if **A** is nonsingular. If **A** is singular, then there is no inverse transformation. Let **J** be the matrix of partial derivatives of the inverse functions:

$$\mathbf{J} = \left[\frac{\partial x_i}{\partial y_j}\right].$$

The absolute value of the determinant of **J**,

$$\text{abs}(|\mathbf{J}|) = \text{abs}\left(\det\left(\left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}'}\right]\right)\right),$$

is the Jacobian determinant of the transformation from **y** to **x**. In the nonsingular case,

$$\text{abs}(|\mathbf{J}|) = \text{abs}(|\mathbf{A}^{-1}|) = \frac{1}{\text{abs}(|\mathbf{A}|)}.$$

In the singular case, the matrix of partial derivatives will be singular and the determinant of the Jacobian will be zero. In this instance, the singular Jacobian implies that **A** is singular or, equivalently, that the transformations from **x** to **y** are functionally dependent. The singular case is analogous to the single-variable case.

Clearly, if the vector **x** is given, then $\mathbf{y} = \mathbf{Ax}$ can be computed from **x**. Whether **x** can be deduced from **y** is another question. Evidently, it depends on the Jacobian. If the Jacobian is not zero, then the inverse transformations exist, and we can obtain **x**. If not, then we cannot obtain **x**.

APPENDIX B

PROBABILITY AND DISTRIBUTION THEORY

B.1 INTRODUCTION

This appendix reviews the distribution theory used later in the book. A previous course in statistics is assumed, so most of the results will be stated without proof. The more advanced results in the later sections will be developed in greater detail.

B.2 RANDOM VARIABLES

We view our observation on some aspect of the economy as the **outcome** or realization of a random process that is almost never under our (the analyst's) control. In the current literature, the descriptive (and perspective laden) term **data generating process (DPG)** is often used for this underlying mechanism. The observed (measured) outcomes of the process are assigned unique numeric values. The assignment is one to one; each outcome gets one value, and no two distinct outcomes receive the same value. This outcome variable, X, is a **random variable** because, until the data are actually observed, it is uncertain what value X will take. Probabilities are associated with outcomes to quantify this uncertainty. We usually use capital letters for the "name" of a random variable and lowercase letters for the values it takes. Thus, the probability that X takes a particular value x might be denoted Prob $(X = x)$.

A random variable is **discrete** if the set of outcomes is either finite in number or countably infinite. The random variable is **continuous** if the set of outcomes is infinitely divisible and, hence, not countable. These definitions will correspond to the types of data we observe in practice. Counts of occurrences will provide observations on discrete random variables, whereas measurements such as time or income will give observations on continuous random variables.

B.2.1 PROBABILITY DISTRIBUTIONS

A listing of the values x taken by a random variable X and their associated probabilities is a **probability distribution**, $f(x)$. For a discrete random variable,

$$f(x) = \text{Prob}(X = x). \qquad \textbf{(B-1)}$$

The **axioms of probability** require that

1. $0 \le \text{Prob}(X = x) \le 1.$ **(B-2)**
2. $\sum_x f(x) = 1.$ **(B-3)**

For the continuous case, the probability associated with any particular point is zero, and we can only assign positive probabilities to intervals in the range (or **support**) of x. The **probability density function (pdf)**, $f(x)$, is defined so that $f(x) \ge 0$ and

1. $\text{Prob}(a \le x \le b) = \int_a^b f(x)\,dx \ge 0.$ **(B-4)**

This result is the area under $f(x)$ in the range from a to b. For a continuous variable,

2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1.$ **(B-5)**

If the range of x is not infinite, then it is understood that $f(x) = 0$ anywhere outside the appropriate range. Because the probability associated with any individual point is 0,

$$\begin{aligned}\text{Prob}(a \le x \le b) &= \text{Prob}(a \le x < b)\\ &= \text{Prob}(a < x \le b)\\ &= \text{Prob}(a < x < b).\end{aligned}$$

B.2.2 CUMULATIVE DISTRIBUTION FUNCTION

For any random variable X, the probability that X is less than or equal to a is denoted $F(a)$. $F(x)$ is the **cumulative density function (cdf)**, or **distribution function**. For a discrete random variable,

$$F(x) = \sum_{X \le x} f(X) = \text{Prob}(X \le x). \qquad \textbf{(B-6)}$$

In view of the definition of $f(x)$,

$$f(x_i) = F(x_i) - F(x_{i-1}). \tag{B-7}$$

For a continuous random variable,

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \tag{B-8}$$

and

$$f(x) = \frac{dF(x)}{dx}. \tag{B-9}$$

In both the continuous and discrete cases, $F(x)$ must satisfy the following properties:

1. $0 \le F(x) \le 1$.
2. If $x > y$, then $F(x) \ge F(y)$.
3. $F(+\infty) = 1$.
4. $F(-\infty) = 0$.

From the definition of the cdf,

$$\text{Prob}(a < x \le b) = F(b) - F(a). \tag{B-10}$$

Any valid pdf will imply a valid cdf, so there is no need to verify these conditions separately.

B.3 EXPECTATIONS OF A RANDOM VARIABLE

DEFINITION B.1 Mean of a Random Variable
*The **mean**, or **expected value**, of a random variable is*

$$E[x] = \begin{cases} \sum_x x f(x) & \text{if } x \text{ is discrete,} \\ \int_x x f(x)\,dx & \text{if } x \text{ is continuous.} \end{cases} \tag{B-11}$$

The notation $\sum_x$ or $\int_x$, used henceforth, means the sum or integral over the entire range of values of x. The mean is usually denoted μ. It is a weighted average of the values taken by x, where the weights are the respective probabilities or densities. It is not necessarily a value actually taken by the random variable. For example, the expected number of heads in one toss of a fair coin is $\frac{1}{2}$.

Other **measures of central tendency** are the **median**, which is the value m such that $\text{Prob}(X \le m) \ge \frac{1}{2}$ and $\text{Prob}(X \ge m) \ge \frac{1}{2}$, and the **mode**, which is the value of x at which $f(x)$ takes its maximum. The first of these measures is more frequently used than the second. Loosely speaking, the median corresponds more closely than the mean to the middle of a distribution. It is unaffected by extreme values. In the discrete case, the modal value of x has the highest probability of occurring. The modal value for a continuous variable will usually not be meaningful.

Let $g(x)$ be a function of x. The function that gives the expected value of $g(x)$ is denoted

$$E[g(x)] = \begin{cases} \sum_x g(x) \operatorname{Prob}(X = x) & \text{if } X \text{ is discrete,} \\ \int_x g(x)f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases} \qquad \textbf{(B-12)}$$

If $g(x) = a + bx$ for constants a and b, then

$$E[a + bx] = a + bE[x].$$

An important case is the expected value of a constant a, which is just a.

DEFINITION B.2 Variance of a Random Variable

The ***variance*** *of a random variable is*

$$\operatorname{Var}[x] = E[(x - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } x \text{ is discrete,} \\ \int_x (x - \mu)^2 f(x)\, dx & \text{if } x \text{ is continuous.} \end{cases} \qquad \textbf{(B-13)}$$

The variance of x, Var[x], which must be positive, is usually denoted σ^2. This function is a measure of the dispersion of a distribution. Computation of the variance is simplified by using the following important result:

$$\operatorname{Var}[x] = E[x^2] - \mu^2. \qquad \textbf{(B-14)}$$

A convenient corollary to (B-14) is

$$E[x^2] = \sigma^2 + \mu^2. \qquad \textbf{(B-15)}$$

By inserting $y = a + bx$ in (B-13) and expanding, we find that

$$\operatorname{Var}[a + bx] = b^2 \operatorname{Var}[x], \qquad \textbf{(B-16)}$$

which implies, for any constant a, that

$$\operatorname{Var}[a] = 0. \qquad \textbf{(B-17)}$$

To describe a distribution, we usually use σ, the positive square root, which is the **standard deviation** of x. The standard deviation can be interpreted as having the same units of measurement as x and μ. For any random variable x and any positive constant k, the **Chebychev inequality** states that

$$\operatorname{Prob}(\mu - k\sigma \leq x \leq \mu + k\sigma) \geq 1 - \frac{1}{k^2}. \qquad \textbf{(B-18)}$$

Two other measures often used to describe a probability distribution are

$$\text{skewness} = E[(x - \mu)^3],$$

and

$$\text{kurtosis} = E[(x - \mu)^4].$$

Skewness is a measure of the asymmetry of a distribution. For symmetric distributions,

$$f(\mu - x) = f(\mu + x),$$

and

$$\text{skewness} = 0.$$

For asymmetric distributions, the skewness will be positive if the "long tail" is in the positive direction. Kurtosis is a measure of the thickness of the tails of the distribution. A shorthand expression for other **central moments** is

$$\mu_r = E[(x - \mu)^r].$$

Because μ_r tends to explode as r grows, the normalized measure, μ_r/σ^r, is often used for description. Two common measures are

$$\text{skewness coefficient} = \frac{\mu_3}{\sigma^3},$$

and

$$\text{degree of excess} = \frac{\mu_4}{\sigma^4} - 3.$$

The second is based on the normal distribution, which has excess of zero. (The value 3 is sometimes labeled the "mesokurtotic" value.)

For any two functions $g_1(x)$ and $g_2(x)$,

$$E[g_1(x) + g_2(x)] = E[g_1(x)] + E[g_2(x)]. \qquad \textbf{(B-19)}$$

For the general case of a possibly nonlinear $g(x)$,

$$E[g(x)] = \int_x g(x)f(x)\,dx, \qquad \textbf{(B-20)}$$

and

$$\text{Var}[g(x)] = \int_x (g(x) - E[g(x)])^2 f(x)\,dx. \qquad \textbf{(B-21)}$$

(For convenience, we shall omit the equivalent definitions for discrete variables in the following discussion and use the integral to mean either integration or summation, whichever is appropriate.)

A device used to approximate $E[g(x)]$ and $\text{Var}[g(x)]$ is the linear Taylor series approximation:

$$g(x) \approx [g(x^0) - g'(x^0)x^0] + g'(x^0)x = \beta_1 + \beta_2 x = g^*(x). \qquad \textbf{(B-22)}$$

If the approximation is reasonably accurate, then the mean and variance of $g^*(x)$ will be approximately equal to the mean and variance of $g(x)$. A natural choice for the expansion point is $x^0 = \mu = E(x)$. Inserting this value in (B-22) gives

$$g(x) \approx [g(\mu) - g'(\mu)\mu] + g'(\mu)x, \tag{B-23}$$

so that

$$E[g(x)] \approx g(\mu), \tag{B-24}$$

and

$$\mathrm{Var}[g(x)] \approx [g'(\mu)]^2\,\mathrm{Var}[x]. \tag{B-25}$$

A point to note in view of (B-22) to (B-24) is that $E[g(x)]$ will generally not equal $g(E[x])$. For the special case in which $g(x)$ is concave—that is, where $g''(x) < 0$—we know from **Jensen's inequality** that $E[g(x)] \le g(E[x])$. For example, $E[\log(x)] \le \log(E[x])$. The result in (B-25) forms the basis for the **delta method**.

B.4 SOME SPECIFIC PROBABILITY DISTRIBUTIONS

Certain experimental situations naturally give rise to specific probability distributions. In the majority of cases in economics, however, the distributions used are merely models of the observed phenomena. Although the normal distribution, which we shall discuss at length, is the mainstay of econometric research, economists have used a wide variety of other distributions. A few are discussed here.[1]

B.4.1 THE NORMAL AND SKEW NORMAL DISTRIBUTIONS

The general form of the normal distribution with mean μ and standard deviation σ is

$$f(x\,|\,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-1/2[(x-\mu)^2/\sigma^2]}. \tag{B-26}$$

This result is usually denoted $x \sim N[\mu, \sigma^2]$. The standard notation $x \sim f(x)$ is used to state that "x has probability distribution $f(x)$." Among the most useful properties of the normal distribution is its preservation under linear transformation.

$$\text{If } x \sim N[\mu,\sigma^2], \qquad \text{then } (a + bx) \sim N[a + b\mu, b^2\sigma^2]. \tag{B-27}$$

One particularly convenient transformation is $a = -\mu/\sigma$ and $b = 1/\sigma$. The resulting variable $z = (x - \mu)/\sigma$ has the **standard normal distribution**, denoted $N[0, 1]$, with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}. \tag{B-28}$$

[1] A much more complete listing appears in Maddala (1977a, Chapters 3 and 18) and in most mathematical statistics textbooks. See also Poirier (1995) and Stuart and Ord (1989). Another useful reference is Evans, Hastings, and Peacock (2010). Johnson et al. (1974, 1993, 1994, 1995, 1997) is an encyclopedic reference on the subject of statistical distributions.

The specific notation $\phi(z)$ is often used for this density and $\Phi(z)$ for its cdf. It follows from the definitions above that if $x \sim N[\mu, \sigma^2]$, then

$$f(x) = \frac{1}{\sigma}\phi\left[\frac{x - \mu}{\sigma}\right].$$

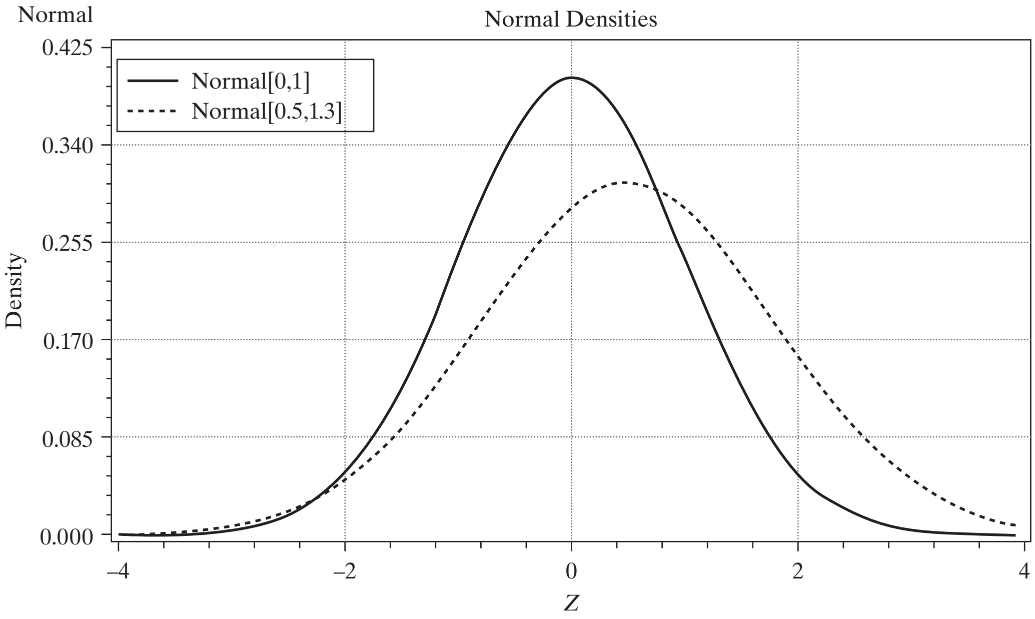
Figure B.1 shows the densities of the standard normal distribution and the normal distribution with mean 0.5, which shifts the distribution to the right, and standard deviation 1.3, which, it can be seen, scales the density so that it is shorter but wider. (The graph is a bit deceiving unless you look closely; both densities are symmetric.)

Tables of the standard normal cdf appear in most statistics and econometrics textbooks. Because the form of the distribution does not change under a linear transformation, it is not necessary to tabulate the distribution for other values of μ and σ. For any normally distributed variable,

$$\text{Prob}(a \leq x \leq b) = \text{Prob}\left(\frac{a - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right), \qquad \textbf{(B-29)}$$

which can always be read from a table of the standard normal distribution. In addition, because the distribution is symmetric, $\Phi(-z) = 1 - \Phi(z)$. Hence, it is not necessary to tabulate both the negative and positive halves of the distribution.

FIGURE B.1 The Normal Distribution.



The centerpiece of the stochastic frontier literture is the skew normal distribution. See Examples 12.2 and 14.8 and Section 19.2.4.) The density of the skew normal random variable is

$$f(x\,|\,\mu, \sigma, \lambda) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(\frac{-\lambda\varepsilon}{\sigma}\right), \varepsilon = (x - \mu).$$

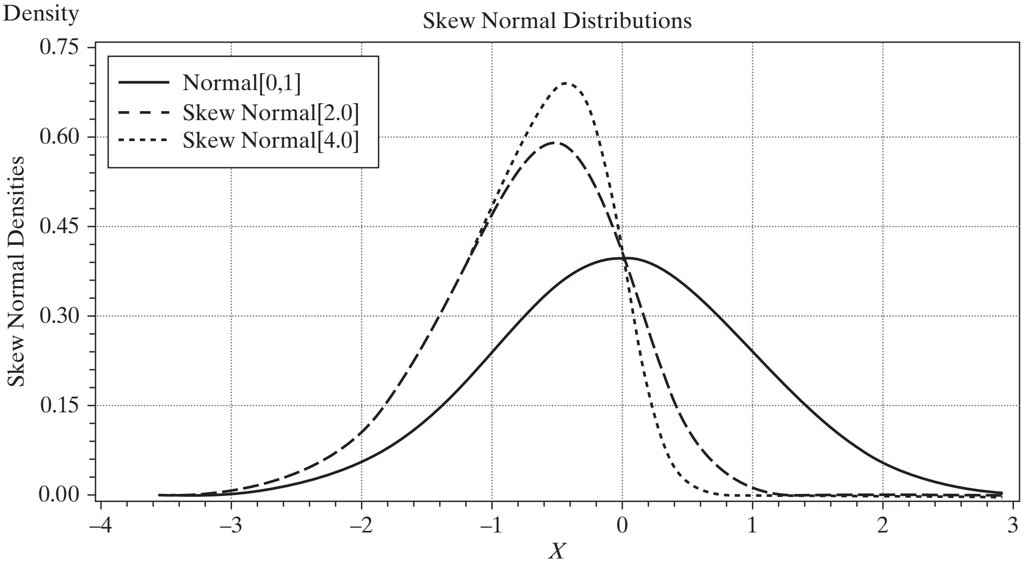
The skew normal reverts to the standard normal if $\lambda = 0$. The random variable arises as the density of $\varepsilon = \sigma_v v - \sigma_u|u|$ where u and v are standard normal variables, in which case $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_{\mathrm{v}}^2 + \sigma_{\mathrm{u}}^2$. (If $\sigma_u|u|$ is added, then $-\lambda$ becomes $+\lambda$ in the density. Figure B.2 shows three cases of the distribution, $\lambda = 0, 2$, and 4. This asymmetric distribution has mean $-\dfrac{\sigma\lambda}{\sqrt{1+\lambda^2}}\sqrt{\dfrac{2}{\pi}}$ and variance $\dfrac{\sigma^2}{1+\lambda^2}\left(1 + \lambda^2\left(\dfrac{\pi - 2}{\pi}\right)\right)$ (which revert to 0 and 1 if $\lambda = 0$).

These are $-\sigma_u(2/\pi)^{1/2}$ and $\sigma_v{}^2 + \sigma_u{}^2(\pi - 2)/\pi$ for the convolution form.

B.4.2 THE CHI-SQUARED, *T*, AND *F* DISTRIBUTIONS

The chi-squared, *t*, and *F* distributions are derived from the normal distribution. They arise in econometrics as sums of n or n_1 and n_2 other variables. These three distributions have associated with them one or two "degrees of freedom" parameters, which for our purposes will be the number of variables in the relevant sum.

FIGURE B.2 Skew Normal Densities.



The first of the essential results is

- If $z \sim N[0, 1]$, then $x = z^2 \sim$ chi-squared[1]—that is, **chi-squared** with one degree of freedom—denoted

$$z^2 \sim \chi^2[1]. \quad \textbf{(B-30)}$$

This distribution is a skewed distribution with mean 1 and variance 2. The second result is

- If $x_1, \ldots, x_n$ are n *independent* chi-squared[1] variables, then

$$\sum_{i=1}^{n} x_i \sim \text{chi-squared}[n]. \quad \textbf{(B-31)}$$

The mean and variance of a chi-squared variable with n degrees of freedom are n and $2n$, respectively. A number of useful corollaries can be derived using (B-30) and (B-31).

- If z_i, $i = 1, \ldots, n$, are independent $N[0, 1]$ variables, then

$$\sum_{i=1}^{n} z_i^2 \sim \chi^2[n]. \quad \textbf{(B-32)}$$

- If z_i, $i = 1, \ldots, n$, are independent $N[0, \sigma^2]$ variables, then

$$\sum_{i=1}^{n} (z_i/\sigma)^2 \sim \chi^2[n]. \quad \textbf{(B-33)}$$

- If x_1 and x_2 are independent chi-squared variables with n_1 and n_2 degrees of freedom, respectively, then

$$x_1 + x_2 \sim \chi^2[n_1 + n_2]. \quad \textbf{(B-34)}$$

This result can be generalized to the sum of an arbitrary number of independent chi-squared variables.

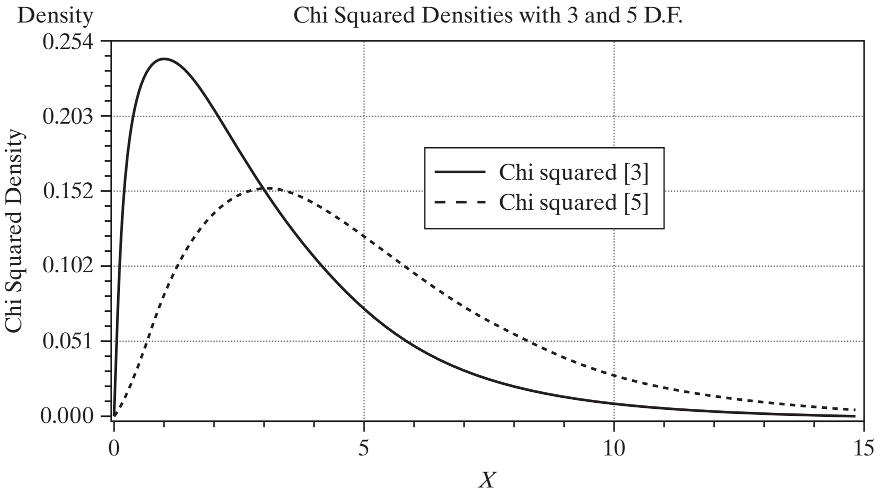
Figure B.3 shows the chi-squared densities for 3 and 5 degrees of freedom. The amount of skewness declines as the number of degrees of freedom rises. Unlike the normal distribution, a separate table is required for the chi-squared distribution for each value of n. Typically, only a few percentage points of the distribution are tabulated for each n.

- The chi-squared[n] random variable has the density of a gamma variable [See (B-39)] with
- parameters $\lambda = {}^1\!/_2$ and $P = n/2$.
- If x_1 and x_2 are two *independent* chi-squared variables with degrees of freedom parameters x_1 and x_1 respectively, then the ratio

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2} \quad \textbf{(B-35)}$$

has the ***F* distribution** with n_1 and n_2 degrees of freedom.

FIGURE B.3 The Chi-Squared[3] Distribution.



The two degrees of freedom parameters n_1 and n_2 are the "numerator and denominator degrees of freedom," respectively. Tables of the F distribution must be computed for each pair of values of (n_1, n_2). As such, only one or two specific values, such as the 95 percent and 99 percent upper tail values, are tabulated in most cases.

- If z is an $N[0, 1]$ variable and x is $\chi^2[n]$ and is independent of z, then the ratio

$$t[n] = \frac{z}{\sqrt{x/n}} \tag{B-36}$$

has the t **distribution** with n degrees of freedom.

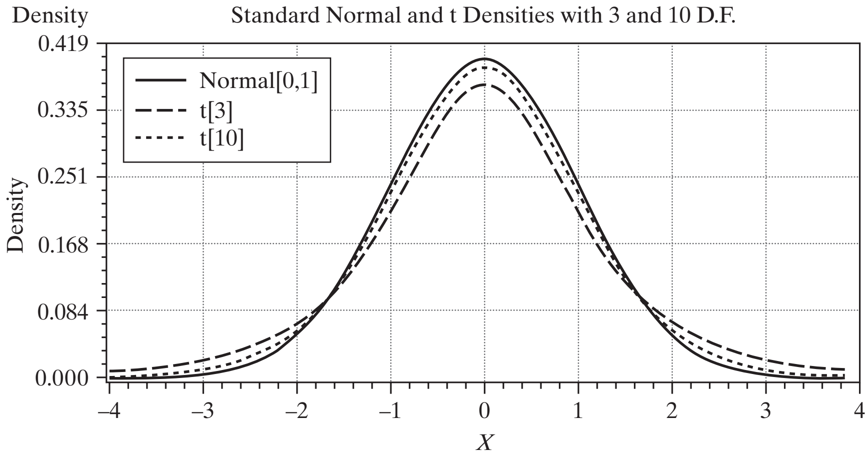
The t distribution has the same shape as the normal distribution but has thicker tails. Figure B.4 illustrates the t distributions with 3 and 10 degrees of freedom with the standard normal distribution. Two effects that can be seen in the figure are how the distribution changes as the degrees of freedom increases, and, overall, the similarity of the t distribution to the standard normal. This distribution is tabulated in the same manner as the chi-squared distribution, with several specific cutoff points corresponding to specified tail areas for various values of the degrees of freedom parameter.

Comparing (B-35) with $n_1 = 1$ and (B-36), we see the useful relationship between the t and F distributions:

- If $t \sim t[n]$, then $t^2 \sim F[1, n]$.

If the numerator in (B-36) has a nonzero mean, then the random variable in (B-36) has a noncentral t distribution and its square has a noncentral F distribution. These

FIGURE B.4 The Standard Normal, t[3], and t [10] Distributions.



distributions arise in the F tests of linear restrictions [see (5-16)] when the restrictions do not hold as follows:

1. *Noncentral chi-squared distribution*. If z has a normal distribution with mean μ and standard deviation 1, then the distribution of z^2 is *noncentral* chi-squared with parameters 1 and $\mu^2/2$.
   - **a.** If $\mathbf{z} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ with J elements, then $\mathbf{z}'\,\boldsymbol{\Sigma}^{-1}\,\mathbf{z}$ has a noncentral chi-squared distribution with J degrees of freedom and noncentrality parameter $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}/2$, which we denote $\boldsymbol{\chi}^2_*[J, \boldsymbol{\mu}'\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{\mu}/2]$.
   - **b.** If $\mathbf{z} \sim N[\boldsymbol{\mu}, \mathbf{I}]$ and $\mathbf{M}$ is an idempotent matrix with rank J, then $\mathbf{z}'\mathbf{M}\mathbf{z} \sim \boldsymbol{\chi}^2_*[J, \boldsymbol{\mu}'\mathbf{M}\boldsymbol{\mu}/2]$.
2. *Noncentral F distribution*. If X_1 has a noncentral chi-squared distribution with noncentrality parameter λ and degrees of freedom n_1 and X_2 has a central chi-squared distribution with degrees of freedom n_2 and is independent of X_1, then

$$F_* = \frac{X_1/n_1}{X_2/n_2}$$

   has a noncentral F distribution with parameters n_1, n_2, and λ. (The denominator chi-squared could also be noncentral, but we shall not use any statistics with doubly noncentral distributions.) In each of these cases, the statistic and the distribution are the familiar ones, except that the effect of the nonzero mean, which induces the noncentrality, is to push the distribution to the right.

B.4.3 DISTRIBUTIONS WITH LARGE DEGREES OF FREEDOM

The chi-squared, t, and F distributions usually arise in connection with sums of sample observations. The degrees of freedom parameter in each case grows with the number of observations. We often deal with larger degrees of freedom than are shown in the tables.

Thus, the standard tables are often inadequate. In all cases, however, there are **limiting distributions** that we can use when the degrees of freedom parameter grows large. The simplest case is the *t* distribution. The *t* distribution with infinite degrees of freedom is equivalent (identical) to the standard normal distribution. Beyond about 100 degrees of freedom, they are almost indistinguishable.

For degrees of freedom greater than 30, a reasonably good approximation for the distribution of the chi-squared variable *x* is

$$z = (2x)^{1/2} - (2n - 1)^{1/2}, \tag{B-37}$$

which is approximately standard normally distributed. Thus,

$$\text{Prob}(\chi^2[n] \leq a) \approx \Phi[(2a)^{1/2} - (2n - 1)^{1/2}].$$

Another simple approximation that relies on the central limit theorem would be $z = (x - n)/(2n)^{1/2}$.

As used in econometrics, the *F* distribution with a large-denominator degrees of freedom is common. As n_2 becomes infinite, the denominator of *F* converges identically to one, so we can treat the variable

$$x = n_1 F \tag{B-38}$$

as a chi-squared variable with n_1 degrees of freedom. The numerator degree of freedom will typically be small, so this approximation will suffice for the types of applications we are likely to encounter.[2] If not, then the approximation given earlier for the chi-squared distribution can be applied to $n_1 F$.

B.4.4 SIZE DISTRIBUTIONS: THE LOGNORMAL DISTRIBUTION

In modeling size distributions, such as the distribution of firm sizes in an industry or the distribution of income in a country, the **lognormal distribution**, denoted $LN[\mu, \sigma^2]$, has been particularly useful.[3] The density is

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} e^{-1/2[(\ln x - \mu)/\sigma]^2}, \quad x > 0.$$

A lognormal variable *x* has

$$E[x] = e^{\mu + \sigma^2/2},$$

and

$$\text{Var}[x] = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1).$$

The relation between the normal and lognormal distributions is

$$\text{If } y \sim LN[\mu, \sigma^2], \quad \ln y \sim N[\mu, \sigma^2].$$

[2] See Johnson, Kotz, and Balakrishnan (1994) for other approximations.

[3] A study of applications of the lognormal distribution appears in Aitchison and Brown (1969).

A useful result for transformations is given as follows:

If x has a lognormal distribution with mean θ and variance λ^2, then

$$\ln x \sim N(\mu, \sigma^2), \quad \text{where } \mu = \ln \theta^2 - \tfrac{1}{2}\ln(\theta^2 + \lambda^2) \quad \text{and} \quad \sigma^2 = \ln(1 + \lambda^2/\theta^2).$$

Because the normal distribution is preserved under linear transformation,

$$\text{if } y \sim LN[\mu, \sigma^2], \quad \text{then } \ln y^r \sim N[r\mu, r^2\sigma^2].$$

If y_1 and y_2 are independent lognormal variables with $y_1 \sim LN[\mu_1, \sigma_1^2]$ and $y_2 \sim LN[\mu_2, \sigma_2^2]$, then

$$y_1 y_2 \sim LN[\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2].$$

B.4.5 THE GAMMA AND EXPONENTIAL DISTRIBUTIONS

The **gamma distribution** has been used in a variety of settings, including the study of income distribution[4] and production functions.[5] The general form of the distribution is

$$f(x) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda x} x^{P-1}, \quad x \geq 0, \lambda > 0, P > 0. \tag{B-39}$$

Many familiar distributions are special cases, including the **exponential distribution** ($P = 1$) and chi-squared ($\lambda = \frac{1}{2}$, $P = \frac{n}{2}$). The **Erlang distribution** results if P is a positive integer. The mean is P/λ, and the variance is P/λ^2. The **inverse gamma distribution** is the distribution of $1/x$, where x has the gamma distribution. Using the change of variable, $y = 1/x$, the Jacobian is $|dx/dy| = 1/y^2$. Making the substitution and the change of variable, we find

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda/y}\, y^{-(P+1)}, y \geq 0, \lambda > 0, P > 0.$$

The density is defined for positive P. However, the mean is $\lambda/(P - 1)$ which is defined only if $P > 1$ and the variance is $\lambda^2/[(P - 1)^2(P - 2)]$ which is defined only for $P > 2$.

B.4.6 THE BETA DISTRIBUTION

Distributions for models are often chosen on the basis of the range within which the random variable is constrained to vary. The lognormal distribution, for example, is sometimes used to model a variable that is always nonnegative. For a variable constrained between 0 and $c > 0$, the **beta distribution** has proved useful. Its density is

$$f(x) = \frac{1}{c}\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{x}{c}\right)^{\alpha-1}\left(1 - \frac{x}{c}\right)^{\beta-1}. \tag{B-40}$$

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$, standard uniform if $\alpha = \beta = c = 1$, asymmetric otherwise, and

[4] Salem and Mount (1974).

[5] Greene (1980a).

can be hump-shaped or U-shaped. The mean is $c\alpha/(\alpha + \beta)$, and the variance is $c^2\alpha\beta/[(\alpha + \beta + 1)(\alpha + \beta)^2]$. The beta distribution has been applied in the study of labor force participation rates.[6]

B.4.7 THE LOGISTIC DISTRIBUTION

The normal distribution is ubiquitous in econometrics. But researchers have found that for some microeconomic applications, there does not appear to be enough mass in the tails of the normal distribution; observations that a model based on normality would classify as "unusual" seem not to be very unusual at all. One approach has been to use thicker-tailed symmetric distributions. The **logistic distribution** is one candidate; the cdf for a logistic random variable is denoted

$$F(x) = \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

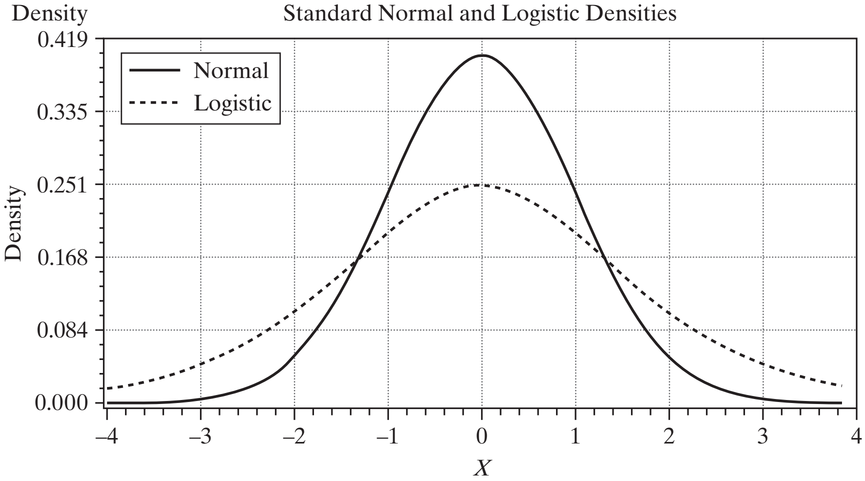
The density is $f(x) = \Lambda(x)[1 - \Lambda(x)]$. The mean and variance of this random variable are zero and $\pi^2/3$. Figure B.5 compares the logistic distribution to the standard normal. The logistic density has a greater variance and thicker tails than the normal. The standardized variable, $z/(\pi/3^{1/2})$ is very close to the t[8] variable.

B.4.8 THE WISHART DISTRIBUTION

The Wishart distribution describes the distribution of a random matrix obtained as

$$\mathbf{W} = \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})',$$

FIGURE B.5 Normal and Logistic Densities.



[6] Heckman and Willis (1976).

where $\mathbf{x}_i$ is the ith of n K element random vectors from the multivariate normal distribution with mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. This is a multivariate counterpart to the chi-squared distribution. The density of the Wishart random matrix is

$$f(\mathbf{W}) = \frac{\exp\left[-\frac{1}{2}trace(\boldsymbol{\Sigma}^{-1}\mathbf{W})\right]|\mathbf{W}|^{-\frac{1}{2}(n-K-1)}}{2^{nK/2}|\boldsymbol{\Sigma}|^{K/2}\pi^{K(K-1)/4}\boldsymbol{\Pi}_{j=1}^{K}\Gamma\left(\frac{n+1-j}{2}\right)}.$$

The mean matrix is $n\boldsymbol{\Sigma}$. For the individual pairs of elements in $\mathbf{W}$,

$$\text{Cov}[w_{ij}, w_{rs}] = n(\sigma_{ir}\sigma_{js} + \sigma_{is}\sigma_{jr}).$$

B.4.9 DISCRETE RANDOM VARIABLES

Modeling in economics frequently involves random variables that take integer values. In these cases, the distributions listed thus far only provide approximations that are sometimes quite inappropriate. We can build up a class of models for discrete random variables from the **Bernoulli distribution** for a single binomial outcome (trial)

$$\text{Prob}(x = 1) = \alpha,$$
$$\text{Prob}(x = 0) = 1 - \alpha,$$

where $0 \leq \alpha \leq 1$. The modeling aspect of this specification would be the assumptions that the success probability α is constant from one trial to the next and that successive trials are independent. If so, then the distribution for x successes in n trials is the **binomial distribution**,

$$\text{Prob}(X = x) = \binom{n}{x}\alpha^x(1-\alpha)^{n-x}, \quad x = 0, 1, \ldots, n.$$

The mean and variance of x are $n\alpha$ and $n\alpha(1-\alpha)$, respectively. If the number of trials becomes large at the same time that the success probability becomes small so that the mean $n\alpha$ is stable, then, the limiting form of the binomial distribution is the **Poisson distribution**,

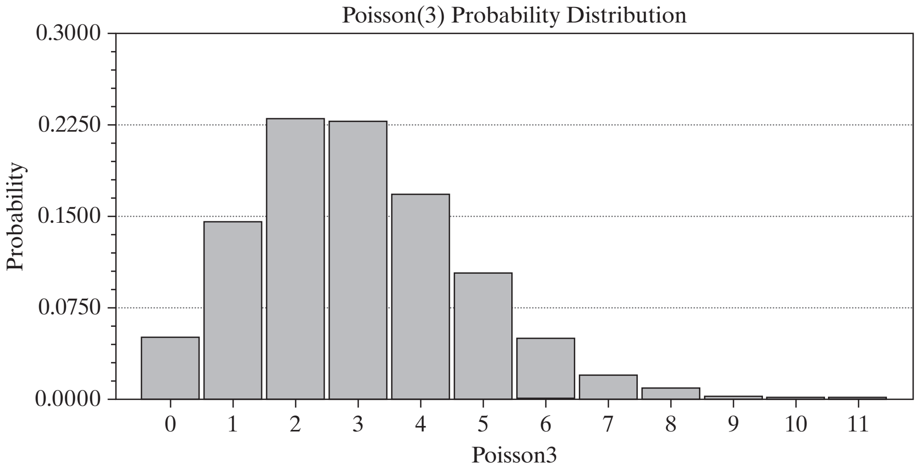
$$\text{Prob}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

The Poisson distribution has seen wide use in econometrics in, for example, modeling patents, crime, recreation demand, and demand for health services. (See Chapter 18.) An example is shown in Figure B.6.

B.5 THE DISTRIBUTION OF A FUNCTION OF A RANDOM VARIABLE

We considered finding the expected value of a function of a random variable. It is fairly common to analyze the random variable itself, which results when we compute a function of some random variable. There are three types of transformation to consider. One discrete random variable may be transformed into another, a continuous variable may be transformed into a discrete one, and one continuous variable may be transformed into another.

FIGURE B.6 The Poisson[3] Distribution.



The simplest case is the first one. The probabilities associated with the new variable are computed according to the laws of probability. If y is derived from x and the function is one to one, then the probability that $Y = y(x)$ equals the probability that $X = x$. If several values of x yield the same value of y, then Prob $(Y = y)$ is the sum of the corresponding probabilities for x.

The second type of transformation is illustrated by the way individual data on income are typically obtained in a survey. Income in the population can be expected to be distributed according to some skewed, continuous distribution such as the one shown in Figure B.7.

Data are often reported categorically, as shown in the lower part of the figure. Thus, the random variable corresponding to observed income is a discrete transformation of the actual underlying continuous random variable. Suppose, for example, that the transformed variable y is the mean income in the respective interval. Then

$$\begin{aligned}\text{Prob}(Y = \mu_1) &= P(-\infty < X \le a),\\ \text{Prob}(Y = \mu_2) &= P(a < X \le b),\\ \text{Prob}(Y = \mu_3) &= P(b < X \le c),\end{aligned}$$

and so on, which illustrates the general procedure.

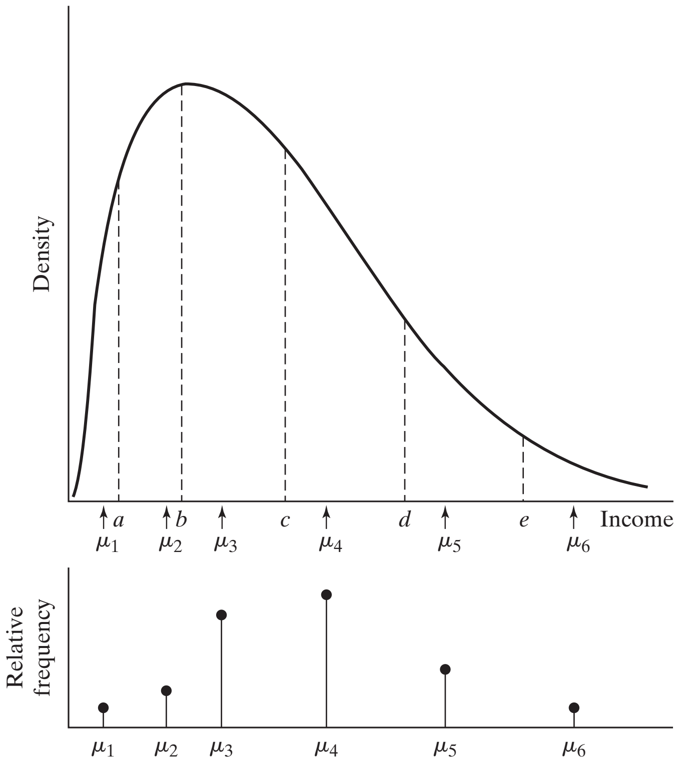
If x is a continuous random variable with pdf $f_x(x)$ and if $y = g(x)$ is a continuous monotonic function of x, then the density of y is obtained by using the change of variable technique to find the cdf of y:

$$\text{Prob}(y \le b) = \int_{-\infty}^{b} f_x|(g^{-1}(y))|g^{-1\prime}(y)|\,dy.$$

This equation can now be written as

$$\text{Prob}(y \le b) = \int_{-\infty}^{b} f_y|(y)\,dy.$$

FIGURE B.7 Censored Distribution.



Hence,

$$f_y(y) = f_x(g^{-1}(y))\left|g^{-1'}(y)\right|. \qquad \textbf{(B-41)}$$

To avoid the possibility of a negative pdf if $g(x)$ is decreasing, we use the absolute value of the derivative in the previous expression. The term $|g^{-1'}(y)|$ must be nonzero for the density of y to be nonzero. In words, the probabilities associated with intervals in the range of y must be associated with intervals in the range of x. If the derivative is zero, the correspondence $y = g(x)$ is vertical, and hence all values of y in the given range are associated with the same value of x. This single point must have probability zero.

One of the most useful applications of the preceding result is the linear transformation of a normally distributed variable. If $x \sim N[\mu, \sigma^2]$, then the distribution of

$$y = \frac{x - \mu}{\sigma}$$

is found using the preceding result. First, the derivative is obtained from the inverse transformation

$$y = \frac{x}{\sigma} - \frac{\mu}{\sigma} \Rightarrow x = \sigma y + \mu \Rightarrow f^{-1'}(y) = \frac{dx}{dy} = \sigma.$$

Therefore,

$$f_y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(\sigma y + \mu) - \mu]^2/(2\sigma^2)} |\sigma| = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

This is the density of a normally distributed variable with mean zero and unit standard deviation one. This is the result which makes it unnecessary to have separate tables for the different normal distributions which result from different means and variances.

B.6 REPRESENTATIONS OF A PROBABILITY DISTRIBUTION

The probability density function (pdf) is a natural and familiar way to formulate the distribution of a random variable. But, there are many other functions that are used to identify or characterize a random variable, depending on the setting. In each of these cases, we can identify some other function of the random variable that has a one-to-one relationship with the density. We have already used one of these quite heavily in the preceding discussion. For a random variable which has density function $f(x)$, the distribution function, $F(x)$, is an equally informative function that identifies the distribution; the relationship between $f(x)$ and $F(x)$ is defined in (B-6) for a discrete random variable and (B-8) for a continuous one. We now consider several other related functions.

For a continuous random variable, the **survival function** is $S(x) = 1 - F(x) = \text{Prob}[X \geq x]$. This function is widely used in epidemiology, where x is time until some transition, such as recovery from a disease. The **hazard function** for a random variable is

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)}.$$

The hazard function is a conditional probability;

$$h(x) = \lim_{t \downarrow 0} \text{Prob}(X \leq x \leq X + t | X \geq x).$$

Hazard functions have been used in econometrics in studying the duration of spells, or conditions, such as unemployment, strikes, time until business failures, and so on. The connection between the hazard and the other functions is $h(x) = -d \ln S(x)/dx$. As an exercise, you might want to verify the interesting special case of $h(x) = 1/\lambda$, a constant—the only distribution which has this characteristic is the exponential distribution noted in Section B.4.5.

For the random variable X, with probability density function $f(x)$, if the function

$$M(t) = E[e^{tx}]$$

exists, then it is the **moment generating function (MGF)**. Assuming the function exists, it can be shown that

$$d^r M(t)/dt^r |_{t=0} = E[x^r].$$

The moment generating function, like the survival and the hazard functions, is a unique characterization of a probability distribution. When it exists, the moment generating function has a one-to-one correspondence with the distribution. Thus, for example, if we begin with some random variable and find that a transformation of it has a particular MGF, then we may infer that the function of the random variable has the distribution associated with that MGF. A convenient application of this result is the MGF for the normal distribution. The MGF for the standard normal distribution is $M_z(t) = e^{t^2/2}$.

A useful feature of MGFs is the following:

If x and y are independent, then the MGF of $x + y$ is $M_x(t)M_y(t)$.

This result has been used to establish the **contagion** property of some distributions, that is, the property that sums of random variables with a given distribution have that same distribution. The normal distribution is a familiar example. This is usually not the case. It is for Poisson and chi-squared random variables.

One qualification of all of the preceding is that in order for these results to hold, the MGF must exist. It will for the distributions that we will encounter in our work, but in at least one important case, we cannot be sure of this. When computing sums of random variables which may have different distributions and whose specific distributions need not be so well behaved, it is likely that the MGF of the sum does not exist. However, the **characteristic function**,

$$\phi(t) = E[e^{itx}],\, i^2 = -1,$$

will always exist, at least for relatively small t. The characteristic function is the device used to prove that certain sums of random variables converge to a normally distributed variable—that is, the characteristic function is a fundamental tool in proofs of the central limit theorem.

B.7 JOINT DISTRIBUTIONS

The **joint density function** for two random variables X and Y denoted $f(x,y)$ is defined so that

$$\text{Prob}(a \le x \le b, c \le y \le d) = \begin{cases} \sum_{a \le x \le b} \sum_{c \le y \le d} f(x, y) & \text{if } x \text{ and } y \text{ are discrete,} \\ \int_a^b \int_c^d f(x, y)\, dy\, dx & \text{if } x \text{ and } y \text{ are continuous.} \end{cases} \tag{B-42}$$

The counterparts of the requirements for a univariate probability density are

$$\begin{aligned} & f(x, y) \ge 0, \\ & \sum_x \sum_y f(x, y) = 1 && \text{if } x \text{ and } y \text{ are discrete,} \\ & \int_x \int_y f(x, y)\, dy\, dx = 1 && \text{if } x \text{ and } y \text{ are continuous.} \end{aligned} \tag{B-43}$$

The cumulative probability is likewise the probability of a joint event:

$$F(x, y) = \text{Prob}(X \le x, Y \le y) = \begin{cases} \sum_{X \le x} \sum_{Y \le y} f(x, y) & \text{in the discrete case} \\ \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s)\, ds\, dt & \text{in the continuous case.} \end{cases} \tag{B-44}$$

B.7.1 MARGINAL DISTRIBUTIONS

A **marginal probability density** or marginal probability distribution is defined with respect to an individual variable. To obtain the marginal distributions from the joint density, it is necessary to sum or integrate out the other variable:

$$f_x(x) = \begin{cases} \sum_y f(x, y) & \text{in the discrete case} \\ \int_y f(x, s)\, ds & \text{in the continuous case,} \end{cases} \qquad \textbf{(B-45)}$$

and similarly for $f_y(y)$.

Two random variables are statistically independent if and only if their joint density is the product of the marginal densities:

$$f(x, y) = f_x(x)f_y(y) \Leftrightarrow x \text{ and } y \text{ are independent.} \qquad \textbf{(B-46)}$$

If (and only if) x and y are independent, then the cdf factors as well as the pdf:

$$F(x, y) = F_x(x)F_y(y), \qquad \textbf{(B-47)}$$

or

$$\text{Prob}(X \leq x, Y \leq y) = \text{Prob}(X \leq x)\text{Prob}(Y \leq y).$$

B.7.2 EXPECTATIONS IN A JOINT DISTRIBUTION

The means, variances, and higher moments of the variables in a joint distribution are defined with respect to the marginal distributions. For the mean of x in a discrete distribution,

$$\begin{aligned} E[x] &= \sum_x x f_x(x) \\ &= \sum_x x\left[\sum_y f(x, y)\right] \\ &= \sum_x \sum_y x f(x, y). \end{aligned} \qquad \textbf{(B-48)}$$

The means of the variables in a continuous distribution are defined likewise, using integration instead of summation:

$$\begin{aligned} E[x] &= \int_x x f_x(x)\, dx \\ &= \int_x \int_y x f(x, y)\, dy\, dx. \end{aligned} \qquad \textbf{(B-49)}$$

Variances are computed in the same manner:

$$\begin{aligned} \text{Var}[x] &= \sum_x (x - E[x])^2 f_x(x) \\ &= \sum_x \sum_y (x - E[x])^2 f(x, y). \end{aligned} \qquad \textbf{(B-50)}$$

B.7.3 COVARIANCE AND CORRELATION

For any function $g(x, y)$,

$$E[g(x,y)] = \begin{cases} \sum_x \sum_y g(x,y)f(x,y) & \text{in the discrete case} \\ \int_x \int_y g(x,y)f(x,y)\,dy\,dx & \text{in the continuous case.} \end{cases} \quad \textbf{(B-51)}$$

The covariance of x and y is a special case:

$$\begin{aligned} \text{Cov}[x,y] &= E[(x-\mu_x),(y-\mu_y)] \\ &= E[xy] - \mu_x\mu_y \\ &= \sigma_{xy}. \end{aligned} \quad \textbf{(B-52)}$$

If x and y are independent, then $f(x, y) = f_x(x)f_y(y)$ and

$$\begin{aligned} \sigma_{xy} &= \sum_x \sum_y f_x(x)f_y(y)(x-\mu_x)(y-\mu_y) \\ &= \sum_x (x-\mu_x)f_x(x)\sum_y (y-\mu_y)f_y(y) \\ &= E[x-\mu_x]E[y-\mu_y] \\ &= 0. \end{aligned}$$

The sign of the covariance will indicate the direction of covariation of X and Y. Its magnitude depends on the scales of measurement, however. In view of this fact, a preferable measure is the correlation coefficient:

$$r[x,y] = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \quad \textbf{(B-53)}$$

where σ_x and σ_y are the standard deviations of x and y, respectively. The correlation coefficient has the same sign as the covariance but is always between -1 and 1 and is thus unaffected by any scaling of the variables.

Variables that are uncorrelated are not necessarily independent. For example, in the discrete distribution $f(-1,1) = f(0,0) = f(1,1) = \frac{1}{3}$, the correlation is zero, but $f(1, 1)$ does not equal $f_x(1)f_y(1) = (\frac{1}{3})(\frac{2}{3})$. An important exception is the joint normal distribution discussed subsequently, in which lack of correlation does imply independence.

Some general results regarding expectations in a joint distribution, which can be verified by applying the appropriate definitions, are

$$E[ax + by + c] = a\,E[x] + bE[y] + c, \quad \textbf{(B-54)}$$

$$\begin{aligned} \text{Var}[ax + by + c] &= a^2\,\text{Var}[x] + b^2\text{Var}[y] + 2ab\,\text{Cov}[x,y] \\ &= \text{Var}[ax + by], \end{aligned} \quad \textbf{(B-55)}$$

and

$$\text{Cov}[ax + by, cx + dy] = ac\,\text{Var}[x] + bd\,\text{Var}[y] + (ad + bc)\text{Cov}[x, y]. \qquad \textbf{(B-56)}$$

If X and Y are uncorrelated, then

$$\begin{aligned}\text{Var}[x + y] &= \text{Var}[x - y]\\ &= \text{Var}[x] + \text{Var}[y]. \qquad \textbf{(B-57)}\end{aligned}$$

For any two functions $g_1(x)$ and $g_2(y)$, if x and y are independent, then

$$E[g_1(x)g_2(y)] = E[g_1(x)]E[g_2(y)]. \qquad \textbf{(B-58)}$$

B.7.4 DISTRIBUTION OF A FUNCTION OF BIVARIATE RANDOM VARIABLES

The result for a function of a random variable in (B-41) must be modified for a joint distribution. Suppose that x_1 and x_2 have a joint distribution $f_x(x_1, x_2)$ and that y_1 and y_2 are two monotonic functions of x_1 and x_2:

$$y_1 = y_1(x_1, x_2), y_2 = y_2(x_1, x_2).$$

Because the functions are monotonic, the inverse transformations,

$$x_1 = x_1(y_1, y_2), x_2 = x_2(y_1, y_2),$$

exist. The Jacobian of the transformations is the matrix of partial derivatives,

$$\mathbf{J} = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} = \left[\frac{\partial \mathbf{x}}{\partial \mathbf{y}'}\right].$$

The joint distribution of y_1 and y_2 is

$$f_y(y_1, y_2) = f_x[x_1(y_1, y_2), x_2(y_1, y_2)]\text{abs}(|\mathbf{J}|).$$

The determinant of the Jacobian must be nonzero for the transformation to exist. A zero determinant implies that the two transformations are functionally dependent.

Certainly the most common application of the preceding in econometrics is the linear transformation of a set of random variables. Suppose that x_1 and x_2 are independently distributed $N[0, 1]$, and the transformations are

$$\begin{aligned} y_1 &= \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2,\\ y_2 &= \alpha_2 + \beta_{21}x_1 + \beta_{22}x_2. \end{aligned}$$

To obtain the joint distribution of y_1 and y_2, we first write the transformations as

$$\mathbf{y} = \mathbf{a} + \mathbf{Bx}.$$

The inverse transformation is

$$\mathbf{x} = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{a}),$$

so the absolute value of the determinant of the Jacobian is

$$\text{abs}\,|\mathbf{J}| = \text{abs}\,|\mathbf{B}^{-1}| = \frac{1}{\text{abs}|\mathbf{B}|}.$$

The joint distribution of **x** is the product of the marginal distributions since they are independent. Thus,

$$f_x(\mathbf{x}) = (2\pi)^{-1}\, e^{-(x_1{}^2 + x_2{}^2)/2} = (2\pi)^{-1} e^{-\mathbf{x}'\mathbf{x}/2}.$$

Inserting the results for $\mathbf{x}(\mathbf{y})$ and J into $f_y(y_1, y_2)$ gives

$$f_y(\mathbf{y}) = (2\pi)^{-1} \frac{1}{\text{abs}\,|\mathbf{B}|} e^{-(\mathbf{y}-\mathbf{a})'(\mathbf{BB}')^{-1}(\mathbf{y}-\mathbf{a})/2}.$$

This **bivariate normal distribution** is the subject of Section B.9. Note that by formulating it as we did earlier, we can generalize easily to the multivariate case, that is, with an arbitrary number of variables.

Perhaps the more common situation is that in which it is necessary to find the distribution of one function of two (or more) random variables. A strategy that often works in this case is to form the joint distribution of the transformed variable and one of the original variables, then integrate (or sum) the latter out of the joint distribution to obtain the marginal distribution. Thus, to find the distribution of $y_1(x_1, x_2)$, we might formulate

$$y_1 = y_1(x_1, x_2)$$
$$y_2 = x_2.$$

The absolute value of the determinant of the Jacobian would then be

$$\mathbf{J} = \text{abs}\begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ 0 & 1 \end{vmatrix} = \text{abs}\left|\frac{\partial x_1}{\partial y_1}\right|.$$

The density of y_1 would then be

$$f_{y_1}(y_1) = \int_{y_2} f_x[x_1(y_1, y_2), y_2]\ \text{abs}\,|\mathbf{J}|\ dy_2.$$

B.8 CONDITIONING IN A BIVARIATE DISTRIBUTION

Conditioning and the use of conditional distributions play a pivotal role in econometric modeling. We consider some general results for a bivariate distribution. (All these results can be extended directly to the multivariate case.)

In a bivariate distribution, there is a **conditional distribution** over y for each value of x. The conditional densities are

$$f(y\,|\,x) = \frac{f(x, y)}{f_x(x)}, \qquad \textbf{(B-59)}$$

and

$$f(x\,|\,y) = \frac{f(x, y)}{f_y(y)}.$$

It follows from (B-46) that.

$$\text{If } x \text{ and } y \text{ are independent, then } f(y\,|\,x) = f_y(y) \text{ and } f(x\,|\,y) = f_x(x). \tag{B-60}$$

The interpretation is that if the variables are independent, the probabilities of events relating to one variable are unrelated to the other. The definition of conditional densities implies the important result

$$f(x, y) = f(y\,|\,x)f_x(x) \;= f(x\,|\,y)f_y(y). \tag{B-61}$$

B.8.1 REGRESSION: THE CONDITIONAL MEAN

A **conditional mean** is the mean of the conditional distribution and is defined by

$$E[y\,|\,x] = \begin{cases} \int_y yf(y\,|\,x)dy & \text{if } y \text{ is continuous} \\ \sum_y yf(y\,|\,x) & \text{if } y \text{ is discrete.} \end{cases} \tag{B-62}$$

The conditional mean function $E[y\,|\,x]$ is called the **regression** of y on x.

A random variable may always be written as

$$\begin{aligned} y &= E[y\,|\,x] + (y - E[y\,|\,x]) \\ &= E[y\,|\,x] + \varepsilon. \end{aligned}$$

B.8.2 CONDITIONAL VARIANCE

A conditional variance is the variance of the conditional distribution:

$$\begin{aligned} \text{Var}[y\,|\,x] &= E[(y - E[y\,|\,x])^2\,|\,x] \\ &= \int_y (y - E[y\,|\,x])^2 f(y\,|\,x)dy, \quad \text{if } y \text{ is continuous,} \end{aligned} \tag{B-63}$$

or

$$\text{Var}[y\,|\,x] = \sum_y (y - E[y\,|\,x])^2 f(y\,|\,x) \quad \text{if } y \text{ is discrete.} \tag{B-64}$$

The computation can be simplified by using

$$\text{Var}[y\,|\,x] = E[y^2\,|\,x] - (E[y\,|\,x])^2. \tag{B-65}$$

The conditional variance is called the **scedastic function** and, like the regression, is generally a function of x. Unlike the conditional mean function, however, it is common for the conditional variance not to vary with x. We shall examine a particular case. This case does not imply, however, that Var$[y\,|\,x]$ equals Var$[y]$, which will usually not be true. It implies only that the conditional variance is a constant. The case in which the conditional variance does not vary with x is called **homoscedasticity** (same variance).

B.8.3 RELATIONSHIPS AMONG MARGINAL AND CONDITIONAL MOMENTS

Some useful results for the moments of a conditional distribution are given in the following theorems.

THEOREM B.1 Law of Iterated Expectations

$$E[y] = E_x[E[y \mid x]]. \qquad \textbf{(B-66)}$$

The notation $E_x[.]$ indicates the expectation over the values of x. Note that $E[y \mid x]$ is a function of x.

THEOREM B.2 Covariance

In any bivariate distribution,

$$\text{Cov}[x, y] = \text{Cov}_x[x, E[y \mid x]] = \int_x (x - E[x])\, E[y \mid x] f_x(x)\, dx. \qquad \textbf{(B-67)}$$

(Note that this is the covariance of x and a function of x.)

The preceding results provide an additional, extremely useful result for the special case in which the conditional mean function is linear in *x*.

THEOREM B.3 Moments in a Linear Regression

If $E[y \mid x] = \alpha + \beta x$, then

$$\alpha = E[y] - \beta E[x]$$

and

$$\beta = \frac{\text{Cov}[x,y]}{\text{Var}[x]}. \qquad \textbf{(B-68)}$$

The proof follows from (*B*-66). *Whether* E[y|x] *is nonlinear or linear, the result in* (*B*-68) *is the linear projection of y on x. The linear projection is developed in Section B*.8.5.

The preceding theorems relate to the conditional mean in a bivariate distribution. The following theorems, which also appear in various forms in regression analysis, describe the conditional variance.

THEOREM B.4 Decomposition of Variance

In a joint distribution,

$$\text{Var}[y] = \text{Var}_x[E[y \mid x]] + E_x[\text{Var}[y \mid x]]. \qquad \textbf{(B-69)}$$

The notation $\text{Var}_x[.]$ indicates the variance over the distribution of x. This equation states that in a bivariate distribution, the variance of y decomposes into the variance of the conditional mean function plus the expected variance around the conditional mean.

THEOREM B.5 Residual Variance in a Regression
In any bivariate distribution,

$$E_x[\text{Var}[y|x]] = \text{Var}[y] - \text{Var}_x[E[y|x]]. \qquad \textbf{(B-70)}$$

On average, conditioning reduces the variance of the variable subject to the conditioning. For example, if y is homoscedastic, then we have the unambiguous result that the variance of the conditional distribution(s) is less than or equal to the unconditional variance of y. Going a step further, we have the result that appears prominently in the bivariate normal distribution (Section B.9).

THEOREM B.6 Linear Regression and Homoscedasticity
In a bivariate distribution, if $E[y|x] = \alpha + \beta x$ *and if* $\text{Var}[y|x]$ *is a constant, then*

$$\text{Var}[y|x] = \text{Var}[y](1 - \text{Corr}^2[y, x]) = \sigma_y^2(1 - \rho_{xy}^2). \qquad \textbf{(B-71)}$$

The proof is straightforward using Theorems B.2 to B.4.

B.8.4 THE ANALYSIS OF VARIANCE

The variance decomposition result implies that in a bivariate distribution, variation in y arises from two sources:

1. Variation because $E[y|x]$ varies with x:

$$\textbf{regression variance} = \text{Var}_x[E[y|x]]. \qquad \textbf{(B-72)}$$

2. Variation because, in each conditional distribution, y varies around the conditional mean:

$$\textbf{residual variance} = E_x[\text{Var}[y|x]]. \qquad \textbf{(B-73)}$$

Thus,

$$\text{Var}[y] = \text{regression variance} + \text{residual variance}. \qquad \textbf{(B-74)}$$

In analyzing a regression, we shall usually be interested in which of the two parts of the total variance, $\text{Var}[y]$, is the larger one. A natural measure is the ratio

$$\textbf{coefficient of determination} = \frac{\text{regression variance}}{\text{total variance}}. \qquad \textbf{(B-75)}$$

In the setting of a linear regression, (B-75) arises from another relationship that emphasizes the interpretation of the correlation coefficient.

$$\text{If } E[y|x] = \alpha + \beta x, \quad \text{then the coefficient of determination} = \text{COD} = \rho^2, \qquad \textbf{(B-76)}$$

where ρ^2 is the squared correlation between x and y. We conclude that the correlation coefficient (squared) is a measure of the proportion of the variance of y accounted for by variation in the mean of y given x. It is in this sense that correlation can be interpreted as a **measure of linear association** between two variables.

B.8.5 LINEAR PROJECTION

Theorems B.3 (Moments in a Linear Regression) and B.6 (Linear Regression and Homoscedasticity) begin with an assumption that $E[y|x] = \alpha + \beta x$. If the conditional mean is not linear, then the results in THEOREM B.6 do not give the slopes in the conditional mean. However, in a bivariate distribution, we can always define the linear projection of y on x, as

$$Proj(y|x) = \gamma_0 + \gamma_1 x$$

where

$$\gamma_0 = E[y] - \gamma_1 E[x] \text{ and } \gamma_1 = \text{Cov}(x,y)/\text{Var}(x).$$

We can see immediately in THEOREM B.3 that if the conditional mean function is linear, then the conditional mean function (the regression of y on x) is also the linear projection. When the conditional mean function is not linear, then the regression and the projection functions will be different. We consider an example that bears some connection to the formulation of loglinear models. If

$$y|x \sim \text{Poisson with conditional mean function } \exp(\beta x), y = 0, 1, \ldots,$$

$$x \sim \text{U}[0,1]; f(x) = 1, 0 \le x \le 1,$$

$$f(x,y) = f(y|x)f(x) = \exp[\text{-}\exp(\beta x)][\exp(\beta x)]^y/y! \times 1,$$

Then, as noted, the conditional mean function is nonlinear; $E[y|x] = \exp(\beta x)$. The slope in the projection of y on x is $\gamma_1 = \text{Cov}(x,y)/\text{Var}[x] = \text{Cov}(x, E[y|x])/\text{Var}[x] = \text{Cov}(x,\exp(\beta x))/\text{Var}[x]$. (THEOREM B.2.) We have $E[x] = 1/2$ and $\text{Var}[x] = 1/12$. To obtain the covariance, we require

$$E[x\exp(\beta x)] = \int_0^1 x \exp(\beta x)dx = \left[\left(\frac{x}{\beta} - \frac{1}{\beta^2}\right)\exp(\beta x)\right]_{x=0}^{x=1}$$

and

$$E[x]E[\exp(\beta x)] = \left(\frac{1}{2}\right)\int_0^1 \exp(\beta x)dx = \left(\frac{1}{2}\right)\left[\frac{\exp(\beta x)}{\beta}\right]_{x=0}^{x=1} = \left(\frac{1}{2}\right)\left[\frac{\exp(\beta) - 1}{\beta}\right].$$

After collecting terms, $\gamma_1 = h(\beta)$. The constant is $\gamma_0 = E[y] - h(\beta)(1/2)$. $E[y] = E[E[y|x]] = [\exp(\beta)\text{-}1]/\beta$. (THEOREM B.1.) Then, the projection is the linear function $\gamma_0 + \gamma_1 x$ while the regression function is the nonlinear function $\exp(\beta x)$. The projection can be viewed as a linear approximation to the conditional mean. (Note, it is not a linear Taylor series approximation.)

In similar fashion to Theorem B.5, we can define the variation around the projection,

$$Proj.Var[y\,|\,x] = E_x[\{y - Proj(y\,|\,x)\}^2\,|\,x].$$

By adding and subtracting the regression, $E[y\,|\,x]$, in the expression, we find

$$Proj.Var[y\,|\,x] = \mathrm{Var}[y\,|\,x] + E_x[\{Proj(y\,|\,x) - E[y\,|\,x]\}^2\,|\,x].$$

This states that the variation of y around the projection consists of the regression variance plus the expected squared approximation error of the projection. As a general observation, we find, not surprisingly, that when the conditional mean is not linear, the projection does not do as well as the regression at prediction of y.

B.9 THE BIVARIATE NORMAL DISTRIBUTION

A bivariate distribution that embodies many of the features described earlier is the **bivariate normal**, which is the joint distribution of two normally distributed variables. The density is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} e^{-1/2[(\varepsilon_x{}^2 + \varepsilon_y{}^2 - 2\rho\varepsilon_x\varepsilon_y)/(1 - \rho^2)]},$$

$$\varepsilon_x = \frac{x - \mu_x}{\sigma_x}, \quad \varepsilon_y = \frac{y - \mu_y}{\sigma_y}. \qquad \textbf{(B-77)}$$

The parameters μ_x, σ_x, μ_y, and σ_y are the means and standard deviations of the marginal distributions of x and y, respectively. The additional parameter ρ is the correlation between x and y. The covariance is

$$\sigma_{xy} = \rho\sigma_x\sigma_y. \qquad \textbf{(B-78)}$$

The density is defined only if ρ is not 1 or -1, which in turn requires that the two variables not be linearly related. If x and y have a bivariate normal distribution, denoted

$$(x, y) \sim N_2[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho],$$

then

- The marginal distributions are normal:

$$f_x(x) = N[\mu_x, \sigma_x^2],$$
$$f_y(y) = N[\mu_y, \sigma_y^2]. \qquad \textbf{(B-79)}$$

- The conditional distributions are normal:

$$f(y\,|\,x) = N[\alpha + \beta x, \sigma_y^2(1 - \rho^2)],$$

$$\alpha = \mu_y - \beta\mu_x \quad \beta = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad \textbf{(B-80)}$$

  and likewise for $f(x\,|\,y)$.
- x and y are independent if and only if $\rho = 0$. The density factors into the product of the two marginal normal distributions if $\rho = 0$.

Two things to note about the conditional distributions beyond their normality are their linear regression functions and their constant conditional variances. The conditional variance is less than the unconditional variance, which is consistent with the results of the previous section.

B.10 MULTIVARIATE DISTRIBUTIONS

The extension of the results for bivariate distributions to more than two variables is direct. It is made much more convenient by using matrices and vectors. The term **random vector** applies to a vector whose elements are random variables. The joint density is $f(\mathbf{x})$, whereas the cdf is

$$F(\mathbf{x}) = \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} f(\mathbf{t})dt_1\cdots dt_{n-1}\,dt_n. \qquad \textbf{(B-81)}$$

Note that the cdf is an n-fold integral. The marginal distribution of any one (or more) of the n variables is obtained by integrating or summing over the other variables.

B.10.1 MOMENTS

The expected value of a vector or matrix is the vector or matrix of expected values. A mean vector is defined as

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} E[x_1] \\ E[x_2] \\ \vdots \\ E[x_n] \end{bmatrix} = E[\mathbf{x}]. \qquad \textbf{(B-82)}$$

Define the matrix

$$(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' = \begin{bmatrix} (x_1-\mu_1)(x_1-\mu_1) & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1),(x_n-\mu_n) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)(x_2-\mu_2) & \cdots & (x_2-\mu_2)(x_n-\mu_n) \\ \vdots & & \vdots & \\ (x_n-\mu_n)(x_1-\mu_1) & (x_n-\mu_n)(x_2-\mu_2) & \cdots & (x_n-\mu_n)(x_n-\mu_n) \end{bmatrix}.$$

The expected value of each element in the matrix is the covariance of the two variables in the product. (The covariance of a variable with itself is its variance.) Thus,

$$E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & & \vdots & \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix} = E[\mathbf{x}\mathbf{x}'] - \boldsymbol{\mu}\boldsymbol{\mu}', \qquad \textbf{(B-83)}$$

which is the **covariance matrix** of the random vector **x**. Henceforth, we shall denote the covariance matrix of a random vector in boldface, as in

$$\text{Var}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

By dividing σ_{ij} by $\sigma_i\sigma_j$, we obtain the **correlation matrix**:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \cdots & 1 \end{bmatrix}.$$

B.10.2 SETS OF LINEAR FUNCTIONS

Our earlier results for the mean and variance of a linear function can be extended to the multivariate case. For the mean,

$$\begin{aligned} E[a_1x_1 + a_2x_2 + \cdots + a_nx_n] &= E[\mathbf{a}'\mathbf{x}] \\ &= a_1\, E[x_1] + a_2E[x_2] + \cdots + a_nE[x_n] \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \\ &= \mathbf{a}', \boldsymbol{\mu}. \qquad \textbf{(B-84)} \end{aligned}$$

For the variance,

$$\begin{aligned} \text{Var}[\mathbf{a}'\mathbf{x}] &= E[(\mathbf{a}'\mathbf{x} - E[\mathbf{a}'\mathbf{x}])^2] \\ &= E[\{\mathbf{a}'(\mathbf{x} - E[\mathbf{x}])\}^2] \\ &= E[\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\,\mathbf{a}] \end{aligned}$$

as $E[\mathbf{x}] = \boldsymbol{\mu}$ and $\mathbf{a}'(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{a}$. Because $\mathbf{a}$ is a vector of constants,

$$\text{Var}[\mathbf{a}'\mathbf{x}] = \mathbf{a}'E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})']\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\sigma_{ij} \qquad \textbf{(B-85)}$$

It is the expected value of a square, so we know that a variance cannot be negative. As such, the preceding quadratic form is nonnegative, and the symmetric matrix $\boldsymbol{\Sigma}$ must be nonnegative definite.

In the set of linear functions $\mathbf{y} = \mathbf{Ax}$, the ith element of $\mathbf{y}$ is $y_i = \mathbf{a}_i\mathbf{x}$, where $\mathbf{a}_i$ is the ith row of $\mathbf{A}$ [see result (A-14)]. Therefore,

$$E[y_i] = \mathbf{a}_i\boldsymbol{\mu}.$$

Collecting the results in a vector, we have

$$E[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\mu}. \qquad \textbf{(B-86)}$$

For two row vectors $\mathbf{a}_i$ and $\mathbf{a}_j$,

$$\text{Cov}[\mathbf{a}_i\mathbf{x}, \mathbf{a}_j\mathbf{x}] = \mathbf{a}_i, \boldsymbol{\Sigma}\mathbf{a}_j'.$$

Because $\mathbf{a}_i\,\boldsymbol{\Sigma}\mathbf{a}_j'$ is the ijth element of $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$,

$$\text{Var}[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \qquad \textbf{(B-87)}$$

This matrix will be either nonnegative definite or positive definite, depending on the column rank of $\mathbf{A}$.

B.10.3 NONLINEAR FUNCTIONS: THE DELTA METHOD

Consider a set of possibly nonlinear functions of $\mathbf{x}$, $\mathbf{y} = \mathbf{g}(\mathbf{x})$. Each element of $\mathbf{y}$ can be approximated with a linear Taylor series. Let $\mathbf{j}^i$ be the row vector of partial derivatives of the i th function with respect to the n elements of $\mathbf{x}$:

$$\mathbf{j}^i(\mathbf{x}) = \frac{\partial g_i(\mathbf{x})}{\partial \mathbf{x}'} = \frac{\partial y_i}{\partial \mathbf{x}'}. \tag{B-88}$$

Then, proceeding in the now familiar way, we use $\boldsymbol{\mu}$, the mean vector of $\mathbf{x}$, as the expansion point, so that $\mathbf{j}^i(\boldsymbol{\mu})$ is the row vector of partial derivatives evaluated at $\boldsymbol{\mu}$. Then

$$g_i(\mathbf{x}) \approx g_i(\boldsymbol{\mu}) + \mathbf{j}^i(\boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}). \tag{B-89}$$

From this we obtain

$$E[g_i(\mathbf{x})] \approx g_i(\boldsymbol{\mu}), \tag{B-90}$$

$$\text{Var}[g_i(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{j}^i(\boldsymbol{\mu})', \tag{B-91}$$

and

$$\text{Cov}[g_i(\mathbf{x}), g_j(\mathbf{x})] \approx \mathbf{j}^i(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{j}^j(\boldsymbol{\mu})'. \tag{B-92}$$

These results can be collected in a convenient form by arranging the row vectors $\mathbf{j}^i(\boldsymbol{\mu})$ in a matrix $\mathbf{J}(\boldsymbol{\mu})$. Then, corresponding to the preceding equations, we have

$$E[\mathbf{g}(\mathbf{x})] \simeq \mathbf{g}(\boldsymbol{\mu}), \tag{B-93}$$

$$\text{Var}[\mathbf{g}(\mathbf{x})] \simeq \mathbf{J}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{J}(\boldsymbol{\mu})'. \tag{B-94}$$

The matrix $\mathbf{J}(\boldsymbol{\mu})$ in the last preceding line is $\partial\mathbf{y}/\partial\mathbf{x}'$ evaluated at $\mathbf{x} = \boldsymbol{\mu}$.

B.11 THE MULTIVARIATE NORMAL DISTRIBUTION

The foundation of most multivariate analysis in econometrics is the multivariate normal distribution. Let the vector $(x_1, x_2, \ldots, x_n)' = \mathbf{x}$ be the set of n random variables, $\boldsymbol{\mu}$ their mean vector, and $\boldsymbol{\Sigma}$ their covariance matrix. The general form of the joint density is

$$f(\mathbf{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}e^{(-1/2)(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{B-95}$$

If $\mathbf{R}$ is the correlation matrix of the variables and $\mathbf{R}_{ij} = \sigma_{ij}/(\sigma_i\sigma_j)$, then

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma_1\sigma_2\cdots\sigma_n)^{-1}|\mathbf{R}|^{-1/2}\,e^{(-1/2)\boldsymbol{\varepsilon}\mathbf{R}^{-1}\boldsymbol{\varepsilon}}, \tag{B-96}$$

where $\varepsilon_i = (x_i - \mu_i)/\sigma_i$.[7]

[7] This result is obtained by constructing $\boldsymbol{\Delta}$, the diagonal matrix with σ_i as its ith diagonal element. Then, $\mathbf{R} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}$, which implies that $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Delta}^{-1}\mathbf{R}^{-1}\boldsymbol{\Delta}^{-1}$. Inserting this in (B-95) yields (B-96). Note that the ith element of $\boldsymbol{\Delta}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is $(x_i - \mu_i)/\sigma_i$.

Two special cases are of interest. If all the variables are uncorrelated, then $\rho_{ij} = 0$ for $i \neq j$. Thus, $\mathbf{R} = \mathbf{I}$, and the density becomes

$$\begin{aligned} f(\mathbf{x}) &= (2\pi)^{-n/2}(\sigma_1\sigma_2 \cdots \sigma_n)^{-1}e^{-\varepsilon'\varepsilon/2} \\ &= f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \end{aligned} \qquad \textbf{(B-97)}$$

As in the bivariate case, if normally distributed variables are uncorrelated, then they are independent. If $\sigma_i = \sigma$ and $\boldsymbol{\mu} = \mathbf{0}$, then $x_i \sim N[0, \sigma^2]$ and $\varepsilon_i = x_i/\sigma$, and the density becomes

$$f(\mathbf{x}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\mathbf{x}'\mathbf{x}/(2\sigma^2)}. \qquad \textbf{(B-98)}$$

Finally, if $\sigma = 1$,

$$f(\mathbf{x}) = (2\pi)^{-n/2}e^{-\mathbf{x}'\mathbf{x}/2}. \qquad \textbf{(B-99)}$$

This distribution is the **multivariate standard normal**, or **spherical normal distribution**.

B.11.1 MARGINAL AND CONDITIONAL NORMAL DISTRIBUTIONS

Let $\mathbf{x}_1$ be any subset of the variables, including a single variable, and let $\mathbf{x}_2$ be the remaining variables. Partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ likewise so that

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Then the marginal distributions are also normal. In particular, we have the following theorem.

THEOREM B.7 Marginal and Conditional Normal Distributions

If $[\mathbf{x}_1, \mathbf{x}_2]$ *have a joint multivariate normal distribution, then the marginal distributions are*

$$\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \qquad \textbf{(B-100)}$$

and

$$\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}). \qquad \textbf{(B-101)}$$

The conditional distribution of $\mathbf{x}_1$ *given* $\mathbf{x}_2$ *is normal as well:*

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N(\boldsymbol{\mu}_{1.2}, \boldsymbol{\Sigma}_{11.2}), \qquad \textbf{(B-102)}$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \qquad \textbf{(B-102a)}$$

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \qquad \textbf{(B-102b)}$$

THEOREM B.7 (continued)

Proof: *We partition* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *as shown earlier and insert the parts in* (*B*-95). *To construct the density, we use* (*A*-72) *to partition the determinant,*

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}|,$$

and (*A*-74) *to partition the inverse,*

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11.2}^{-1} & -\boldsymbol{\Sigma}_{11.2}^{-1}\mathbf{B} \\ -\mathbf{B}'\boldsymbol{\Sigma}_{11.2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \mathbf{B}'\boldsymbol{\Sigma}_{11.2}^{-1}\mathbf{B} \end{bmatrix}.$$

For simplicity, we let

$$\mathbf{B} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

Inserting these in (*B*-95) *and collecting terms produces the joint density as a product of two terms:*

$$f(\mathbf{x}_1, \mathbf{x}_2) = f_{1.2}(\mathbf{x}_1|\mathbf{x}_2)f_2(\mathbf{x}_2).$$

The first of these is a normal distribution with mean $\boldsymbol{\mu}_{1.2}$ *and variance* $\boldsymbol{\Sigma}_{11.2}$, *whereas the second is the marginal distribution of* $\mathbf{x}_2$.

The conditional mean vector in the multivariate normal distribution is a linear function of the unconditional mean and the conditioning variables, and the conditional covariance matrix is constant and is smaller (in the sense discussed in Section A.7.3) than the unconditional covariance matrix. Notice that the conditional covariance matrix is the inverse of the upper left block of $\boldsymbol{\Sigma}^{-1}$; that is, this matrix is of the form shown in (A-74) for the partitioned inverse of a matrix.

B.11.2 THE CLASSICAL NORMAL LINEAR REGRESSION MODEL

An important special case of the preceding is that in which $\mathbf{x}_1$ is a single variable, y, and $\mathbf{x}_2$ is K variables, $\mathbf{x}$. Then the conditional distribution is a multivariate version of that in (B-80) with $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\sigma_{\mathbf{xy}}$, where $\sigma_{\mathbf{xy}}$ is the vector of covariances of y with $\mathbf{x}_2$. Recall that any random variable, y, can be written as its mean plus the deviation from the mean. If we apply this tautology to the multivariate normal, we obtain

$$y = E[y|\mathbf{x}] + (y - E[y|\mathbf{x}]) = \alpha + \boldsymbol{\beta}'\mathbf{x} + \varepsilon,$$

where $\boldsymbol{\beta}$ is given earlier, $\alpha = \mu_y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{x}}$, and ε has a normal distribution. We thus have, in this multivariate normal distribution, the **classical normal linear regression model**.

B.11.3 LINEAR FUNCTIONS OF A NORMAL VECTOR

Any linear function of a vector of joint normally distributed variables is also normally distributed. The mean vector and covariance matrix of $\mathbf{Ax}$, where $\mathbf{x}$ is normally distributed, follow the general pattern given earlier. Thus,

$$\text{If } \mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}], \quad \text{then } \mathbf{Ax} + \mathbf{b} \sim N[\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}']. \qquad \textbf{(B-103)}$$

If $\mathbf{A}$ does not have full rank, then $\mathbf{A\Sigma A}'$ is singular and the density does not exist in the full dimensional space of $\mathbf{x}$ although it does exist in the subspace of dimension equal to the rank of $\mathbf{\Sigma}$. Nonetheless, the individual elements of $\mathbf{Ax} + \mathbf{b}$ will still be normally distributed, and the joint *distribution* of the full vector is still a multivariate normal.

B.11.4 QUADRATIC FORMS IN A STANDARD NORMAL VECTOR

The earlier discussion of the chi-squared distribution gives the distribution of $\mathbf{x}'\mathbf{x}$ if $\mathbf{x}$ has a standard normal distribution. It follows from (A-36) that

$$\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}^2. \qquad \textbf{(B-104)}$$

We know from (B-32) that $\mathbf{x}'\mathbf{x}$ has a chi-squared distribution. It seems natural, therefore, to invoke (B-34) for the two parts on the right-hand side of (B-104). It is not yet obvious, however, that either of the two terms has a chi-squared distribution or that the two terms are independent, as required. To show these conditions, it is necessary to derive the distributions of **idempotent quadratic forms** and to show when they are independent.

To begin, the second term is the square of $\sqrt{n}\,\bar{x}$, which can easily be shown to have a standard normal distribution. Thus, the second term is the square of a standard normal variable and has chi-squared distribution with one degree of freedom. But the first term is the sum of n nonindependent variables, and it remains to be shown that the two terms are independent.

DEFINITION B.3 Orthonormal Quadratic Form
A particular case of (*B*-103) *is the following:*

If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{C}$ *is a square matrix such that* $\mathbf{C}'\mathbf{C} = \mathbf{I}$, *then* $\mathbf{C}'\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$.

Consider, then, a quadratic form in a standard normal vector $\mathbf{x}$ with symmetric matrix $\mathbf{A}$:

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}. \qquad \textbf{(B-105)}$$

Let the characteristic roots and vectors of $\mathbf{A}$ be arranged in a diagonal matrix $\mathbf{\Lambda}$ and an orthogonal matrix $\mathbf{C}$, as in Section A.6.3. Then

$$q = \mathbf{x}'\mathbf{C\Lambda C}'\mathbf{x}. \qquad \textbf{(B-106)}$$

By definition, $\mathbf{C}$ satisfies the requirement that $\mathbf{C}'\mathbf{C} = \mathbf{I}$. Thus, the vector $\mathbf{y} = \mathbf{C}'\mathbf{x}$ has a standard normal distribution. Consequently,

$$q = \mathbf{y}'\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{n}\lambda_i y_i^2. \qquad \textbf{(B-107)}$$

If λ_i is always one or zero, then

$$q = \sum_{j=1}^{J} y_j^2, \qquad \textbf{(B-108)}$$

which has a chi-squared distribution. The sum is taken over the $j = 1, \dots, J$ elements associated with the roots that are equal to one. A matrix whose characteristic roots are all zero or one is idempotent. Therefore, we have proved the next theorem.

THEOREM B.8 Distribution of an Idempotent Quadratic Form in a Standard Normal Vector
If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{A}$ *is idempotent, then* $\mathbf{x}'\mathbf{A}\mathbf{x}$ *has a chi-squared distribution with degrees of freedom equal to the number of unit roots of* $\mathbf{A}$*, which is equal to the rank of* $\mathbf{A}$.

The rank of a matrix is equal to the number of nonzero characteristic roots it has. Therefore, the degrees of freedom in the preceding chi-squared distribution equals J, the rank of $\mathbf{A}$.

We can apply this result to the earlier sum of squares. The first term is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathbf{x}'\mathbf{M}^0\mathbf{x},$$

where $\mathbf{M}^0$ was defined in (A-34) as the matrix that transforms data to mean deviation form:

$$\mathbf{M}^0 = \mathbf{I} - \frac{1}{n}\mathbf{i}\mathbf{i}'.$$

Because $\mathbf{M}^0$ is idempotent, the sum of squared deviations from the mean has a chi-squared distribution. The degrees of freedom equals the rank $\mathbf{M}^0$, which is not obvious except for the useful result in (A-108), that

- The rank of an idempotent matrix is equal to its trace. **(B-109)**
  Each diagonal element of $\mathbf{M}^0$ is $1 - (1/n)$; hence, the trace is $n[1 - (1/n)] = n - 1$. Therefore, we have an application of Theorem B.8.

$$\text{If } \mathbf{x} \sim N(\mathbf{0}, \mathbf{I}),\ \sum_{i=1}^{n}(x_i - \bar{x})^2 \sim \chi^2[n - 1]. \qquad \textbf{(B-110)}$$

We have already shown that the second term in (B-104) has a chi-squared distribution with one degree of freedom. It is instructive to set this up as a quadratic form as well:

$$n\bar{x}^2 = \mathbf{x}'\left[\frac{1}{n}\mathbf{i}\mathbf{i}'\right]\mathbf{x} = \mathbf{x}'[\mathbf{j}\mathbf{j}']\mathbf{x}, \quad \text{where } \mathbf{j} = \left(\frac{1}{\sqrt{n}}\right)\mathbf{i}. \qquad \textbf{(B-111)}$$

The matrix in brackets is the outer product of a nonzero vector, which always has rank one. You can verify that it is idempotent by multiplication. Thus, $\mathbf{x}'\mathbf{x}$ is the sum of two chi-squared variables, one with $n - 1$ degrees of freedom and the other with one. It is now necessary to show that the two terms are independent. To do so, we will use the next theorem.

THEOREM B.9 Independence of Idempotent Quadratic Forms

If $\mathbf{x} \sim N[\mathbf{0}, \mathbf{I}]$ *and* $\mathbf{x}'\mathbf{A}\mathbf{x}$ *and* $\mathbf{x}'\mathbf{B}\mathbf{x}$ *are two idempotent quadratic forms in* $\mathbf{x}$, *then* $\mathbf{x}'\mathbf{A}\mathbf{x}$ *and* $\mathbf{x}'\mathbf{B}\mathbf{x}$ *are independent if* $\mathbf{AB} = \mathbf{0}$. **(B-112)**

As before, we show the result for the general case and then specialize it for the example. Because both $\mathbf{A}$ and $\mathbf{B}$ are symmetric and idempotent, $\mathbf{A} = \mathbf{A}'\mathbf{A}$ and $\mathbf{B} = \mathbf{B}'\mathbf{B}$. The quadratic forms are therefore

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{A}\mathbf{x} = \mathbf{x}_1'\mathbf{x}_1, \quad \text{where } \mathbf{x}_1 = \mathbf{A}\mathbf{x}, \quad \text{and } \mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{x}_2'\mathbf{x}_2, \quad \text{where } \mathbf{x}_2 = \mathbf{B}\mathbf{x}. \tag{B-113}$$

Both vectors have zero mean vectors, so the covariance matrix of $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$E(\mathbf{x}_1\mathbf{x}_2') = \mathbf{A}\mathbf{I}\mathbf{B}' = \mathbf{AB} = \mathbf{0}.$$

Because $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are linear functions of a normally distributed random vector, they are, in turn, normally distributed. Their zero covariance matrix implies that they are statistically independent,[8] which establishes the independence of the two quadratic forms. For the case of $\mathbf{x}'\mathbf{x}$, the two matrices are $\mathbf{M}^0$ and $[\mathbf{I} - \mathbf{M}^0]$. You can show that $\mathbf{M}^0[\mathbf{I} - \mathbf{M}^0] = \mathbf{0}$ just by multiplying it out.

B.11.5 THE *F* DISTRIBUTION

The normal family of distributions (chi-squared, F, and t) can all be derived as functions of idempotent quadratic forms in a standard normal vector. The F distribution is the ratio of two independent chi-squared variables, each divided by its respective degrees of freedom. Let $\mathbf{A}$ and $\mathbf{B}$ be two idempotent matrices with ranks r_a and r_b, and let $\mathbf{AB} = \mathbf{0}$. Then

$$\frac{\mathbf{x}'\mathbf{A}\mathbf{x}/r_a}{\mathbf{x}'\mathbf{B}\mathbf{x}/r_b} \sim F[r_a, r_b]. \tag{B-114}$$

If $\text{Var}[\mathbf{x}] = \sigma^2\mathbf{I}$ instead, then this is modified to

$$\frac{(\mathbf{x}'\mathbf{A}\mathbf{x}/\sigma^2)/r_a}{(\mathbf{x}'\mathbf{B}\mathbf{x}/\sigma^2)/r_b} \sim F[r_a, r_b]. \tag{B-115}$$

B.11.6 A FULL RANK QUADRATIC FORM

Finally, consider the general case,

$$\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}].$$

[8] Note that both $\mathbf{x}_1 = \mathbf{A}\mathbf{x}$ and $\mathbf{x}_2 = \mathbf{B}\mathbf{x}$ have singular covariance matrices. Nonetheless, every element of $\mathbf{x}_1$ is independent of every element $\mathbf{x}_2$, so the vectors are independent.

We are interested in the distribution of

$$q = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \qquad \textbf{(B-116)}$$

First, the vector can be written as $\mathbf{z} = \mathbf{x} - \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{z}$ as well as of $\mathbf{x}$. Therefore, we seek the distribution of

$$q = \mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} = \mathbf{z}'(\text{Var}[\mathbf{z}])^{-1}\mathbf{z}, \qquad \textbf{(B-117)}$$

where $\mathbf{z}$ is normally distributed with mean $\mathbf{0}$. This equation is a quadratic form, but not necessarily in an idempotent matrix.[9] Because $\boldsymbol{\Sigma}$ is positive definite, it has a square root. Define the symmetric matrix $\boldsymbol{\Sigma}^{1/2}$ so that $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}$. Then

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2},$$

and

$$\begin{aligned}\mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} &= \mathbf{z}'\boldsymbol{\Sigma}^{-1/2\prime}\boldsymbol{\Sigma}^{-1/2}\mathbf{z} \\ &= (\boldsymbol{\Sigma}^{-1/2}\mathbf{z})'(\boldsymbol{\Sigma}^{-1/2}\mathbf{z}) \\ &= \mathbf{w}'\mathbf{w}.\end{aligned}$$

Now $\mathbf{w} = \mathbf{A}\mathbf{z}$, so

$$E(\mathbf{w}) = \mathbf{A}E[\mathbf{z}] = \mathbf{0},$$

and

$$\text{Var}[\mathbf{w}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{0} = \mathbf{I}.$$

This provides the following important result:

THEOREM B.10 Distribution of a Standardized Normal Vector

If $\mathbf{x} \sim N[\mu, \boldsymbol{\Sigma}]$, *then* $\boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N[\mathbf{0}, \mathbf{I}]$.

The simplest special case is that in which $\mathbf{x}$ has only one variable, so that the transformation is just $(x - \mu)/\sigma$. Combining this case with (B-32) concerning the sum of squares of standard normals, we have the following theorem.

THEOREM B.11 Distribution of $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ When x Is Normal

If $\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, *then* $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2[n]$.

[9] It will be idempotent only in the special case of $\boldsymbol{\Sigma} = \mathbf{I}$.

B.11.7 INDEPENDENCE OF A LINEAR AND A QUADRATIC FORM

The t distribution is used in many forms of hypothesis tests. In some situations, it arises as the ratio of a linear to a quadratic form in a normal vector. To establish the distribution of these statistics, we use the following result.

THEOREM B.12 Independence of a Linear and a Quadratic Form
A linear function $\mathbf{Lx}$ *and a symmetric idempotent quadratic form* $\mathbf{x'Ax}$ *in a standard normal vector are statistically independent if* $\mathbf{LA} = \mathbf{0}$.

The proof follows the same logic as that for two quadratic forms. Write $\mathbf{x'Ax}$ as $\mathbf{x'A'Ax} = (\mathbf{Ax})'(\mathbf{Ax})$. The covariance matrix of the variables $\mathbf{Lx}$ and $\mathbf{Ax}$ is $\mathbf{LA} = \mathbf{0}$, which establishes the independence of these two random vectors. The independence of the linear function and the quadratic form follows because functions of independent random vectors are also independent.

The t distribution is defined as the ratio of a standard normal variable to the square root of an independent chi-squared variable divided by its degrees of freedom:

$$t[J] = \frac{N[0,1]}{\{\chi^2[J]/J\}^{1/2}}.$$

A particular case is

$$t[n-1] = \frac{\sqrt{n}\,\bar{x}}{\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{1/2}} = \frac{\sqrt{n}\bar{x}}{s},$$

where s is the standard deviation of the values of $\mathbf{x}$. The distribution of the two variables in $t[n-1]$ was shown earlier; we need only show that they are independent. But

$$\sqrt{n}\bar{x} = \frac{1}{\sqrt{n}}\mathbf{i'x} = \mathbf{j'x},$$

and

$$s^2 = \frac{\mathbf{x'M^0x}}{n-1}.$$

It suffices to show that $\mathbf{M^0j} = \mathbf{0}$, which follows from

$$\mathbf{M^0i} = [\mathbf{I} - \mathbf{i}(\mathbf{i'i})^{-1}\mathbf{i'}]\mathbf{i} = \mathbf{i} - \mathbf{i}(\mathbf{i'i})^{-1}(\mathbf{i'i}) = \mathbf{0}.$$

APPENDIX C

ESTIMATION AND INFERENCE

C.1 INTRODUCTION

The probability distributions discussed in Appendix B serve as models for the underlying data generating processes that produce our observed data. The goal of statistical inference in econometrics is to use the principles of mathematical statistics to combine these theoretical distributions and the observed data into an empirical model of the economy. This analysis takes place in one of two frameworks, classical or Bayesian. The overwhelming majority of empirical study in econometrics has been done in the classical framework. Our focus, therefore, will be on classical methods of inference. Bayesian methods are discussed in Chapter 16.[1]

C.2 SAMPLES AND RANDOM SAMPLING

The classical theory of statistical inference centers on rules for using the sampled data effectively. These rules, in turn, are based on the properties of samples and sampling distributions.

A sample of n observations on one or more variables, denoted $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is a **random sample** if the n observations are drawn independently from the same population, or probability distribution, $f(\mathbf{x}_i, \boldsymbol{\theta})$. The sample may be univariate if $\mathbf{x}_i$ is a single random variable or multivariate if each observation contains several variables. A random sample of observations, denoted $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ or $\{\mathbf{x}_i\}_{i=1,\ldots,n}$, is said to be **independent, identically distributed,** which we denote *i. i. d*. The vector $\boldsymbol{\theta}$ contains one or more unknown parameters. Data are generally drawn in one of two settings. A **cross section** is a sample of a number of observational units all drawn at the same point in time. A **time series** is a set of observations drawn on the same observational unit at a number of (usually evenly spaced) points in time. Many recent studies have been based on time-series cross sections, which generally consist of the same cross-sectional units observed at several points in time. Because the typical data set of this sort consists of a large number of cross-sectional units observed at a few points in time, the common term **panel data set** is usually more fitting for this sort of study.

[1] An excellent reference is Leamer (1978). A summary of the results as they apply to econometrics is contained in Zellner (1971) and in Judge et al. (1985). See, as well, Poirier (1991, 1995). Recent textbooks on Bayesian econometrics include Koop (2003), Lancaster (2004) and Geweke (2005).

C.3 DESCRIPTIVE STATISTICS

Before attempting to estimate parameters of a population or fit models to data, we normally examine the data themselves. In raw form, the sample data are a disorganized mass of information, so we will need some organizing principles to distill the information into something meaningful. Consider, first, examining the data on a single variable. In most cases, and particularly if the number of observations in the sample is large, we shall use some summary **statistics** to describe the sample data. Of most interest are measures of **location**—that is, the center of the data—and **scale**, or the dispersion of the data. A few measures of central tendency are as follows:

$$\textbf{mean: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\textbf{median: } M = \text{middle ranked observation},$$

$$\textbf{sample midrange: } \text{midrange} = \frac{\text{maximum} + \text{minimum}}{2}. \qquad \textbf{(C-1)}$$

The dispersion of the sample observations is usually measured by the

$$\textbf{standard deviation: } s_x = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right]^{1/2}. \qquad \textbf{(C-2)}$$

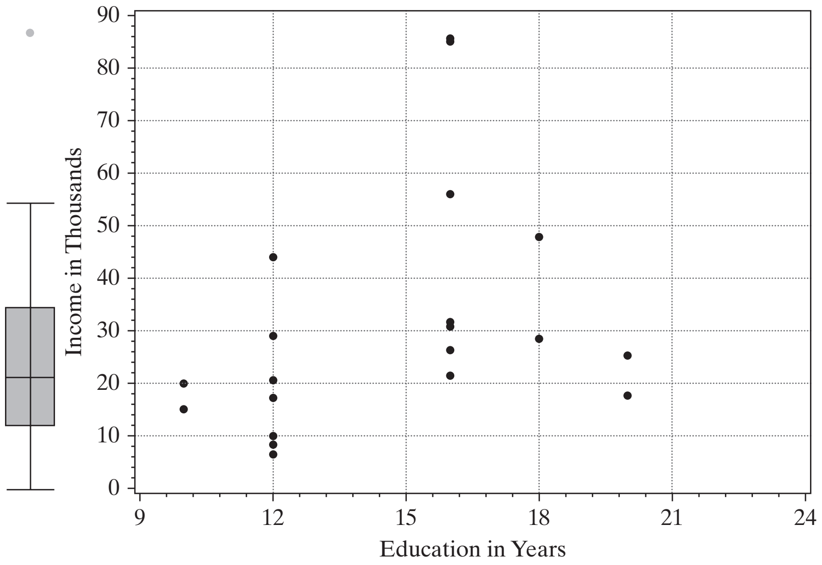
Other measures, such as the average absolute deviation from the sample mean, are also used, although less frequently than the standard deviation. The shape of the distribution of values is often of interest as well. Samples of income or expenditure data, for example, tend to be highly skewed while financial data such as asset returns and exchange rate movements are relatively more symmetrically distributed but are also more widely dispersed than other variables that might be observed. Two measures used to quantify these effects are the

$$\textbf{skewness} = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3(n-1)}\right], \quad \text{and} \quad \textbf{kurtosis} = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4(n-1)}\right].$$

(Benchmark values for these two measures are zero for a symmetric distribution, and three for one which is "normally" dispersed.) The skewness coefficient has a bit less of the intuitive appeal of the mean and standard deviation, and the kurtosis measure has very little at all. The **box and whisker plot** is a graphical device which is often used to capture a large amount of information about the sample in a simple visual display. This plot shows in a figure the median, the range of values contained in the 25th and 75th percentile, some limits that show the normal range of values expected, such as the median plus and minus two standard deviations, and in isolation values that could be viewed as outliers. A box and whisker plot is shown in Figure C.1 for the income variable in Example C.1.

If the sample contains data on more than one variable, we will also be interested in measures of association among the variables. A **scatter diagram** is useful in a bivariate sample if the sample contains a reasonable number of observations. Figure C.1 shows an

FIGURE C.1 Box and Whisker Plot for Income and Scatter Diagram for Income and Education.



example for a small data set. If the sample is a multivariate one, then the degree of linear association among the variables can be measured by the pairwise measures

$$\textbf{covariance:}\; s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

$$\textbf{correlation:}\; r_{xy} = \frac{s_{xy}}{s_x s_y}. \qquad \textbf{(C-3)}$$

If the sample contains data on several variables, then it is sometimes convenient to arrange the covariances or correlations in a

$$\textbf{covariance matrix: S} = [s_{ij}], \qquad \textbf{(C-4)}$$

or

$$\textbf{correlation matrix: R} = [r_{ij}].$$

Some useful algebraic results for any two variables (x_i, y_i), $i = 1, \ldots, n$, and constants a and b are

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n - 1}, \qquad \textbf{(C-5)}$$

$$s_{xy} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\,\bar{y}}{n - 1}, \qquad \textbf{(C-6)}$$

$$-1 \leq r_{xy} \leq 1,$$

$$r_{ax,\,by} = \frac{ab}{|ab|} r_{xy}, \quad a, b \neq 0, \tag{C-7}$$

$$s_{ax} = |a|\; s_x, \tag{C-8}$$

$$s_{ax,\,by} = (ab)s_{xy}.$$

Note that these algebraic results parallel the theoretical results for bivariate probability distributions. [We note in passing, while the formulas in (C-2) and (C-5) are algebraically the same, (C-2) will generally be more accurate in practice, especially when the values in the sample are very widely dispersed.]

Example C.1 Descriptive Statistics for a Random Sample

Appendix Table FC.1 contains a (hypothetical) sample of observations on income and education (The observations all appear in the calculations of the means below.) A scatter diagram appears in Figure C.1. It suggests a weak positive association between income and education in these data. The box and whisker plot for income at the left of the scatter plot shows the distribution of the income data as well.

$$\textit{Means}: \bar{I} = \frac{1}{20}\begin{bmatrix} 20.5 + 31.5 + 47.7 + 26.2 + 44.0 + 8.28 + 30.8 + \\ 17.2 + 19.9 + 9.96 + 55.8 + 25.2 + 29.0 + 85.5 + \\ 15.1 + 28.5 + 21.4 + 17.7 + 6.42 + 84.9 \end{bmatrix} = 31.278,$$

$$\bar{E} = \frac{1}{20}\begin{bmatrix} 12 + 16 + 18 + 16 + 12 + 12 + 16 + 12 + 10 + 12 + \\ 16 + 20 + 12 + 16 + 10 + 18 + 16 + 20 + 12 + 16 \end{bmatrix} = 14.600.$$

Standard deviations:

$$s_I = \sqrt{\tfrac{1}{19}[(20.5 - 31.278)^2 + \cdots + (84.9 - 31.278)^2]} = 22.376,$$

$$s_E = \sqrt{\tfrac{1}{19}[(12 - 14.6)^2 + \cdots + (16 - 14.6)^2]} = 3.119.$$

Covariance: $s_{IE} = \frac{1}{19}[20.5(12) + \cdots + 84.9(16) - 20(31.28)(14.6)] = 23.597,$

Correlation: $r_{IE} = \dfrac{23.597}{(22.376)(3.119)} = 0.3382.$

The positive correlation is consistent with our observation in the scatter diagram.

The statistics just described will provide the analyst with a more concise description of the data than a raw tabulation. However, we have not, as yet, suggested that these measures correspond to some underlying characteristic of the process that generated the data. We do assume that there is an underlying mechanism, the data generating process that produces the data in hand. Thus, these serve to do more than describe the data; they characterize that process, or population. Because we have assumed that there is an underlying probability distribution, it might be useful to produce a statistic that gives a broader view of the DGP. The **histogram** is a simple graphical device that produces this result—see Examples C.3 and C.4 for applications. For small samples or widely dispersed data, however, histograms tend to be rough and difficult to make informative. A burgeoning literature[2] has demonstrated the usefulness of the **kernel density estimator** as a substitute for the histogram as a descriptive tool for the underlying distribution that produced a sample of data. The underlying theory of the kernel density estimator is fairly complicated, but the computations are surprisingly simple. The estimator is computed using

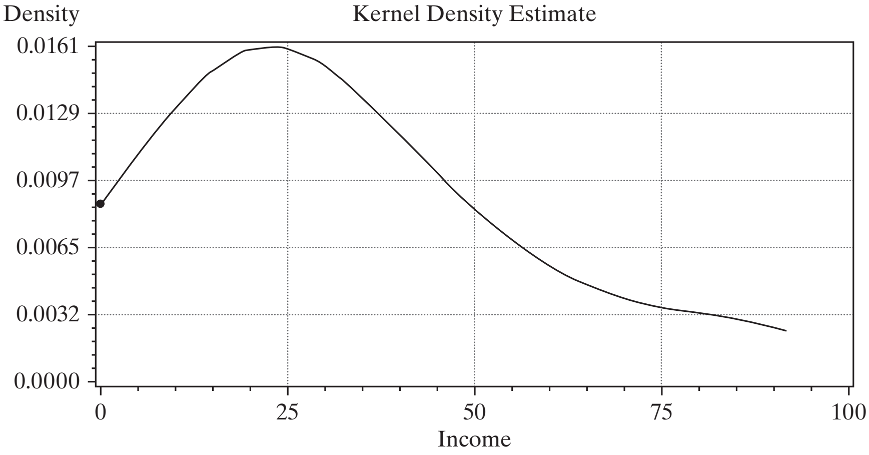
$$\hat{f}(x^*) = \frac{1}{nh}\sum_{i=1}^{n} K\left[\frac{x_i - x^*}{h}\right],$$

where $x_1, \ldots, x_n$ are the n observations in the sample, $\hat{f}(x^*)$ denotes the estimated density function, x^* is the value at which we wish to evaluate the density, and h and $K[\cdot]$ are the "bandwidth" and "kernel function" that we now consider. The density estimator is rather like a histogram, in which the bandwidth is the width of the intervals. The kernel function is a weight function which is generally chosen so that it takes large values when x^* is close to x_i and tapers off to zero in as they diverge in either direction. The weighting function used in the following example is the logistic density discussed in Section B.4.7. The bandwidth is chosen to be a function of $1/n$ so that the intervals can become narrower as the sample becomes larger (and richer). The one used for Figure C.2 is $h = 0.9 \text{ Min } (s, \text{range}/3)/n^{.2}$. (We will revisit this method of estimation in Chapter 12.) Example C.2 illustrates the computation for the income data used in Example C.1.

Example C.2 Kernel Density Estimator for the Income Data

Figure C.2 suggests the large skew in the income data that is also suggested by the box and whisker plot (and the scatter plot in Example C.1.)

FIGURE C.2 Kernel Density Estimate for Income.



[2] See for example, Pagan and Ullah (1999), Li and Racine (2007) and Henderson and Parmeter (2015).

C.4 STATISTICS AS ESTIMATORS—SAMPLING DISTRIBUTIONS

The measures described in the preceding section summarize the data in a random sample. Each measure has a counterpart in the population, that is, the distribution from which the data were drawn. Sample quantities such as the means and the correlation coefficient correspond to population expectations, whereas the kernel density estimator and the values in Table C.1 parallel the population pdf and cdf. In the setting of a random sample, we expect these quantities to mimic the population, although not perfectly. The precise manner in which these quantities reflect the population values defines the sampling distribution of a sample statistic.

DEFINITION C.1 Statistic
A statistic is any function computed from the data in a sample.

If another sample were drawn under identical conditions, different values would be obtained for the observations, as each one is a random variable. Any statistic is a function of these random values, so it is also a random variable with a probability distribution called a **sampling distribution**. For example, the following shows an exact result for the sampling behavior of a widely used statistic.

THEOREM C.1 Sampling Distribution of the Sample Mean
If $x_1, \ldots, x_n$ are a random sample from a population with mean μ and variance σ^2, then $\bar{x}$ is a random variable with mean μ and variance σ^2/n.
Proof: $\bar{x} = (1/n)\Sigma_i x_i$. $E[\bar{x}] = (1/n)\Sigma_i\mu = \mu$. *The observations are independent, so* $\text{Var}[\bar{x}] = (1/n)^2 \text{ Var}[\Sigma_i x_i] = (1/n^2)\Sigma_i\sigma^2 = \sigma^2/n$.

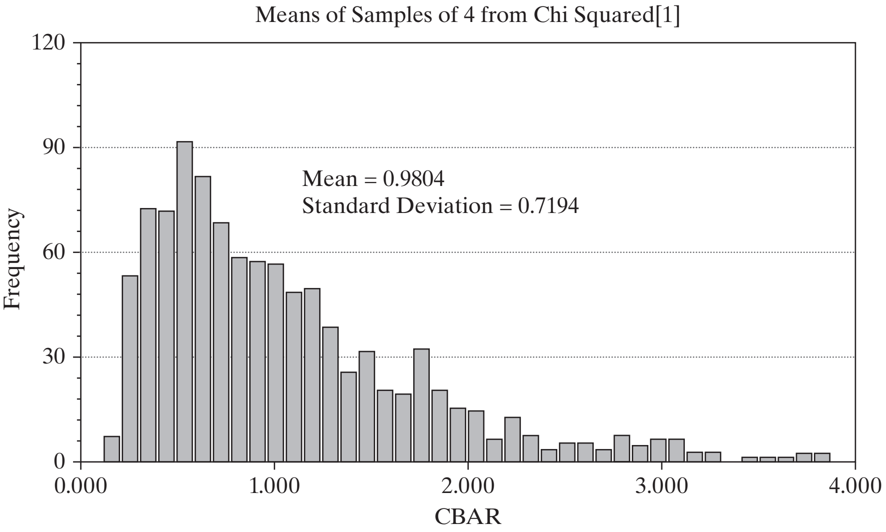
Example C.3 illustrates the behavior of the sample mean in samples of four observations drawn from a chi-squared population with one degree of freedom. The crucial concepts illustrated in this example are, first, the mean and variance results in Theorem C.1 and, second, the phenomenon of **sampling variability**.

Notice that the fundamental result in Theorem C.1 does not assume a distribution for x_i. Indeed, looking back at Section C.3, nothing we have done so far has required any assumption about a particular distribution.

TABLE C.1 Income Distribution

Range	*Relative Frequency*	*Cumulative Frequency*
<\$10,000	0.15	0.15
10,000–25,000	0.30	0.45
25,000–50,000	0.40	0.85
>50,000	0.15	1.00

FIGURE C.3 Sampling Distribution of Means of 1,000 Samples of Size 4 from Chi-Squared[1].



Example C.3 Sampling Distribution of a Sample Mean

Figure C.3 shows a frequency plot of the means of 1,000 random samples of four observations drawn from a chi-squared distribution with one degree of freedom, which has mean 1 and variance 2.

We are often interested in how a statistic behaves as the sample size increases. Example C.4 illustrates one such case. Figure C.4 shows two sampling distributions, one based on samples of three and a second, of the same statistic, but based on samples of six. The effect of increasing sample size in this figure is unmistakable. It is easy to visualize the behavior of this statistic if we extrapolate the experiment in Example C.4 to samples of, say, 100.

Example C.4 Sampling Distribution of the Sample Minimum

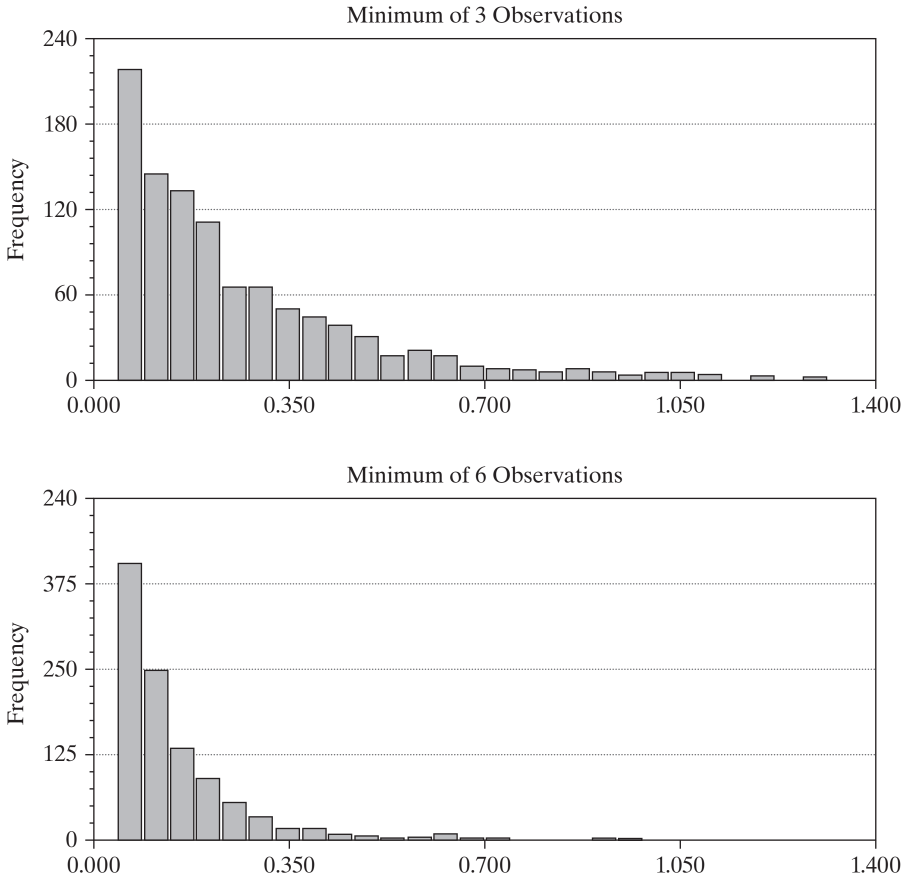
If $x_1, \ldots, x_n$ are a random sample from an exponential distribution with $f(x) = \theta e^{-\theta x}$, then the sampling distribution of the sample minimum in a sample of n observations, denoted $x_{(1)}$, is

$$f(x_{(1)}) = (n\theta)e^{-(n\theta)x_{(1)}}.$$

Because $E[x] = 1/\theta$ and $\text{Var}[x] = 1/\theta^2$, by analogy $E[x_{(1)}] = 1/(n\theta)$ and $\text{Var}[x_{(1)}] = 1/(n\theta)^2$. Thus, in increasingly larger samples, the minimum will be arbitrarily close to 0. [The Chebychev inequality in Theorem D.2 can be used to prove this intuitively appealing result.]

Figure C.4 shows the results of a simple sampling experiment you can do to demonstrate this effect. It requires software that will allow you to produce pseudorandom numbers uniformly distributed in the range zero to one and that will let you plot a histogram and control the axes. (We used NLOGIT. This can be done with Stata, Excel, or several other packages.) The experiment consists of drawing 1,000 sets of nine random values,

FIGURE C.4 Histograms of the Sample Minimum of 3 and 6 Observations.



U_{ij}, $i = 1, \ldots 1{,}000$, $j = 1, \ldots, 9$. To transform these uniform draws to exponential with parameter θ—we used $\theta = 1.5$, use the inverse probability transform—see Section E.2.3. For an exponentially distributed variable, the transformation is $z_{ij} = -(1/\theta)\log(1 - U_{ij})$. We then created $z_{(1)}|3$ from the first three draws and $z_{(1)}|6$ from the other six. The two histograms show clearly the effect on the sampling distribution of increasing sample size from just 3 to 6.

Sampling distributions are used to make inferences about the population. To consider a perhaps obvious example, because the sampling distribution of the mean of a set of normally distributed observations has mean μ, the sample mean is a natural candidate for an estimate of μ. The observation that the sample "mimics" the population is a statement about the sampling distributions of the sample statistics. Consider, for example, the sample data collected in Figure C.3. The sample mean of four observations clearly has a sampling distribution, which appears to have a mean roughly equal to the population mean. Our theory of parameter estimation departs from this point.

C.5 POINT ESTIMATION OF PARAMETERS

Our objective is to use the sample data to infer the value of a parameter or set of parameters, which we denote θ. A **point estimate** is a statistic computed from a sample that gives a single value for θ. The **standard error** of the estimate is the standard deviation of the sampling distribution of the statistic; the square of this quantity is the **sampling variance**. An **interval estimate** is a range of values that will contain the true parameter with a preassigned probability. There will be a connection between the two types of estimates; generally, if $\hat{\theta}$ is the point estimate, then the interval estimate will be $\hat{\theta} \pm$ a measure of sampling error.

An **estimator** is a rule or strategy for using the data to estimate the parameter. It is defined before the data are drawn. Obviously, some estimators are better than others. To take a simple example, your intuition should convince you that the sample mean would be a better estimator of the population mean than the sample minimum; the minimum is almost certain to underestimate the mean. Nonetheless, the minimum is not entirely without virtue; it is easy to compute, which is occasionally a relevant criterion. The search for good estimators constitutes much of econometrics. Estimators are compared on the basis of a variety of attributes. **Finite sample properties** of estimators are those attributes that can be compared regardless of the sample size. Some estimation problems involve characteristics that are not known in finite samples. In these instances, estimators are compared on the basis on their large sample, or **asymptotic properties**. We consider these in turn.

C.5.1 ESTIMATION IN A FINITE SAMPLE

The following are some finite sample estimation criteria for estimating a single parameter. The extensions to the multiparameter case are direct. We shall consider them in passing where necessary.

DEFINITION C.2 Unbiased Estimator
An estimator of a parameter θ is unbiased if the mean of its sampling distribution is θ. Formally,

$$E[\hat{\theta}] = \theta$$

or

$$E[\hat{\theta} - \theta] = \text{Bias}[\hat{\theta} \mid \theta] = 0$$

implies that $\hat{\theta}$ is unbiased. Note that this implies that the expected sampling error is zero. If $\boldsymbol{\theta}$ is a vector of parameters, then the estimator is unbiased if the expected value of every element of $\hat{\boldsymbol{\theta}}$ equals the corresponding element of $\boldsymbol{\theta}$.

If samples of size n are drawn repeatedly and $\hat{\theta}$ is computed for each one, then the average value of these estimates will tend to equal θ. For example, the average of the 1,000 sample means underlying Figure C.3 is 0.9804, which is reasonably close to the population mean of one. The sample minimum is clearly a biased estimator of the mean; it will almost always underestimate the mean, so it will do so on average as well.

Unbiasedness is a desirable attribute, but it is rarely used by itself as an estimation criterion. One reason is that there are many unbiased estimators that are poor uses of the data. For example, in a sample of size n, the first observation drawn is an unbiased estimator of the mean that clearly wastes a great deal of information. A second criterion used to choose among unbiased estimators is efficiency.

DEFINITION C.3 Efficient Unbiased Estimator

An unbiased estimator $\hat{\theta}_1$ is more efficient than another unbiased estimator $\hat{\theta}_2$ if the sampling variance of $\hat{\theta}_1$ is less than that of $\hat{\theta}_2$. That is,

$$\text{Var}[\hat{\theta}_1] < \text{Var}[\hat{\theta}_2].$$

In the multiparameter case, the comparison is based on the covariance matrices of the two estimators; $\hat{\boldsymbol{\theta}}_1$ is more efficient than $\hat{\boldsymbol{\theta}}_2$ if $\text{Var}[\hat{\boldsymbol{\theta}}_2] - \text{Var}[\hat{\boldsymbol{\theta}}_1]$ *is a positive definite matrix.*

By this criterion, the sample mean is obviously to be preferred to the first observation as an estimator of the population mean. If σ^2 is the population variance, then

$$\text{Var}[x_1] = \sigma^2 > \text{Var}[\overline{x}] = \frac{\sigma^2}{n}.$$

In discussing efficiency, we have restricted the discussion to unbiased estimators. Clearly, there are biased estimators that have smaller variances than the unbiased ones we have considered. Any constant has a variance of zero. Of course, using a constant as an estimator is not likely to be an effective use of the sample data. Focusing on unbiasedness may still preclude a tolerably biased estimator with a much smaller variance, however. A criterion that recognizes this possible tradeoff is the mean squared error. Figure C.5 illustrates the effect. In this example, on average, the biased estimator will be closer to the true parameter than will the unbiased estimator.

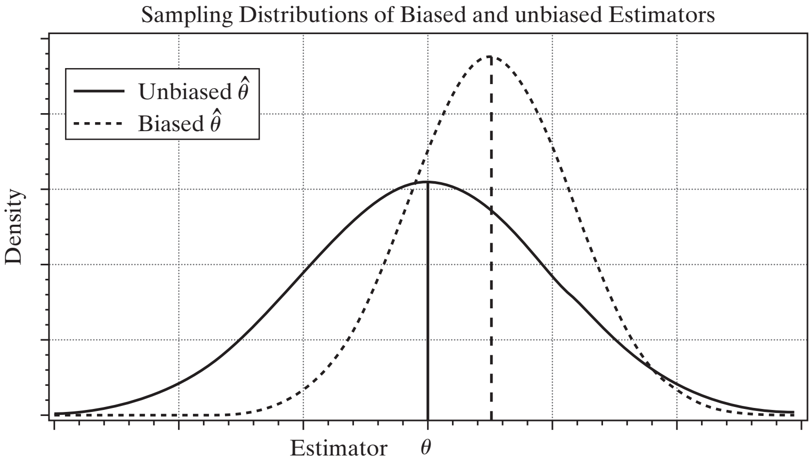
DEFINITION C.4 Mean Squared Error

The mean squared error of an estimator is

$$\begin{aligned} \text{MSE}[\hat{\theta}\,|\,\theta] &= E[(\hat{\theta} - \theta)^2] \\ &= \text{Var}[\hat{\theta}] + (\text{Bias}[\hat{\theta}\,|\,\theta])^2 && \text{if } \theta \text{ is a scalar,} \\ \text{MSE}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}] &= \text{Var}[\hat{\boldsymbol{\theta}}] + \text{Bias}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}]\text{Bias}[\hat{\boldsymbol{\theta}}\,|\,\boldsymbol{\theta}]' && \text{if } \boldsymbol{\theta} \text{ is a vector.} \end{aligned} \quad \textbf{(C-9)}$$

Which of these criteria should be used in a given situation depends on the particulars of that setting and our objectives in the study. Unfortunately, the MSE criterion is rarely

FIGURE C.5 Sampling Distributions.



operational; minimum mean squared error estimators, when they exist at all, usually depend on unknown parameters. Thus, we are usually less demanding. A commonly used criterion is **minimum variance unbiasedness**.

Example C.5 Mean Squared Error of the Sample Variance

In sampling from a normal distribution, the most frequently used estimator for $\boldsymbol{\sigma}^2$ is

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

It is straightforward to show that s^2 is unbiased, so

$$\text{Var}[s^2] = \frac{2\sigma^4}{n - 1} = \text{MSE}[s^2 \,|\, \sigma^2].$$

A proof is based on the distribution of the idempotent quadratic form $(\mathbf{x} - \mathbf{i}\mu)'\mathbf{M}^0(\mathbf{x} - \mathbf{i}\mu)$, which we discussed in Section B.11.4. A less frequently used estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = [(n - 1)/n]s^2.$$

This estimator is slightly biased downward:

$$E[\hat{\sigma}^2] = \frac{(n - 1)E(s^2)}{n} = \frac{(n - 1)\sigma^2}{n},$$

so its bias is

$$E[\hat{\sigma}^2 - \sigma^2] = \text{Bias}[\hat{\sigma}^2 \,|\, \sigma^2] = \frac{-1}{n}\sigma^2.$$

But it has a smaller variance than s^2:

$$\text{Var}[\hat{\sigma}^2] = \left[\frac{n - 1}{n}\right]^2\left[\frac{2\sigma^4}{n - 1}\right] < \text{Var}[s^2].$$

To compare the two estimators, we can use the difference in their mean squared errors:

$$\text{MSE}[\hat{\sigma}^2 \,|\, \sigma^2] - \text{MSE}[s^2 \,|\, \sigma^2] = \sigma^4\left[\frac{2n-1}{n^2} - \frac{2}{n-1}\right] < 0.$$

The biased estimator is a bit more precise. The difference will be negligible in a large sample, but, for example, it is about 1.2 percent in a sample of 16.

C.5.2 EFFICIENT UNBIASED ESTIMATION

In a random sample of n observations, the density of each observation is $f(x_i, \theta)$. Because the n observations are independent, their joint density is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n, \theta) &= f(x_1, \theta)f(x_2, \theta)\cdots f(x_n, \theta) \\ &= \prod_{i=1}^{n} f(x_i, \theta) = L(\theta \,|\, x_1, x_2, \ldots, x_n). \end{aligned} \qquad \textbf{(C-10)}$$

This function, denoted $L(\theta \,|\, \mathbf{X})$, is called the likelihood function for θ given the data $\mathbf{X}$. It is frequently abbreviated to $L(\theta)$. Where no ambiguity can arise, we shall abbreviate it further to L.

Example C.6 Likelihood Functions for Exponential and Normal Distributions

If $x_1, \ldots, x_n$ are a sample of n observations from an exponential distribution with parameter θ, then

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \Sigma^{n}{}_{i=1} x_i}.$$

If $x_1, \ldots, x_n$ are a sample of n observations from a normal distribution with mean μ and standard deviation σ, then

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-[1/(2\sigma^2)](x_i - \mu)^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-[1/(2\sigma^2)]\Sigma_i (x_i - \mu)^2}. \end{aligned} \qquad \textbf{(C-11)}$$

The likelihood function is the cornerstone for most of our theory of parameter estimation. An important result for efficient estimation is the following.

THEOREM C.2 Cramér–Rao Lower Bound
Assuming that the density of x satisfies certain regularity conditions, the variance of an unbiased estimator of a parameter θ will always be at least as large as

$$[I(\theta)]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1} = \left(E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]\right)^{-1}. \qquad \textbf{(C-12)}$$

The quantity $I(\theta)$ is the information number for the sample. We will prove the result that the negative of the expected second derivative equals the expected square of the first derivative in Chapter 14. Proof of the main result of the theorem is quite involved. See, for example, Stuart and Ord (1989).

The regularity conditions are technical. (See Section 14.4.1.) Loosely, they are conditions imposed on the density of the random variable that appears in the likelihood function; these conditions will ensure that the Lindeberg–Levy central limit theorem will apply to moments of the sample of observations on the random vector $\mathbf{y} = \partial \ln f(x_i|\theta)/\partial\theta,\ i = 1, \ldots, n$. Among the conditions are finite moments of x up to order 3. An additional condition usually included in the set is that the range of the random variable be independent of the parameters.

In some cases, the second derivative of the log likelihood is a constant, so the Cramér–Rao bound is simple to obtain. For instance, in sampling from an exponential distribution, from Example C.6,

$$\ln L = n \ln \theta - \theta \sum_{i=1}^{n} x_i,$$

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i,$$

so $\partial^2 \ln L/\partial\theta^2 = -n/\theta^2$ and the variance bound is $[I(\theta)]^{-1} = \theta^2/n$. In many situations, the second derivative is a random variable with a distribution of its own. The following examples show two such cases.

Example C.7 Variance Bound for the Poisson Distribution

For the Poisson distribution,

$$f(x) = \frac{e^{-\theta}\theta^x}{x!},$$

$$\ln L = -n\theta + \left(\sum_{i=1}^{n} x_i\right)\ln\theta - \sum_{i=1}^{n} \ln(x_i\,!),$$

$$\frac{\partial \ln L}{\partial\theta} = -n + \frac{\sum_{i=1}^{n} x_i}{\theta},$$

$$\frac{\partial^2 \ln L}{\partial\theta^2} = \frac{-\sum_{i=1}^{n} x_i}{\theta^2}.$$

The sum of n identical Poisson variables has a Poisson distribution with parameter equal to n times the parameter of the individual variables. Therefore, the actual distribution of the first derivative will be that of a linear function of a Poisson distributed variable. Because $E[\sum_{i=1}^{n} x_i] = nE[x_i] = n\theta$, the variance bound for the Poisson distribution is $[I(\theta)]^{-1} = \theta/n$. (Note also that the same result implies that $E[\partial \ln L/\partial\theta] = 0$, which is a result we will use in Chapter 14. The same result holds for the exponential distribution.)

Consider, finally, a multivariate case. If $\boldsymbol{\theta}$ is a vector of parameters, then $\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix**. The Cramér–Rao theorem states that the difference between the covariance matrix of any unbiased estimator and the inverse of the information matrix,

$$[\mathbf{I}(\boldsymbol{\theta})]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right)^{-1} = \left\{E\left[\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right)\right]\right\}^{-1}, \quad \textbf{(C-13)}$$

will be a nonnegative definite matrix.

In some settings, numerous estimators are available for the parameters of a distribution. The usefulness of the Cramér–Rao bound is that if one of these is known to attain the variance bound, then there is no need to consider any other to seek a more efficient estimator. Regarding the use of the variance bound, we emphasize that if an unbiased estimator attains it, then that estimator is efficient. If a given estimator does not attain the variance bound, however, then we do not know, except in a few special cases, whether this estimator is efficient or not. It may be that no unbiased estimator can attain the Cramér–Rao bound, which can leave the question of whether a given unbiased estimator is efficient or not unanswered.

We note, finally, that in some cases we further restrict the set of estimators to linear functions of the data.

DEFINITION C.5 Minimum Variance Linear Unbiased Estimator (MVLUE)
An estimator is the minimum variance linear unbiased estimator or best linear unbiased estimator (BLUE) if it is a linear function of the data and has minimum variance among linear unbiased estimators.

In a few instances, such as the normal mean, there will be an efficient linear unbiased estimator; $\overline{x}$ is efficient among all unbiased estimators, both linear and nonlinear. In other cases, such as the normal variance, there is no linear unbiased estimator. This criterion is useful because we can sometimes find an MVLUE without having to specify the distribution at all. Thus, by limiting ourselves to a somewhat restricted class of estimators, we free ourselves from having to assume a particular distribution.

C.6 INTERVAL ESTIMATION

Regardless of the properties of an estimator, the estimate obtained will vary from sample to sample, and there is some probability that it will be quite erroneous. A point estimate will not provide any information on the likely range of error. The logic behind an **interval estimate** is that we use the sample data to construct an interval, [lower (**X**), upper (**X**)], such that we can expect this interval to contain the true parameter in some specified proportion of samples, or equivalently, with some desired level of confidence. Clearly, the wider the interval, the more confident we can be that it will, in any given sample, contain the parameter being estimated.

The theory of interval estimation is based on a **pivotal quantity**, which is a function of both the parameter and a point estimate that has a known distribution. Consider the following examples.

Example C.8 Confidence Intervals for the Normal Mean

In sampling from a normal distribution with mean μ and standard deviation σ,

$$z = \frac{\sqrt{n}(\overline{x} - \mu)}{s} \sim t[n - 1],$$

and

$$c = \frac{(n - 1)s^2}{\sigma^2} \sim \chi^2[n - 1].$$

Given the pivotal quantity, we can make probability statements about events involving the parameter and the estimate. Let $p(g, \theta)$ be the constructed random variable, for example, z or c. Given a prespecified **confidence level**, $1 - \alpha$, we can state that

$$\text{Prob}(\text{lower} \leq p(g, \theta) \leq \text{upper}) = 1 - \alpha, \tag{C-14}$$

where lower and upper are obtained from the appropriate table. This statement is then manipulated to make equivalent statements about the endpoints of the intervals. For example, the following statements are equivalent:

$$\text{Prob}(-z \leq \frac{\sqrt{n}(\bar{x} - \mu)}{s} \leq z) = 1 - \alpha,$$

$$\text{Prob}\left(\bar{x} - \frac{zs}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{zs}{\sqrt{n}}\right) = 1 - \alpha.$$

The second of these is a statement about the interval, not the parameter; that is, it is the interval that is random, not the parameter. We attach a probability, or $100(1 - \alpha)$ percent confidence level, to the interval itself; in repeated sampling, an interval constructed in this fashion will contain the true parameter $100(1 - \alpha)$ percent of the time.

In general, the interval constructed by this method will be of the form

$$\text{lower}(\mathbf{X}) = \hat{\theta} - e_1,$$

$$\text{upper}(\mathbf{X}) = \hat{\theta} + e_2,$$

where $\mathbf{X}$ is the sample data, e_1 and e_2 are sampling errors, and $\hat{\theta}$ is a point estimate of θ. It is clear from the preceding example that if the sampling distribution of the pivotal quantity is either t or standard normal, which will be true in the vast majority of cases we encounter in practice, then the confidence interval will be

$$\hat{\theta} \pm C_{1-\alpha/2}[\text{se}(\hat{\theta})], \tag{C-15}$$

where se (.) is the (known or estimated) standard error of the parameter estimate and $C_{1-\alpha/2}$ is the value from the t or standard normal distribution that is exceeded with probability $1 - \alpha/2$. The usual values for α are 0.10, 0.05, or 0.01. The theory does not prescribe exactly how to choose the endpoints for the confidence interval. An obvious criterion is to minimize the width of the interval. If the sampling distribution is symmetric, then the symmetric interval is the best one. If the sampling distribution is not symmetric, however, then this procedure will not be optimal.

Example C.9 Estimated Confidence Intervals for a Normal Mean and Variance

In a sample of 25, $\bar{x} = 1.63$ and $s = 0.51$. Construct a 95 percent confidence interval for μ.

Assuming that the sample of 25 is from a normal distribution,

$$\text{Prob}\left(-2.064 \leq \frac{5(\bar{x} - \mu)}{s} \leq 2.064\right) = 0.95,$$

where 2.064 is the critical value from a t distribution with 24 degrees of freedom. Thus, the confidence interval is $1.63 \pm [2.064(0.51)/5]$ or [1.4195, 1.8405].

Remark: Had the parent distribution not been specified, it would have been natural to use the standard normal distribution instead, perhaps relying on the central limit theorem. But a sample size of 25 is small enough that the more conservative t distribution might still be preferable.

The chi-squared distribution is used to construct a confidence interval for the variance of a normal distribution. Using the data from Example C.9, we find that the usual procedure would use

$$\text{Prob}\left(12.4 \leq \frac{24s^2}{\sigma^2} \leq 39.4\right) = 0.95,$$

where 12.4 and 39.4 are the 0.025 and 0.975 cutoff points from the chi-squared (24) distribution. This procedure leads to the 95 percent confidence interval [0.1581, 0.5032]. By making use of the asymmetry of the distribution, a narrower interval can be constructed. Allocating 4 percent to the left-hand tail and 1 percent to the right instead of 2.5 percent to each, the two cutoff points are 13.4 and 42.9, and the resulting 95 percent confidence interval is [0.1455, 0.4659].

Finally, the confidence interval can be manipulated to obtain a confidence interval for a function of a parameter. For example, based on the preceding, a 95 percent confidence interval for σ would be $[\sqrt{0.1581}, \sqrt{0.5032}] = [0.3976, 0.7094]$.

C.7 HYPOTHESIS TESTING

The second major group of statistical inference procedures is hypothesis tests. The classical testing procedures are based on constructing a statistic from a random sample that will enable the analyst to decide, with reasonable confidence, whether or not the data in the sample would have been generated by a hypothesized population. The formal procedure involves a statement of the hypothesis, usually in terms of a "null" or maintained hypothesis and an "alternative," conventionally denoted H_0 and H_1, respectively. The procedure itself is a rule, stated in terms of the data, that dictates whether the null hypothesis should be rejected or not. For example, the hypothesis might state a parameter is equal to a specified value. The decision rule might state that the hypothesis should be rejected if a sample estimate of that parameter is too far away from that value (where "far" remains to be defined). The classical, or Neyman–Pearson, methodology involves partitioning the sample space into two regions. If the observed data (i.e., the test statistic) fall in the **rejection region** (sometimes called the **critical region**), then the null hypothesis is rejected; if they fall in the **acceptance region**, then it is not.

C.7.1 CLASSICAL TESTING PROCEDURES

Because the sample is random, the test statistic, however defined, is also random. The same test procedure can lead to different conclusions in different samples. As such, there are two ways such a procedure can be in error:

1. **Type I error.** The procedure may lead to rejection of the null hypothesis when it is true.
2. **Type II error.** The procedure may fail to reject the null hypothesis when it is false.

To continue the previous example, there is some probability that the estimate of the parameter will be quite far from the hypothesized value, even if the hypothesis is true. This outcome might cause a type I error.

DEFINITION C.6 Size of a Test
The probability of a type I error is the **size** *of the test. This is conventionally denoted α and is also called the* **significance level**.

The size of the test is under the control of the analyst. It can be changed just by changing the decision rule. Indeed, the type I error could be eliminated altogether just by making the rejection region very small, but this would come at a cost. By eliminating the probability of a type I error—that is, by making it unlikely that the hypothesis is rejected—we must increase the probability of a type II error. Ideally, we would like both probabilities to be as small as possible. It is clear, however, that there is a tradeoff between the two. The best we can hope for is that for a given probability of type I error, the procedure we choose will have as small a probability of type II error as possible.

DEFINITION C.7 Power of a Test
The ***power*** *of a test is the probability that it will correctly lead to rejection of a false null hypothesis:*

$$\text{power} = 1 - \beta = 1 - \text{Prob(type II error)}. \qquad \textbf{(C-16)}$$

For a given significance level α, we would like β to be as small as possible. Because β is defined in terms of the alternative hypothesis, it depends on the value of the parameter.

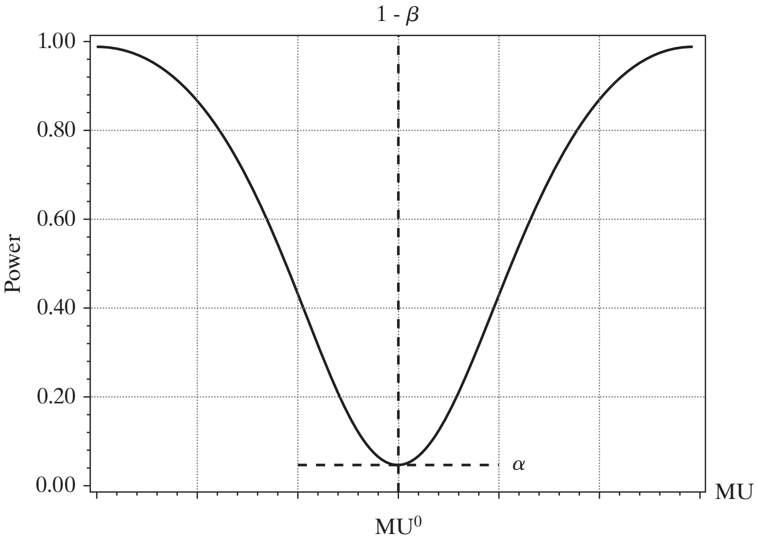
Example C.10 Testing a Hypothesis About a Mean

For testing H_0: $\mu = \mu^0$ in a normal distribution with known variance σ^2, the decision rule is to reject the hypothesis if the absolute value of the z statistic, $\sqrt{n}(\bar{x} - \mu^0)/\sigma$, exceeds the predetermined critical value. For a test at the 5 percent significance level, we set the critical value at 1.96. The power of the test, therefore, is the probability that the absolute value of the test statistic will exceed 1.96 given that the true value of μ is, in fact, not μ^0. This value depends on the alternative value of μ, as shown in Figure C.6. Notice that for this test the power is equal to the size at the point where μ equals μ^0. As might be expected, the test becomes more powerful the farther the true mean is from the hypothesized value.

Testing procedures, like estimators, can be compared using a number of criteria.

DEFINITION C.8 Most Powerful Test
A test is ***most powerful*** *if it has greater power than any other test of the same size.*

FIGURE C.6 Power Function for a Test.



This requirement is very strong. Because the power depends on the alternative hypothesis, we might require that the test be **uniformly most powerful (UMP)**, that is, have greater power than any other test of the same size for all admissible values of the parameter. There are few situations in which a UMP test is available. We usually must be less stringent in our requirements. Nonetheless, the criteria for comparing hypothesis testing procedures are generally based on their respective power functions. A common and very modest requirement is that the test be unbiased.

DEFINITION C.9 Unbiased Test
*A test is **unbiased** if its power* $(1 - \beta)$ *is greater than or equal to its size* α *for all values of the parameter.*

If a test is **biased**, then, for some values of the parameter, we are more likely to retain the null hypothesis when it is false than when it is true.

The use of the term *unbiased* here is unrelated to the concept of an unbiased estimator. Fortunately, there is little chance of confusion. Tests and estimators are clearly connected, however. The following criterion derives, in general, from the corresponding attribute of a parameter estimate.

DEFINITION C.10 Consistent Test
*A test is **consistent** if its power goes to one as the sample size grows to infinity.*

Example C.11 Consistent Test About a Mean

A confidence interval for the mean of a normal distribution is $\bar{x} \pm t_{1-\alpha/2}(s/\sqrt{n})$, where $\bar{x}$ and s are the usual consistent estimators for μ and σ (see Section D.2.1), n is the sample size, and $t_{1-\alpha/2}$ is the correct critical value from the t distribution with $n - 1$ degrees of freedom. For testing H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$, let the procedure be to reject H_0 if the confidence interval does not contain μ_0. Because $\bar{x}$ is consistent for μ, one can discern if H_0 is false as $n \to \infty$, with probability 1, because $\bar{x}$ will be arbitrarily close to the true μ. Therefore, this test is consistent.

As a general rule, a test will be consistent if it is based on a consistent estimator of the parameter.

C.7.2 TESTS BASED ON CONFIDENCE INTERVALS

There is an obvious link between interval estimation and the sorts of hypothesis tests we have been discussing here. The confidence interval gives a range of plausible values for the parameter. Therefore, it stands to reason that if a hypothesized value of the parameter does not fall in this range of plausible values, then the data are not consistent with the hypothesis, and it should be rejected. Consider, then, testing

$$H_0: \theta = \theta_0,\ H_1: \theta \neq \theta_0.$$

We form a confidence interval based on $\hat{\theta}$ as described earlier:

$$\hat{\theta} - C_{1-\alpha/2}[\text{se}(\hat{\theta})] < \theta < \hat{\theta} + C_{1-\alpha/2}[\text{se}(\hat{\theta})].$$

H_0 is rejected if θ_0 exceeds the upper limit or is less than the lower limit. Equivalently, H_0 is rejected if

$$\left| \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \right| > C_{1-\alpha/2}.$$

In words, the hypothesis is rejected if the estimate is too far from θ_0, where the distance is measured in standard error units. The critical value is taken from the t or standard normal distribution, whichever is appropriate.

Example C.12 Testing a Hypothesis About a Mean with a Confidence Interval

For the results in Example C.8, test H_0: $\mu = 1.98$ versus H_1: $\mu \neq 1.98$, assuming sampling from a normal distribution:

$$t = \left| \frac{\bar{x} - 1.98}{s/\sqrt{n}} \right| = \left| \frac{1.63 - 1.98}{0.102} \right| = 3.43.$$

The 95 percent critical value for $t(24)$ is 2.064. Therefore, reject H_0. If the critical value for the standard normal table of 1.96 is used instead, then the same result is obtained.

If the test is one-sided, as in

$$H_0: \theta \geq \theta_0,$$
$$H_1: \theta < \theta_0,$$

then the critical region must be adjusted. Thus, for this test, H_0 will be rejected if a point estimate of θ falls sufficiently below θ_0. (Tests can usually be set up by departing from the decision criterion, “What sample results are inconsistent with the hypothesis?”)

Example C.13 One-Sided Test About a Mean

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

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

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

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

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

C.7.3 SPECIFICATION TESTS

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

APPENDIX D

LARGE-SAMPLE DISTRIBUTION THEORY

D.1 INTRODUCTION

Most of this book is about parameter estimation. In studying that subject, we will usually be interested in determining how best to use the observed data when choosing among competing estimators. That, in turn, requires us to examine the sampling behavior of estimators. In a