Example C.13 One-Sided Test About a Mean

A sample of 25 from a normal distribution yields $\bar{x} = 1.63$ and $s = 0.51$. Test

$$H_0\colon \mu \le 1.5,$$
$$H_1\colon \mu > 1.5.$$

Clearly, no observed $\bar{x}$ less than or equal to 1.5 will lead to rejection of H_0. Using the borderline value of 1.5 for μ, we obtain

$$\text{Prob}\left(\frac{\sqrt{n}(\bar{x} - 1.5)}{s} > \frac{5(1.63 - 1.5)}{0.51}\right) = \text{Prob}(t_{24} > 1.27).$$

This is approximately 0.11. This value is not unlikely by the usual standards. Hence, at a significant level of 0.11, we would not reject the hypothesis.

C.7.3 SPECIFICATION TESTS

The hypothesis testing procedures just described are known as classical testing procedures. In each case, the null hypothesis tested came in the form of a restriction on the alternative. You can verify that in each application we examined, the parameter space assumed under the null hypothesis is a subspace of that described by the alternative. For that reason, the models implied are said to be *nested*. The null hypothesis is contained within the alternative. This approach suffices for most of the testing situations encountered in practice, but there are common situations in which two competing models cannot be viewed in these terms. For example, consider a case in which there are two completely different, competing theories to explain the same observed data. Many models for censoring and truncation discussed in Chapter 19 rest upon a fragile assumption of normality, for example. Testing of this nature requires a different approach from the classical procedures discussed here. These are discussed at various points throughout the book, for example, in Chapter 19, where we study the difference between fixed and random effects models.

APPENDIX D

LARGE-SAMPLE DISTRIBUTION THEORY

D.1 INTRODUCTION

Most of this book is about parameter estimation. In studying that subject, we will usually be interested in determining how best to use the observed data when choosing among competing estimators. That, in turn, requires us to examine the sampling behavior of estimators. In a few cases, such as those presented in Appendix C and the least squares estimator considered in Chapter 4, we can make broad statements about sampling distributions that will apply regardless of the size of the sample. But, in most situations, it will only be possible to make approximate statements about estimators, such as whether they improve as the sample size increases and what can be said about their sampling distributions in large samples as an approximation to the finite samples we actually observe. This appendix will collect most of the formal, fundamental theorems and results needed for this analysis. A few additional results will be developed in the discussion of time-series analysis later in the book.

D.2 LARGE-SAMPLE DISTRIBUTION THEORY[1]

In most cases, whether an estimator is exactly unbiased or what its exact sampling variance is in samples of a given size will be unknown. But we may be able to obtain approximate results about the behavior of the distribution of an estimator as the sample becomes large. For example, it is well known that the distribution of the mean of a sample tends to approximate normality as the sample size grows, regardless of the distribution of the individual observations. Knowledge about the limiting behavior of the distribution of an estimator can be used to infer an approximate distribution for the estimator in a finite sample. To describe how this is done, it is necessary, first, to present some results on convergence of random variables.

D.2.1 CONVERGENCE IN PROBABILITY

Limiting arguments in this discussion will be with respect to the sample size n. Let x_n be a sequence random variable indexed by the sample size.

DEFINITION D.1 Convergence in Probability
The random variable x_n **converges in probability** *to a constant* c *if* $\lim_{n\to\infty}\text{Prob}(|x_n - c| > \varepsilon) = 0$ *for any positive* ε.

Convergence in probability implies that the values that the variable may take that are not close to c become increasingly unlikely as n increases. To consider one example, suppose that the random variable x_n takes two values, zero and n, with probabilities $1 - (1/n)$ and $(1/n)$, respectively. As n increases, the second point will become ever more remote from any constant but, at the same time, will become increasingly less probable. In this example, x_n converges in probability to zero. The crux of this form of convergence is that all the mass of the probability distribution becomes concentrated at points close to c. If x_n converges in probability to c, then we write

$$\text{plim}\, x_n = c. \qquad \textbf{(D-1)}$$

[1] A comprehensive summary of many results in large-sample theory appears in White (2001). The results discussed here will apply to samples of independent observations. Time-series cases in which observations are correlated are analyzed in Chapters 20 and 21.

We will make frequent use of a special case of convergence in probability, **convergence in mean square** or **convergence in quadratic mean**.

THEOREM D.1 Convergence in Quadratic Mean
If x_n has mean μ_n and variance σ_n^2 such that the ordinary limits of μ_n and σ_n^2 are c and 0, respectively, then x_n converges in mean square to c , and

$$\text{plim } x_n = c.$$

A proof of Theorem D.1 can be based on another useful theorem.

THEOREM D.2 Chebychev's Inequality
If x_n is a random variable and c and ε are constants, then $\text{Prob}(|x_n - c| > \varepsilon) \leq E[(x_n - c)^2]/\varepsilon^2$.

To establish the Chebychev inequality, we use another result [see Goldberger (1991, p. 31)].

THEOREM D.3 Markov's Inequality
If y_n is a nonnegative random variable and δ is a positive constant, then $\text{Prob}[y_n \geq \delta] \leq E[y_n]/\delta$.
Proof: $E[y_n] = \text{Prob}[y_n < \delta]E[y_n | y_n < \delta] + \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$.
Because y_n is non-negative, both terms must be nonnegative, so
$E[y_n] \geq \text{Prob}[y_n \geq \delta]E[y_n | y_n \geq \delta]$. *Because $E[y_n | y_n \geq \delta]$ must be greater than or equal to δ,* $E[y_n] \geq \text{Prob}[y_n \geq \delta]\delta$, *which is the result.*

Now, to prove Theorem D.1, let y_n be $(x_n - c)^2$ and δ be ε^2 in Theorem D.3. Then, $(x_n - c)^2 > \delta$ implies that $|x_n - c| > \varepsilon$. Finally, we will use a special case of the Chebychev inequality, where $c = \mu_n$, so that we have

$$\text{Prob}(|x_n - \mu_n| > \varepsilon) \leq \sigma_n^2/\varepsilon^2. \tag{D-2}$$

Taking the limits of μ_n and σ_n^2 in (D-2), we see that if

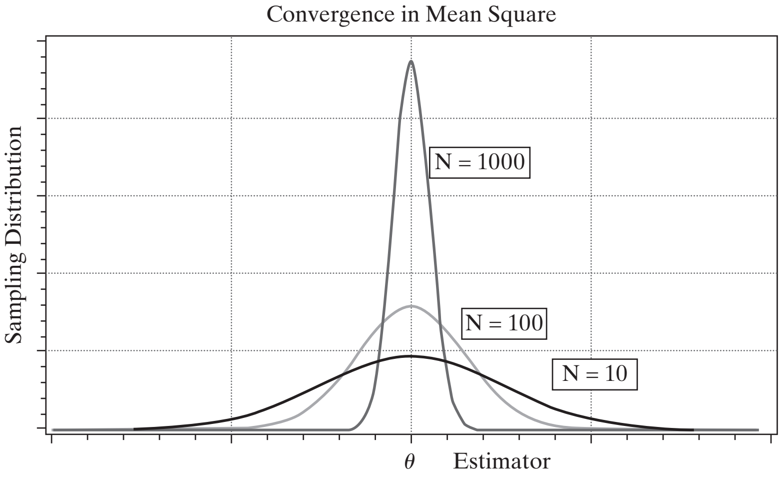
$$\lim_{n\to\infty} E[x_n] = c, \quad \text{and} \quad \lim_{n\to\infty} \text{Var}[x_n] = 0, \tag{D-3}$$

then

$$\text{plim } x_n = c.$$

We have shown that convergence in mean square implies convergence in probability. Mean-square convergence implies that the distribution of x_n collapses to a spike at plim x_n, as shown in Figure D.1.

FIGURE D.1 Quadratic Convergence to a Constant, θ.



Example D.1 Mean Square Convergence of the Sample Minimum in Exponential Sampling

As noted in Example C.4, in sampling of n observations from an exponential distribution, for the sample minimum $x_{(1)}$,

$$\lim_{n\to\infty} E[x_{(1)}] = \lim_{n\to\infty} \frac{1}{n\theta} = 0$$

and

$$\lim_{n\to\infty} \text{Var}[x_{(1)}] = \lim_{n\to\infty} \frac{1}{(n\theta)^2} = 0.$$

Therefore,

$$\text{plim}\, x_{(1)} = 0.$$

Note, in particular, that the variance is divided by n^2. This estimator converges very rapidly to 0.

Convergence in probability does not imply convergence in mean square. Consider the simple example given earlier in which x_n equals either zero or n with probabilities $1 - (1/n)$ and $(1/n)$. The exact expected value of x_n is 1 for all n, which is not the probability limit. Indeed, if we let $\text{Prob}(x_n = n^2) = (1/n)$ instead, the mean of the distribution explodes, but the probability limit is still zero. Again, the point $x_n = n^2$ becomes ever more extreme but, at the same time, becomes ever less likely.

The conditions for convergence in mean square are usually easier to verify than those for the more general form. Fortunately, we shall rarely encounter circumstances in which it will be necessary to show convergence in probability in which we cannot rely upon convergence in mean square. Our most frequent use of this concept will be in formulating consistent estimators.

DEFINITION D.2 Consistent Estimator
An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator of θ if and only if

$$\text{plim } \hat{\theta}_n = \theta. \qquad \textbf{(D-4)}$$

THEOREM D.4 Consistency of the Sample Mean
The mean of a random sample from any population with finite mean μ and finite variance σ^2 is a consistent estimator of μ.
Proof: *$E[\bar{x}_n] = \mu$ and $\text{Var}[\bar{x}_n] = \sigma^2/n$. Therefore, $\bar{x}_n$ converges in mean square to μ, or plim $\bar{x}_n = \mu$.*

Theorem D.4 is broader than it might appear at first.

COROLLARY TO THEOREM D.4 Consistency of a Mean of Functions
In random sampling, for any function $g(x)$, if $E[g(x)]$ and $\text{Var}[g(x)]$ are finite constants, then

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n} g(x_i) = E[g(x)]. \qquad \textbf{(D-5)}$$

Proof: *Define $y_i = g(x_i)$ and use Theorem D.4.*

Example D.2 Estimating a Function of the Mean

In sampling from a normal distribution with mean μ and variance 1, $E[e^x] = e^{\mu+1/2}$ and $\text{Var}[e^x] = e^{2\mu+2} - e^{2\mu+1}$. (See Section B.4.4 on the lognormal distribution.) Hence,

$$\text{plim } \frac{1}{n}\sum_{i=1}^{n} e^{x_i} = e^{\mu+1/2}.$$

D.2.2 OTHER FORMS OF CONVERGENCE AND LAWS OF LARGE NUMBERS

Theorem D.4 and the corollary just given are particularly narrow forms of a set of results known as **laws of large numbers** that are fundamental to the theory of parameter estimation. Laws of large numbers come in two forms depending on the type of convergence considered. The simpler of these are “weak laws of large numbers” which rely on convergence in probability as we defined it above. “Strong laws” rely on a broader type of convergence called **almost sure convergence**. Overall, the law of large numbers is a statement about the behavior of an average of a large number of random variables.

THEOREM D.5 Khinchine's Weak Law of Large Numbers
If x_i, $i = 1, \ldots, n$ is a random (i.i.d.) sample from a distribution with finite mean $E[x_i] = \mu$, then

$$\text{plim}\, \bar{x}_n = \mu.$$

Proofs of this and the theorem below are fairly intricate. Rao (1973) provides one.

Notice that this is already broader than Theorem D.4, as it does not require that the variance of the distribution be finite. On the other hand, it is not broad enough, because most of the situations we encounter where we will need a result such as this will not involve i.i.d. random sampling. A broader result is

THEOREM D.6 Chebychev's Weak Law of Large Numbers
If $x_i, i = 1, \ldots, n$ is a sample of observations such that $E[x_i] = \mu_i < \infty$ and $\text{Var}[x_i] = \sigma_i^2 < \infty$ such that $\bar{\sigma}_n^2/n = (1/n^2)\Sigma_i\sigma_i^2 \rightarrow 0$ as $n \rightarrow \infty$, then $\text{plim}(\bar{x}_n - \bar{\mu}_n) = 0$.

There is a subtle distinction between these two theorems that you should notice. The Chebychev theorem does not state that $\bar{x}_n$ converges to $\bar{\mu}_n$, or even that it converges to a constant at all. That would require a precise statement about the behavior of $\bar{\mu}_n$. The theorem states that as n increases without bound, these two quantities will be arbitrarily close to each other—that is, the difference between them converges to a constant, zero. This is an important notion that enters the derivation when we consider statistics that converge to random variables, instead of to constants. What we do have with these two theorems are extremely broad conditions under which a sample mean will converge in probability to its population counterpart. The more important difference between the Khinchine and Chebychev theorems is that the second allows for heterogeneity in the distributions of the random variables that enter the mean.

In analyzing time-series data, the sequence of outcomes is itself viewed as a random event. Consider, then, the sample mean, $\bar{x}_n$. The preceding results concern the behavior of this statistic as $n \rightarrow \infty$ for a particular realization of the sequence $\bar{x}_1, \ldots, \bar{x}_n$. But, if the sequence, itself, is viewed as a random event, then the limit to which $\bar{x}_n$ converges may be also. The stronger notion of almost sure convergence relates to this possibility.

DEFINITION D.3 Almost Sure Convergence
The random variable x_n converges almost surely to the constant c if and only if

$$\text{Prob}\left(\lim_{n\rightarrow\infty} x_n = c\right) = 1.$$

This is denoted $x_n \xrightarrow{a.s.} c$. It states that the probability of observing a sequence that does not converge to c ultimately vanishes. Intuitively, it states that once the sequence x_n becomes close to c, it stays close to c.

Almost sure convergence is used in a stronger form of the law of large numbers:

THEOREM D.7 Kolmogorov's Strong Law of Large Numbers
If x_i, $i = 1, \ldots, n$ is a sequence of independently distributed random variables such that $E[x_i] = \mu_i < \infty$ and $\operatorname{Var}[x_i] = \sigma_i^2 < \infty$ such that $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$ as $n \rightarrow \infty$ then $\bar{x}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$.

THEOREM D.8 Markov's Strong Law of Large Numbers
If $\{z_i\}$ is a sequence of independent random variables with $E[z_i] = \mu_i < \infty$ and if for some $0 < \delta < 1$, $\sum_{i=1}^{\infty} E[|z_i - \mu_i|^{1+\delta}]/i^{1+\delta} < \infty$, then $\bar{z}_n - \bar{\mu}_n$ converges almost surely to 0, which we denote $\bar{z}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$.[2]

The variance condition is satisfied if every variance in the sequence is finite, but this is not strictly required; it only requires that the variances in the sequence increase at a slow enough rate that the sequence of variances as defined is bounded. The theorem allows for heterogeneity in the means and variances. If we return to the conditions of the Khinchine theorem, i.i.d. sampling, we have a corollary:

COROLLARY TO THEOREM D.8 (Kolmogorov)
If x_i, $i = 1, \ldots, n$ is a sequence of independent and identically distributed random variables such that $E[x_i] = \mu < \infty$ and $E[|x_i|] < \infty$, then $\bar{x}_n - \mu \xrightarrow{a.s.} 0$.

Note that the corollary requires identically distributed observations while the theorem only requires independence. Finally, another form of convergence encountered in the analysis of time-series data is convergence in rth mean:

[2] The use of the expected absolute deviation differs a bit from the expected squared deviation that we have used heretofore to characterize the spread of a distribution. Consider two examples. If $z \sim N[0, \sigma^2]$, then $E[|z|] = \text{Prob}[z < 0]E[-z|z < 0] + \text{Prob}[z \geq 0]E[z|z \geq 0] = 0.7979\sigma$. (See Theorem 18.2.) So, finite expected absolute value is the same as finite second moment for the normal distribution. But if z takes values $[0, n]$ with probabilities $[1 - 1/n, 1/n]$, then the variance of z is $(n - 1)$, but $E[|z - \mu_z|]$ is $2 - 2/n$. For this case, finite expected absolute value occurs without finite expected second moment. These are different characterizations of the spread of the distribution.

DEFINITION D.4 Convergence in rth Mean
If x_n is a sequence of random variables such that $E[|x_n|^r] < \infty$ and $\lim_{n\to\infty} E[|x_n - c|^r] = 0$, then x_n converges in rth mean to c. This is denoted $x_n \xrightarrow{r.m.} c$.

Surely the most common application is the one we met earlier, convergence in means square, which is convergence in the second mean. Some useful results follow from this definition:

THEOREM D.9 Convergence in Lower Powers
If x_n converges in rth mean to c, then x_n converges in sth mean to c for any $s < r$. The proof uses Jensen's Inequality, Theorem D.13. Write $E[|x_n - c|^s] = E[(|x_n - c|^r)^{s/r}] \leq E[(|x_n - c|^r)]\}^{s/r}$ and the inner term converges to zero so the full function must also.

THEOREM D.10 Generalized Chebychev's Inequality
If x_n is a random variable and c is a constant such that with $E[|x_n - c|^r] < \infty$ and ε is a positive constant, then $\text{Prob}(|x_n - c| > \varepsilon) \leq E[|x_n - c|^r]/\varepsilon^r$.

We have considered two cases of this result already, when $r = 1$ which is the Markov inequality, Theorem D.3, and when $r = 2$, which is the Chebychev inequality we looked at first in Theorem D.2.

THEOREM D.11 Convergence in rth mean and Convergence in Probability
If $x_n \xrightarrow{r.m.} c$, for some $r > 0$, then $x_n \xrightarrow{p} c$. The proof relies on Theorem D.10. By assumption, $\lim_{n\to\infty} E[|x_n - c|^r] = 0$ so for some n sufficiently large, $E[|x_n - c|^r] < \infty$. By Theorem D.10, then, $\text{Prob}(|x_n - c| > \varepsilon) \leq E[|x_n - c|^r]/\varepsilon^r$ for any $\varepsilon > 0$. The denominator of the fraction is a fixed constant and the numerator converges to zero by our initial assumption, so $\lim_{n\to\infty} \text{Prob}(|x_n - c| > \varepsilon) = 0$, which completes the proof.

One implication of Theorem D.11 is that although convergence in mean square is a convenient way to prove convergence in probability, it is actually stronger than necessary, as we get the same result for any positive *r*.

Finally, we note that we have now shown that both almost sure convergence and convergence in *r*th mean are stronger than convergence in probability; each implies the latter. But they, themselves, are different notions of convergence, and neither implies the other.

DEFINITION D.5 Convergence of a Random Vector or Matrix
Let $\mathbf{x}_n$ *denote a random vector and* $\mathbf{X}_n$ *a random matrix, and* $\mathbf{c}$ *and* $\mathbf{C}$ *denote a vector and matrix of constants with the same dimensions as* $\mathbf{x}_n$ *and* $\mathbf{X}_n$*, respectively. All of the preceding notions of convergence can be extended to* $(\mathbf{x}_n, \mathbf{c})$ *and* $(\mathbf{X}_n, \mathbf{C})$ *by applying the results to the respective corresponding elements.*

D.2.3 CONVERGENCE OF FUNCTIONS

A particularly convenient result is the following.

THEOREM D.12 Slutsky Theorem
For a continuous function $g(x_n)$ *that is not a function of* n,

$$\text{plim } g(x_n) = g(\text{plim } x_n). \tag{D-6}$$

The generalization of Theorem D.12 to a function of several random variables is direct, as illustrated in the next example.

Example D.3 Probability Limit of a Function of $\bar{x}$ and s^2

In random sampling from a population with mean μ and variance σ^2, the exact expected value of $\bar{x}_n^2/s_n^2$ will be difficult, if not impossible, to derive. But, by the Slutsky theorem,

$$\text{plim } \frac{\bar{x}_n^2}{s_n^2} = \frac{\mu^2}{\sigma^2}.$$

An application that highlights the difference between expectation and probability limit is suggested by the following useful relationships.

THEOREM D.13 Inequalities for Expectations
Jensen's Inequality. *If* $g(x_n)$ *is a concave function of* x_n*, then* $g(E[x_n]) \geq E[g(x_n)]$. **Cauchy–Schwarz Inequality**. *For two random variables,* $E[|xy|] \leq \{E[x^2]\}^{1/2}\{E[y^2]\}^{1/2}$.

Although the expected value of a function of x_n may not equal the function of the expected value—it exceeds it if the function is concave—the probability limit of the function *is* equal to the function of the probability limit.

The Slutsky theorem highlights a comparison between the expectation of a random variable and its probability limit. Theorem D.12 extends directly in two important directions. First, though stated in terms of convergence in probability, the same set of results applies to convergence in *r*th mean and almost sure convergence. Second, so long as the functions are continuous, the Slutsky theorem can be extended to vector or matrix valued functions of random scalars, vectors, or matrices. The following describe some specific applications. Some implications of the Slutsky theorem are now summarized.

THEOREM D.14 Rules for Probability Limits
If x_n and y_n are random variables with $\text{plim}\, x_n = c$ *and* $\text{plim}\, y_n = d$, *then*

$$\text{plim}(x_n + y_n) = c + d, \quad \textbf{(sum rule)} \tag{D-7}$$

$$\text{plim}\, x_n y_n = cd, \quad \textbf{(product rule)} \tag{D-8}$$

$$\text{plim}\, x_n/y_n = c/d \quad \text{if} \quad d \neq 0. \quad \textbf{(ratio rule)} \tag{D-9}$$

If $\mathbf{W}_n$ *is a matrix whose elements are random variables and if* $\text{plim}\, \mathbf{W}_n = \boldsymbol{\Omega}$, *then*

$$\text{plim}\, \mathbf{W}_n^{-1} = \boldsymbol{\Omega}^{-1}. \quad \textbf{(matrix inverse rule)} \tag{D-10}$$

If $\mathbf{X}_n$ *and* $\mathbf{Y}_n$ *are random matrices with* $\text{plim}\, \mathbf{X}_n = \mathbf{A}$ *and* $\text{plim}\, \mathbf{Y}_n = \mathbf{B}$, *then*

$$\text{plim}\, \mathbf{X}_n\mathbf{Y}_n = \mathbf{AB}. \quad \textbf{(matrix product rule)} \tag{D-11}$$

D.2.4 CONVERGENCE TO A RANDOM VARIABLE

The preceding has dealt with conditions under which a random variable converges to a constant, for example, the way that a sample mean converges to the population mean. To develop a theory for the behavior of estimators, as a prelude to the discussion of limiting distributions, we now consider cases in which a random variable converges not to a constant, but to another random variable. These results will actually subsume those in the preceding section, as a constant may always be viewed as a degenerate random variable, that is one with zero variance.

DEFINITION D.6 Convergence in Probability to a Random Variable
The random variable x_n converges in probability to the random variable x if $\lim_{n\to\infty} \text{Prob}(|x_n - x| > \varepsilon) = 0$ *for any positive* ε.

As before, we write $\text{plim}\, x_n = x$ to denote this case. The interpretation (at least the intuition) of this type of convergence is different when x is a random variable. The notion of closeness defined here relates not to the concentration of the mass of the probability mechanism generating x_n at a point c, but to the closeness of that probability mechanism to that of x. One can think of this as a convergence of the CDF of x_n to that of x.

DEFINITION D.7 Almost Sure Convergence to a Random Variable
The random variable x_n converges almost surely to the random variable x if and only if $\lim_{n\to\infty}\text{Prob}(|x_i - x| > \varepsilon$ *for all* $i \geq n) = 0$ *for all* $\varepsilon > 0$.

DEFINITION D.8 Convergence in rth Mean to a Random Variable
The random variable x_n converges in rth mean to the random variable x if and only if $\lim_{n\to\infty}E[|x_n - x|^r] = 0$. *This is labeled* $x_n \xrightarrow{r.m.} x$. *As before, the case $r = 2$ is labeled convergence in mean square.*

Once again, we have to revise our understanding of convergence when convergence is to a random variable.

THEOREM D.15 Convergence of Moments
Suppose $x_n \xrightarrow{r.m.} x$ and $E[|x|^r]$ is finite. Then, $\lim_{n\to\infty}E[|x_n|^r] = E[|x|^r]$.

Theorem D.15 raises an interesting question. Suppose we let r grow, and suppose that $x_n \xrightarrow{r.m.} x$ and, in addition, all moments are finite. If this holds for any r, do we conclude that these random variables have the same distribution? The answer to this longstanding problem in probability theory—the problem of the sequence of moments—is no. The sequence of moments does not uniquely determine the distribution. Although convergence in rth mean and almost surely still both imply convergence in probability, it remains true, even with convergence to a random variable instead of a constant, that these are different forms of convergence.

D.2.5 CONVERGENCE IN DISTRIBUTION: LIMITING DISTRIBUTIONS

A second form of convergence is **convergence in distribution**. Let x_n be a sequence of random variables indexed by the sample size, and assume that x_n has cdf $F_n(x_n)$.

DEFINITION D.9 Convergence in Distribution
x_n converges in distribution to a random variable x with CDF $F(x)$ if $\lim_{n\to\infty}|F_n(x_n) - F(x)| = 0$ *at all continuity points of* $F(x)$.

This statement is about the probability distribution associated with x_n; it does not imply that x_n converges at all. To take a trivial example, suppose that the exact distribution of the random variable x_n is

$$\text{Prob}(x_n = 1) = \frac{1}{2} + \frac{1}{n+1}, \quad \text{Prob}(x_n = 2) = \frac{1}{2} - \frac{1}{n+1}.$$

As n increases without bound, the two probabilities converge to $\frac{1}{2}$, but x_n does not converge to a constant.

DEFINITION D.10 Limiting Distribution
If x_n converges in distribution to x, where $F_n(x_n)$ is the CDF of x_n, then $F(x)$ is the **limiting distribution** *of x_n. This is written $x_n \xrightarrow{d} x$.*

The limiting distribution is often given in terms of the pdf, or simply the parametric family. For example, "the limiting distribution of x_n is standard normal."

Convergence in distribution can be extended to random vectors and matrices, although not in the element by element manner that we extended the earlier convergence forms. The reason is that convergence in distribution is a property of the CDF of the random variable, not the variable itself. Thus, we can obtain a convergence result analogous to that in Definition D.9 for vectors or matrices by applying definition to the joint CDF for the elements of the vector or matrices. Thus, $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ if $\lim_{n\to\infty}|F_n(\mathbf{x}_n) - F(\mathbf{x})| = 0$ and likewise for a random matrix.

Example D.4 Limiting Distribution of t_{n-1}

Consider a sample of size n from a standard normal distribution. A familiar inference problem is the test of the hypothesis that the population mean is zero. The test statistic usually used is the t statistic:

$$t_{n-1} = \frac{\overline{x}_n}{s_n/\sqrt{n}},$$

where

$$s_n^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{n-1}.$$

The exact distribution of the random variable t_{n-1} is t with $n - 1$ degrees of freedom. The density is different for every n:

$$f(t_{n-1}) = \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}[(n-1)\pi]^{-1/2}\left[1 + \frac{t_{n-1}^2}{n-1}\right]^{-n/2}, \qquad \textbf{(D-12)}$$

as is the CDF, $F_{n-1}(t) = \int_{-\infty}^{t} f_{n-1}(x)\,dx$. This distribution has mean zero and variance $(n-1)/(n-3)$. As n grows to infinity, t_{n-1} converges to the standard normal, which is written

$$t_{n-1} \xrightarrow{d} N[0, 1].$$

DEFINITION D.11 Limiting Mean and Variance
The **limiting mean** *and* **variance** *of a random variable are the mean and variance of the limiting distribution, assuming that the limiting distribution and its moments exist.*

For the random variable with $t[n]$ distribution, the exact mean and variance are zero and $n/(n - 2)$, whereas the limiting mean and variance are zero and one. The example might suggest that the limiting mean and variance are zero and one; that is, that the moments of the limiting distribution are the ordinary limits of the moments of the finite sample distributions. This situation is almost always true, but it need not be. It is possible to construct examples in which the exact moments do not even exist, even though the moments of the limiting distribution are well defined.[3] Even in such cases, we can usually derive the mean and variance of the limiting distribution.

Limiting distributions, like probability limits, can greatly simplify the analysis of a problem. Some results that combine the two concepts are as follows.[4]

THEOREM D.16 Rules for Limiting Distributions

1. *If* $x_n \xrightarrow{d} x$ *and* $\text{plim } y_n = c$, *then*

$$x_n y_n \xrightarrow{d} cx, \qquad \textbf{(D-13)}$$

which means that the limiting distribution of $x_n y_n$ *is the distribution of* cx. *Also,*

$$x_n + y_n \xrightarrow{d} x + c, \qquad \textbf{(D-14)}$$

$$x_n/y_n \xrightarrow{d} x/c, \quad \text{if } c \neq 0. \qquad \textbf{(D-15)}$$

2. *If* $x_n \xrightarrow{d} x$ *and* $g(x_n)$ *is a continuous function, then*

$$g(x_n) \xrightarrow{d} g(x). \qquad \textbf{(D-16)}$$

This result is analogous to the Slutsky theorem for probability limits. For an example, consider the t_n *random variable discussed earlier. The exact distribution of* t_n^2 *is* $F[1, n]$. *But as* $n \longrightarrow \infty$, t_n *converges to a standard normal variable. According to this result, the limiting distribution of* t_n^2 *will be that of the square of a standard normal, which is chi-squared with one degree of freedom. We conclude, therefore, that*

$$F[1, n] \xrightarrow{d} \textit{chi-squared}[1]. \qquad \textbf{(D-17)}$$

We encountered this result in our earlier discussion of limiting forms of the standard normal family of distributions.

3. *If* y_n *has a limiting distribution and* $\text{plim } (x_n - y_n) = 0$, *then* x_n *has the same limiting distribution as* y_n.

[3] See, for example, Maddala (1977a, p. 150).

[4] For proofs and further discussion, see, for example, Greenberg and Webster (1983).

The third result in Theorem D.16 combines convergence in distribution and in probability. The second result can be extended to vectors and matrices.

Example D.5 The F Distribution

Suppose that $\boldsymbol{t}_{1,n}$ and $\boldsymbol{t}_{2,n}$ are a $K\times 1$ and an $M\times 1$ random vector of variables whose components are independent with each distributed as t with n degrees of freedom. Then, as we saw in the preceding, for any component in either random vector, the limiting distribution is standard normal, so for the entire vector, $\boldsymbol{t}_{j,n} \xrightarrow{d} \boldsymbol{z}_j$, a vector of independent standard normally distributed variables. The results so far show that $\dfrac{(\boldsymbol{t}'_{1,n}\,\boldsymbol{t}_{1,n})/K}{(\boldsymbol{t}'_{2,n}\,t_{2,n})/M} \xrightarrow{d} F[K, M]$.

Finally, a specific case of result 2 in Theorem D.16 produces a tool known as the Cramér–Wold device.

THEOREM D.17 Cramer–Wold Device

If $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, *then* $\mathbf{c}'\mathbf{x}_n \xrightarrow{d} \mathbf{c}'\mathbf{x}$ *for all conformable vectors* $\mathbf{c}$ *with real valued elements.*

By allowing $\mathbf{c}$ to be a vector with just a one in a particular position and zeros elsewhere, we see that convergence in distribution of a random vector $\mathbf{x}_n$ to $\mathbf{x}$ does imply that each component does likewise.

D.2.6 CENTRAL LIMIT THEOREMS

We are ultimately interested in finding a way to describe the statistical properties of estimators when their exact distributions are unknown. The concepts of consistency and convergence in probability are important. But the theory of limiting distributions given earlier is not yet adequate. We rarely deal with estimators that are not consistent for something, though perhaps not always the parameter we are trying to estimate. As such,

$$\text{if plim } \hat{\theta}_n = \theta, \quad \text{then } \hat{\theta}_n \xrightarrow{d} \theta.$$

That is, the limiting distribution of $\hat{\theta}_n$ is a spike. This is not very informative, nor is it at all what we have in mind when we speak of the statistical properties of an estimator. (To endow our finite sample estimator $\hat{\theta}_n$ with the zero sampling variance of the spike at θ would be optimistic in the extreme.)

As an intermediate step, then, to a more reasonable description of the statistical properties of an estimator, we use a **stabilizing transformation** of the random variable to one that does have a well-defined limiting distribution. To jump to the most common application, whereas

$$\text{plim } \hat{\theta}_n = \theta,$$

we often find that

$$z_n = \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} f(z),$$

where $f(z)$ is a well-defined distribution with a mean and a positive variance. An estimator which has this property is said to be **root-*n* consistent**. The single most important theorem in econometrics provides an application of this proposition. A basic form of the theorem is as follows.

THEOREM D.18 Lindeberg–Levy Central Limit Theorem (Univariate)
If $x_1, \ldots, x_n$ are a random sample from a probability distribution with finite mean μ and finite variance σ^2 and $\bar{x}_n = (1/n)\sum_{i=1}^{n} x_i$, then $\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[0, \sigma^2]$. A proof appears in Rao (1973, p. 127).

The result is quite remarkable as it holds regardless of the form of the parent distribution. For a striking example, return to Figure C.3. The distribution from which the data were drawn in that figure does not even remotely resemble a normal distribution. In samples of only four observations the force of the central limit theorem is clearly visible in the sampling distribution of the means. The sampling experiment Example D.6 shows the effect in a systematic demonstration of the result.

The Lindeberg–Levy theorem is one of several forms of this extremely powerful result. For our purposes, an important extension allows us to relax the assumption of equal variances. The Lindeberg–Feller form of the central limit theorem is the centerpiece of most of our analysis in econometrics.

THEOREM D.19 Lindeberg–Feller Central Limit Theorem (with Unequal Variances)
Suppose that $\{x_i\}, i = 1, \ldots, n$, is a sequence of independent random variables with finite means μ_i and finite positive variances σ_i^2. Let

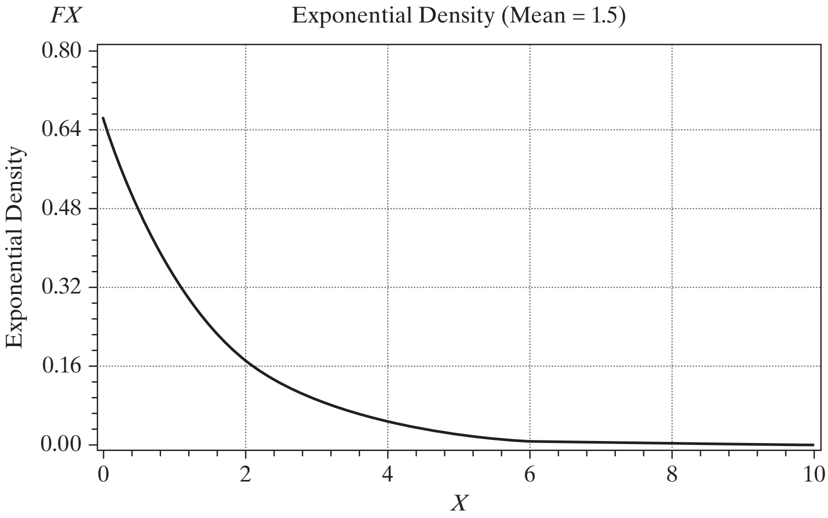
$$\bar{\mu}_n = \frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n), \quad \text{and} \quad \bar{\sigma}_n^2 = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 + \cdots, \sigma_n^2).$$

If no single term dominates this average variance, which we could state as $\lim_{n\to\infty}\max(\sigma_i)/(\sqrt{n}\bar{\sigma}_n) = 0$, and if the average variance converges to a finite constant, $\bar{\sigma}^2 = \lim_{n\to\infty}\bar{\sigma}_n^2$, then $\sqrt{n}(\bar{x}_n - \bar{\mu}_n) \xrightarrow{d} N[0, \bar{\sigma}^2]$.

In practical terms, the theorem states that sums of random variables, regardless of their form, will tend to be normally distributed. The result is yet more remarkable in that *it does not require the variables in the sum to come from the same underlying distribution. It requires, essentially, only that the mean be a mixture of many random variables, none of which is large compared with their sum.* Because nearly all the estimators we construct in econometrics fall under the purview of the central limit theorem, it is obviously an important result.

Proof of the Lindeberg–Feller theorem requires some quite intricate mathematics [see, e.g., Loeve (1977)] that are well beyond the scope of our work here. We do note an important consideration in this theorem. The result rests on a condition known as the *Lindeberg condition*. The sample mean computed in the theorem is a mixture of random

FIGURE D.2 The Exponential Distribution.



variables from possibly different distributions. The Lindeberg condition, in words, states that the contribution of the tail areas of these underlying distributions to the variance of the sum must be negligible in the limit. The condition formalizes the assumption in Theorem D.19 that the average variance be positive and not be dominated by any single term. [For an intuitively crafted mathematical discussion of this condition, see White (2001, pp. 117–118).] The condition is essentially impossible to verify in practice, so it is useful to have a simpler version of the theorem that encompasses it.

Example D.6 The Lindeberg–Levy Central Limit Theorem

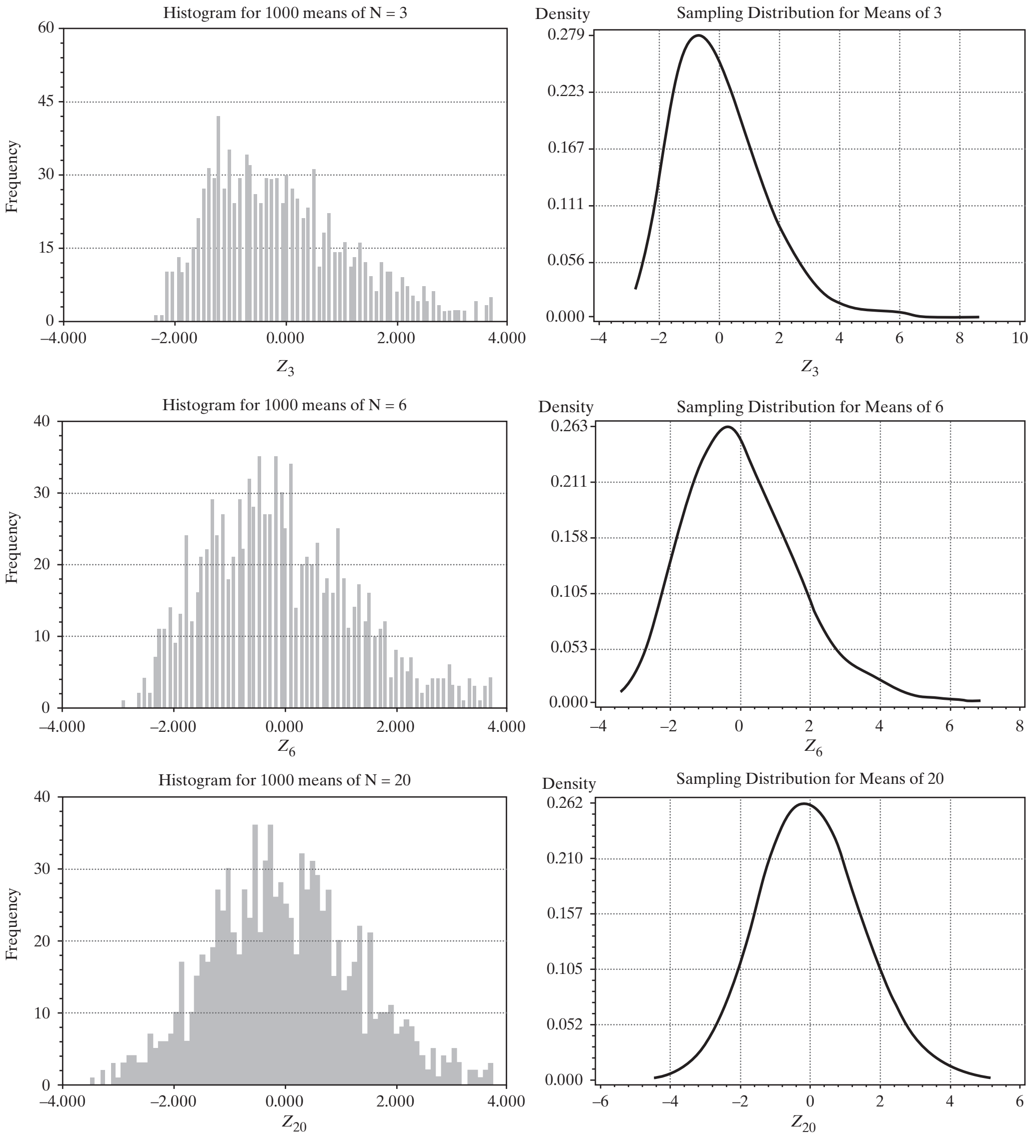
We'll use a sampling experiment to demonstrate the operation of the central limit theorem. Consider random sampling from the exponential distribution with mean 1.5—this is the setting used in Example C.4. The density is shown in Figure D.2.

We've drawn 1,000 samples of 3, 6, and 20 observations from this population and computed the sample means for each. For each mean, we then computed $z_{in} = \sqrt{n}(\bar{x}_{in} - \mu)$, where $i = 1, \ldots, 1{,}000$ and n is 3, 6, or 20. The three rows of figures in Figure D.3 show histograms of the observed samples of sample means and kernel density estimates of the underlying distributions for the three samples of transformed means. The force of the central limit is clearly visible in the shapes of the distributions.

THEOREM D.20 Liapounov Central Limit Theorem

Suppose that $\{x_i\}$ is a sequence of independent random variables with finite means μ_i and finite positive variances σ_i^2 such that $E[|x_i - \mu_i|^{2+\delta}]$ is finite for some $\delta > 0$. If $\bar{\sigma}_n$ is positive and finite for all n sufficiently large, then $\sqrt{n}(\bar{x}_n - \bar{\mu}_n)/\bar{\sigma}_n \xrightarrow{d} N[0, 1]$.

FIGURE D.3 The Central Limit Theorem.



This version of the central limit theorem requires only that moments slightly larger than two be finite.

Note the distinction between the laws of large numbers in Theorems D.5 and D.6 and the central limit theorems. Neither asserts that sample means tend to normality. Sample means (i.e., the distributions of them) converge to spikes at the true mean. It is the transformation of the mean, $\sqrt{n}(\bar{x}_n - \mu)/\sigma$, that converges to standard normality. To see this at work, if you have access to the necessary software, you might try reproducing Example D.6 using the raw means, $\bar{x}_{in}$. What do you expect to observe?

For later purposes, we will require multivariate versions of these theorems. Proofs of the following may be found, for example, in Greenberg and Webster (1983) or Rao (1973) and references cited there.

THEOREM D.18A Multivariate Lindeberg–Levy Central Limit Theorem
If $\mathbf{x}_1, \dots, \mathbf{x}_n$ are a random sample from a multivariate distribution with finite mean vector $\boldsymbol{\mu}$ and finite positive definite covariance matrix $\mathbf{Q}$, then

$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}],$$

where

$$\bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

To get from D.18 to D.18A (and D.19 to D.19A) we need to add a step. Theorem D.18 applies to the individual elements of the vector. A vector has a multivariate normal distribution if the individual elements are normally distributed and if every linear combination is normally distributed. We can use Theorem D.18 (D.19) for the individual terms and Theorem D.17 to establish that linear combinations behave likewise. This establishes the extensions.

The extension of the Lindeberg–Feller theorem to unequal covariance matrices requires some intricate mathematics. The following is an informal statement of the relevant conditions. Further discussion and references appear in Fomby, Hill, and Johnson (1984) and Greenberg and Webster (1983).

THEOREM D.19A Multivariate Lindeberg–Feller Central Limit Theorem
Suppose that $\mathbf{x}_1, \dots, \mathbf{x}_n$ are a sample of random vectors such that $E[\mathbf{x}_i] = \boldsymbol{\mu}_i$, $\text{Var}[\mathbf{x}_i] = \mathbf{Q}_i$, and all mixed third moments of the multivariate distribution are finite. Let

$$\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\mu_i \text{ and } \overline{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}_i.$$

We assume that

$$\lim_{n\to\infty}\overline{\mathbf{Q}}_n = \mathbf{Q},$$

where $\mathbf{Q}$ is a finite, positive definite matrix, and that for every i,

$$\lim_{n\to\infty}(n\overline{\mathbf{Q}}_n)^{-1}\mathbf{Q}_i = \lim_{n\to\infty}\left(\sum_{i=1}^{n}\mathbf{Q}_i\right)^{-1}\mathbf{Q}_i = \mathbf{0}.$$

We allow the means of the random vectors to differ, although in the cases that we will analyze, they will generally be identical. The second assumption states that individual components of the sum must be finite and diminish in significance. There is also an implicit assumption that the sum of matrices is nonsingular. Because the limiting matrix is nonsingular, the assumption must hold for large enough n, which is all that concerns us here. With these in place, the result is

$$\sqrt{n}(\bar{\mathbf{x}}_n - \bar{\mu}_n) \xrightarrow{d} N[\mathbf{0}, \mathbf{Q}].$$

D.2.7 THE DELTA METHOD

At several points in Appendix C, we used a linear Taylor series approximation to analyze the distribution and moments of a random variable. We are now able to justify this usage. We complete the development of Theorem D.12 (probability limit of a function of a random variable), Theorem D.16 (2) (limiting distribution of a function of a random variable), and the central limit theorems, with a useful result that is known as the **delta method**. For a single random variable (sample mean or otherwise), we have the following theorem.

THEOREM D.21 Limiting Normal Distribution of a Function
If $\sqrt{n}(z_n - \mu) \xrightarrow{d} N[0, \sigma^2]$ and if $g(z_n)$ is a continuous and continuously differentiable function with $g'(\mu)$ not equal to zero and not involving n, then

$$\sqrt{n}[g(z_n) - g(\mu)] \xrightarrow{d} N[0, \{g'(\mu)\}^2 \sigma^2]. \qquad \textbf{(D-18)}$$

Notice that the mean and variance of the limiting distribution are the mean and variance of the linear Taylor series approximation:

$$g(z_n) \simeq g(\mu) + g'(\mu)(z_n - \mu).$$

The multivariate version of this theorem will be used at many points in the text.

THEOREM D.21A Limiting Normal Distribution of a Set of Functions
If $\mathbf{z}_n$ is a $K \times 1$ sequence of vector-valued random variables such that $\sqrt{n}(\mathbf{z}_n - \boldsymbol{\mu}) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Sigma}]$ and if $\mathbf{c}(\mathbf{z}_n)$ is a set of J continuous and continuously differentiable functions of $\mathbf{z}_n$ with $\mathbf{C}(\boldsymbol{\mu})$ not equal to zero, not involving n, then

$$\sqrt{n}[\mathbf{c}(\mathbf{z}_n) - \mathbf{c}(\boldsymbol{\mu})] \xrightarrow{d} N[\mathbf{0}, \mathbf{C}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{C}(\boldsymbol{\mu})'], \qquad \textbf{(D-19)}$$

where $\mathbf{C}(\boldsymbol{\mu})$ is the $J \times K$ matrix $\partial \mathbf{c}(\boldsymbol{\mu})/\partial \boldsymbol{\mu}'$. The jth row of $\mathbf{C}(\boldsymbol{\mu})$ is the vector of partial derivatives of the jth function with respect to $\boldsymbol{\mu}'$.

D.3 ASYMPTOTIC DISTRIBUTIONS

The theory of limiting distributions is only a means to an end. We are interested in the behavior of the estimators themselves. The limiting distributions obtained through the central limit theorem all involve unknown parameters, generally the ones we are trying to estimate. Moreover, our samples are always finite. Thus, we depart from the limiting distributions to derive the asymptotic distributions of the estimators.

DEFINITION D.12 Asymptotic Distribution
An asymptotic distribution is a distribution that is used to approximate the true finite sample distribution of a random variable.[5]

By far the most common means of formulating an asymptotic distribution (at least by econometricians) is to construct it from the known limiting distribution of a function of the random variable. If

$$\sqrt{n}[(\bar{x}_n - \mu)/\sigma] \xrightarrow{d} N[0, 1],$$

then approximately, or asymptotically, $\bar{x}_n \sim N[\mu, \sigma^2/n]$, which we write as

$$\bar{x}_n \overset{a}{\sim} N[\mu, \sigma^2/n].$$

The statement "$\bar{x}_n$ is asymptotically normally distributed with mean μ and variance σ^2/n" says only that this normal distribution provides an approximation to the true distribution, not that the true distribution is exactly normal.

Example D.7 Asymptotic Distribution of the Mean of an Exponential Sample

In sampling from an exponential distribution with parameter θ, the *exact* distribution of $\bar{x}_n$ is that of $\theta/(2n)$ times a chi-squared variable with $2n$ degrees of freedom. The *asymptotic* distribution is $N[\theta, \theta^2/n]$. The exact and asymptotic distributions are shown in Figure D.4 for the case of $\theta = 1$ and $n = 16$.

Extending the definition, suppose that $\hat{\boldsymbol{\theta}}_n$ is an estimator of the parameter vector $\boldsymbol{\theta}$. The asymptotic distribution of the vector $\hat{\boldsymbol{\theta}}_n$ is obtained from the limiting distribution:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \mathbf{V}] \qquad \textbf{(D-20)}$$

implies that

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\left[\boldsymbol{\theta}, \frac{1}{n}\mathbf{V}\right]. \qquad \textbf{(D-21)}$$

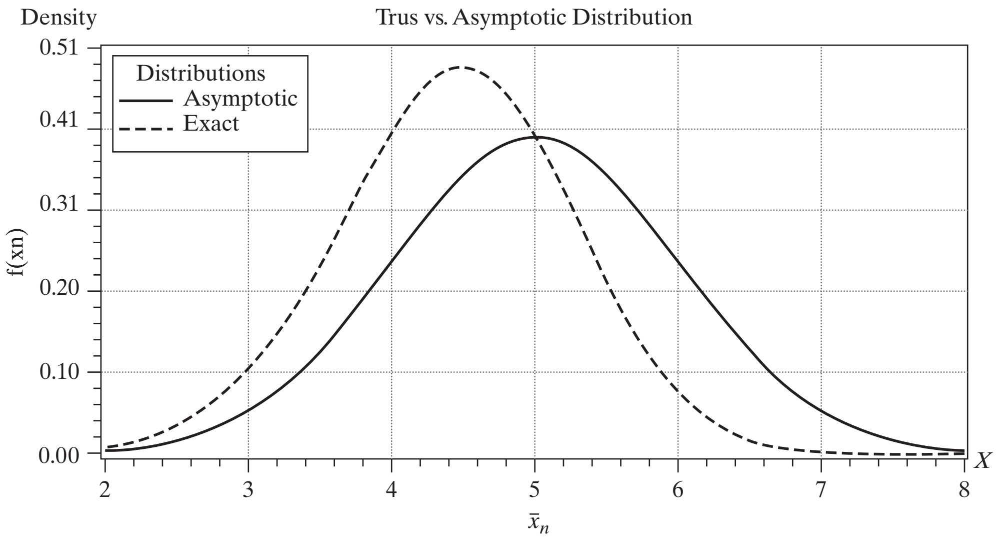
This notation is read "$\hat{\boldsymbol{\theta}}_n$ is asymptotically normally distributed, with mean vector $\boldsymbol{\theta}$ and covariance matrix $(1/n)\mathbf{V}$." The covariance matrix of the asymptotic distribution is the **asymptotic covariance matrix** and is denoted

$$\text{Asy. Var}[\hat{\boldsymbol{\theta}}_n] = \frac{1}{n}\mathbf{V}.$$

Note, once again, the logic used to reach the result; (D-20) holds exactly as $n \rightarrow \infty$. We assume that it holds approximately for finite n, which leads to (D-21).

[5] We differ a bit from some other treatments—for example, White (2001), Hayashi (2000, p. 90) —at this point, because they make no distinction between an asymptotic distribution and the limiting distribution, although the treatments are largely along the lines discussed here. In the interest of maintaining consistency of the discussion, we prefer to retain the sharp distinction and derive the asymptotic distribution of an estimator, **t** by first obtaining the *limiting* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$. By our construction, the *limiting* distribution of **t** is degenerate, whereas the *asymptotic* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$ is not useful.

FIGURE D.4 True Versus Asymptotic Distribution.



DEFINITION D.13 Asymptotic Normality and Asymptotic Efficiency
An estimator $\hat{\boldsymbol{\theta}}_n$ is asymptotically normal if (D-20) holds. The estimator is asymptotically efficient if the covariance matrix of any other consistent, asymptotically normally distributed estimator exceeds $(1/n)\mathbf{V}$ by a nonnegative definite matrix.

For most estimation problems, these are the criteria used to choose an estimator.

Example D.8 Asymptotic Inefficiency of the Median in Normal Sampling

In sampling from a normal distribution with mean μ and variance σ^2, both the mean $\bar{x}_n$ and the median M_n of the sample are consistent estimators of μ. The limiting distributions of both estimators are spikes at μ, so they can only be compared on the basis of their asymptotic properties. The necessary results are

$$\bar{x}_n \overset{a}{\sim} N[\mu, \sigma^2/n], \quad \text{and} \quad M_n \overset{a}{\sim} N[\mu, (\pi/2)\sigma^2/n]. \tag{D-22}$$

Therefore, the mean is more efficient by a factor of $\pi/2$. (But, see Example 15.7 for a finite sample result.)

D.3.1 ASYMPTOTIC DISTRIBUTION OF A NONLINEAR FUNCTION

Theorems D.12 and D.14 for functions of a random variable have counterparts in asymptotic distributions.

THEOREM D.22 Asymptotic Distribution of a Nonlinear Function
If $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N[0, \sigma^2]$ *and if* $g(\theta)$ *is a continuous and continuously differentiable function with* $g'(\theta)$ *not equal to zero and not involving* n*, then* $g(\hat{\theta}_n) \overset{a}{\sim} N[g(\theta), (1/n)\{g'(\theta)\}^2 \sigma^2]$*. If* $\hat{\boldsymbol{\theta}}_n$ *is a vector of parameter estimators such that* $\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N[\boldsymbol{\theta}, (1/n)\mathbf{V}]$ *and if* $\mathbf{c}(\boldsymbol{\theta})$ *is a set of J continuous functions not involving* n*, then* $\mathbf{c}(\hat{\boldsymbol{\theta}}_n) \overset{a}{\sim} N[\mathbf{c}(\boldsymbol{\theta}), (1/n)\mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})']$*, where* $\mathbf{C}(\boldsymbol{\theta}) = \partial\mathbf{c}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$.

Example D.9 Asymptotic Distribution of a Function of Two Estimators

Suppose that b_n and t_n are estimators of parameters β and θ such that

$$\begin{bmatrix} b_n \\ t_n \end{bmatrix} \overset{a}{\sim} N\left[\begin{pmatrix} \beta \\ \theta \end{pmatrix}, \begin{pmatrix} \sigma_{\beta\beta} & \sigma_{\beta\theta} \\ \sigma_{\theta\beta} & \sigma_{\theta\theta} \end{pmatrix}\right].$$

Find the asymptotic distribution of $c_n = b_n/(1 - t_n)$. Let $\gamma = \beta/(1 - \theta)$. By the Slutsky theorem, c_n is consistent for γ. We shall require

$$\frac{\partial\gamma}{\partial\beta} = \frac{1}{1-\theta} = \gamma_\beta, \quad \frac{\partial\gamma}{\partial\theta} = \frac{\beta}{(1-\theta)^2} = \gamma_\theta.$$

Let $\boldsymbol{\Sigma}$ be the 2×2 asymptotic covariance matrix given previously. Then the asymptotic variance of c_n is

$$\text{Asy. Var}[c_n] = (\gamma_\beta \ \gamma_\theta)\boldsymbol{\Sigma}\begin{pmatrix} \gamma_\beta \\ \gamma_\theta \end{pmatrix} = \gamma_\beta^2\sigma_{\beta\beta} + \gamma_\theta^2\sigma_{\theta\theta} + 2\gamma_\beta\gamma_\theta\sigma_{\beta\theta},$$

which is the variance of the linear Taylor series approximation:

$$\hat{\gamma}_n \simeq \gamma + \gamma_\beta(b_n - \beta) + \gamma_\theta(t_n - \theta).$$

D.3.2 ASYMPTOTIC EXPECTATIONS

The asymptotic mean and variance of a random variable are usually the mean and variance of the asymptotic distribution. Thus, for an estimator with the limiting distribution defined in

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N[\mathbf{0}, \mathbf{V}],$$

the asymptotic expectation is $\boldsymbol{\theta}$ and the asymptotic variance is $(1/n)\ \mathbf{V}$. This statement implies, among other things, that the estimator is "asymptotically unbiased."

At the risk of clouding the issue a bit, it is necessary to reconsider one aspect of the previous description. We have deliberately avoided the use of consistency even though, in most instances, that is what we have in mind. The description thus far might suggest that consistency and asymptotic unbiasedness are the same. Unfortunately (because it is a source of some confusion), they are not. They are if the estimator is consistent and asymptotically normally distributed, or CAN. They may differ in other settings, however. There are at least three possible definitions of asymptotic unbiasedness:

1. The mean of the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ is 0.

$$\lim_{n\to\infty} E[\hat{\theta}_n] = \theta. \qquad \textbf{(D-23)}$$

2. plim $\theta_n = \theta$.

In most cases encountered in practice, the estimator in hand will have all three properties, so there is no ambiguity. It is not difficult to construct cases in which the left-hand sides of all three definitions are different, however.[6] There is no general agreement among authors as to the precise meaning of asymptotic unbiasedness, perhaps because the term is misleading at the outset; *asymptotic* refers to an approximation, whereas *unbiasedness* is an exact result.[7] Nonetheless, the majority view seems to be that (2) is the proper definition of asymptotic unbiasedness.[8] Note, though, that this definition relies on quantities that are generally unknown and that may not exist.

A similar problem arises in the definition of the asymptotic variance of an estimator. One common definition is[9]

$$\text{Asy. Var}[\hat{\theta}_n] = \frac{1}{n}\lim_{n\to\infty} E\,[\{\sqrt{n}(\hat{\theta}_n - \lim_{n\to\infty} E[\hat{\theta}_n])\}^2]. \qquad \textbf{(D-24)}$$

This result is a **leading term approximation**, and it will be sufficient for nearly all applications. Note, however, that like definition 2 of asymptotic unbiasedness, it relies on unknown and possibly nonexistent quantities.

Example D.10 Asymptotic Moments of the Normal Sample Variance

The exact expected value and variance of the variance estimator in a normal sample

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textbf{(D-25)}$$

are

$$E[m_2] = \frac{(n-1)\sigma^2}{n}, \qquad \textbf{(D-26)}$$

and

$$\text{Var}[m_2] = \frac{\mu_4 - \sigma^4}{n} - \frac{2(\mu_4 - 2\sigma^4)}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}, \qquad \textbf{(D-27)}$$

where $\mu_4 = E[(x - \mu)^4]$. [See Goldberger (1964, pp. 97–99).] The leading term approximation would be

$$\text{Asy. Var}[m_2] = \frac{1}{n}(\mu_4 - \sigma^4).$$

[6] See, for example, Maddala (1977a, p. 150).

[7] See, for example, Theil (1971, p. 377).

[8] Many studies of estimators analyze the "asymptotic bias" of, say, $\hat{\theta}_n$ as an estimator of a parameter θ. In most cases, the quantity of interest is actually plim $[\hat{\theta}_n - \theta]$. See, for example, Greene (1980b) and another example in Johnston (1984, p. 312).

[9] Kmenta (1986, p.165).

D.4 SEQUENCES AND THE ORDER OF A SEQUENCE

This section has been concerned with sequences of constants, denoted, for example, c_n, and random variables, such as x_n, that are indexed by a sample size, n. An important characteristic of a sequence is the rate at which it converges (or diverges). For example, as we have seen, the mean of a random sample of n observations from a distribution with finite mean, μ, and finite variance, σ^2, is itself a random variable with variance $\gamma_n^2 = \sigma^2/n$. We see that as long as σ^2 is a finite constant, γ_n^2 is a sequence of constants that converges to zero. Another example is the random variable $x_{(1),n}$, the minimum value in a random sample of n observations from the exponential distribution with mean $1/\theta$ defined in Example C.4. It turns out that $x_{(1),n}$ has variance $1/(n\theta)^2$. Clearly, this variance also converges to zero, but, intuition suggests, faster than σ^2/n does. On the other hand, the sum of the integers from one to n, $S_n = n(n + 1)/2$, obviously diverges as $n \rightarrow \infty$, albeit faster (one might expect) than the log of the likelihood function for the exponential distribution in Example C.6, which is $\ln L(\theta) = n(\ln \theta - \theta\bar{x}_n)$. As a final example, consider the downward bias of the maximum likelihood estimator of the variance of the normal distribution, $c_n = (n - 1)/n$, which is a constant that converges to one. (See Example C.5.)

We will define the rate at which a sequence converges or diverges in terms of the **order of the sequence**.

DEFINITION D.14 Order n^δ
A sequence c_n is of order n^δ, denoted $O(n^\delta)$, if and only if $\operatorname{plim}(1/n^\delta)c_n$ *is a finite nonzero constant.*

DEFINITION D.15 Order less than n^δ
A sequence c_n, is of order less than n^δ, denoted $o(n^\delta)$, if and only if $\operatorname{plim}(1/n^\delta)c_n$ *equals zero.*

Thus, in our examples, γ_n^2 is $O(n^{-1})$, $\text{Var}[x_{(1),n}]$ is $O(n^{-2})$ and $o(n^{-1})$, S_n is $O(n^2)$ (δ equals $+2$ in this case), $\ln L(\theta)$ is $\text{O}(n)$ (δ equals $+1$), and c_n is $O(1)(\delta = 0)$. Important particular cases that we will encounter repeatedly in our work are sequences for which $\delta = 1$ or -1.

The notion of order of a sequence is often of interest in econometrics in the context of the variance of an estimator. Thus, we see in Section D.3 that an important element of our strategy for forming an asymptotic distribution is that the variance of the limiting distribution of $\sqrt{n}(\bar{x}_n - \mu)/\sigma$ is $O(1)$. In Example D.10 the variance of m_2 is the sum of three terms that are $O(n^{-1})$, $O(n^{-2})$, and $O(n^{-3})$. The sum is $O(n^{-1})$, because $n\,\text{Var}[m_2]$ converges to $\mu_4 - \sigma^4$, the numerator of the first, or *leading term,* whereas the second and third terms converge to zero. This term is also the *dominant term* of the sequence. Finally, consider the two divergent examples in the preceding list. S_n is simply a deterministic function of n that explodes. However, $\ln L(\theta) = n \ln \theta - \theta \Sigma_i x_i$ is the sum of a constant that is $O(n)$ and a random variable with variance equal to n/θ. The random variable "diverges" in the sense that its variance grows without bound as n increases.

APPENDIX E

COMPUTATION AND OPTIMIZATION

E.1 INTRODUCTION

The computation of empirical estimates by econometricians involves using digital computers and software written either by the researchers themselves or by others.[1] It is also a surprisingly balanced mix of art and science. It is important for software users to be aware of how results are obtained, not only to understand routine computations, but also to be able to explain the occasional strange and contradictory results that do arise. This appendix will describe some of the basic elements of computing and a number of tools that are used by econometricians.[2] Section E.2 describes some techniques for computing certain integrals and derivatives that are recurrent in econometric applications. Section E.3 presents methods of optimization of functions. Some examples are given in Section E.4.

E.2 COMPUTATION IN ECONOMETRICS

This section will discuss some methods of computing integrals that appear frequently in econometrics.

[1] It is one of the interesting aspects of the development of econometric methodology that the adoption of certain classes of techniques has proceeded in discrete jumps with the development of software. Noteworthy examples include the appearance, both around 1970, of G. K. Joreskog's LISREL [Joreskog and Sorbom (1981)] program, which spawned a still-growing industry in linear structural modeling, and TSP [Hall (1982, 1984)], which was among the first computer programs to accept symbolic representations of econometric models and which provided a significant advance in econometric practice with its LSQ procedure for systems of equations. An extensive survey of the evolution of econometric software is given in Renfro (2007, 2009).

[2] This discussion is not intended to teach the reader how to write computer programs. For those who expect to do so, there are whole libraries of useful sources. Three very useful works are Kennedy and Gentle (1980), Abramovitz and Stegun (1971), and especially Press et al. (2007). The third of these provides a wealth of expertly written programs and a large amount of information about how to do computation efficiently and accurately. A recent survey of many areas of computation is Judd (1998).