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

$$H_0: \mu \leq 1.5,$$

$$H_1: \mu > 1.5.$$

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

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

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

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

[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 through 23.

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: *$E[y_n] = \text{Prob}[y_n < \delta]E[y_n \mid y_n < \delta] + \text{Prob}[y_n \geq \delta]E[y_n \mid y_n \geq \delta]$. Because y_n is nonnegative, both terms must be nonnegative, so $E[y_n] \geq \text{Prob}[y_n \geq \delta]E[y_n \mid y_n \geq \delta]$. Because $E[y_n \mid y_n \geq \delta]$ 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.

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 $x_{(1)}$,

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

and

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

Therefore,

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

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

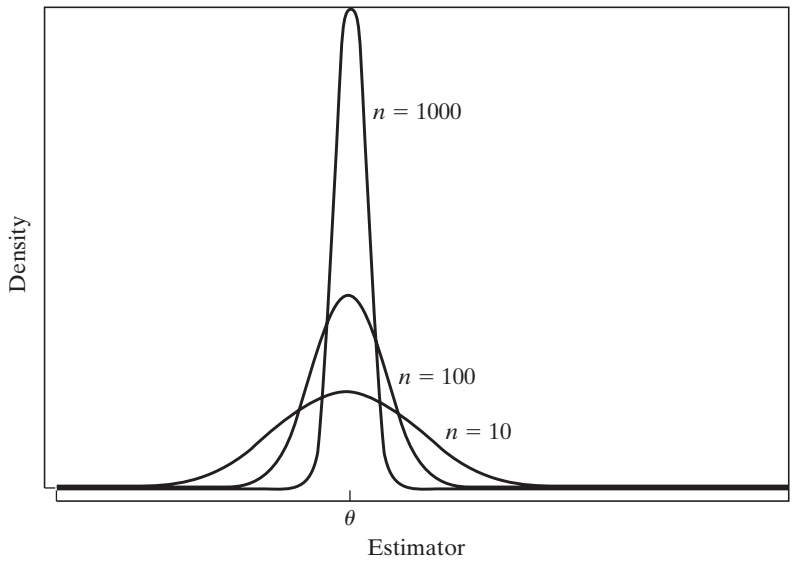


FIGURE D.1 Quadratic Convergence to a Constant, θ.

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

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

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 $\bar{\sigma}_n^2/n = (1/n^2)\Sigma_i\sigma_i^2 \to 0$ as $n \to \infty$, 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 $n \to \infty$ 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 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

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

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 $\text{Var}[x_i] = \sigma_i^2 < \infty$ such that $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$ as $n \to \infty$ 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 $\delta > 0$, $\sum_{i=1}^{\infty} E[|z_i - \mu_i|^{1+\delta}]/i^{1+\delta} < \infty$, 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:

DEFINITION D.4 Convergence in *r*th 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 $E[|x_n - c|^s] = E[(|x_n - c|^r)^{s/r}] \leq \left\{E[(|x_n - c|^r)]\right\}^{s/r}$ and the inner term converges to zero so the full function must also.

[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 $E[|z|] = \text{Prob}[z < 0]E[-z \mid z < 0] + \text{Prob}[z \geq 0]E[z \mid z \geq 0] = 0.7979\sigma$. (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.

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 *r*th 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 rth 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,

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

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 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 $g\big(E[x_n]\big) \geq E[g(x_n)]$.*

Cauchy–Schwarz Inequality. *For two random variables,*

$$E[|xy|] \leq \big\{E[x^2]\big\}^{1/2}\big\{E[y^2]\big\}^{1/2}.$$

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 rth 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*

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

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

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

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 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 $\lim_{n\to\infty} \text{Prob}(|x_i - x| > \varepsilon \text{ for all } i \geq n) = 0$ *for all* $\varepsilon > 0$.

DEFINITION D.8 Convergence in *r*th 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:

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

where

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

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* $\operatorname{plim} y_n = c$, *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*

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

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

THEOREM D.16 (Continued)

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* plim $(x_n - y_n) = 0$, *then x_n has the same limiting distribution as y_n.*

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 $\mathbf{t}_{1,n}$ and $\mathbf{t}_{2,n}$ 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, $\mathbf{t}_{j,n} \xrightarrow{d} \mathbf{z}_j$, a vector of independent standard normally distributed variables. The results so far show that $\frac{(\mathbf{t}_{1,n}'\mathbf{t}_{1,n})/K}{(\mathbf{t}_{2,n}'\mathbf{t}_{2,n})/M} \xrightarrow{d} F[K, M]$.

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.2. 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

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

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

$$\sqrt{n}(\bar{x}_n - \bar{\mu}_n) \xrightarrow{d} N[0, \bar{\sigma}^2].$$

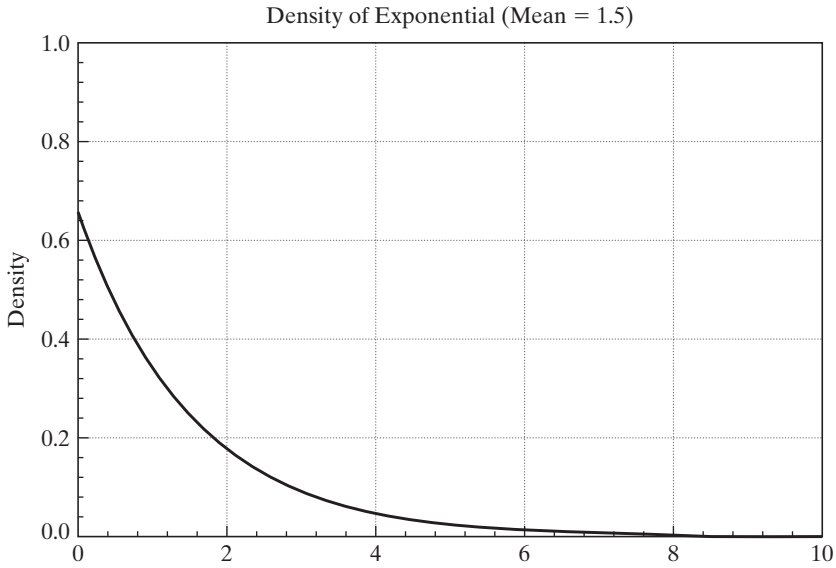


FIGURE D.2 The Exponential Distribution.

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.

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.

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 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.

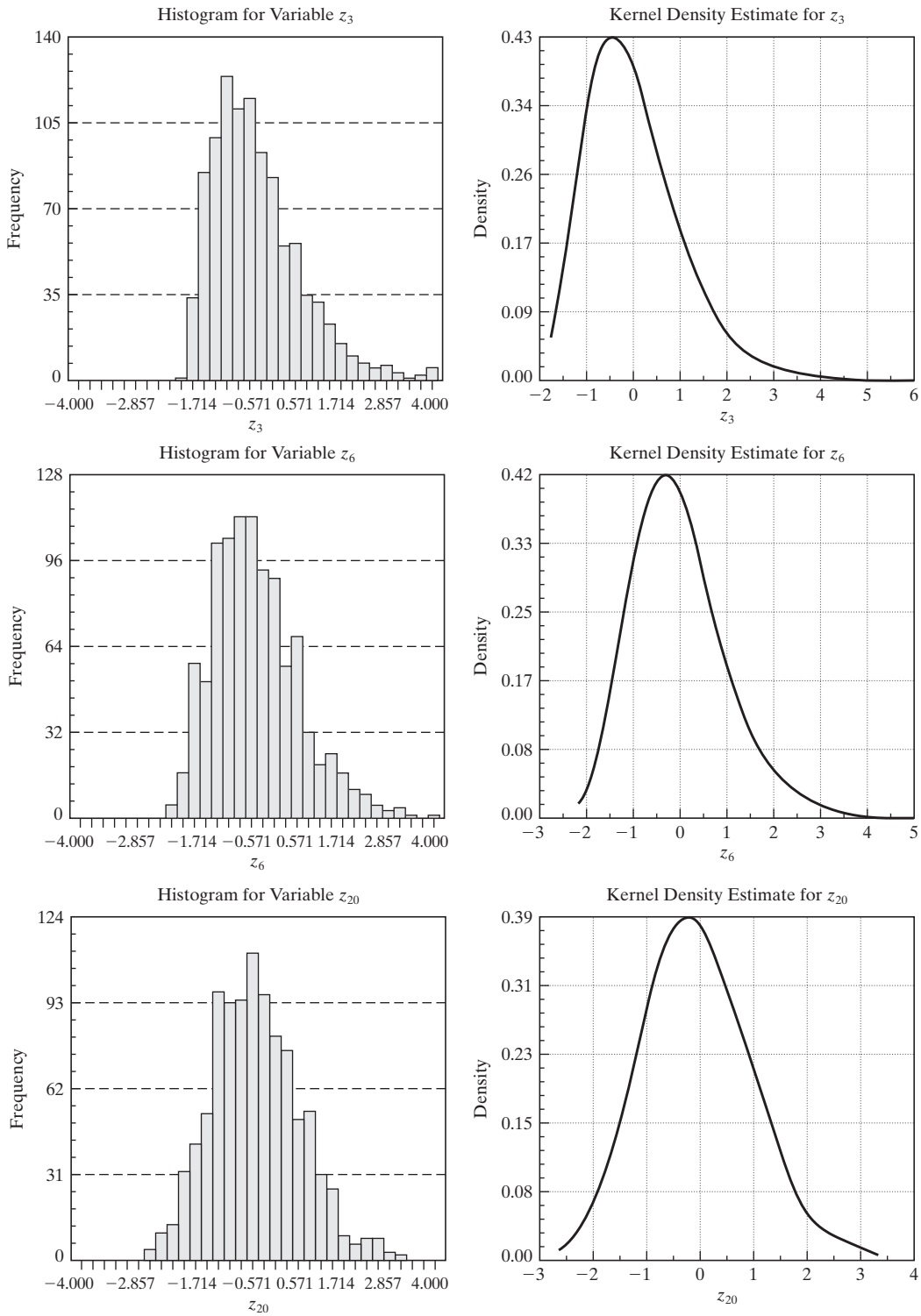


FIGURE D.3 The Central Limit Theorem.

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

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a random sample from a multivariate distribution with finite mean vector $\boldsymbol{\mu}$ and finite positive definite covariance matrix $\mathbf{Q}$, then

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

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *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*

$$\bar{\boldsymbol{\mu}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}_i,$$

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

We assume that

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

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

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

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

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

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*

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

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

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

[5] We depart somewhat from some other treatments [e.g., 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, $\mathbf{t}$ by first obtaining the *limiting* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$. By our construction, the *limiting* distribution of $\mathbf{t}$ is degenerate, whereas the *asymptotic* distribution of $\sqrt{n}(\mathbf{t} - \boldsymbol{\theta})$ is not useful.

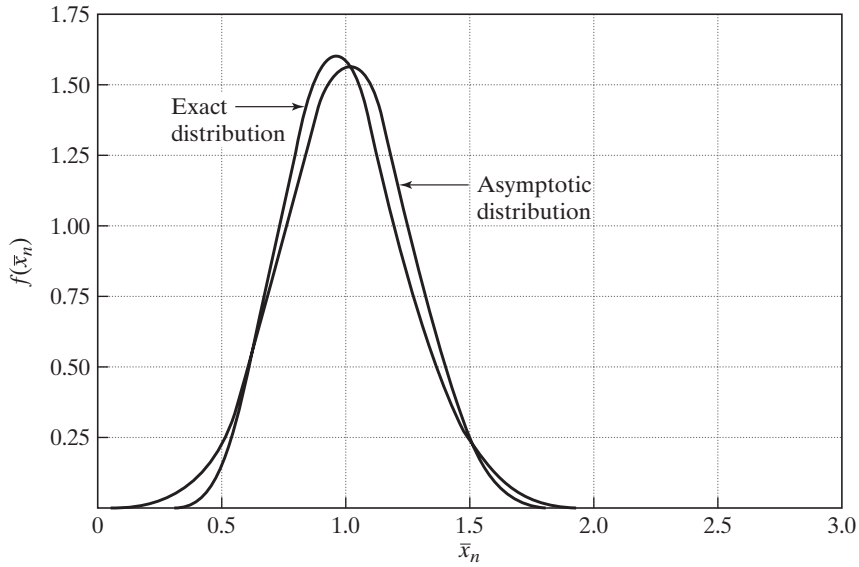


FIGURE D.4 True Versus Asymptotic Distribution.

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

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

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 $n \to \infty$. We assume that it holds approximately for finite n, which leads to (D-21).

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

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

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.
2. $\lim_{n\to\infty} E[\hat{\theta}_n] = \theta$. **(D-23)**
3. 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]

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

[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).

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 Sample Variance

The exact expected value and variance of the variance estimator

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

are

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

and

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

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

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

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 $n \to \infty$, 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 $\text{plim}(1/n^\delta)c_n$ *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 $\text{plim}(1/n^\delta)c_n$ *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 $O(n)$(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

[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)], 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).

[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. (1986). 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).

then 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.

E.2.1 COMPUTING INTEGRALS

One advantage of computers is their ability rapidly to compute approximations to complex functions such as logs and exponents. The basic functions, such as these, trigonometric functions, and so forth, are standard parts of the libraries of programs that accompany all scientific computing installations.[3] But one of the very common applications that often requires some high-level creativity by econometricians is the evaluation of integrals that do not have simple closed forms and that do not typically exist in "system libraries." We will consider several of these in this section. We will not go into detail on the nuts and bolts of how to compute integrals with a computer; rather, we will turn directly to the most common applications in econometrics.

E.2.2 THE STANDARD NORMAL CUMULATIVE DISTRIBUTION FUNCTION

The standard normal cumulative distribution function (cdf) is ubiquitous in econometric models. Yet this most homely of applications must be computed by approximation. There are a number of ways to do so.[4] Recall that what we desire is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt, \quad \text{where } \phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

One way to proceed is to use a Taylor series:

$$\Phi(x) \approx \sum_{i=0}^{M} \frac{1}{i!} \frac{d^i \Phi(x_0)}{dx_0^i} (x - x_0)^i.$$

The normal cdf has some advantages for this approach. First, the derivatives are simple and not integrals. Second, the function is **analytic;** as $M \longrightarrow \infty$, the approximation converges to the true value. Third, the derivatives have a simple form; they are the **Hermite polynomials** and they can be computed by a simple recursion. The 0th term in the preceding expansion is $\Phi(x)$ evaluated at the expansion point. The first derivative of the cdf is the pdf, so the terms from 2 onward are the derivatives of $\phi(x)$, once again evaluated at x_0. The derivatives of the standard normal pdf obey the recursion

$$\phi^i/\phi(x) = -x\phi^{i-1}/\phi(x) - (i-1)\phi^{i-2}/\phi(x),$$

where ϕ^i is $d^i\phi(x)/dx^i$. The zero and one terms in the sequence are one and $-x$. The next term is $x^2 - 1$, followed by $3x - x^3$ and $x^4 - 6x^2 + 3$, and so on. The approximation can be made

[3] Of course, at some level, these must have been programmed as approximations by someone.

[4] Many system libraries provide a related function, the *error function,* $\text{erf}(x) = (2/\sqrt{\pi}) \int_0^x e^{-t^2} dt$. If this is available, then the normal cdf can be obtained from $\Phi(x) = \frac{1}{2} + \frac{1}{2}\text{erf}(x/\sqrt{2})$, $x \geq 0$ and $\Phi(x) = 1 - \Phi(-x)$, $x \leq 0$.

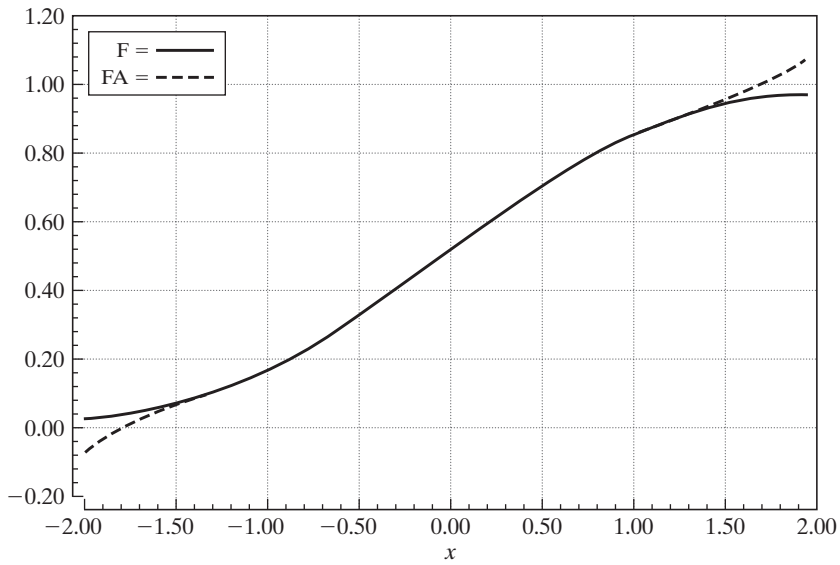


FIGURE E.1 Approximation to Normal cdf.

more accurate by adding terms. Consider using a fifth-order Taylor series approximation around the point $x = 0$, where $\Phi(0) = 0.5$ and $\phi(0) = 0.3989423$. Evaluating the derivatives at zero and assembling the terms produces the approximation

$$\Phi(x) \approx \tfrac{1}{2} + 0.3989423[x - x^3/6 + x^5/40].$$

[Some of the terms (every other one, in fact) will conveniently drop out.] Figure E.1 shows the actual values (F) and approximate values (FA) over the range -2 to 2. The figure shows two important points. First, the approximation is remarkably good over most of the range. Second, as is usually true for Taylor series approximations, the quality of the approximation deteriorates as one gets far from the expansion point.

Unfortunately, it is the tail areas of the standard normal distribution that are usually of interest, so the preceding is likely to be problematic. An alternative approach that is used much more often is a polynomial approximation reported by Abramovitz and Stegun (1971, p. 932):

$$\Phi(-|x|) = \phi(x)\sum_{i=1}^{5} a_i t^i + \varepsilon(x), \quad \text{where } t = 1/[1 + a_0|x|].$$

(The complement is taken if x is positive.) The error of approximation is less than $\pm 7.5 \times 10^{-8}$ for all x. (Note that the error exceeds the function value at $|x| > 5.7$, so this is the operational limit of this approximation.)

E.2.3 THE GAMMA AND RELATED FUNCTIONS

The standard normal cdf is probably the most common application of numerical integration of a function in econometrics. Another very common application is the class of gamma functions. For

positive constant P, the gamma function is

$$\Gamma(P) = \int_0^\infty t^{P-1}e^{-t}dt.$$

The gamma function obeys the recursion $\Gamma(P) = (P-1)\Gamma(P-1)$, so for integer values of P, $\Gamma(P) = (P-1)!$ This result suggests that the gamma function can be viewed as a generalization of the factorial function for noninteger values. Another convenient value is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. By making a change of variable, it can be shown that for positive constants a, c, and P,

$$\int_0^\infty t^{P-1}e^{-at^c}\,dt = \int_0^\infty t^{-(P+1)}e^{-a/t^c}\,dt = \left(\frac{1}{c}\right)a^{-P/c}\Gamma\left(\frac{P}{c}\right). \quad \textbf{(E-1)}$$

As a generalization of the factorial function, the gamma function will usually overflow for the sorts of values of P that normally appear in applications. The log of the function should normally be used instead. The function $\ln\Gamma(P)$ can be approximated remarkably accurately with only a handful of terms and is very easy to program. A number of approximations appear in the literature; they are generally modifications of **Stirling's approximation** to the factorial function $P! \approx (2\pi P)^{1/2}P^Pe^{-P}$, so

$$\ln\Gamma(P) \approx (P-0.5)\ln P - P + 0.5\ln(2\pi) + C + \varepsilon(P),$$

where C is the correction term [see, e.g., Abramovitz and Stegun (1971, p. 257), Press et al. (1986, p. 157), or Rao (1973, p. 59)] and $\varepsilon(P)$ is the approximation error.[5]

The derivatives of the gamma function are

$$\frac{d^r\Gamma(P)}{dP^r} = \int_0^\infty (\ln t)^r t^{P-1}e^{-t}\,dt.$$

The first two derivatives of $\ln\Gamma(P)$ are denoted $\Psi(P) = \Gamma'/\Gamma$ and $\Psi'(P) = (\Gamma\Gamma'' - \Gamma'^2)/\Gamma^2$ and are known as the **digamma** and **trigamma** functions.[6] The **beta function,** denoted $\beta(a, b)$,

$$\beta(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

is related.

E.2.4 APPROXIMATING INTEGRALS BY QUADRATURE

The digamma and trigamma functions, and the gamma function for noninteger values of P and values that are not integers plus $\frac{1}{2}$, do not exist in closed form and must be approximated. Most other applications will also involve integrals for which no simple computing function exists. The simplest approach to approximating

$$F(x) = \int_{L(x)}^{U(x)} f(t)\,dt$$

[5]For example, one widely used formula is $C = z^{-1}/12 - z^{-3}/360 - z^{-5}/1260 + z^{-7}/1680 - q$, where $z = P$ and $q = 0$ if $P > 18$, or $z = P + J$ and $q = \ln[P(P+1)(P+2)\cdots(P+J-1)]$, where $J = 18 - \text{INT}(P)$, if not. Note, in the approximation, we write $\Gamma(P) = (P!)/P +$ a correction.

[6]Tables of specific values for the gamma, digamma, and trigamma functions appear in Abramovitz and Stegun (1971). Most contemporary econometric programs have built-in functions for these common integrals, so the tables are not generally needed.

is likely to be a variant of Simpson's rule, or the trapezoid rule. For example, one approximation [see Press et al. (1986, p. 108)] is

$$F(x) \approx \Delta\left[\tfrac{1}{3}f_1 + \tfrac{4}{3}f_2 + \tfrac{2}{3}f_3 + \tfrac{4}{3}f_4 + \cdots + \tfrac{2}{3}f_{N-2} + \tfrac{4}{3}f_{N-1} + \tfrac{1}{3}f_N\right],$$

where f_j is the function evaluated at N equally spaced points in $[L, U]$ including the endpoints and $\Delta = (L-U)/(N-1)$. There are a number of problems with this method, most notably that it is difficult to obtain satisfactory accuracy with a moderate number of points.

Gaussian quadrature is a popular method of computing integrals. The general approach is to use an approximation of the form

$$\int_L^U W(x)\,f(x)\,dx \approx \sum_{j=1}^{M} w_j\,f(a_j),$$

where $W(x)$ is viewed as a "weighting" function for integrating $f(x)$, w_j is the **quadrature weight,** and a_j is the **quadrature abscissa.** Different weights and abscissas have been derived for several weighting functions. Two weighting functions common in econometrics are

$$W(x) = x^c e^{-x}, \quad x \in [0, \infty),$$

for which the computation is called **Gauss–Laguerre quadrature,** and

$$W(x) = e^{-x^2}, \quad x \in (-\infty, \infty),$$

for which the computation is called **Gauss–Hermite quadrature.** The theory for deriving weights and abscissas is given in Press et al. (1986, pp. 121–125). Tables of weights and abscissas for many values of M are given by Abramovitz and Stegun (1971). Applications of the technique appear in Chapters 14 and 17.

E.3 OPTIMIZATION

Nonlinear optimization (e.g., maximizing log-likelihood functions) is an intriguing practical problem. Theory provides few hard and fast rules, and there are relatively few cases in which it is obvious how to proceed. This section introduces some of the terminology and underlying theory of nonlinear optimization.[7] We begin with a general discussion on how to search for a solution to a nonlinear optimization problem and describe some specific commonly used methods. We then consider some practical problems that arise in optimization. An example is given in the final section.

Consider maximizing the quadratic function

$$F(\boldsymbol{\theta}) = a + \mathbf{b}'\boldsymbol{\theta} - \tfrac{1}{2}\boldsymbol{\theta}'\mathbf{C}\boldsymbol{\theta},$$

where $\mathbf{C}$ is a positive definite matrix. The first-order condition for a maximum is

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{b} - \mathbf{C}\boldsymbol{\theta} = \mathbf{0}. \tag{E-2}$$

This set of *linear* equations has the unique solution

$$\boldsymbol{\theta} = \mathbf{C}^{-1}\mathbf{b}. \tag{E-3}$$

[7]There are numerous excellent references that offer a more complete exposition. Among these are Quandt (1983), Bazaraa and Shetty (1979), Fletcher (1980), and Judd (1998).

This is a linear optimization problem. Note that it has a **closed-form solution;** for any a, $\mathbf{b}$, and $\mathbf{C}$, the solution can be computed directly.[8] In the more typical situation,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{E-4}$$

is a set of nonlinear equations that cannot be solved explicitly for $\boldsymbol{\theta}$.[9] The techniques considered in this section provide systematic means of searching for a solution.

We now consider the general problem of maximizing a function of several variables:

$$\text{maximize}_{\theta} F(\boldsymbol{\theta}), \tag{E-5}$$

where $F(\boldsymbol{\theta})$ may be a log-likelihood or some other function. Minimization of $F(\boldsymbol{\theta})$ is handled by maximizing $-F(\boldsymbol{\theta})$. Two special cases are

$$F(\boldsymbol{\theta}) = \sum_{i=1}^{n} f_i(\boldsymbol{\theta}), \tag{E-6}$$

which is typical for maximum likelihood problems, and the **least squares problem,**[10]

$$f_i(\boldsymbol{\theta}) = -(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2. \tag{E-7}$$

We treated the nonlinear least squares problem in detail in Chapter 7. An obvious way to search for the $\boldsymbol{\theta}$ that maximizes $F(\boldsymbol{\theta})$ is by trial and error. If $\boldsymbol{\theta}$ has only a single element and it is known approximately where the optimum will be found, then a **grid search** will be a feasible strategy. An example is a common time-series problem in which a one-dimensional search for a correlation coefficient is made in the interval $(-1, 1)$. The grid search can proceed in the obvious fashion—that is, $\ldots, -0.1, 0, 0.1, 0.2, \ldots$, then $\hat{\theta}_{\max} - 0.1$ to $\hat{\theta}_{\max} + 0.1$ in increments of 0.01, and so on—until the desired precision is achieved.[11] If $\boldsymbol{\theta}$ contains more than one parameter, then a grid search is likely to be extremely costly, particularly if little is known about the parameter vector at the outset. Nonetheless, relatively efficient methods have been devised. Quandt (1983) and Fletcher (1980) contain further details.

There are also systematic, derivative-free methods of searching for a function optimum that resemble in some respects the algorithms that we will examine in the next section. The **downhill simplex** (and other simplex) methods[12] have been found to be very fast and effective for some problems. A recent entry in the econometrics literature is the method of **simulated annealing.**[13] These derivative-free methods, particularly the latter, are often very effective in problems with many variables in the objective function, but they usually require far more function evaluations than the methods based on derivatives that are considered below. Because the problems typically analyzed in econometrics involve relatively few parameters but often quite complex functions involving large numbers of terms in a summation, on balance, the gradient methods are usually going to be preferable.[14]

[8] Notice that the constant a is irrelevant to the solution. Many maximum likelihood problems are presented with the preface "neglecting an irrelevant constant." For example, the log-likelihood for the normal linear regression model contains a term $-(n/2)\ln(2\pi)$—that can be discarded.

[9] See, for example, the normal equations for the nonlinear least squares estimators of Chapter 7.

[10] Least squares is, of course, a minimization problem. The negative of the criterion is used to maintain consistency with the general formulation.

[11] There are more efficient methods of carrying out a one-dimensional search, for example, the **golden section** method. See Press et al. (1986, Chap. 10).

[12] See Nelder and Mead (1965) and Press et al. (1986).

[13] See Goffe, Ferrier, and Rodgers (1994) and Press et al. (1986, pp. 326–334).

[14] Goffe, Ferrier, and Rodgers (1994) did find that the method of simulated annealing was quite adept at finding the best among multiple solutions. This problem is common for derivative-based methods, because they usually have no method of distinguishing between a local optimum and a global one.

E.3.1 ALGORITHMS

A more effective means of solving most nonlinear maximization problems is by an **iterative algorithm:**

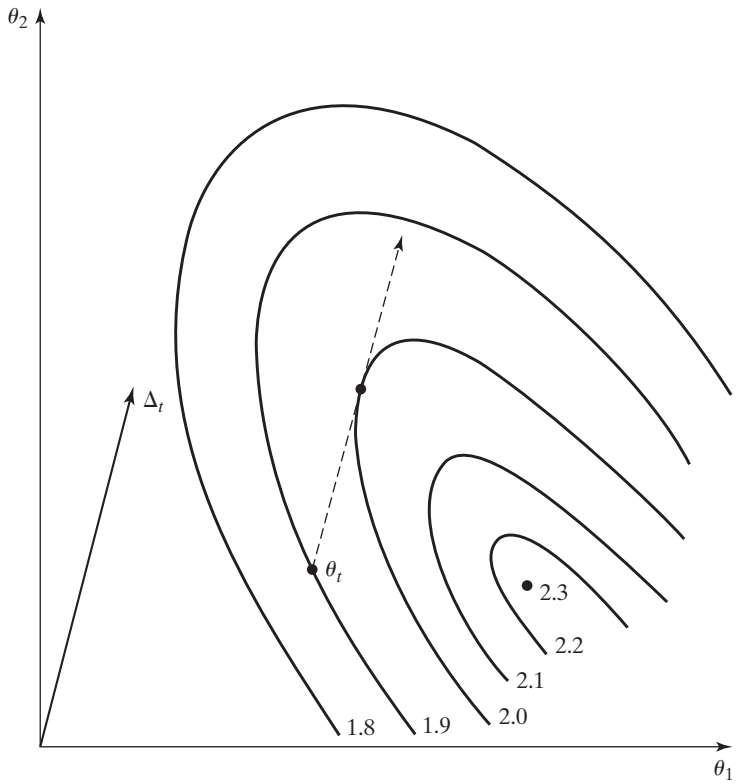
> Beginning from initial value $\boldsymbol{\theta}_0$, at entry to iteration t, if $\boldsymbol{\theta}_t$ is not the optimal value for $\boldsymbol{\theta}$, compute direction vector $\boldsymbol{\Delta}_t$, step size λ_t, then
>
> $$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t. \qquad \textbf{(E-8)}$$

Figure E.2 illustrates the structure of an iteration for a hypothetical function of two variables. The direction vector $\boldsymbol{\Delta}_t$ is shown in the figure with $\boldsymbol{\theta}_t$. The dashed line is the set of points $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$. Different values of λ_t lead to different contours; for this $\boldsymbol{\theta}_t$ and $\boldsymbol{\Delta}_t$, the best value of λ_t is about 0.5.

Notice in Figure E.2 that for a given direction vector $\boldsymbol{\Delta}_t$ and current parameter vector $\boldsymbol{\theta}_t$, a secondary optimization is required to find the best λ_t. Translating from Figure E.2, we obtain the form of this problem as shown in Figure E.3. This subsidiary search is called a **line search,** as we search along the line $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$ for the optimal value of $F(.)$. The formal solution to the line search problem would be the λ_t that satisfies

$$\frac{\partial F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)}{\partial \lambda_t} = \mathbf{g}(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)' \boldsymbol{\Delta}_t = 0, \qquad \textbf{(E-9)}$$

FIGURE E.2 Iteration.



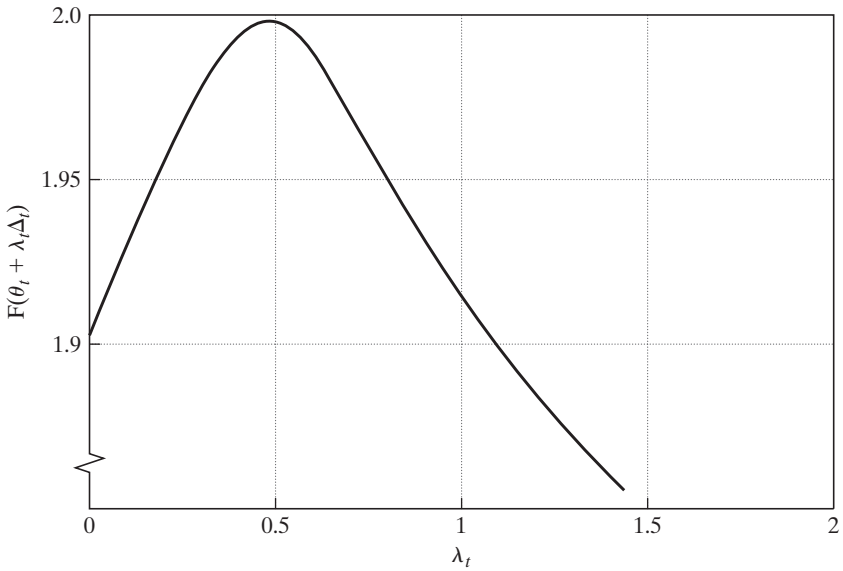


FIGURE E.3 Line Search.

where **g** is the vector of partial derivatives of $F(.)$ evaluated at $\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t$. In general, this problem will also be a nonlinear one. In most cases, adding a formal search for λ_t will be too expensive, as well as unnecessary. Some approximate or ad hoc method will usually be chosen. It is worth emphasizing that finding the λ_t that maximizes $F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)$ at a given iteration does not generally lead to the overall solution in that iteration. This situation is clear in Figure E.3, where the optimal value of λ_t leads to $F(.) = 2.0$, at which point we reenter the iteration.

E.3.2 COMPUTING DERIVATIVES

For certain functions, the programming of derivatives may be quite difficult. Numeric approximations can be used, although it should be borne in mind that analytic derivatives obtained by formally differentiating the functions involved are to be preferred. First derivatives can be approximated by using

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \theta_i} \approx \frac{F(\cdots \theta_i + \varepsilon \cdots) - F(\cdots \theta_i - \varepsilon \cdots)}{2\varepsilon}.$$

The choice of ε is a remaining problem. Extensive discussion may be found in Quandt (1983).

There are three drawbacks to this means of computing derivatives compared with using the analytic derivatives. A possible major consideration is that it may substantially increase the amount of computation needed to obtain a function and its gradient. In particular, $K+1$ function evaluations (the criterion and K derivatives) are replaced with $2K+1$ functions. The latter may be more burdensome than the former, depending on the complexity of the partial derivatives compared with the function itself. The comparison will depend on the application. But in most settings, careful programming that avoids superfluous or redundant calculation can make the advantage of the analytic derivatives substantial. Second, the choice of ε can be problematic. If it is chosen too large, then the approximation will be inaccurate. If it is chosen too small, then there may be insufficient variation in the function to produce a good estimate of the derivative.

A compromise that is likely to be effective is to compute ε_i separately for each parameter, as in

$$\varepsilon_i = \text{Max}[\alpha|\theta_i|, \gamma]$$

[see Goldfeld and Quandt (1971)]. The values α and γ should be relatively small, such as 10^{-5}. Third, although numeric derivatives computed in this fashion are likely to be reasonably accurate, in a sum of a large number of terms, say, several thousand, enough approximation error can accumulate to cause the numerical derivatives to differ significantly from their analytic counterparts. Second derivatives can also be computed numerically. In addition to the preceding problems, however, it is generally not possible to ensure negative definiteness of a Hessian computed in this manner. Unless the choice of ε is made extremely carefully, an indefinite matrix is a possibility. In general, the use of numeric derivatives should be avoided if the analytic derivatives are available.

E.3.3 GRADIENT METHODS

The most commonly used algorithms are **gradient methods,** in which

$$\boldsymbol{\Delta}_t = \mathbf{W}_t \mathbf{g}_t, \tag{E-10}$$

where $\mathbf{W}_t$ is a positive definite matrix and $\mathbf{g}_t$ is the **gradient** of $F(\boldsymbol{\theta}_t)$:

$$\mathbf{g}_t = \mathbf{g}(\boldsymbol{\theta}_t) = \frac{\partial F(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}. \tag{E-11}$$

These methods are motivated partly by the following. Consider a linear Taylor series approximation to $F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)$ around $\lambda_t = 0$:

$$F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t) \simeq F(\boldsymbol{\theta}_t) + \lambda_t \mathbf{g}(\boldsymbol{\theta}_t)' \boldsymbol{\Delta}_t. \tag{E-12}$$

Let $F(\boldsymbol{\theta}_t + \lambda_t \boldsymbol{\Delta}_t)$ equal F_{t+1}. Then,

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \boldsymbol{\Delta}_t.$$

If $\boldsymbol{\Delta}_t = \mathbf{W}_t \mathbf{g}_t$, then

$$F_{t+1} - F_t \simeq \lambda_t \mathbf{g}_t' \mathbf{W}_t \mathbf{g}_t.$$

If $\mathbf{g}_t$ is not $\mathbf{0}$ and λ_t is small enough, then $F_{t+1} - F_t$ must be positive. Thus, if $F(\boldsymbol{\theta})$ is not already at its maximum, then we can always find a step size such that a gradient-type iteration will lead to an increase in the function. (Recall that $\mathbf{W}_t$ is assumed to be positive definite.)

In the following, we will omit the iteration index t, except where it is necessary to distinguish one vector from another. The following are some commonly used algorithms.[15]

Steepest Ascent The simplest algorithm to employ is the **steepest ascent** method, which uses

$$\mathbf{W} = \mathbf{I} \text{ so that } \boldsymbol{\Delta} = \mathbf{g}. \tag{E-13}$$

As its name implies, the direction is the one of greatest increase of $F(.)$. Another virtue is that the line search has a straightforward solution; at least near the maximum, the optimal λ is

$$\lambda = \frac{-\mathbf{g}'\mathbf{g}}{\mathbf{g}'\mathbf{H}\mathbf{g}}, \tag{E-14}$$

[15] A more extensive catalog may be found in Judge et al. (1985, Appendix B). Those mentioned here are some of the more commonly used ones and are chosen primarily because they illustrate many of the important aspects of nonlinear optimization.

where

$$\mathbf{H} = \frac{\partial^2 F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}.$$

Therefore, the steepest ascent iteration is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\mathbf{g}_t'\mathbf{g}_t}{\mathbf{g}_t'\mathbf{H}_t\mathbf{g}_t}\mathbf{g}_t. \tag{E-15}$$

Computation of the second derivatives matrix may be extremely burdensome. Also, if $\mathbf{H}_t$ is not negative definite, which is likely if $\boldsymbol{\theta}_t$ is far from the maximum, the iteration may diverge. A systematic line search can bypass this problem. This algorithm usually converges very slowly, however, so other techniques are usually used.

Newton's Method The template for most gradient methods in common use is Newton's method. The basis for **Newton's method** is a linear Taylor series approximation. Expanding the first-order conditions,

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0},$$

equation by equation, in a linear Taylor series around an arbitrary $\boldsymbol{\theta}^0$ yields

$$\frac{\partial F(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \simeq \mathbf{g}^0 + \mathbf{H}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) = \mathbf{0}, \tag{E-16}$$

where the superscript indicates that the term is evaluated at $\boldsymbol{\theta}^0$. Solving for $\boldsymbol{\theta}$ and then equating $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{t+1}$ and $\boldsymbol{\theta}^0$ to $\boldsymbol{\theta}_t$, we obtain the iteration

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1}\mathbf{g}_t. \tag{E-17}$$

Thus, for Newton's method,

$$\mathbf{W} = -\mathbf{H}^{-1}, \qquad \boldsymbol{\Delta} = -\mathbf{H}^{-1}\mathbf{g}, \qquad \lambda = 1. \tag{E-18}$$

Newton's method will converge very rapidly in many problems. If the function is quadratic, then this method will reach the optimum in one iteration from any starting point. If the criterion function is globally concave, as it is in a number of problems that we shall examine in this text, then it is probably the best algorithm available. This method is very well suited to maximum likelihood estimation.

Alternatives to Newton's Method Newton's method is very effective in some settings, but it can perform very poorly in others. If the function is not approximately quadratic or if the current estimate is very far from the maximum, then it can cause wide swings in the estimates and even fail to converge at all. A number of algorithms have been devised to improve upon Newton's method. An obvious one is to include a line search at each iteration rather than use $\lambda = 1$. Two problems remain, however. At points distant from the optimum, the second derivatives matrix may not be negative definite, and, in any event, the computational burden of computing $\mathbf{H}$ may be excessive.

The **quadratic hill-climbing method** proposed by Goldfeld, Quandt, and Trotter (1966) deals directly with the first of these problems. In any iteration, if $\mathbf{H}$ is not negative definite, then it is replaced with

$$\mathbf{H}_\alpha = \mathbf{H} - \alpha\mathbf{I}, \tag{E-19}$$

where α is a positive number chosen large enough to ensure the negative definiteness of $\mathbf{H}_\alpha$. Another suggestion is that of Greenstadt (1967), which uses, at every iteration,

$$\mathbf{H}_\pi = -\sum_{i=1}^{n} |\pi_i|\, \mathbf{c}_i \mathbf{c}_i', \tag{E-20}$$

where π_i is the ith characteristic root of $\mathbf{H}$ and $\mathbf{c}_i$ is its associated characteristic vector. Other proposals have been made to ensure the negative definiteness of the required matrix at each iteration.[16]

Quasi-Newton Methods: Davidon–Fletcher–Powell A very effective class of algorithms has been developed that eliminates second derivatives altogether and has excellent convergence properties, even for ill-behaved problems. These are the **quasi-Newton methods,** which form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{E}_t,$$

where $\mathbf{E}_t$ is a positive definite matrix.[17] As long as $\mathbf{W}_0$ is positive definite—$\mathbf{I}$ is commonly used—$\mathbf{W}_t$ will be positive definite at every iteration. In the **Davidon–Fletcher–Powell (DFP) method,** after a sufficient number of iterations, $\mathbf{W}_{t+1}$ will be an approximation to $-\mathbf{H}^{-1}$. Let

$$\boldsymbol{\delta}_t = \lambda_t \boldsymbol{\Delta}_t, \quad \text{and} \quad \boldsymbol{\gamma}_t = \mathbf{g}(\boldsymbol{\theta}_{t+1}) - \mathbf{g}(\boldsymbol{\theta}_t). \tag{E-21}$$

The DFP **variable metric algorithm** uses

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\boldsymbol{\delta}_t \boldsymbol{\delta}_t'}{\boldsymbol{\delta}_t' \boldsymbol{\gamma}_t} + \frac{\mathbf{W}_t \boldsymbol{\gamma}_t \boldsymbol{\gamma}_t' \mathbf{W}_t}{\boldsymbol{\gamma}_t' \mathbf{W}_t \boldsymbol{\gamma}_t}. \tag{E-22}$$

Notice that in the DFP algorithm, the change in the first derivative vector is used in $\mathbf{W}$; an estimate of the inverse of the second derivatives matrix is being accumulated.

The variable metric algorithms are those that update $\mathbf{W}$ at each iteration while preserving its definiteness. For the DFP method, the accumulation of $\mathbf{W}_{t+1}$ is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{a}\mathbf{a}' + \mathbf{b}\mathbf{b}' = \mathbf{W}_t + [\mathbf{a} \quad \mathbf{b}][\mathbf{a} \quad \mathbf{b}]'.$$

The two-column matrix $[\mathbf{a} \quad \mathbf{b}]$ will have rank two; hence, DFP is called a **rank two update** or **rank two correction.** The **Broyden–Fletcher–Goldfarb–Shanno (BFGS)** method is a rank three correction that subtracts $v\mathbf{d}\mathbf{d}'$ from the **DFP** update, where $v = (\boldsymbol{\gamma}_t' \mathbf{W}_t \boldsymbol{\gamma}_t)$ and

$$\mathbf{d}_t = \left(\frac{1}{\boldsymbol{\delta}_t' \boldsymbol{\gamma}_t}\right)\boldsymbol{\delta}_t - \left(\frac{1}{\boldsymbol{\gamma}_t' \mathbf{W}_t \boldsymbol{\gamma}_t}\right)\mathbf{W}_t \boldsymbol{\gamma}_t.$$

There is some evidence that this method is more efficient than DFP. Other methods, such as **Broyden's method,** involve a rank one correction instead. Any method that is of the form

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{Q}\mathbf{Q}'$$

will preserve the definiteness of $\mathbf{W}$ regardless of the number of columns in $\mathbf{Q}$.

The DFP and BFGS algorithms are extremely effective and are among the most widely used of the gradient methods. An important practical consideration to keep in mind is that although $\mathbf{W}_t$ accumulates an estimate of the negative inverse of the second derivatives matrix for both algorithms, in maximum likelihood problems it rarely converges to a very good estimate of the covariance matrix of the estimator and should generally not be used as one.

[16] See, for example, Goldfeld and Quandt (1971).

[17] See Fletcher (1980).

E.3.4 ASPECTS OF MAXIMUM LIKELIHOOD ESTIMATION

Newton's method is often used for maximum likelihood problems. For solving a maximum likelihood problem, the **method of scoring** replaces $\mathbf{H}$ with

$$\bar{\mathbf{H}} = E[\mathbf{H}(\boldsymbol{\theta})], \tag{E-23}$$

which will be recognized as the asymptotic covariance of the maximum likelihood estimator. There is some evidence that where it can be used, this method performs better than Newton's method. The exact form of the expectation of the Hessian of the log likelihood is rarely known, however.[18] Newton's method, which uses actual instead of expected second derivatives, is generally used instead.

One-Step Estimation A convenient variant of Newton's method is the **one-step maximum likelihood estimator.** It has been shown that if $\boldsymbol{\theta}^0$ is *any* consistent initial estimator of $\boldsymbol{\theta}$ and $\mathbf{H}^*$ is $\mathbf{H}$, $\bar{\mathbf{H}}$, or any other asymptotically equivalent estimator of $\text{Var}[\mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MLE}})]$, then

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 - (\mathbf{H}^*)^{-1}\mathbf{g}^0 \tag{E-24}$$

is an estimator of $\boldsymbol{\theta}$ that has the same asymptotic properties as the maximum likelihood estimator.[19] (Note that it is *not* the maximum likelihood estimator. As such, for example, it should not be used as the basis for likelihood ratio tests.)

Covariance Matrix Estimation In computing maximum likelihood estimators, a commonly used method of estimating $\mathbf{H}$ simultaneously simplifies the calculation of $\mathbf{W}$ and solves the occasional problem of indefiniteness of the Hessian. The method of Berndt et al. (1974) replaces $\mathbf{W}$ with

$$\hat{\mathbf{W}} = \left[\sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i'\right]^{-1} = (\mathbf{G}'\mathbf{G})^{-1}, \tag{E-25}$$

where

$$\mathbf{g}_i = \frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{E-26}$$

Then, $\mathbf{G}$ is the $n \times K$ matrix with ith row equal to $\mathbf{g}_i'$. Although $\hat{\mathbf{W}}$ and other suggested estimators of $(-\mathbf{H})^{-1}$ are asymptotically equivalent, $\hat{\mathbf{W}}$ has the additional virtues that it is always nonnegative definite, and it is only necessary to differentiate the log-likelihood once to compute it.

The Lagrange Multiplier Statistic The use of $\hat{\mathbf{W}}$ as an estimator of $(-\mathbf{H})^{-1}$ brings another intriguing convenience in maximum likelihood estimation. When testing restrictions on parameters estimated by maximum likelihood, one approach is to use the **Lagrange multiplier** statistic. We will examine this test at length at various points in this book, so we need only sketch it briefly here. The logic of the LM test is as follows. The gradient $\mathbf{g}(\boldsymbol{\theta})$ of the log-likelihood function equals $\mathbf{0}$ at the unrestricted maximum likelihood estimators (that is, at least to within the precision of the computer program in use). If $\hat{\boldsymbol{\theta}}_r$ is an MLE that is computed subject to some restrictions on $\boldsymbol{\theta}$, then we know that $\mathbf{g}(\hat{\boldsymbol{\theta}}_r) \neq \mathbf{0}$. The LM test is used to test whether, at $\hat{\boldsymbol{\theta}}_r$, $\mathbf{g}_r$ is *significantly* different from $\mathbf{0}$ or whether the deviation of $\mathbf{g}_r$ from $\mathbf{0}$ can be viewed as sampling variation. The covariance matrix of the gradient of the log-likelihood is $-\mathbf{H}$, so the Wald statistic for testing this hypothesis is $W = \mathbf{g}'(-\mathbf{H})^{-1}\mathbf{g}$. Now, suppose that we use $\hat{\mathbf{W}}$ to estimate $-\mathbf{H}^{-1}$. Let $\mathbf{G}$ be the $n \times K$ matrix with ith row equal to $\mathbf{g}_i'$, and let $\mathbf{i}$ denote an $n \times 1$ column of ones. Then the LM statistic can be

[18] Amemiya (1981) provides a number of examples.

[19] See, for example, Rao (1973).

computed as

$$\text{LM} = \mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}.$$

Because $\mathbf{i}'\mathbf{i} = n$,

$$\text{LM} = n[\mathbf{i}'\mathbf{G}(\mathbf{G}'\mathbf{G})^{-1}\mathbf{G}'\mathbf{i}/n] = nR_i^2,$$

where R_i^2 is the *uncentered* R^2 in a regression of a column of ones on the derivatives of the log-likelihood function.

The Concentrated Log-Likelihood Many problems in maximum likelihood estimation can be formulated in terms of a partitioning of the parameter vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$ such that at the solution to the optimization problem, $\boldsymbol{\theta}_{2,\text{ML}}$, can be written as an explicit function of $\boldsymbol{\theta}_{1,\text{ML}}$. When the solution to the likelihood equation for $\boldsymbol{\theta}_2$ produces

$$\boldsymbol{\theta}_{2,\text{ML}} = \mathbf{t}(\boldsymbol{\theta}_{1,\text{ML}}),$$

then, if it is convenient, we may "concentrate" the log-likelihood function by writing

$$F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = F[\boldsymbol{\theta}_1, \mathbf{t}(\boldsymbol{\theta}_1)] = F_c(\boldsymbol{\theta}_1).$$

The unrestricted solution to the problem $\text{Max}_{\boldsymbol{\theta}_1} F_c(\boldsymbol{\theta}_1)$ provides the full solution to the optimization problem. Once the optimizing value of $\boldsymbol{\theta}_1$ is obtained, the optimizing value of $\boldsymbol{\theta}_2$ is simply $\mathbf{t}(\hat{\boldsymbol{\theta}}_{1,\text{ML}})$. Note that $F^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a subset of the set of values of the log-likelihood function, namely those values at which the second parameter vector satisfies the first-order conditions.[20]

E.3.5 OPTIMIZATION WITH CONSTRAINTS

Occasionally, some of or all the parameters of a model are constrained, for example, to be positive in the case of a variance or to be in a certain range, such as a correlation coefficient. Optimization subject to constraints is often yet another art form. The elaborate literature on the general problem provides some guidance—see, for example, Appendix B in Judge et al. (1985)—but applications still, as often as not, require some creativity on the part of the analyst. In this section, we will examine a few of the most common forms of constrained optimization as they arise in econometrics.

Parametric constraints typically come in two forms, which may occur simultaneously in a problem. Equality constraints can be written $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$, where $c_j(\boldsymbol{\theta})$ is a continuous and differentiable function. Typical applications include linear constraints on slope vectors, such as a requirement that a set of elasticities in a log-linear model add to one; exclusion restrictions, which are often cast in the form of interesting hypotheses about whether or not a variable should appear in a model (i.e., whether a coefficient is zero or not); and equality restrictions, such as the symmetry restrictions in a translog model, which require that parameters in two different equations be equal to each other. Inequality constraints, in general, will be of the form $a_j \leq c_j(\boldsymbol{\theta}) \leq b_j$, where a_j and b_j are known constants (either of which may be infinite). Once again, the typical application in econometrics involves a restriction on a single parameter, such as $\sigma > 0$ for a variance parameter, $-1 \leq \rho \leq 1$ for a correlation coefficient, or $\beta_j \geq 0$ for a particular slope coefficient in a model. We will consider the two cases separately.

In the case of equality constraints, for practical purposes of optimization, there are usually two strategies available. One can use a Lagrangean multiplier approach. The new optimization problem is

$$\text{Max}_{\boldsymbol{\theta},\boldsymbol{\lambda}} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = F(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{c}(\boldsymbol{\theta}).$$

[20] A formal proof that this is a valid way to proceed is given by Amemiya (1985, pp. 125–127).

The necessary conditions for an optimum are

$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} = \mathbf{g}(\boldsymbol{\theta}) + \mathbf{C}(\boldsymbol{\theta})'\boldsymbol{\lambda} = \mathbf{0},$$
$$\frac{\partial L(\boldsymbol{\theta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0},$$

where $\mathbf{g}(\boldsymbol{\theta})$ is the familiar gradient of $F(\boldsymbol{\theta})$ and $\mathbf{C}(\boldsymbol{\theta})$ is a $J \times K$ matrix of derivatives with jth row equal to $\partial c_j/\partial \boldsymbol{\theta}'$. The joint solution will provide the constrained optimizer, as well as the Lagrange multipliers, which are often interesting in their own right. The disadvantage of this approach is that it increases the dimensionality of the optimization problem. An alternative strategy is to eliminate some of the parameters by either imposing the constraints directly on the function or by solving out the constraints. For exclusion restrictions, which are usually of the form $\theta_j = 0$, this step usually means dropping a variable from a model. Other restrictions can often be imposed just by building them into the model. For example, in a function of θ_1, θ_2, and θ_3, if the restriction is of the form $\theta_3 = \theta_1\theta_2$, then θ_3 can be eliminated from the model by a direct substitution.

Inequality constraints are more difficult. For the general case, one suggestion is to transform the constrained problem into an unconstrained one by imposing some sort of penalty function into the optimization criterion that will cause a parameter vector that violates the constraints, or nearly does so, to be an unattractive choice. For example, to force a parameter θ_j to be nonzero, one might maximize the augmented function $F(\boldsymbol{\theta}) - |1/\theta_j|$. This approach is feasible, but it has the disadvantage that because the penalty is a function of the parameters, different penalty functions will lead to different solutions of the optimization problem. For the most common problems in econometrics, a simpler approach will usually suffice. One can often reparameterize a function so that the new parameter is unconstrained. For example, the "method of squaring" is sometimes used to force a parameter to be positive. If we require θ_j to be positive, then we can define $\theta_j = \alpha^2$ and substitute α^2 for θ_j wherever it appears in the model. Then an unconstrained solution for α is obtained. An alternative reparameterization for a parameter that must be positive that is often used is $\theta_j = \exp(\alpha)$. To force a parameter to be between zero and one, we can use the function $\theta_j = 1/[1 + \exp(\alpha)]$. The range of α is now unrestricted. Experience suggests that a third, less orthodox approach works very well for many problems. When the constrained optimization is begun, there is a starting value $\boldsymbol{\theta}^0$ that begins the iterations. Presumably, $\boldsymbol{\theta}^0$ obeys the restrictions. (If not, and none can be found, then the optimization process must be terminated immediately.) The next iterate, $\boldsymbol{\theta}^1$, is a step away from $\boldsymbol{\theta}^0$, by $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \lambda_0\boldsymbol{\delta}^0$. Suppose that $\boldsymbol{\theta}^1$ violates the constraints. By construction, we know that there is some value $\boldsymbol{\theta}^1_*$ between $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}^1$ that does not violate the constraint, where "between" means only that a shorter step is taken. Therefore, the next value for the iteration can be $\boldsymbol{\theta}^1_*$. The logic is true at every iteration, so a way to proceed is to alter the iteration so that the step length is shortened when necessary when a parameter violates the constraints.

E.3.6 SOME PRACTICAL CONSIDERATIONS

The reasons for the good performance of many algorithms, including DFP, are unknown. Moreover, different algorithms may perform differently in given settings. Indeed, for some problems, one algorithm may fail to converge whereas another will succeed in finding a solution without great difficulty. In view of this, computer programs such as GQOPT,[21] Gauss, and MatLab that offer a menu of different preprogrammed algorithms can be particularly useful. It is sometimes worth the effort to try more than one algorithm on a given problem.

[21] Goldfeld and Quandt (1972).

Step Sizes Except for the steepest ascent case, an optimal line search is likely to be infeasible or to require more effort than it is worth in view of the potentially large number of function evaluations required. In most cases, the choice of a step size is likely to be rather ad hoc. But within limits, the most widely used algorithms appear to be robust to inaccurate line searches. For example, one method employed by the widely used TSP computer program[22] is the method of *squeezing,* which tries $\lambda = 1, \frac{1}{2}, \frac{1}{4}$, and so on until an improvement in the function results. Although this approach is obviously a bit unorthodox, it appears to be quite effective when used with the Gauss–Newton method for nonlinear least squares problems. (See Chapter 7.) A somewhat more elaborate rule is suggested by Berndt et al. (1974). Choose an ε between 0 and $\frac{1}{2}$, and then find a λ such that

$$\varepsilon < \frac{F(\boldsymbol{\theta} + \lambda\boldsymbol{\Delta}) - F(\boldsymbol{\theta})}{\lambda \mathbf{g}'\boldsymbol{\Delta}} < 1 - \varepsilon. \qquad \textbf{(E-27)}$$

Of course, which value of ε to choose is still open, so the choice of λ remains ad hoc. Moreover, in neither of these cases is there any optimality to the choice; we merely find a λ that leads to a function improvement. Other authors have devised relatively efficient means of searching for a step size without doing the full optimization at each iteration.[23]

Assessing Convergence Ideally, the iterative procedure should terminate when the gradient is zero. In practice, this step will not be possible, primarily because of accumulated rounding error in the computation of the function and its derivatives. Therefore, a number of alternative convergence criteria are used. Most of them are based on the relative changes in the function or the parameters. There is considerable variation in those used in different computer programs, and there are some pitfalls that should be avoided. A critical absolute value for the elements of the gradient or its norm will be affected by any scaling of the function, such as normalizing it by the sample size. Similarly, stopping on the basis of small absolute changes in the parameters can lead to premature convergence when the parameter vector approaches the maximizer. It is probably best to use several criteria simultaneously, such as the proportional change in both the function and the parameters. Belsley (1980) discusses a number of possible stopping rules. One that has proved useful and is immune to the scaling problem is to base convergence on $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$.

Multiple Solutions It is possible for a function to have several local extrema. It is difficult to know a priori whether this is true of the one at hand. But if the function is not globally concave, then it may be a good idea to attempt to maximize it from several starting points to ensure that the maximum obtained is the global one. Ideally, a starting value near the optimum can facilitate matters; in some settings, this can be obtained by using a consistent estimate of the parameter for the starting point. The method of moments, if available, is sometimes a convenient device for doing so.

No Solution Finally, it should be noted that in a nonlinear setting the iterative algorithm can break down, even in the absence of constraints, for at least two reasons. The first possibility is that the problem being solved may be so numerically complex as to defy solution. The second possibility, which is often neglected, is that the proposed model may simply be inappropriate for the data. In a linear setting, a low R^2 or some other diagnostic test may suggest that the model and data are mismatched, but as long as the full rank condition is met by the regressor matrix, a linear regression can *always* be computed. Nonlinear models are not so forgiving. The failure of an iterative algorithm to find a maximum of the criterion function may be a warning that the model is not appropriate for this body of data.

[22] Hall (1982, p. 147).

[23] See, for example, Joreskog and Gruvaeus (1970), Powell (1964), Quandt (1983), and Hall (1982).

E.3.7 THE EM ALGORITHM

The latent class model can be characterized as a **missing data model.** Consider the mixture model we used for DocVis in Chapter 14, which we will now generalize to allow more than two classes:

$$f(y_{it} \mid \mathbf{x}_{it}, class_i = j) = \theta_{it,j}(1-\theta_{it,j})^{y_{it}}, \theta_{it,j} = 1/(1+\lambda_{it,j}), \lambda_{it,j} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j),\ y_{it} = 0, 1, \ldots.$$

$$\text{Prob}(class_i = j \mid \mathbf{z}_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}{\sum_{j=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)},\ j = 1, 2, \ldots, J.$$

With all parts incorporated, the log-likelihood for this latent class model is

$$\begin{aligned}\ln L_M &= \sum_{i=1}^{n} \ln L_{i,M} \\ &= \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{J} \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}{\sum_{m=1}^{J}\exp(\mathbf{z}_i'\boldsymbol{\alpha}_m)} \prod_{t=1}^{T_i} \left(\frac{1}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{y_{it}} \right\}.\end{aligned} \tag{E-28}$$

Suppose the actual class memberships were known (i.e., observed). Then, the class probabilities in $\ln L_M$ would be unnecessary. The appropriate **complete data log-likelihood** for this case would be

$$\begin{aligned}\ln L_C &= \sum_{i=1}^{n} \ln L_{i,C} \\ &= \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{J} D_{ij} \prod_{t=1}^{T_i} \left(\frac{1}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{y_{it}} \right\},\end{aligned} \tag{E-29}$$

where D_{ij} is an observed dummy variable that equals one if individual i is from class j, and zero otherwise. With this specification, the log-likelihood breaks into J separate log-likelihoods, one for each (now known) class. The maximum likelihood estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ would be obtained simply by separating the sample into the respective subgroups and estimating the appropriate model for each group using maximum likelihood. The method we have used to estimate the parameters of the full model is to replace the D_{ij} variables with their unconditional espectations, $\text{Prob}(class_i = j|\mathbf{z}_i)$, then maximize the resulting log-likelihood function. This is the essential logic of the **EM** (expectation–maximization) **algorithm** [Dempster et al. (1977)]; however, the method uses the conditional (posterior) class probabilities instead of the unconditional probabilities. The iterative steps of the EM algorithm are

(E step) Form the expectation of the missing data log-likelihood, conditional on the previous parameter estimates and the data in the sample;

(M step) Maximize the expected log-likelihood function. Then either return to the E step or exit if the estimates have converged.

The EM algorithm can be used in a variety of settings. [See McLachlan and Krishnan (1997).] It has a particularly appealing form for estimating latent class models. The iterative steps for the latent class model are as follows:

(E step) Form the conditional (posterior) class probabilities, $\pi_{ij}|\mathbf{z}_i$, based on the current estimates. These are based on the likelihood function.

(M step) For each class, estimate the class-specific parameters by maximizing a weighted log-likelihood,

$$\ln L_{M\ step,j} = \sum_{i=1}^{n_c} \pi_{ij} \ln\ L_i \mid class = j.$$

The parameters of the class probability model are also reestimated, as shown later, when there are variables in $\mathbf{z}_i$ other than a constant term.

This amounts to a simple weighted estimation. For example, in the latent class linear regression model, the M step would amount to nothing more than weighted least squares. For nonlinear models such as the geometric model above, the M step involves maximizing a weighted log-likelihood function.

For the preceding geometric model, the precise steps are as follows: First, obtain starting values for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J$. Recall, $\boldsymbol{\alpha}_J = \mathbf{0}$. Then;

1. Form the contributions to the likelihood function using (E-28),

$$\begin{aligned} L_i &= \sum_{j=1}^{J} \pi_{ij} \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_j, class_i = j) \\ &= \sum_{j=1}^{J} L_i \mid class = j. \end{aligned} \quad \textbf{(E-30)}$$

2. Form the conditional probabilities, $w_{ij} = \dfrac{L_i \mid class = j}{\sum_{m=1}^{J} L_i \mid class = m}$. **(E-31)**

3. For each j, now maximize the weighted log likelihood functions (one at a time),

$$\ln\ L_{j,M}(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} w_{ij} \ln \prod_{t=1}^{T_i} \left(\frac{1}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{(1-y_{it})} \left(\frac{\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)}{1+\exp(\mathbf{x}_{it}'\boldsymbol{\beta}_j)} \right)^{y_{it}} \quad \textbf{(E-32)}$$

4. To update the $\boldsymbol{\alpha}_j$ parameters, maximize the following log-likelihood function

$$\ln L(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J) = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij} \ln \frac{\exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}{\sum_{j=1}^{J} \exp(\mathbf{z}_i'\boldsymbol{\alpha}_j)}, \quad \boldsymbol{\alpha}_J = \mathbf{0}. \quad \textbf{(E-33)}$$

Step 4 defines a multinomial logit model (with "grouped") data. If the class probability model does not contain any variables in $\mathbf{z}_i$, other than a constant, then the solutions to this optimization will be

$$\hat{\pi}_j = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}}, \text{ then } \hat{\alpha}_j = \ln \frac{\hat{\pi}_j}{\hat{\pi}_J}. \quad \textbf{(E-34)}$$

(Note that this preserves the restriction $\hat{\alpha}_J = 0$.) With these in hand, we return to steps 1 and 2 to rebuild the weights, then perform steps 3 and 4. The process is iterated until the estimates of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ converge. Step 1 is constructed in a generic form. For a different model, it is necessary only to change the density that appears at the end of the expresssion in (E-32). For a cross section instead of a panel, the product term in step 1 becomes simply the log of the single term.

The EM algorithm has an intuitive appeal in this (and other) settings. In practical terms, it is often found to be a very slow algorithm. It can take many iterations to converge. (The estimates in Example 14.17 were computed using a gradient method, not the EM algorithm.) In its favor,

the EM method is very stable. It has been shown [Dempster, Laird, and Rubin (1977)] that the algorithm always climbs uphill. The log-likelihood improves with each iteration. Applications differ widely in the methods used to estimate latent class models. Adding to the variety are the very many Bayesian applications, none of which use either of the methods discussed here.

E.4 EXAMPLES

To illustrate the use of gradient methods, we consider some simple problems.

E.4.1 FUNCTION OF ONE PARAMETER

First, consider maximizing a function of a single variable, $f(\theta) = \ln(\theta) - 0.1\theta^2$. The function is shown in Figure E.4. The first and second derivatives are

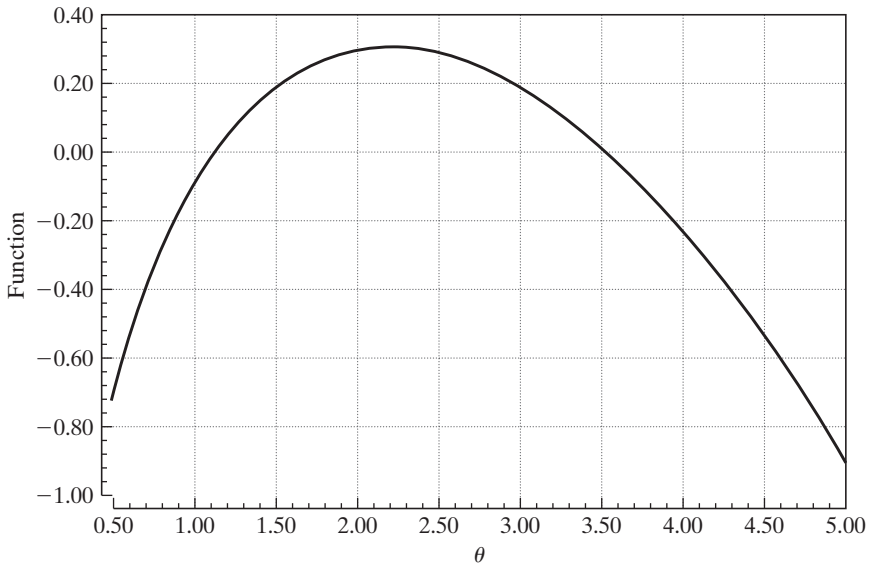
$$f'(\theta) = \frac{1}{\theta} - 0.2\,\theta,$$

$$f''(\theta) = \frac{-1}{\theta^2} - 0.2.$$

Equating f' to zero yields the solution $\theta = \sqrt{5} = 2.236$. At the solution, $f'' = -0.4$, so this solution is indeed a maximum. To demonstrate the use of an iterative method, we solve this problem using Newton's method. Observe, first, that the second derivative is always negative for any admissible (positive) θ.[24] Therefore, it should not matter where we start the iterations; we shall eventually find the maximum. For a single parameter, Newton's method is

$$\theta_{t+1} = \theta_t - [f'_t / f''_t].$$

FIGURE E.4 Function of One Variable Parameter.



[24] In this problem, an inequality restriction, $\theta > 0$, is required. As is common, however, for our first attempt we shall neglect the constraint.

TABLE E.1 Iterations for Newton's Method

Iteration	θ	f	f'	f''
0	5.00000	−0.890562	−0.800000	−0.240000
1	1.66667	0.233048	0.266667	−0.560000
2	2.14286	0.302956	0.030952	−0.417778
3	2.23404	0.304718	0.000811	−0.400363
4	2.23607	0.304719	0.0000004	−0.400000

The sequence of values that results when 5 is used as the starting value is given in Table E.1. The path of the iterations is also shown in the table.

E.4.2 FUNCTION OF TWO PARAMETERS: THE GAMMA DISTRIBUTION

For random sampling from the gamma distribution,

$$f(y_i, \beta, \rho) = \frac{\beta^\rho}{\Gamma(\rho)} e^{-\beta y_i} y_i^{\rho-1}.$$

The log-likelihood is $\ln L(\beta, \rho) = n\rho \ln \beta - n \ln \Gamma(\rho) - \beta \sum_{i=1}^n y_i + (\rho - 1) \sum_{i=1}^n \ln y_i$. (See Section 14.6.4 and Example 13.5.) It is often convenient to scale the log-likelihood by the sample size. Suppose, as well, that we have a sample with $\bar{y} = 3$ and $\overline{\ln y} = 1$. Then the function to be maximized is $F(\beta, \rho) = \rho \ln \beta - \ln \Gamma(\rho) - 3\beta + \rho - 1$. The derivatives are

$$\frac{\partial F}{\partial \beta} = \frac{\rho}{\beta} - 3, \qquad \frac{\partial F}{\partial \rho} = \ln \beta - \frac{\Gamma'}{\Gamma} + 1 = \ln \beta - \Psi(\rho) + 1,$$

$$\frac{\partial^2 F}{\partial \beta^2} = \frac{-\rho}{\beta^2}, \qquad \frac{\partial^2 F}{\partial \rho^2} = \frac{-(\boldsymbol{\Gamma}\boldsymbol{\Gamma}'' - \Gamma'^2)}{\boldsymbol{\Gamma}^2} = -\boldsymbol{\Psi}'(\rho), \qquad \frac{\partial^2 F}{\partial \beta\, \partial \rho} = \frac{1}{\beta}.$$

Finding a good set of starting values is often a difficult problem. Here we choose three starting points somewhat arbitrarily: $(\rho^0, \beta^0) = (4, 1)$, $(8, 3)$, and $(2, 7)$. The solution to the problem is $(5.233, 1.7438)$. We used Newton's method and DFP with a line search to maximize this function.[25] For Newton's method, $\lambda = 1$. The results are shown in Table E.2. The two methods were essentially the same when starting from a good starting point (trial 1), but they differed substantially when starting from a poorer one (trial 2). Note that DFP and Newton approached the solution from different directions in trial 2. The third starting point shows the value of a line search. At this

TABLE E.2 Iterative Solutions to Max$(\rho, \beta)\rho \ln \beta - \ln \boldsymbol{\Gamma}(\rho) - 3\beta + \rho - 1$

	Trial 1				*Trial 2*				*Trial 3*			
	DFP		*Newton*		*DFP*		*Newton*		*DFP*		*Newton*	
Iter.	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β	ρ	β
0	4.000	1.000	4.000	1.000	8.000	3.000	8.000	3.000	2.000	7.000	2.000	7.000
1	3.981	1.345	3.812	1.203	7.117	2.518	2.640	0.615	6.663	2.027	−47.7	−233.
2	4.005	1.324	4.795	1.577	7.144	2.372	3.203	0.931	6.195	2.075	—	—
3	5.217	1.743	5.190	1.728	7.045	2.389	4.257	1.357	5.239	1.731	—	—
4	5.233	1.744	5.231	1.744	5.114	1.710	5.011	1.656	5.251	1.754	—	—
5	—	—	—	—	5.239	1.747	5.219	1.740	5.233	1.744	—	—
6	—	—	—	—	5.233	1.744	5.233	1.744	—	—	—	—

[25] The one used is described in Joreskog and Gruvaeus (1970).

starting value, the Hessian is extremely large, and the second value for the parameter vector with Newton's method is (−47.671, −233.35), at which point F cannot be computed and this method must be abandoned. Beginning with $\mathbf{H} = \mathbf{I}$ and using a line search, DFP reaches the point (6.63, 2.03) at the first iteration, after which convergence occurs routinely in three more iterations. At the solution, the Hessian is $[(-1.72038, 0.191153)', (0.191153, -0.210579)']$. The diagonal elements of the Hessian are negative and its determinant is 0.32574, so it is negative definite. (The two characteristic roots are −1.7442 and −0.18675). Therefore, this result is indeed the maximum of the function.

E.4.3 A CONCENTRATED LOG-LIKELIHOOD FUNCTION

There is another way that the preceding problem might have been solved. The first of the necessary conditions implies that at the joint solution for (β, ρ), β will equal $\rho/3$. Suppose that we impose this requirement on the function we are maximizing. The **concentrated** (over β) **log-likelihood function** is then produced:

$$\begin{aligned} F_c(\rho) &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 3(\rho/3) + \rho - 1 \\ &= \rho \ln(\rho/3) - \ln \Gamma(\rho) - 1. \end{aligned}$$

This function could be maximized by an iterative search or by a simple one-dimensional grid search. Figure E.5 shows the behavior of the function. As expected, the maximum occurs at $\rho = 5.233$. The value of β is found as $5.23/3 = 1.743$.

The concentrated log-likelihood is a useful device in many problems. (See Section 14.9.6.d for an application.) Note the interpretation of the function plotted in Figure E.5. The original function of ρ and β is a surface in three dimensions. The curve in Figure E.5 is a projection of that function; it is a plot of the function values above the line $\beta = \rho/3$. By virtue of the first-order condition, we know that one of these points will be the maximizer of the function. Therefore, we may restrict our search for the overall maximum of $F(\beta, \rho)$ to the points on this line.

FIGURE E.5 Concentrated Log-Likelihood.