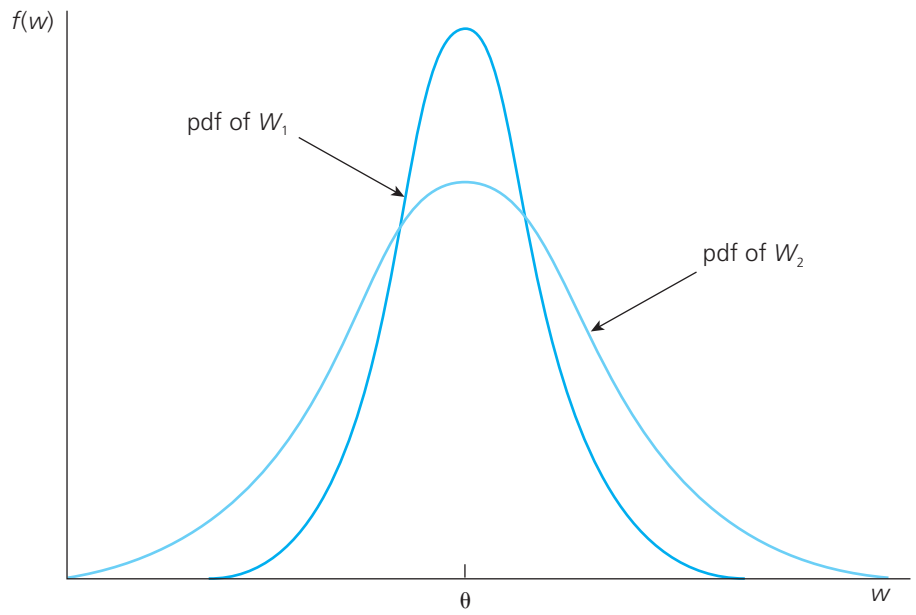
To summarize the situation shown in Figure C.2, we rely on the variance (or standard deviation) of an estimator. Recall that this gives a single measure of the dispersion in the distribution. The variance of an estimator is often called its **sampling variance** because it is the variance associated with a sampling distribution. Remember, the sampling variance is not a random variable; it is a constant, but it might be unknown.

We now obtain the variance of the sample average for estimating the mean μ from a population:

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \mathrm{Var}\left((1/n)\sum_{i=1}^{n}Y_i\right) = (1/n^2)\mathrm{Var}\left(\sum_{i=1}^{n}Y_i\right) = (1/n^2)\left(\sum_{i=1}^{n}\mathrm{Var}(Y_i)\right)\\ &= (1/n^2)\left(\sum_{i=1}^{n}\sigma^2\right) = (1/n^2)(n\sigma^2) = \sigma^2/n. \qquad \text{[C.6]}\end{aligned}$$

Notice how we used the properties of variance from Sections B-3 and B-4 (VAR.2 and VAR.4), as well as the independence of the Y_i. To summarize: If $\{Y_i: i = 1, 2, \ldots, n\}$ is a random sample from a population with mean μ and variance σ^2, then $\bar{Y}$ has the same mean as the population, but its sampling variance equals the population variance, σ^2, divided by the sample size.

FIGURE C.2 The sampling distributions of two unbiased estimators of θ.



An important implication of $\text{Var}(\bar{Y}) = \sigma^2/n$ is that it can be made very close to zero by increasing the sample size n. This is a key feature of a reasonable estimator, and we return to it in Section C-3.

As suggested by Figure C.2, among unbiased estimators, we prefer the estimator with the smallest variance. This allows us to eliminate certain estimators from consideration. For a random sample from a population with mean μ and variance σ^2, we know that $\bar{Y}$ is unbiased and $\text{Var}(\bar{Y}) = \sigma^2/n$. What about the estimator Y_1, which is just the first observation drawn? Because Y_1 is a random draw from the population, $\text{Var}(Y_1) = \sigma^2$. Thus, the difference between $\text{Var}(Y_1)$ and $\text{Var}(\bar{Y})$ can be large even for small sample sizes. If $n = 10$, then $\text{Var}(Y_1)$ is 10 times as large as $\text{Var}(\bar{Y}) = \sigma^2/10$. This gives us a formal way of excluding Y_1 as an estimator of μ.

To emphasize this point, Table C.1 contains the outcome of a small simulation study. Using the statistical package Stata®, 20 random samples of size 10 were generated from a normal distribution, with $\mu = 2$ and $\sigma^2 = 1$; we are interested in estimating μ here. For each of the 20 random samples, we compute two estimates, y_1 and $\bar{y}$; these values are listed in Table C.1. As can be seen from the table, the values for y_1 are much more spread out than those for $\bar{y}$: y_1 ranges from -0.64 to 4.27, while $\bar{y}$ ranges only from 1.16 to 2.58. Further, in 16 out of 20 cases, $\bar{y}$ is closer than y_1 to $\mu = 2$. The average of y_1 across the simulations is about 1.89, while that for $\bar{y}$ is 1.96. The fact that these averages are close to 2 illustrates the unbiasedness of both estimators (and we could get these averages closer to 2 by doing more than 20 replications). But comparing just the average outcomes across random draws masks the fact that the sample average $\bar{Y}$ is far superior to Y_1 as an estimator of μ.

C-2e Efficiency

Comparing the variances of $\bar{Y}$ and Y_1 in the previous subsection is an example of a general approach to comparing different unbiased estimators.

Relative Efficiency. If W_1 and W_2 are two unbiased estimators of θ, W_1 is efficient relative to W_2 when $\text{Var}(W_1) \leq \text{Var}(W_2)$ for all θ, with strict inequality for at least one value of θ.

TABLE C.1 Simulation of Estimators for a Normal(μ, 1) Distribution with $\mu = 2$

Replication	y_1	$\bar{y}$
1	−0.64	1.98
2	1.06	1.43
3	4.27	1.65
4	1.03	1.88
5	3.16	2.34
6	2.77	2.58
7	1.68	1.58
8	2.98	2.23
9	2.25	1.96
10	2.04	2.11
11	0.95	2.15
12	1.36	1.93
13	2.62	2.02
14	2.97	2.10
15	1.93	2.18
16	1.14	2.10
17	2.08	1.94
18	1.52	2.21
19	1.33	1.16
20	1.21	1.75

Earlier, we showed that, for estimating the population mean μ, $\text{Var}(\bar{Y}) < \text{Var}(Y_1)$ for any value of σ^2 whenever $n > 1$. Thus, $\bar{Y}$ is efficient relative to Y_1 for estimating μ. We cannot always choose between unbiased estimators based on the smallest variance criterion: given two unbiased estimators of θ, one can have smaller variance from some values of θ, while the other can have smaller variance for other values of θ.

If we restrict our attention to a certain class of estimators, we can show that the sample average has the smallest variance. Problem C.2 asks you to show that $\bar{Y}$ has the smallest variance among all unbiased estimators that are also linear functions of $Y_1, Y_2, \ldots, Y_n$. The assumptions are that the Y_i have common mean and variance, and that they are pairwise uncorrelated.

If we do not restrict our attention to unbiased estimators, then comparing variances is meaningless. For example, when estimating the population mean μ, we can use a trivial estimator that is equal to zero, regardless of the sample that we draw. Naturally, the variance of this estimator is zero (since it is the same value for every random sample). But the bias of this estimator is $-\mu$, so it is a very poor estimator when $|\mu|$ is large.

One way to compare estimators that are not necessarily unbiased is to compute the **mean squared error (MSE)** of the estimators. If W is an estimator of θ, then the MSE of W is defined as $\text{MSE}(W) = \text{E}[(W - \theta)^2]$. The MSE measures how far, on average, the estimator is away from θ. It can be shown that $\text{MSE}(W) = \text{Var}(W) + [\text{Bias}(W)]^2$, so that $\text{MSE}(W)$ depends on the variance and bias (if any is present). This allows us to compare two estimators when one or both are biased.

C-3 Asymptotic or Large Sample Properties of Estimators

In Section C-2, we encountered the estimator Y_1 for the population mean μ, and we saw that, even though it is unbiased, it is a poor estimator because its variance can be much larger than that of the sample mean. One notable feature of Y_1 is that it has the same variance for any sample size. It seems reasonable to require any estimation procedure to improve as the sample size increases. For estimating a population mean μ, $\overline{Y}$ improves in the sense that its variance gets smaller as n gets larger; Y_1 does not improve in this sense.

We can rule out certain silly estimators by studying the *asymptotic* or *large sample* properties of estimators. In addition, we can say something positive about estimators that are not unbiased and whose variances are not easily found.

Asymptotic analysis involves approximating the features of the sampling distribution of an estimator. These approximations depend on the size of the sample. Unfortunately, we are necessarily limited in what we can say about how "large" a sample size is needed for asymptotic analysis to be appropriate; this depends on the underlying population distribution. But large sample approximations have been known to work well for sample sizes as small as $n = 20$.

C-3a Consistency

The first asymptotic property of estimators concerns how far the estimator is likely to be from the parameter it is supposed to be estimating as we let the sample size increase indefinitely.

Consistency. Let W_n be an estimator of θ based on a sample $Y_1, Y_2, \ldots, Y_n$ of size n. Then, W_n is a **consistent estimator** of θ if for every $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty. \qquad [C.7]$$

If W_n is not consistent for θ, then we say it is **inconsistent**.

When W_n is consistent, we also say that θ is the **probability limit** of W_n, written as $\text{plim}(W_n) = \theta$.

Unlike unbiasedness—which is a feature of an estimator for a given sample size—consistency involves the behavior of the sampling distribution of the estimator as the sample size n gets large. To emphasize this, we have indexed the estimator by the sample size in stating this definition, and we will continue with this convention throughout this section.

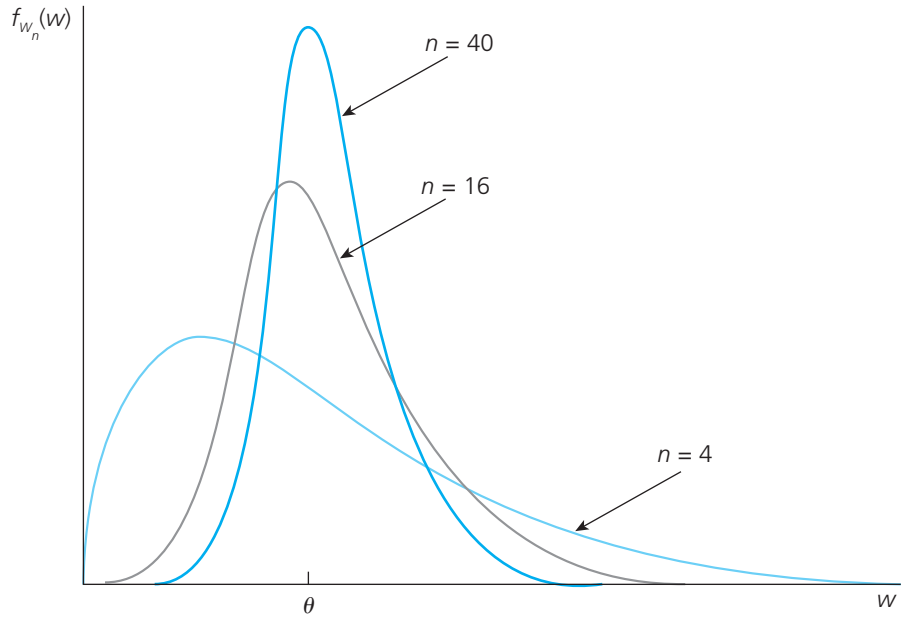
Equation (C.7) looks technical, and it can be rather difficult to establish based on fundamental probability principles. By contrast, interpreting (C.7) is straightforward. It means that the distribution of W_n becomes more and more concentrated about θ, which roughly means that for larger sample sizes, W_n is less and less likely to be very far from θ. This tendency is illustrated in Figure C.3.

If an estimator is not consistent, then it does not help us to learn about θ, even with an unlimited amount of data. For this reason, consistency is a minimal requirement of an estimator used in statistics or econometrics. We will encounter estimators that are consistent under certain assumptions and inconsistent when those assumptions fail. When estimators are inconsistent, we can usually find their probability limits, and it will be important to know how far these probability limits are from θ.

As we noted earlier, unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows *are* consistent. This can be stated formally: If W_n is an unbiased estimator of θ and $\text{Var}(W_n) \to 0$ as $n \to \infty$, then $\text{plim}(W_n) = \theta$. Unbiased estimators that use the entire data sample will usually have a variance that shrinks to zero as the sample size grows, thereby being consistent.

A good example of a consistent estimator is the average of a random sample drawn from a population with mean μ and variance σ^2. We have already shown that the sample average is unbiased for μ.

FIGURE C.3 The sampling distributions of a consistent estimator for three sample sizes.



In Equation (C.6), we derived $\text{Var}(\bar{Y}_n) = \sigma^2/n$ for any sample size n. Therefore, $\text{Var}(\bar{Y}_n) \to 0$ as $n \to \infty$, so $\bar{Y}_n$ is a consistent estimator of μ (in addition to being unbiased).

The conclusion that $\bar{Y}_n$ is consistent for μ holds even if $\text{Var}(\bar{Y}_n)$ does not exist. This classic result is known as the **law of large numbers (LLN)**.

Law of Large Numbers. Let $Y_1, Y_2, \ldots, Y_n$ be independent, identically distributed random variables with mean μ. Then,

$$\text{plim}(\bar{Y}_n) = \mu. \qquad \text{[C.8]}$$

The law of large numbers means that, if we are interested in estimating the population average μ, we can get arbitrarily close to μ by choosing a sufficiently large sample. This fundamental result can be combined with basic properties of plims to show that fairly complicated estimators are consistent.

Property PLIM.1: Let θ be a parameter and define a new parameter, $\gamma = g(\theta)$, for some *continuous function* $g(\theta)$. Suppose that $\text{plim}(W_n) = \theta$. Define an estimator of γ by $G_n = g(W_n)$. Then,

$$\text{plim}(G_n) = \gamma. \qquad \text{[C.9]}$$

This is often stated as

$$\text{plim } g(W_n) = g(\text{plim } W_n) \qquad \text{[C.10]}$$

for a continuous function $g(\theta)$.

The assumption that $g(\theta)$ is continuous is a technical requirement that has often been described nontechnically as "a function that can be graphed without lifting your pencil from the paper." Because all the functions we encounter in this text are continuous, we do not provide a formal definition of a continuous function. Examples of continuous functions are $g(\theta) = a + b\theta$ for constants a and b, $g(\theta) = \theta^2$, $g(\theta) = 1/\theta$, $g(\theta) = \sqrt{\theta}$, $g(\theta) = \exp(\theta)$, and many variants on these. We will not need to mention the continuity assumption again.

As an important example of a consistent but biased estimator, consider estimating the standard deviation, σ, from a population with mean μ and variance σ^2. We already claimed that the sample variance $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$ is unbiased for σ^2. Using the law of large numbers and some algebra, S_n^2 can also be shown to be consistent for σ^2. The natural estimator of $\sigma = \sqrt{\sigma^2}$ is $S_n = \sqrt{S_n^2}$ (where the square root is always the positive square root). S_n, which is called the **sample standard deviation**, is *not* an unbiased estimator because the expected value of the square root is *not* the square root of the expected value (see Section B-3). Nevertheless, by PLIM.1, $\text{plim}\, S_n = \sqrt{\text{plim}\, S_n^2} = \sqrt{\sigma^2} = \sigma$, so S_n is a consistent estimator of σ.

Here are some other useful properties of the probability limit:

Property PLIM.2: If $\text{plim}(T_n) = \alpha$ and $\text{plim}(U_n) = \beta$, then

(i) $\text{plim}(T_n + U_n) = \alpha + \beta$;
(ii) $\text{plim}(T_n U_n) = \alpha\beta$;
(iii) $\text{plim}(T_n/U_n) = \alpha/\beta$, provided $\beta \neq 0$.

These three facts about probability limits allow us to combine consistent estimators in a variety of ways to get other consistent estimators. For example, let $\{Y_1, \ldots, Y_n\}$ be a random sample of size n on annual earnings from the population of workers with a high school education and denote the population mean by μ_Y. Let $\{Z_1, \ldots, Z_n\}$ be a random sample on annual earnings from the population of workers with a college education and denote the population mean by μ_Z. We wish to estimate the percentage difference in annual earnings between the two groups, which is $\gamma = 100\cdot(\mu_Z - \mu_Y)/\mu_Y$. (This is the percentage by which average earnings for college graduates differs from average earnings for high school graduates.) Because $\overline{Y}_n$ is consistent for μ_Y and $\overline{Z}_n$ is consistent for μ_Z, it follows from PLIM.1 and part (iii) of PLIM.2 that

$$G_n \equiv 100\cdot(\overline{Z}_n - \overline{Y}_n)/\overline{Y}_n$$

is a consistent estimator of γ. G_n is just the percentage difference between $\overline{Z}_n$ and $\overline{Y}_n$ in the sample, so it is a natural estimator. G_n is not an unbiased estimator of γ, but it is still a good estimator except possibly when n is small.

C-3b Asymptotic Normality

Consistency is a property of point estimators. Although it does tell us that the distribution of the estimator is collapsing around the parameter as the sample size gets large, it tells us essentially nothing about the *shape* of that distribution for a given sample size. For constructing interval estimators and testing hypotheses, we need a way to approximate the distribution of our estimators. Most econometric estimators have distributions that are well approximated by a normal distribution for large samples, which motivates the following definition.

Asymptotic Normality. Let $\{Z_n: n = 1, 2, \ldots\}$ be a sequence of random variables, such that for all numbers z,

$$\text{P}(Z_n \leq z) \to \Phi(z) \text{ as } n \to \infty, \qquad \text{[C.11]}$$

where $\Phi(z)$ is the standard normal cumulative distribution function. Then, Z_n is said to have an *asymptotic standard normal distribution*. In this case, we often write $Z_n \overset{a}{\sim} \text{Normal}(0, 1)$. (The "$a$" above the tilde stands for "asymptotically" or "approximately.")

Property (C.11) means that the cumulative distribution function for Z_n gets closer and closer to the cdf of the standard normal distribution as the sample size n gets large. When **asymptotic normality** holds, for large n we have the approximation $\text{P}(Z_n \leq z) \approx \Phi(z)$. Thus, probabilities concerning Z_n can be approximated by standard normal probabilities.

The **central limit theorem (CLT)** is one of the most powerful results in probability and statistics. It states that the average from a random sample for *any* population (with finite variance), when standardized, has an asymptotic standard normal distribution.

Central Limit Theorem. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample with mean μ and variance σ^2. Then,

$$Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \qquad [C.12]$$

has an asymptotic standard normal distribution.

The variable Z_n in (C.12) is the standardized version of $\bar{Y}_n$: we have subtracted off $\mathrm{E}(\bar{Y}_n) = \mu$ and divided by $\mathrm{sd}(\bar{Y}_n) = \sigma/\sqrt{n}$. Thus, regardless of the population distribution of Y, Z_n has mean zero and variance one, which coincides with the mean and variance of the standard normal distribution. Remarkably, the entire distribution of Z_n gets arbitrarily close to the standard normal distribution as n gets large.

We can write the standardized variable in equation (C.12) as $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, which shows that we must multiply the difference between the sample mean and the population mean by the square root of the sample size in order to obtain a useful limiting distribution. Without the multiplication by $\sqrt{n}$, we would just have $(\bar{Y}_n - \mu)/\sigma$, which converges in probability to zero. In other words, the distribution of $(\bar{Y}_n - \mu)/\sigma$ simply collapses to a single point as $n \to \infty$, which we know cannot be a good approximation to the distribution of $(\bar{Y}_n - \mu)/\sigma$ for reasonable sample sizes. Multiplying by $\sqrt{n}$ ensures that the variance of Z_n remains constant. Practically, we often treat $\bar{Y}_n$ as being approximately normally distributed with mean μ and variance σ^2/n, and this gives us the correct statistical procedures because it leads to the standardized variable in equation (C.12).

Most estimators encountered in statistics and econometrics can be written as functions of sample averages, in which case we can apply the law of large numbers and the central limit theorem. When two consistent estimators have asymptotic normal distributions, we choose the estimator with the smallest asymptotic variance.

In addition to the standardized sample average in (C.12), many other statistics that depend on sample averages turn out to be asymptotically normal. An important one is obtained by replacing σ with its consistent estimator S_n in equation (C.12):

$$\frac{\bar{Y}_n - \mu}{S_n/\sqrt{n}} \qquad [C.13]$$

also has an approximate standard normal distribution for large n. The exact (finite sample) distributions of (C.12) and (C.13) are definitely not the same, but the difference is often small enough to be ignored for large n.

Throughout this section, each estimator has been subscripted by n to emphasize the nature of asymptotic or large sample analysis. Continuing this convention clutters the notation without providing additional insight, once the fundamentals of asymptotic analysis are understood. Henceforth, we drop the n subscript and rely on you to remember that estimators depend on the sample size, and properties such as consistency and asymptotic normality refer to the growth of the sample size without bound.

C-4 General Approaches to Parameter Estimation

Until this point, we have used the sample average to illustrate the finite and large sample properties of estimators. It is natural to ask: Are there general approaches to estimation that produce estimators with good properties, such as unbiasedness, consistency, and efficiency?

The answer is yes. A detailed treatment of various approaches to estimation is beyond the scope of this text; here, we provide only an informal discussion. A thorough discussion is given in Larsen and Marx (1986, Chapter 5).

C-4a Method of Moments

Given a parameter θ appearing in a population distribution, there are usually many ways to obtain unbiased and consistent estimators of θ. Trying all different possibilities and comparing them on the basis of the criteria in Sections C-2 and C-3 is not practical. Fortunately, some methods have been shown to have good general properties, and, for the most part, the logic behind them is intuitively appealing.

In the previous sections, we have studied the sample average as an unbiased estimator of the population average and the sample variance as an unbiased estimator of the population variance. These estimators are examples of **method of moments** estimators. Generally, method of moments estimation proceeds as follows. The parameter θ is shown to be related to some expected value in the distribution of Y, usually $E(Y)$ or $E(Y^2)$ (although more exotic choices are sometimes used). Suppose, for example, that the parameter of interest, θ, is related to the population mean as $\theta = g(\mu)$ for some function g. Because the sample average $\bar{Y}$ is an unbiased and consistent estimator of μ, it is natural to replace μ with $\bar{Y}$, which gives us the estimator $g(\bar{Y})$ of θ. The estimator $g(\bar{Y})$ is consistent for θ, and if $g(\mu)$ is a linear function of μ, then $g(\bar{Y})$ is unbiased as well. What we have done is replace the population moment, μ, with its sample counterpart, $\bar{Y}$. This is where the name "method of moments" comes from.

We cover two additional method of moments estimators that will be useful for our discussion of regression analysis. Recall that the covariance between two random variables X and Y is defined as $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$. The method of moments suggests estimating σ_{XY} by $n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. This is a consistent estimator of σ_{XY}, but it turns out to be biased for essentially the same reason that the sample variance is biased if n, rather than $n - 1$, is used as the divisor. The **sample covariance** is defined as

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \qquad \text{[C.14]}$$

It can be shown that this is an unbiased estimator of σ_{XY}. (Replacing n with $n - 1$ makes no difference as the sample size grows indefinitely, so this estimator is still consistent.)

As we discussed in Section B-4, the covariance between two variables is often difficult to interpret. Usually, we are more interested in correlation. Because the population correlation is $\rho_{XY} = \sigma_{XY}/(\sigma_X\sigma_Y)$, the method of moments suggests estimating ρ_{XY} as

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)^{1/2}}, \qquad \text{[C.15]}$$

which is called the **sample correlation coefficient** (or *sample correlation* for short). Notice that we have canceled the division by $n - 1$ in the sample covariance and the sample standard deviations. In fact, we could divide each of these by n, and we would arrive at the same final formula.

It can be shown that the sample correlation coefficient is always in the interval $[-1,1]$, as it should be. Because S_{XY}, S_X, and S_Y are consistent for the corresponding population parameter, R_{XY} is a consistent estimator of the population correlation, ρ_{XY}. However, R_{XY} is a biased estimator for two reasons. First, S_X and S_Y are biased estimators of σ_X and σ_Y, respectively. Second, R_{XY} is a ratio of estimators, so it would not be unbiased, even if S_X and S_Y were. For our purposes, this is not important, although the fact that no unbiased estimator of ρ_{XY} exists is a classical result in mathematical statistics.

C-4b Maximum Likelihood

Another general approach to estimation is the method of *maximum likelihood*, a topic covered in many introductory statistics courses. A brief summary in the simplest case will suffice here. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample from the population distribution $f(y; \theta)$. Because of the random

sampling assumption, the joint distribution of $\{Y_1, Y_2, \ldots, Y_n\}$ is simply the product of the densities: $f(y_1; \theta)f(y_2; \theta) \cdots f(y_n; \theta)$. In the discrete case, this is $P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$. Now, define the *likelihood function* as

$$L(\theta; Y_1, \ldots, Y_n) = f(Y_1; \theta)f(Y_2; \theta) \cdots f(Y_n; \theta),$$

which is a random variable because it depends on the outcome of the random sample $\{Y_1, Y_2, \ldots, Y_n\}$. The **maximum likelihood estimator** of θ, call it W, is the value of θ that maximizes the likelihood function. (This is why we write L as a function of θ, followed by the random sample.) Clearly, this value depends on the random sample. The maximum likelihood principle says that, out of all the possible values for θ, the value that makes the likelihood of the observed data largest should be chosen. Intuitively, this is a reasonable approach to estimating θ.

Usually, it is more convenient to work with the *log-likelihood function*, which is obtained by taking the natural log of the likelihood function:

$$\log[L(\theta; Y_1, \ldots, Y_n)] = \sum_{i=1}^{n} \log[f(Y_i; \theta)], \qquad \text{[C.16]}$$

where we use the fact that the log of the product is the sum of the logs. Because (C.16) is the sum of independent, identically distributed random variables, analyzing estimators that come from (C.16) is relatively easy.

Maximum likelihood estimation (MLE) is usually consistent and sometimes unbiased. But so are many other estimators. The widespread appeal of MLE is that it is generally the most asymptotically efficient estimator when the population model $f(y; \theta)$ is correctly specified. In addition, the MLE is sometimes the **minimum variance unbiased estimator**; that is, it has the smallest variance among all unbiased estimators of θ. [See Larsen and Marx (1986, Chapter 5) for verification of these claims.]

In Chapter 17, we will need maximum likelihood to estimate the parameters of more advanced econometric models. In econometrics, we are almost always interested in the distribution of Y conditional on a set of explanatory variables, say, $X_1, X_2, \ldots, X_k$. Then, we replace the density in (C.16) with $f(Y_i|X_{i1}, \ldots, X_{ik}; \theta_1, \ldots, \theta_p)$, where this density is allowed to depend on p parameters, $\theta_1, \ldots, \theta_p$. Fortunately, for successful application of maximum likelihood methods, we do not need to delve much into the computational issues or the large-sample statistical theory. Wooldridge (2010, Chapter 13) covers the theory of MLE.

C-4c Least Squares

A third kind of estimator, and one that plays a major role throughout the text, is called a **least squares estimator**. We have already seen an example of least squares: the sample mean, $\overline{Y}$, is a least squares estimator of the population mean, μ. We already know $\overline{Y}$ is a method of moments estimator. What makes it a least squares estimator? It can be shown that the value of m that makes the sum of squared deviations

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

as small as possible is $m = \overline{Y}$. Showing this is not difficult, but we omit the algebra.

For some important distributions, including the normal and the Bernoulli, the sample average $\overline{Y}$ is also the maximum likelihood estimator of the population mean μ. Thus, the principles of least squares, method of moments, and maximum likelihood often result in the *same* estimator. In other cases, the estimators are similar but not identical.

C-5 Interval Estimation and Confidence Intervals

C-5a The Nature of Interval Estimation

A point estimate obtained from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but, by its nature, it provides no information about how close the estimate is "likely" to be to the population parameter. As an example, suppose a researcher reports, on the basis of a random sample of workers, that job training grants increase hourly wage by 6.4%. How are we to know whether or not this is close to the effect in the population of workers who could have been trained? Because we do not know the population value, we cannot know how close an estimate is for a particular sample. However, we can make statements involving probabilities, and this is where interval estimation comes in.

We already know one way of assessing the uncertainty in an estimator: find its sampling standard deviation. Reporting the standard deviation of the estimator, along with the point estimate, provides some information on the accuracy of our estimate. However, even if the problem of the standard deviation's dependence on unknown population parameters is ignored, reporting the standard deviation along with the point estimate makes no direct statement about where the population value is likely to lie in relation to the estimate. This limitation is overcome by constructing a **confidence interval**.

We illustrate the concept of a confidence interval with an example. Suppose the population has a Normal$(\mu, 1)$ distribution and let $\{Y_1, \ldots, Y_n\}$ be a random sample from this population. (We assume that the variance of the population is known and equal to unity for the sake of illustration; we then show what to do in the more realistic case that the variance is unknown.) The sample average, $\overline{Y}$, has a normal distribution with mean μ and variance $1/n$: $\overline{Y} \sim \text{Normal}(\mu, 1/n)$. From this, we can standardize $\overline{Y}$, and, because the standardized version of $\overline{Y}$ has a standard normal distribution, we have

$$\text{P}\left(-1.96 < \frac{\overline{Y} - \mu}{1/\sqrt{n}} < 1.96\right) = .95.$$

The event in parentheses is identical to the event $\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96/\sqrt{n}$, so

$$\text{P}(\overline{Y} - 1.96/\sqrt{n} < \mu < \overline{Y} + 1.96\sqrt{n}) = .95. \qquad \text{[C.17]}$$

Equation (C.17) is interesting because it tells us that the probability that the random interval $[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}]$ contains the population mean μ is .95, or 95%. This information allows us to construct an *interval estimate* of μ, which is obtained by plugging in the sample outcome of the average, $\bar{y}$. Thus,

$$[\bar{y} - 1.96/\sqrt{n}, \bar{y} + 1.96/\sqrt{n}] \qquad \text{[C.18]}$$

is an example of an interval estimate of μ. It is also called a 95% confidence interval. A shorthand notation for this interval is $\bar{y} \pm 1.96/\sqrt{n}$.

The confidence interval in equation (C.18) is easy to compute, once the sample data $\{y_1, y_2, \ldots, y_n\}$ are observed; $\bar{y}$ is the only factor that depends on the data. For example, suppose that $n = 16$ and the average of the 16 data points is 7.3. Then, the 95% confidence interval for μ is $7.3 \pm 1.96/\sqrt{16} = 7.3 \pm .49$, which we can write in interval form as [6.81,7.79]. By construction, $\bar{y} = 7.3$ is in the center of this interval.

Unlike its computation, the meaning of a confidence interval is more difficult to understand. When we say that equation (C.18) is a 95% confidence interval for μ, we mean that the *random* interval

$$[\overline{Y} - 1.96/\sqrt{n}, \overline{Y} + 1.96/\sqrt{n}] \qquad \text{[C.19]}$$

contains μ with probability .95. In other words, *before* the random sample is drawn, there is a 95% chance that (C.19) contains μ. Equation (C.19) is an example of an **interval estimator**. It is a random interval, since the endpoints change with different samples.

A confidence interval is often interpreted as follows: "The probability that μ is in the interval (C.18) is .95." This is incorrect. Once the sample has been observed and $\bar{y}$ has been computed, the limits of the confidence interval are simply numbers (6.81 and 7.79 in the example just given). The population parameter, μ, though unknown, is also just some number. Therefore, μ either is or is not in the interval (C.18) (and we will never know with certainty which is the case). Probability plays no role once the confidence interval is computed for the particular data at hand. The probabilistic interpretation comes from the fact that for 95% of all random samples, the constructed confidence interval will contain μ.

To emphasize the meaning of a confidence interval, Table C.2 contains calculations for 20 random samples (or replications) from the Normal(2,1) distribution with sample size $n = 10$. For each of the 20 samples, $\bar{y}$ is obtained, and (C.18) is computed as $\bar{y} \pm 1.96/\sqrt{10} = \bar{y} \pm .62$ (each rounded to two decimals). As you can see, the interval changes with each random sample. Nineteen of the twenty intervals contain the population value of μ. Only for replication number 19 is μ not in the confidence interval. In other words, 95% of the samples result in a confidence interval that contains μ. This did not have to be the case with only 20 replications, but it worked out that way for this particular simulation.

TABLE C.2 Simulated Confidence Intervals from a Normal(μ, 1) Distribution with $\mu = 2$

Replication	$\bar{y}$	95% Interval	Contains μ?
1	1.98	(1.36,2.60)	Yes
2	1.43	(0.81,2.05)	Yes
3	1.65	(1.03,2.27)	Yes
4	1.88	(1.26,2.50)	Yes
5	2.34	(1.72,2.96)	Yes
6	2.58	(1.96,3.20)	Yes
7	1.58	(.96,2.20)	Yes
8	2.23	(1.61,2.85)	Yes
9	1.96	(1.34,2.58)	Yes
10	2.11	(1.49,2.73)	Yes
11	2.15	(1.53,2.77)	Yes
12	1.93	(1.31,2.55)	Yes
13	2.02	(1.40,2.64)	Yes
14	2.10	(1.48,2.72)	Yes
15	2.18	(1.56,2.80)	Yes
16	2.10	(1.48,2.72)	Yes
17	1.94	(1.32,2.56)	Yes
18	2.21	(1.59,2.83)	Yes
19	1.16	(.54,1.78)	No
20	1.75	(1.13,2.37)	Yes

C-5b Confidence Intervals for the Mean from a Normally Distributed Population

The confidence interval derived in equation (C.18) helps illustrate how to construct and interpret confidence intervals. In practice, equation (C.18) is not very useful for the mean of a normal population because it assumes that the variance is known to be unity. It is easy to extend (C.18) to the case where the standard deviation σ is known to be any value: the 95% confidence interval is

$$[\bar{y} - 1.96\sigma/\sqrt{n}, \bar{y} + 1.96\sigma/\sqrt{n}]. \qquad \text{[C.20]}$$

Therefore, provided σ is known, a confidence interval for μ is readily constructed. To allow for unknown σ, we must use an estimate. Let

$$s = \left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2} \qquad \text{[C.21]}$$

denote the sample standard deviation. Then, we obtain a confidence interval that depends entirely on the observed data by replacing σ in equation (C.20) with its estimate, s. Unfortunately, this does not preserve the 95% level of confidence because s depends on the particular sample. In other words, the random interval $[\bar{Y} \pm 1.96(S/\sqrt{n})]$ no longer contains μ with probability .95 because the constant σ has been replaced with the random variable S.

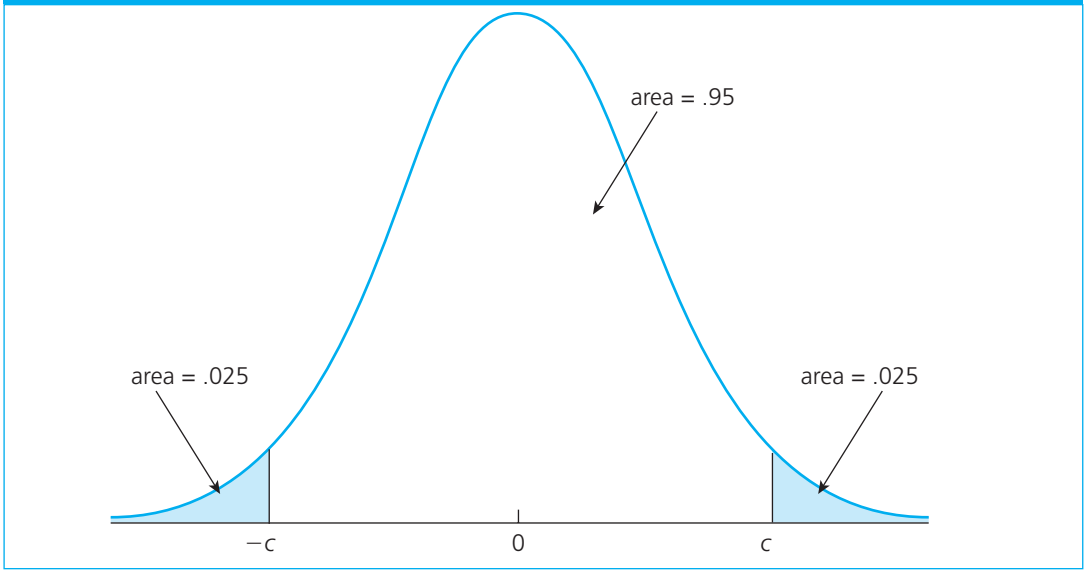
How should we proceed? Rather than using the standard normal distribution, we must rely on the t distribution. The t distribution arises from the fact that

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \qquad \text{[C.22]}$$

where $\bar{Y}$ is the sample average and S is the sample standard deviation of the random sample $\{Y_1, \ldots, Y_n\}$. We will not prove (C.22); a careful proof can be found in a variety of places [for example, Larsen and Marx (1986, Chapter 7)].

To construct a 95% confidence interval, let c denote the 97.5$^{\text{th}}$ percentile in the t_{n-1} distribution. In other words, c is the value such that 95% of the area in the t_{n-1} is between $-c$ and c: $P(-c < t_{n-1} < c) = .95$. (The value of c depends on the degrees of freedom $n - 1$, but we do not

FIGURE C.4 The 97.5$^{\text{th}}$ percentile, c, in a t distribution.



make this explicit.) The choice of c is illustrated in Figure C.4. Once c has been properly chosen, the random interval $[\bar{Y} - c \cdot S/\sqrt{n}, \bar{Y} + c \cdot S/\sqrt{n}]$ contains μ with probability .95. For a particular sample, the 95% confidence interval is calculated as

$$[\bar{y} - c \cdot s/\sqrt{n}, \bar{y} + c \cdot s/\sqrt{n}]. \qquad [C.23]$$

The values of c for various degrees of freedom can be obtained from Table G.2 in Appendix G. For example, if $n = 20$, so that the df is $n - 1 = 19$, then $c = 2.093$. Thus, the 95% confidence interval is $[\bar{y} \pm 2.093(s/\sqrt{20})]$, where $\bar{y}$ and s are the values obtained from the sample. Even if $s = \sigma$ (which is very unlikely), the confidence interval in (C.23) is wider than that in (C.20) because $c > 1.96$. For small degrees of freedom, (C.23) is much wider.

More generally, let c_α denote the $100(1 - \alpha)$ percentile in the t_{n-1} distribution. Then, a $100(1 - \alpha)\%$ confidence interval is obtained as

$$[\bar{y} - C_{\alpha/2}S/\sqrt{n}, \bar{y} + C_{\alpha/2}S/\sqrt{n}]. \qquad [C.24]$$

Obtaining $c_{\alpha/2}$ requires choosing α and knowing the degrees of freedom $n - 1$; then, Table G.2 can be used. For the most part, we will concentrate on 95% confidence intervals.

There is a simple way to remember how to construct a confidence interval for the mean of a normal distribution. Recall that $\text{sd}(\bar{Y}) = \sigma/\sqrt{n}$. Thus, $s/\sqrt{n}$ is the point estimate of $\text{sd}(\bar{Y})$. The associated random variable, $S/\sqrt{n}$, is sometimes called the **standard error** of $\bar{Y}$. Because what shows up in formulas is the point estimate $s/\sqrt{n}$, we define the standard error of $\bar{y}$ as $\text{se}(\bar{y}) = s/\sqrt{n}$. Then, (C.24) can be written in shorthand as

$$[\bar{y} \pm c_{\alpha/2} \cdot \text{se}(\bar{y})]. \qquad [C.25]$$

This equation shows why the notion of the standard error of an estimate plays an important role in econometrics.

EXAMPLE C.2 Effect of Job Training Grants on Worker Productivity

Holzer, Block, Cheatham, and Knott (1993) studied the effects of job training grants on worker productivity by collecting information on "scrap rates" for a sample of Michigan manufacturing firms receiving job training grants in 1988. Table C.3 lists the scrap rates—measured as number of items per 100 produced that are not usable and therefore need to be scrapped—for 20 firms. Each of these firms received a job training grant in 1988; there were no grants awarded in 1987. We are interested in constructing a confidence interval for the change in the scrap rate from 1987 to 1988 for the population of all manufacturing firms that could have received grants.

We assume that the change in scrap rates has a normal distribution. Since $n = 20$, a 95% confidence interval for the mean change in scrap rates μ is $[\bar{y} \pm 2.093 \cdot \text{se}(\bar{y})]$, where $\text{se}(\bar{y}) = s/\sqrt{n}$. The value 2.093 is the 97.5^{th} percentile in a t_{19} distribution. For the particular sample values, $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$ (each rounded to two decimals), so the 95% confidence interval is $[-2.28, -.02]$. The value zero is excluded from this interval, so we conclude that, with 95% confidence, the average change in scrap rates in the population is not zero.

TABLE C.3 Scrap Rates for 20 Michigan Manufacturing Firms

Firm	1987	1988	Change
1	10	3	−7
2	1	1	0
3	6	5	−1
4	.45	.5	.05
5	1.25	1.54	.29
6	1.3	1.5	.2
7	1.06	.8	−.26
8	3	2	−1
9	8.18	.67	−7.51
10	1.67	1.17	−.5
11	.98	.51	−.47
12	1	.5	−.5
13	.45	.61	.16
14	5.03	6.7	1.67
15	8	4	−4
16	9	7	−2
17	18	19	1
18	.28	.2	−.08
19	7	5	−2
20	3.97	3.83	−.14
Average	4.38	3.23	−1.15

At this point, Example C.2 is mostly illustrative because it has some potentially serious flaws as an econometric analysis. Most importantly, it assumes that any systematic reduction in scrap rates is due to the job training grants. But many things can happen over the course of the year to change worker productivity. From this analysis, we have no way of knowing whether the fall in average scrap rates is attributable to the job training grants or if, at least partly, some external force is responsible.

C.5c A Simple Rule of Thumb for a 95% Confidence Interval

The confidence interval in (C.25) can be computed for any sample size and any confidence level. As we saw in Section B-5, the t distribution approaches the standard normal distribution as the degrees of freedom gets large. In particular, for $\alpha = .05$, $c_{\alpha/2} \rightarrow 1.96$ as $n \rightarrow \infty$, although $c_{\propto/2}$ is always greater than 1.96 for each n. A *rule of thumb* for an approximate 95% confidence interval is

$$[\bar{y} \pm 2 \cdot \text{se}(\bar{y})]. \qquad \text{[C.26]}$$

In other words, we obtain $\bar{y}$ and its standard error and then compute $\bar{y}$ plus or minus twice its standard error to obtain the confidence interval. This is slightly too wide for very large n, and it is too narrow for small n. As we can see from Example C.2, even for n as small as 20, (C.26) is in the ballpark for a 95% confidence interval for the mean from a normal distribution. This means we can get pretty close to a 95% confidence interval without having to refer to t tables.

C.5d Asymptotic Confidence Intervals for Nonnormal Populations

In some applications, the population is clearly nonnormal. A leading case is the Bernoulli distribution, where the random variable takes on only the values zero and one. In other cases, the nonnormal population has no standard distribution. This does not matter, provided the sample size is sufficiently large for the central limit theorem to give a good approximation for the distribution of the sample average $\overline{Y}$. For large n, an *approximate* 95% confidence interval is

$$[\bar{y} \pm 1.96 \cdot \text{se}(\bar{y})], \qquad \text{[C.27]}$$

where the value 1.96 is the 97.5$^{\text{th}}$ percentile in the standard normal distribution. Mechanically, computing an approximate confidence interval does not differ from the normal case. A slight difference is that the number multiplying the standard error comes from the standard normal distribution, rather than the t distribution, because we are using asymptotics. Because the t distribution approaches the standard normal as the df increases, equation (C.25) is also perfectly legitimate as an approximate 95% interval; some prefer this to (C.27) because the former is exact for normal populations.

EXAMPLE C.3 Race Discrimination in Hiring

The Urban Institute conducted a study in 1988 in Washington, D.C., to examine the extent of race discrimination in hiring. Five pairs of people interviewed for several jobs. In each pair, one person was black and the other person was white. They were given résumés indicating that they were virtually the same in terms of experience, education, and other factors that determine job qualification. The idea was to make individuals as similar as possible with the exception of race. Each person in a pair interviewed for the same job, and the researchers recorded which applicant received a job offer. This is an example of a *matched pairs analysis*, where each trial consists of data on two people (or two firms, two cities, and so on) that are thought to be similar in many respects but different in one important characteristic.

Let θ_B denote the probability that the black person is offered a job and let θ_W be the probability that the white person is offered a job. We are primarily interested in the difference, $\theta_B - \theta_W$. Let B_i denote a Bernoulli variable equal to one if the black person gets a job offer from employer i, and zero otherwise. Similarly, $W_i = 1$ if the white person gets a job offer from employer i, and zero otherwise. Pooling across the five pairs of people, there were a total of $n = 241$ trials (pairs of interviews with employers). Unbiased estimators of θ_B and θ_W are $\overline{B}$ and $\overline{W}$, the fractions of interviews for which blacks and whites were offered jobs, respectively.

To put this into the framework of computing a confidence interval for a population mean, define a new variable $Y_i = B_i - W_i$. Now, Y_i can take on three values: -1 if the black person did not get the job but the white person did, 0 if both people either did or did not get the job, and 1 if the black person got the job and the white person did not. Then, $\mu \equiv \text{E}(Y_i) = \text{E}(B_i) - \text{E}(W_i) = \theta_B - \theta_W$.

The distribution of Y_i is certainly not normal—it is discrete and takes on only three values. Nevertheless, an approximate confidence interval for $\theta_B - \theta_W$ can be obtained by using large sample methods.

The data from the Urban Institute audit study are in the file AUDIT. Using the 241 observed data points, $\overline{b} = .224$ and $\overline{w} = .357$, so $\bar{y} = .224 - .357 = -.133$. Thus, 22.4% of black applicants were offered jobs, while 35.7% of white applicants were offered jobs. This is *prima facie* evidence of discrimination against blacks, but we can learn much more by computing a confidence interval for μ. To compute an approximate 95% confidence interval, we need the sample standard deviation. This turns out to be $s = .482$ [using equation (C.21)]. Using (C.27), we obtain a 95% CI for $\mu = \theta_B - \theta_W$ as $-.133 \pm 1.96(.482/\sqrt{241}) = -.133 \pm .031 = [-.164, -.102]$. The approximate 99% CI is $-.133 \pm 2.58(.482/\sqrt{241}) = [-.213, -.053]$. Naturally, this contains a wider range of values than the 95% CI. But even the 99% CI does not contain the value zero. Thus, we are very confident that the population difference $\theta_B - \theta_W$ is not zero.

Before we turn to hypothesis testing, it is useful to review the various population and sample quantities that measure the spreads in the population distributions and the sampling distributions of the estimators. These quantities appear often in statistical analysis, and extensions of them are important for the regression analysis in the main text. The quantity σ is the (unknown) population standard deviation; it is a measure of the spread in the distribution of Y. When we divide σ by $\sqrt{n}$, we obtain the **sampling standard deviation** of $\overline{Y}$ (the sample average). While σ is a fixed feature of the population, $\text{sd}(\overline{Y}) = \sigma/\sqrt{n}$ shrinks to zero as $n \to \infty$: our estimator of μ gets more and more precise as the sample size grows.

The estimate of σ for a particular sample, s, is called the sample standard deviation because it is obtained from the sample. (We also call the underlying random variable, S, which changes across different samples, the sample standard deviation.) Like $\bar{y}$ as an estimate of μ, s is our "best guess" at σ given the sample at hand. The quantity $s/\sqrt{n}$ is what we call the standard error of $\bar{y}$, and it is our best estimate of $\sigma/\sqrt{n}$. Confidence intervals for the population parameter μ depend directly on $\text{se}(\bar{y}) = s/\sqrt{n}$. Because this standard error shrinks to zero as the sample size grows, a larger sample size generally means a smaller confidence interval. Thus, we see clearly that one benefit of more data is that they result in narrower confidence intervals. The notion of the standard error of an estimate, which in the vast majority of cases shrinks to zero at the rate $1/\sqrt{n}$, plays a fundamental role in hypothesis testing (as we will see in the next section) and for confidence intervals and testing in the context of multiple regression (as discussed in Chapter 4).

C.6 Hypothesis Testing

So far, we have reviewed how to evaluate point estimators, and we have seen—in the case of a population mean—how to construct and interpret confidence intervals. But sometimes the question we are interested in has a definite yes or no answer. Here are some examples: (1) Does a job training program effectively increase average worker productivity? (see Example C.2); (2) Are blacks discriminated against in hiring? (see Example C.3); (3) Do stiffer state drunk driving laws reduce the number of drunk driving arrests? Devising methods for answering such questions, using a sample of data, is known as hypothesis testing.

C.6a Fundamentals of Hypothesis Testing

To illustrate the issues involved with hypothesis testing, consider an election example. Suppose there are two candidates in an election, Candidates A and B. Candidate A is reported to have received 42% of the popular vote, while Candidate B received 58%. These are supposed to represent the true percentages in the voting population, and we treat them as such.

Candidate A is convinced that more people must have voted for him, so he would like to investigate whether the election was rigged. Knowing something about statistics, Candidate A hires a consulting agency to randomly sample 100 voters to record whether or not each person voted for him. Suppose that, for the sample collected, 53 people voted for Candidate A. This sample estimate of 53% clearly exceeds the reported population value of 42%. Should Candidate A conclude that the election was indeed a fraud?

While it appears that the votes for Candidate A were undercounted, we cannot be certain. Even if only 42% of the population voted for Candidate A, it is possible that, in a sample of 100, we observe 53 people who did vote for Candidate A. The question is: How *strong* is the sample evidence against the officially reported percentage of 42%?

One way to proceed is to set up a **hypothesis test**. Let θ denote the true proportion of the population voting for Candidate A. The hypothesis that the reported results are accurate can be stated as

$$H_0: \theta = .42 \qquad \text{[C.28]}$$

This is an example of a **null hypothesis**. We always denote the null hypothesis by H_0. In hypothesis testing, the null hypothesis plays a role similar to that of a defendant on trial in many judicial systems: just as a defendant is presumed to be innocent until proven guilty, the null hypothesis is presumed to be true until the data strongly suggest otherwise. In the current example, Candidate A must present fairly strong evidence against (C.28) in order to win a recount.

The **alternative hypothesis** in the election example is that the true proportion voting for Candidate A in the election is greater than .42:

$$H_1: \theta > .42. \qquad \text{[C.29]}$$

In order to conclude that H_0 is false and that H_1 is true, we must have evidence "beyond reasonable doubt" against H_0. How many votes out of 100 would be needed before we feel the evidence is strongly against H_0? Most would agree that observing 43 votes out of a sample of 100 is not enough to overturn the original election results; such an outcome is well within the expected sampling variation. On the other hand, we do not need to observe 100 votes for Candidate A to cast doubt on H_0. Whether 53 out of 100 is enough to reject H_0 is much less clear. The answer depends on how we quantify "beyond reasonable doubt."

Before we turn to the issue of quantifying uncertainty in hypothesis testing, we should head off some possible confusion. You may have noticed that the hypotheses in equations (C.28) and (C.29) do not exhaust all possibilities: it could be that θ is less than .42. For the application at hand, we are not particularly interested in that possibility; it has nothing to do with overturning the results of the election. Therefore, we can just state at the outset that we are ignoring alternatives θ with $\theta < .42$. Nevertheless, some authors prefer to state null and alternative hypotheses so that they are exhaustive, in which case our null hypothesis should be $H_0: \theta \leq .42$. Stated in this way, the null hypothesis is a *composite* null hypothesis because it allows for more than one value under H_0. [By contrast, equation (C.28) is an example of a *simple* null hypothesis.] For these kinds of examples, it does not matter whether we state the null as in (C.28) or as a composite null: the most difficult value to reject if $\theta \leq .42$ is $\theta = .42$. (That is, if we reject the value $\theta = .42$, against $\theta > .42$, then logically we must reject any value less than .42.) Therefore, our testing procedure based on (C.28) leads to the same test as if $H_0: \theta \leq .42$. In this text, we always state a null hypothesis as a simple null hypothesis.

In hypothesis testing, we can make two kinds of mistakes. First, we can reject the null hypothesis when it is in fact true. This is called a **Type I error**. In the election example, a Type I error occurs if we reject H_0 when the true proportion of people voting for Candidate A is in fact .42. The second kind of error is failing to reject H_0 when it is actually false. This is called a **Type II error**. In the election example, a Type II error occurs if $\theta > .42$ but we fail to reject H_0.

After we have made the decision of whether or not to reject the null hypothesis, we have either decided correctly or we have committed an error. We will never know with certainty whether an error was committed. However, we can compute the *probability* of making either a Type I or a Type II error. Hypothesis testing rules are constructed to make the probability of committing a Type I error fairly small. Generally, we define the **significance level** (or simply the *level*) of a test as the probability of a Type I error; it is typically denoted by α. Symbolically, we have

$$\alpha = \mathrm{P}(\text{Reject } H_0 | H_0). \qquad \text{[C.30]}$$

The right-hand side is read as: "The probability of rejecting H_0 given that H_0 is true."

Classical hypothesis testing requires that we initially specify a significance level for a test. When we specify a value for α, we are essentially quantifying our tolerance for a Type I error. Common values for α are .10, .05, and .01. If $\alpha = .05$, then the researcher is willing to falsely reject H_0 5% of the time, in order to detect deviations from H_0.

Once we have chosen the significance level, we would then like to minimize the probability of a Type II error. Alternatively, we would like to maximize the **power of a test** against all relevant alternatives. The power of a test is just one minus the probability of a Type II error. Mathematically,

$$\pi(\theta) = \mathrm{P}(\text{Reject } H_0 | \theta) = 1 - \mathrm{P}(\text{Type II} | \theta),$$

where θ denotes the actual value of the parameter. Naturally, we would like the power to equal unity whenever the null hypothesis is false. But this is impossible to achieve while keeping the significance level small. Instead, we choose our tests to maximize the power for a given significance level.

C.6b Testing Hypotheses about the Mean in a Normal Population

In order to test a null hypothesis against an alternative, we need to choose a test statistic (or statistic, for short) and a critical value. The choices for the statistic and critical value are based on convenience and on the desire to maximize power given a significance level for the test. In this subsection, we review how to test hypotheses for the mean of a normal population.

A **test statistic**, denoted T, is some function of the random sample. When we compute the statistic for a particular outcome, we obtain an outcome of the test statistic, which we will denote by t.

Given a test statistic, we can define a rejection rule that determines when H_0 is rejected in favor of H_1. In this text, all rejection rules are based on comparing the value of a test statistic, t, to a **critical value**, c. The values of t that result in rejection of the null hypothesis are collectively known as the **rejection region**. To determine the critical value, we must first decide on a significance level of the test. Then, given α, the critical value associated with α is determined by the distribution of T, *assuming* that H_0 is true. We will write this critical value as c, suppressing the fact that it depends on α.

Testing hypotheses about the mean μ from a Normal(μ, σ^2) population is straightforward. The null hypothesis is stated as

$$H_0: \mu = \mu_0, \tag{C.31}$$

where μ_0 is a value that we specify. In the majority of applications, $\mu_0 = 0$, but the general case is no more difficult.

The rejection rule we choose depends on the nature of the alternative hypothesis. The three alternatives of interest are

$$H_1: \mu > \mu_0, \tag{C.32}$$

$$H_1: \mu < \mu_0, \tag{C.33}$$

and

$$H_1: \mu \neq \mu_0. \tag{C.34}$$

Equation (C.32) gives a **one-sided alternative**, as does (C.33). When the alternative hypothesis is (C.32), the null is effectively $H_0: \mu \leq \mu_0$, since we reject H_0 only when $\mu > \mu_0$. This is appropriate when we are interested in the value of μ only when μ is at least as large as μ_0. Equation (C.34) is a **two-sided alternative**. This is appropriate when we are interested in any departure from the null hypothesis.

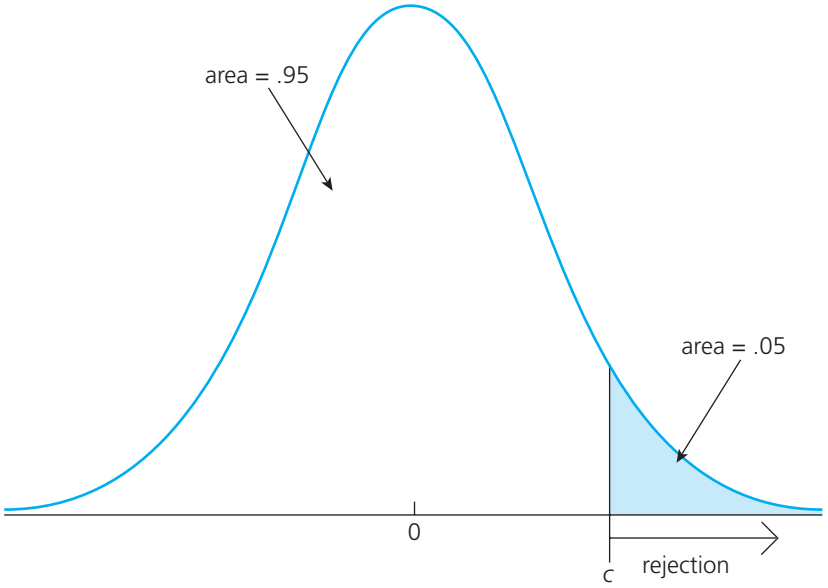
Consider first the alternative in (C.32). Intuitively, we should reject H_0 in favor of H_1 when the value of the sample average, $\bar{y}$, is "sufficiently" greater than μ_0. But how should we determine when $\bar{y}$ is large enough for H_0 to be rejected at the chosen significance level? This requires knowing the probability of rejecting the null hypothesis when it is true. Rather than working directly with $\bar{y}$, we use its standardized version, where σ is replaced with the sample standard deviation, s:

$$t = \sqrt{n}(\bar{y} - \mu_0)/s = (\bar{y} - \mu_0)/\text{se}(\bar{y}), \tag{C.35}$$

where $\text{se}(\bar{y}) = s/\sqrt{n}$ is the standard error of $\bar{y}$. Given the sample of data, it is easy to obtain t. We work with t because, under the null hypothesis, the random variable

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S$$

FIGURE C.5 Rejection region for a 5% significance level test against the one-sided alternative $\mu > \mu_0$.



has a t_{n-1} distribution. Now, suppose we have settled on a 5% significance level. Then, the critical value c is chosen so that $P(T > c|H_0) = .05$; that is, the probability of a Type I error is 5%. Once we have found c, the rejection rule is

$$t > c, \qquad \text{[C.36]}$$

where c is the $100(1 - \alpha)$ percentile in a t_{n-1} distribution; as a percent, the significance level is $100 \cdot \alpha\%$. This is an example of a **one-tailed test** because the rejection region is in one tail of the t distribution. For a 5% significance level, c is the 95th percentile in the t_{n-1} distribution; this is illustrated in Figure C.5. A different significance level leads to a different critical value.

The statistic in equation (C.35) is often called the t **statistic** for testing H_0: $\mu = \mu_0$. The t statistic measures the distance from $\bar{y}$ to μ_0 *relative* to the standard error of $\bar{y}$, $se(\bar{y})$.

EXAMPLE C.4 Effect of Enterprise Zones on Business Investments

In the population of cities granted enterprise zones in a particular state [see Papke (1994) for Indiana], let Y denote the percentage change in investment from the year before to the year after a city became an enterprise zone. Assume that Y has a Normal(μ, σ^2) distribution. The null hypothesis that enterprise zones have no effect on business investment is H_0: $\mu = 0$; the alternative that they have a positive effect is H_1: $\mu > 0$. (We assume that they do not have a negative effect.) Suppose that we wish to test H_0 at the 5% level. The test statistic in this case is

$$t = \frac{\bar{y}}{s/\sqrt{n}} = \frac{\bar{y}}{se(\bar{y})}. \qquad \text{[C.37]}$$

Suppose that we have a sample of 36 cities that are granted enterprise zones. Then, the critical value is $c = 1.69$ (see Table G.2), and we reject H_0 in favor of H_1 if $t > 1.69$. Suppose that the sample yields $\bar{y} = 8.2$ and $s = 23.9$. Then, $t \approx 2.06$, and H_0 is therefore rejected at the 5% level. Thus, we conclude

that, at the 5% significance level, enterprise zones have an effect on average investment. The 1% critical value is 2.44, so H_0 is not rejected at the 1% level. The same caveat holds here as in Example C.2: we have not controlled for other factors that might affect investment in cities over time, so we cannot claim that the effect is causal.

The rejection rule is similar for the one-sided alternative (C.33). A test with a significance level of $100 \cdot \alpha\%$ rejects H_0 against (C.33) whenever

$$t < -c; \qquad [C.38]$$

in other words, we are looking for negative values of the t statistic—which implies $\bar{y} < \mu_0$—that are sufficiently far from zero to reject H_0.

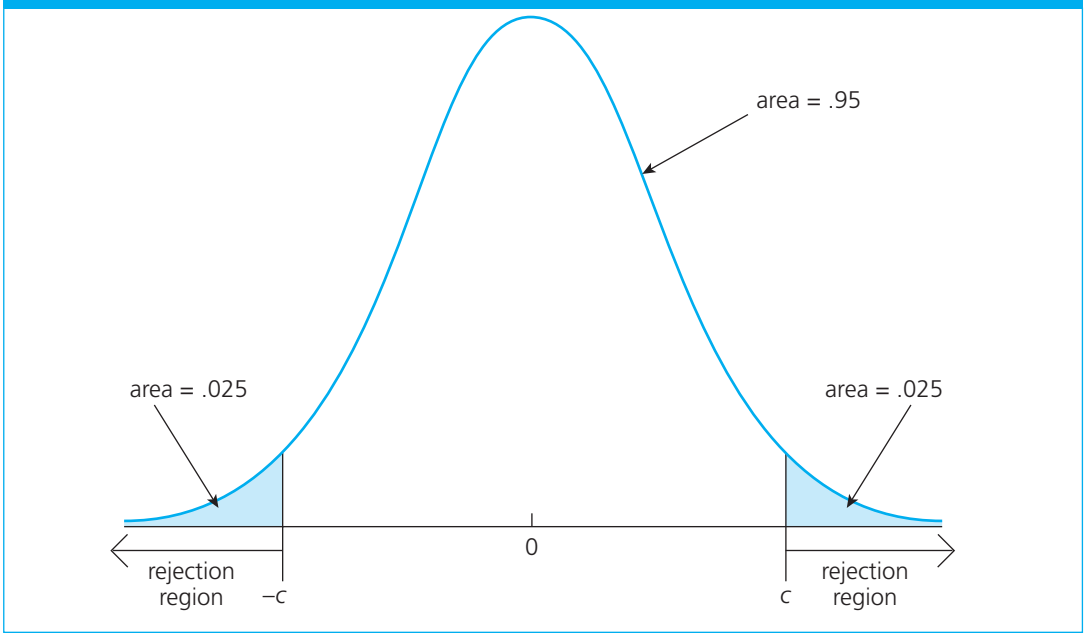
For two-sided alternatives, we must be careful to choose the critical value so that the significance level of the test is still α. If H_1 is given by $H_1: \mu \neq \mu_0$, then we reject H_0 if $\bar{y}$ is far from μ_0 in *absolute value*: a $\bar{y}$ much larger or much smaller than μ_0 provides evidence against H_0 in favor of H_1. A $100 \cdot \alpha\%$ level test is obtained from the rejection rule

$$|t| > c, \qquad [C.39]$$

where $|t|$ is the absolute value of the t statistic in (C.35). This gives a **two-tailed test**. We must now be careful in choosing the critical value: c is the $100(1 - \alpha/2)$ percentile in the t_{n-1} distribution. For example, if $\alpha = .05$, then the critical value is the 97.5^{th} percentile in the t_{n-1} distribution. This ensures that H_0 is rejected only 5% of the time when it is true (see Figure C.6). For example, if $n = 22$, then the critical value is $c = 2.08$, the 97.5^{th} percentile in a t_{21} distribution (see Table G.2). The absolute value of the t statistic must exceed 2.08 in order to reject H_0 against H_1 at the 5% level.

It is important to know the proper language of hypothesis testing. Sometimes, the appropriate phrase "we fail to reject H_0 in favor of H_1 at the 5% significance level" is replaced with "we accept H_0 at the 5% significance level." The latter wording is incorrect. With the same set of data, there are

FIGURE C.6 Rejection region for a 5% significance level test against the two-sided alternative $H_1: \mu \neq \mu_0$.



usually many hypotheses that cannot be rejected. In the earlier election example, it would be logically inconsistent to say that H_0: $\theta = .42$ and H_0: $\theta = .43$ are both "accepted," since only one of these can be true. But it is entirely possible that neither of these hypotheses is rejected. For this reason, we always say "fail to reject H_0" rather than "accept H_0."

C.6c Asymptotic Tests for Nonnormal Populations

If the sample size is large enough to invoke the central limit theorem (see Section C-3), the mechanics of hypothesis testing for population means are the *same* whether or not the population distribution is normal. The theoretical justification comes from the fact that, under the null hypothesis,

$$T = \sqrt{n}(\bar{Y} - \mu_0)/S \overset{a}{\sim} \text{Normal}(0,1).$$

Therefore, with large n, we can compare the t statistic in (C.35) with the critical values from a standard normal distribution. Because the t_{n-1} distribution converges to the standard normal distribution as n gets large, the t and standard normal critical values will be very close for extremely large n. Because asymptotic theory is based on n increasing without bound, it cannot tell us whether the standard normal or t critical values are better. For moderate values of n, say, between 30 and 60, it is traditional to use the t distribution because we know this is correct for normal populations. For $n > 120$, the choice between the t and standard normal distributions is largely irrelevant because the critical values are practically the same.

Because the critical values chosen using either the standard normal or t distribution are only approximately valid for nonnormal populations, our chosen significance levels are also only approximate; thus, for nonnormal populations, our significance levels are really *asymptotic* significance levels. Thus, if we choose a 5% significance level, but our population is nonnormal, then the actual significance level will be larger or smaller than 5% (and we cannot know which is the case). When the sample size is large, the actual significance level will be very close to 5%. Practically speaking, the distinction is not important, so we will now drop the qualifier "asymptotic."

EXAMPLE C.5 Race Discrimination in Hiring

In the Urban Institute study of discrimination in hiring (see Example C.3) using the data in AUDIT, we are primarily interested in testing H_0: $\mu = 0$ against H_1: $\mu < 0$ where $\mu = \theta_B - \theta_W$ is the difference in probabilities that blacks and whites receive job offers. Recall that μ is the population mean of the variable $Y = B - W$, where B and W are binary indicators. Using the $n = 241$ paired comparisons in the data file AUDIT, we obtained $\bar{y} = -.133$ and $\text{se}(\bar{y}) = .482/\sqrt{241} \approx .031$. The t statistic for testing H_0: $\mu = 0$ is $t = -.133/.031 \approx -4.29$. You will remember from Appendix B that the standard normal distribution is, for practical purposes, indistinguishable from the t distribution with 240 degrees of freedom. The value -4.29 is so far out in the left tail of the distribution that we reject H_0 at any reasonable significance level. In fact, the .005 (one-half of a percent) critical value (for the one-sided test) is about -2.58. A t value of -4.29 is *very* strong evidence against H_0 in favor of H_1. Hence, we conclude that there is discrimination in hiring.

C.6d Computing and Using *p*-Values

The traditional requirement of choosing a significance level ahead of time means that different researchers, using the same data and same procedure to test the same hypothesis, could wind up with different conclusions. Reporting the significance level at which we are carrying out the test solves this problem to some degree, but it does not completely remove the problem.

To provide more information, we can ask the following question: What is the *largest* significance level at which we could carry out the test and still fail to reject the null hypothesis? This value is known as the ***p*-value** of a test (sometimes called the *prob-value*). Compared with choosing a significance level ahead of time and obtaining a critical value, computing a p-value is somewhat more difficult. But with the advent of quick and inexpensive computing, p-values are now fairly easy to obtain.

As an illustration, consider the problem of testing H_0: $\mu = 0$ in a Normal(μ, σ^2) population. Our test statistic in this case is $T = \sqrt{n}\cdot\bar{Y}/S$, and we assume that n is large enough to treat T as having a standard normal distribution under H_0. Suppose that the observed value of T for our sample is $t = 1.52$. (Note how we have skipped the step of choosing a significance level.) Now that we have seen the value t, we can find the largest significance level at which we would fail to reject H_0. This is the significance level associated with using t as our critical value. Because our test statistic T has a standard normal distribution under H_0, we have

$$p\text{-value} = P(T > 1.52|H_0) = 1 - \Phi(1.52) = .065, \qquad \text{[C.40]}$$

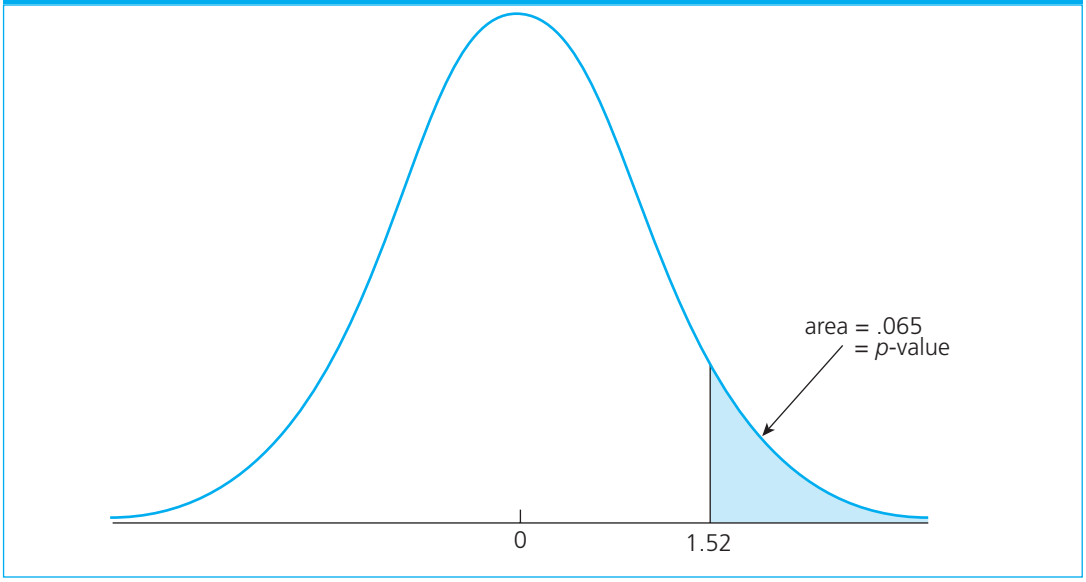
where $\Phi(\cdot)$ denotes the standard normal cdf. In other words, the p-value in this example is simply the area to the right of 1.52, the observed value of the test statistic, in a standard normal distribution. See Figure C.7 for illustration.

Because the p-value $= .065$, the largest significance level at which we can carry out this test and fail to reject is 6.5%. If we carry out the test at a level below 6.5% (such as at 5%), we fail to reject H_0. If we carry out the test at a level larger than 6.5% (such as 10%), we reject H_0. With the p-value at hand, we can carry out the test at any level.

The p-value in this example has another useful interpretation: it is the probability that we observe a value of T as large as 1.52 when the null hypothesis is true. If the null hypothesis is actually true, we would observe a value of T as large as 1.52 due to chance only 6.5% of the time. Whether this is small enough to reject H_0 depends on our tolerance for a Type I error. The p-value has a similar interpretation in all other cases, as we will see.

Generally, small p-values are evidence *against* H_0, since they indicate that the outcome of the data occurs with small probability if H_0 is true. In the previous example, if t had been a larger value, say, $t = 2.85$, then the p-value would be $1 - \Phi(2.85) \approx .002$. This means that, if the null hypothesis were true, we would observe a value of T as large as 2.85 with probability .002. How do we

FIGURE C.7 The *p*-value when $t = 1.52$ for the one-sided alternative $\mu\cdot\mu_0$.



interpret this? Either we obtained a very unusual sample or the null hypothesis is false. Unless we have a *very* small tolerance for Type I error, we would reject the null hypothesis. On the other hand, a large p-value is weak evidence against H_0. If we had gotten $t = .47$ in the previous example, then the p-value $= 1 - \Phi(.47) = .32$. Observing a value of T larger than .47 happens with probability .32, even when H_0 is true; this is large enough so that there is insufficient doubt about H_0, unless we have a very high tolerance for Type I error.

For hypothesis testing about a population mean using the t distribution, we need detailed tables in order to compute p-values. Table G.2 only allows us to put bounds on p-values. Fortunately, many statistics and econometrics packages now compute p-values routinely, and they also provide calculation of cdfs for the t and other distributions used for computing p-values.

EXAMPLE C.6 Effect of Job Training Grants on Worker Productivity

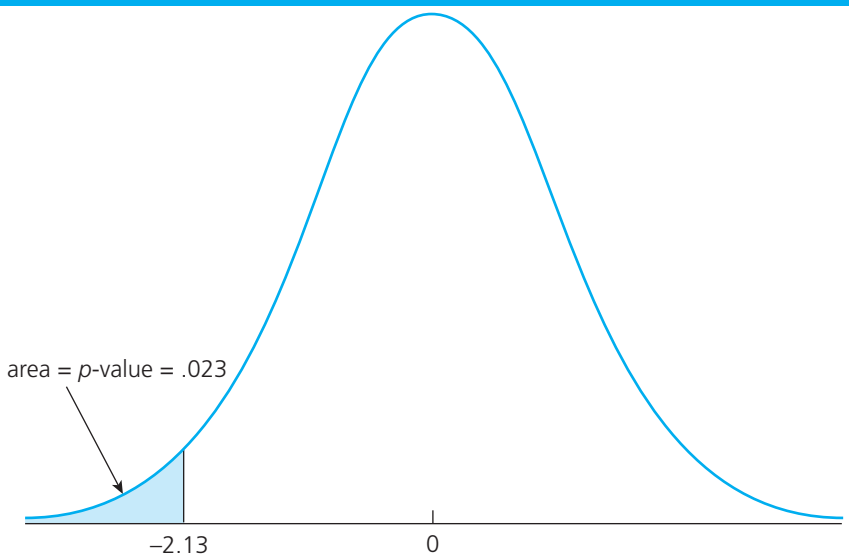
Consider again the Holzer et al. (1993) data in Example C.2. From a policy perspective, there are two questions of interest. First, what is our best estimate of the mean change in scrap rates, μ? We have already obtained this for the sample of 20 firms listed in Table C.3: the sample average of the change in scrap rates is -1.15. Relative to the initial average scrap rate in 1987, this represents a fall in the scrap rate of about 26.3% $(-1.15/4.38 \approx -.263)$, which is a nontrivial effect.

We would also like to know whether the sample provides strong evidence for an effect in the population of manufacturing firms that could have received grants. The null hypothesis is H_0: $\mu = 0$, and we test this against H_1: $\mu < 0$, where μ is the average change in scrap rates. Under the null, the job training grants have no effect on average scrap rates. The alternative states that there is an effect. We do not care about the alternative $\mu > 0$, so the null hypothesis is effectively H_0: $\mu \geq 0$.

Since $\bar{y} = -1.15$ and $\text{se}(\bar{y}) = .54$, $t = -1.15/.54 = -2.13$. This is below the 5% critical value of -1.73 (from a t_{19} distribution) but above the 1% critical value, -2.54. The p-value in this case is computed as

$$p\text{-value} = \text{P}(T_{19} < -2.13), \qquad \text{[C.41]}$$

FIGURE C.8 The p-value when $t = -2.13$ with 19 degrees of freedom for the one-sided alternative $\mu < 0$.



where T_{19} represents a t distributed random variable with 19 degrees of freedom. The inequality is reversed from (C.40) because the alternative has the form in (C.33). The probability in (C.41) is the area to the left of -2.13 in a t_{19} distribution (see Figure C.8).

Using Table G.2, the most we can say is that the p-value is between .025 and .01, but it is closer to .025 (since the 97.5th percentile is about 2.09). Using a statistical package, such as Stata, we can compute the exact p-value. It turns out to be about .023, which is reasonable evidence against H_0. This is certainly enough evidence to reject the null hypothesis that the training grants had no effect at the 2.5% significance level (and therefore at the 5% level).

Computing a p-value for a two-sided test is similar, but we must account for the two-sided nature of the rejection rule. For t testing about population means, the p-value is computed as

$$P(|T_{n-1}| > |t|) = 2P(T_{n-1} > |t|), \qquad [C.42]$$

where t is the value of the test statistic and T_{n-1} is a t random variable. (For large n, replace T_{n-1} with a standard normal random variable.) Thus, compute the absolute value of the t statistic, find the area to the right of this value in a t_{n-1} distribution, and multiply the area by two.

For nonnormal populations, the exact p-value can be difficult to obtain. Nevertheless, we can find *asymptotic* p-values by using the same calculations. These p-values are valid for large sample sizes. For n larger than, say, 120, we might as well use the standard normal distribution. Table G.1 is detailed enough to get accurate p-values, but we can also use a statistics or econometrics program.

EXAMPLE C.7 **Race Discrimination in Hiring**

Using the matched pairs data from the Urban Institute in the AUDIT data file ($n = 241$), we obtained $t = -4.29$. If Z is a standard normal random variable, $P(Z < -4.29)$ is, for practical purposes, zero. In other words, the (asymptotic) p-value for this example is essentially zero. This is very strong evidence against H_0.

Summary of How to Use *p*-Values:

(i) Choose a test statistic T and decide on the nature of the alternative. This determines whether the rejection rule is $t > c$, $t < -c$, or $|t| > c$.

(ii) Use the observed value of the t statistic as the critical value and compute the corresponding significance level of the test. This is the p-value. If the rejection rule is of the form $t > c$, then p-value $= P(T > t)$. If the rejection rule is $t < -c$, then p-value $= P(T < t)$; if the rejection rule is $|t| > c$, then p-value $= P(|T| > |t|)$.

(iii) If a significance level α has been chosen, then we reject H_0 at the $100 \cdot \alpha\%$ level if p-value $< \alpha$. If p-value $\geq \alpha$, then we fail to reject H_0 at the $100 \cdot \alpha\%$ level. Therefore, it is a small p-value that leads to rejection of the null hypothesis.

C.6e The Relationship between Confidence Intervals and Hypothesis Testing

Because constructing confidence intervals and hypothesis tests both involve probability statements, it is natural to think that they are somehow linked. It turns out that they are. After a confidence interval has been constructed, we can carry out a variety of hypothesis tests.

The confidence intervals we have discussed are all two-sided by nature. (In this text, we will have no need to construct one-sided confidence intervals.) Thus, confidence intervals can be used to

test against *two-sided* alternatives. In the case of a population mean, the null is given by (C.31), and the alternative is (C.34). Suppose we have constructed a 95% confidence interval for μ. Then, if the hypothesized value of μ under H_0, μ_0, is not in the confidence interval, then H_0: $\mu = \mu_0$ is rejected against H_1: $\mu \neq \mu_0$ at the 5% level. If μ_0 lies in this interval, then we fail to reject H_0 at the 5% level. Notice how any value for μ_0 can be tested once a confidence interval is constructed, and since a confidence interval contains more than one value, there are many null hypotheses that will not be rejected.

EXAMPLE C.8 Training Grants and Worker Productivity

In the Holzer et al. example, we constructed a 95% confidence interval for the mean change in scrap rate μ as $[-2.28, -.02]$. Since zero is excluded from this interval, we reject H_0: $\mu = 0$ against H_1: $\mu \neq 0$ at the 5% level. This 95% confidence interval also means that we fail to reject H_0: $\mu = -2$ at the 5% level. In fact, there is a continuum of null hypotheses that are not rejected given this confidence interval.

C.6f Practical versus Statistical Significance

In the examples covered so far, we have produced three kinds of evidence concerning population parameters: point estimates, confidence intervals, and hypothesis tests. These tools for learning about population parameters are equally important. There is an understandable tendency for students to focus on confidence intervals and hypothesis tests because these are things to which we can attach confidence or significance levels. But in any study, we must also interpret the *magnitudes* of point estimates.

The sign and magnitude of $\bar{y}$ determine its **practical significance** and allow us to discuss the direction of an intervention or policy effect, and whether the estimated effect is "large" or "small." On the other hand, **statistical significance** of $\bar{y}$ depends on the magnitude of its t statistic. For testing H_0: $\mu = 0$, the t statistic is simply $t = \bar{y}/\text{se}(\bar{y})$. In other words, statistical significance depends on the ratio of $\bar{y}$ to its standard error. Consequently, a t statistic can be large because $\bar{y}$ is large or $\text{se}(\bar{y})$ is small. In applications, it is important to discuss both practical and statistical significance, being aware that an estimate can be statistically significant without being especially large in a practical sense. Whether an estimate is practically important depends on the context as well as on one's judgment, so there are no set rules for determining practical significance.

EXAMPLE C.9 Effect of Freeway Width on Commute Time

Let Y denote the change in commute time, measured in minutes, for commuters in a metropolitan area from before a freeway was widened to after the freeway was widened. Assume that $Y \sim \text{Normal}(\mu,\sigma^2)$. The null hypothesis that the widening did not reduce average commute time is H_0: $\mu = 0$; the alternative that it reduced average commute time is H_1: $\mu < 0$. Suppose a random sample of commuters of size $n = 900$ is obtained to determine the effectiveness of the freeway project. The average change in commute time is computed to be $\bar{y} = -3.6$, and the sample standard deviation is $s = 32.7$; thus, $\text{se}(\bar{y}) = 32.7/\sqrt{900} = 1.09$. The t statistic is $t = -3.6/1.09 \approx -3.30$, which is very statistically significant; the p-value is about .0005. Thus, we conclude that the freeway widening had a statistically significant effect on average commute time.

If the outcome of the hypothesis test is all that were reported from the study, it would be misleading. Reporting only statistical significance masks the fact that the estimated reduction in average commute time, 3.6 minutes, seems pretty meager, although this depends to some extent on what the average commute time was prior to widening the freeway. To be up front, we should report the point estimate of -3.6, along with the significance test.

Finding point estimates that are statistically significant without being practically significant can occur when we are working with large samples. To discuss why this happens, it is useful to have the following definition.

Test Consistency. A **consistent test** rejects H_0 with probability approaching one as the sample size grows whenever H_1 is true.

Another way to say that a test is consistent is that, as the sample size tends to infinity, the power of the test gets closer and closer to unity whenever H_1 is true. All of the tests we cover in this text have this property. In the case of testing hypotheses about a population mean, test consistency follows because the variance of $\overline{Y}$ converges to zero as the sample size gets large. The t statistic for testing H_0: $\mu = 0$ is $T = \overline{Y}/(S/\sqrt{n})$. Since $\text{plim}(\overline{Y}) = \mu$ and $\text{plim}(S) = \sigma$, it follows that if, say, $\mu > 0$, then T gets larger and larger (with high probability) as $n \to \infty$. In other words, no matter how close m is to zero, we can be almost certain to reject H_0: $\mu = 0$ given a large enough sample size. This says nothing about whether μ is large in a practical sense.

C.7 Remarks on Notation

In our review of probability and statistics here and in Appendix B, we have been careful to use standard conventions to denote random variables, estimators, and test statistics. For example, we have used W to indicate an estimator (random variable) and w to denote a particular estimate (outcome of the random variable W). Distinguishing between an estimator and an estimate is important for understanding various concepts in estimation and hypothesis testing. However, making this distinction quickly becomes a burden in econometric analysis because the models are more complicated: many random variables and parameters will be involved, and being true to the usual conventions from probability and statistics requires many extra symbols.

In the main text, we use a simpler convention that is widely used in econometrics. If θ is a population parameter, the notation $\hat{\theta}$ ("theta hat") will be used to denote both an estimator and an estimate of θ. This notation is useful in that it provides a simple way of attaching an estimator to the population parameter it is supposed to be estimating. Thus, if the population parameter is β, then $\hat{\beta}$ denotes an estimator or estimate of β; if the parameter is σ^2, $\hat{\sigma}^2$ is an estimator or estimate of σ^2; and so on. Sometimes, we will discuss two estimators of the same parameter, in which case we will need a different notation, such as $\tilde{\theta}$ ("theta tilde").

Although dropping the conventions from probability and statistics to indicate estimators, random variables, and test statistics puts additional responsibility on you, it is not a big deal once the difference between an estimator and an estimate is understood. If we are discussing *statistical* properties of $\hat{\theta}$—such as deriving whether or not it is unbiased or consistent—then we are necessarily viewing $\hat{\theta}$ as an estimator. On the other hand, if we write something like $\hat{\theta} = 1.73$, then we are clearly denoting a point estimate from a given sample of data. The confusion that can arise by using $\hat{\theta}$ to denote both should be minimal once you have a good understanding of probability and statistics.

Summary

We have discussed topics from mathematical statistics that are heavily relied upon in econometric analysis. The notion of an estimator, which is simply a rule for combining data to estimate a population parameter, is fundamental. We have covered various properties of estimators. The most important small sample properties are unbiasedness and efficiency, the latter of which depends on comparing variances when estimators are unbiased. Large sample properties concern the sequence of estimators

obtained as the sample size grows, and they are also depended upon in econometrics. Any useful estimator is consistent. The central limit theorem implies that, in large samples, the sampling distribution of most estimators is approximately normal.

The sampling distribution of an estimator can be used to construct confidence intervals. We saw this for estimating the mean from a normal distribution and for computing approximate confidence intervals in nonnormal cases. Classical hypothesis testing, which requires specifying a null hypothesis, an alternative hypothesis, and a significance level, is carried out by comparing a test statistic to a critical value. Alternatively, a *p*-value can be computed that allows us to carry out a test at any significance level.

Key Terms

Alternative Hypothesis
Asymptotic Normality
Bias
Biased Estimator
Central Limit Theorem (CLT)
Confidence Interval
Consistent Estimator
Consistent Test
Critical Value
Estimate
Estimator
Hypothesis Test
Inconsistent
Interval Estimator
Law of Large Numbers (LLN)
Least Squares Estimator
Maximum Likelihood Estimator
Mean Squared Error (MSE)
Method of Moments
Minimum Variance Unbiased Estimator
Null Hypothesis
One-Sided Alternative
One-Tailed Test
Population
Power of a Test
Practical Significance
Probability Limit
p-Value
Random Sample
Rejection Region
Sample Average
Sample Correlation Coefficient
Sample Covariance
Sample Standard Deviation
Sample Variance
Sampling Distribution
Sampling Standard Deviation
Sampling Variance
Significance Level
Standard Error
Statistical Significance
t Statistic
Test Statistic
Two-Sided Alternative
Two-Tailed Test
Type I Error
Type II Error
Unbiased Estimator

Problems

1 Let Y_1, Y_2, Y_3, and Y_4 be independent, identically distributed random variables from a population with mean μ and variance σ^2. Let $\bar{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$ denote the average of these four random variables.

(i) What are the expected value and variance of $\bar{Y}$ in terms of μ and σ^2?

(ii) Now, consider a different estimator of μ:

$$W = \frac{1}{8}Y_1 + \frac{1}{8}Y_2 + \frac{1}{4}Y_3 + \frac{1}{2}Y_4.$$

This is an example of a *weighted* average of the Y_i. Show that W is also an unbiased estimator of μ. Find the variance of W.

(iii) Based on your answers to parts (i) and (ii), which estimator of m do you prefer, $\bar{Y}$ or W?

2 This is a more general version of Problem C.1. Let $Y_1, Y_2, \ldots, Y_n$ be n pairwise uncorrelated random variables with common mean m and common variance σ^2. Let $\bar{Y}$ denote the sample average.

(i) Define the class of *linear estimators* of μ by

$$W_a = a_1Y_1 + a_2Y_2 + \ \ldots \ + a_nY_n,$$

where the a_i are constants. What restriction on the a_i is needed for W_a to be an unbiased estimator of μ?

(ii) Find $\text{Var}(W_a)$.

(iii) For any numbers $a_1, a_2, \ldots, a_n$, the following inequality holds: $(a_1 + a_2 + \ldots + a_n)^2/n \leq a_1^2 + a_2^2 + \ldots + a_n^2$. Use this, along with parts (i) and (ii), to show that $\text{Var}(W_a) \geq \text{Var}(\bar{Y})$ whenever W_a is unbiased, so that $\bar{Y}$ is the *best linear unbiased estimator*. [*Hint*: What does the inequality become when the a_i satisfy the restriction from part (i)?]

3 Let $\bar{Y}$ denote the sample average from a random sample with mean μ and variance σ^2. Consider two alternative estimators of μ: $W_1 = [(n-1)/n]\bar{Y}$ and $W_2 = \bar{Y}/2$.

(i) Show that W_1 and W_2 are both biased estimators of μ and find the biases. What happens to the biases as $n \to \infty$? Comment on any important differences in bias for the two estimators as the sample size gets large.

(ii) Find the probability limits of W_1 and W_2. {*Hint*: Use Properties PLIM.1 and PLIM.2; for W_1, note that plim $[(n-1)/n] = 1$.} Which estimator is consistent?

(iii) Find $\text{Var}(W_1)$ and $\text{Var}(W_2)$.

(iv) Argue that W_1 is a better estimator than $\bar{Y}$ if μ is "close" to zero. (Consider both bias and variance.)

4 For positive random variables X and Y, suppose the expected value of Y given X is $\text{E}(Y|X) = \theta X$. The unknown parameter θ shows how the expected value of Y changes with X.

(i) Define the random variable $Z = Y/X$. Show that $\text{E}(Z) = \theta$. [*Hint*: Use Property CE.2 along with the law of iterated expectations, Property CE.4. In particular, first show that $\text{E}(Z|X) = \theta$ and then use CE.4.]

(ii) Use part (i) to prove that the estimator $W_1 = n^{-1}\sum_{i=1}^{n}(Y_i/X_i)$ is unbiased for θ, where $\{(X_i, Y_i): i = 1, 2, \ldots, n\}$ is a random sample.

(iii) Explain why the estimator $W_2 = \bar{Y}/\bar{X}$, where the overbars denote sample averages, is not the same as W_1. Nevertheless, show that W_2 is also unbiased for θ.

(iv) The following table contains data on corn yields for several counties in Iowa. The USDA predicts the number of hectares of corn in each county based on satellite photos. Researchers count the number of "pixels" of corn in the satellite picture (as opposed to, for example, the number of pixels of soybeans or of uncultivated land) and use these to predict the actual number of hectares. To develop a prediction equation to be used for counties in general, the USDA surveyed farmers in selected counties to obtain corn yields in hectares. Let Y_i = corn yield in county i and let X_i = number of corn pixels in the satellite picture for county i. There are $n = 17$ observations for eight counties. Use this sample to compute the estimates of θ devised in parts (ii) and (iii). Are the estimates similar?

Plot	Corn Yield	Corn Pixels
1	165.76	374
2	96.32	209
3	76.08	253
4	185.35	432
5	116.43	367
6	162.08	361
7	152.04	288
8	161.75	369
9	92.88	206
10	149.94	316
11	64.75	145
12	127.07	355
13	133.55	295
14	77.70	223
15	206.39	459
16	108.33	290
17	118.17	307

5 Let Y denote a Bernoulli(θ) random variable with $0 < \theta < 1$. Suppose we are interested in estimating the *odds ratio*, $\gamma = \theta/(1 - \theta)$, which is the probability of success over the probability of failure. Given a random sample $\{Y_1, \ldots, Y_n\}$, we know that an unbiased and consistent estimator of θ is $\overline{Y}$, the proportion of successes in n trials. A natural estimator of γ is $G = \overline{Y}/(1 - \overline{Y})$, the proportion of successes over the proportion of failures in the sample.

(i) Why is G not an unbiased estimator of γ?

(ii) Use PLIM.2 (iii) to show that G is a consistent estimator of γ.

6 You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a Normal(μ, σ^2) distribution.

(i) The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

(ii) The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

(iii) Now, suppose your sample size is $n = 900$ and you obtain the estimates $\bar{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? At the 1% level?

(iv) Would you say that the estimated fall in consumption is large in magnitude? Comment on the practical versus statistical significance of this estimate.

(v) What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

7 The new management at a bakery claims that workers are now more productive than they were under old management, which is why wages have "generally increased." Let W_i^b be Worker i's wage under the old management and let W_i^a be Worker i's wage after the change. The difference is $D_i \equiv W_i^a - W_i^b$. Assume that the D_i are a random sample from a Normal (μ, σ^2) distribution.

(i) Using the following data on 15 workers, construct an exact 95% confidence interval for μ.

(ii) Formally state the null hypothesis that there has been no change in average wages. In particular, what is $E(D_i)$ under H_0? If you are hired to examine the validity of the new management's claim, what is the relevant alternative hypothesis in terms of $\mu = E(D_i)$?

(iii) Test the null hypothesis from part (ii) against the stated alternative at the 5% and 1% levels.

(iv) Obtain the p-value for the test in part (iii).

Worker	Wage Before	Wage After
1	8.30	9.25
2	9.40	9.00
3	9.00	9.25
4	10.50	10.00
5	11.40	12.00
6	8.75	9.50
7	10.00	10.25
8	9.50	9.50
9	10.80	11.50
10	12.55	13.10
11	12.00	11.50
12	8.65	9.00
13	7.75	7.75
14	11.25	11.50
15	12.65	13.00

8 The *New York Times* (2/5/90) reported three-point shooting performance for the top 10 three-point shooters in the NBA. The following table summarizes these data:

Player	FGA-FGM
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Note: *FGA* = field goals attempted and *FGM* = field goals made.

For a given player, the outcome of a particular shot can be modeled as a Bernoulli (zero-one) variable: if Y_i is the outcome of shot i, then $Y_i = 1$ if the shot is made, and $Y_i = 0$ if the shot is missed. Let θ denote the probability of making any particular three-point shot attempt. The natural estimator of θ is $\overline{Y} = FGM/FGA$.

(i) Estimate θ for Mark Price.

(ii) Find the standard deviation of the estimator $\overline{Y}$ in terms of θ and the number of shot attempts, n.

(iii) The asymptotic distribution of $(\overline{Y} - \theta)/\text{se}(\overline{Y})$ is standard normal, where $\text{se}(\overline{Y}) = \sqrt{\overline{Y}(1 - \overline{Y})/n}$. Use this fact to test H_0: $\theta = .5$ against H_1: $\theta < .5$ for Mark Price. Use a 1% significance level.

9 Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. A human rights group suspects foul play and hires you to test the validity of the dictator's claim. You have a budget that allows you to randomly sample 200 voters from the country.

(i) Let X be the number of yes votes obtained from a random sample of 200 out of the entire voting population. What is the expected value of X if, in fact, 65% of all voters supported the dictator?

(ii) What is the standard deviation of X, again assuming that the true fraction voting yes in the plebiscite is .65?

(iii) Now, you collect your sample of 200, and you find that 115 people actually voted yes. Use the CLT to approximate the probability that you would find 115 or fewer yes votes from a random sample of 200 if, in fact, 65% of the entire population voted yes.

(iv) How would you explain the relevance of the number in part (iii) to someone who does not have training in statistics?

10 Before a strike prematurely ended the 1994 major league baseball season, Tony Gwynn of the San Diego Padres had 165 hits in 419 at bats, for a .394 batting average. There was discussion about whether Gwynn was a potential .400 hitter that year. This issue can be couched in terms of Gwynn's probability of getting a hit on a particular at bat, call it θ. Let Y_i be the Bernoulli(θ) indicator equal to unity if Gwynn gets a hit during his i^{th} at bat, and zero otherwise. Then, $Y_1, Y_2, \ldots, Y_n$ is a random sample from a Bernoulli(θ) distribution, where θ is the probability of success, and $n = 419$.

Our best point estimate of θ is Gwynn's batting average, which is just the proportion of successes: $\bar{y} = .394$. Using the fact that $\text{se}(\bar{y}) = \sqrt{\bar{y}(1 - \bar{y})/n}$, construct an approximate 95% confidence interval for θ, using the standard normal distribution. Would you say there is strong evidence against Gwynn's being a potential .400 hitter? Explain.

11 Suppose that between their first and second years in college, 400 students are randomly selected and given a university grant to purchase a new computer. For student i, y_i denotes the change in GPA from the first year to the second year. If the average change is $\bar{y} = .132$ with standard deviation $s = 1.27$, is the average change in GPAs statistically greater than zero?