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 $\mathbf{A}$ is singular or, equivalently, that the transformations from $\mathbf{x}$ to $\mathbf{y}$ are functionally dependent. The singular case is analogous to the single-variable case.

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

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** 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**, or DGP 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 $\text{Prob}(X = x)$.

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,

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

The axioms of probability require that

1. $0 \leq \text{Prob}(X = x) \leq 1.$ **(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 of x. The **probability density function (pdf)** is defined so that $f(x) \geq 0$ and

1. $$\text{Prob}(a \leq x \leq b) = \int_a^b f(x)\,dx \geq 0.$$ **(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,

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

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 distribution function (cdf)**. For a discrete random variable,

$$F(x) = \sum_{X \leq x} f(X) = \text{Prob}(X \leq x).$$ **(B-6)**

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

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

For a continuous random variable,

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

and

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

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

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

From the definition of the cdf,

$$\text{Prob}(a < x \leq b) = F(b) - F(a).$$ **(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. 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 $\text{Prob}(X \leq m) \geq \frac{1}{2}$ and $\text{Prob}(X \geq m) \geq \frac{1}{2}$, 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.

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

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

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

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

$\text{Var}[x]$, 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:

$$\text{Var}[x] = E[x^2] - \mu^2. \tag{B-14}$$

A convenient corollary to (B-14) is

$$E[x^2] = \sigma^2 + \mu^2. \tag{B-15}$$

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

$$\text{Var}[a + bx] = b^2 \text{Var}[x], \tag{B-16}$$

which implies, for any constant a, that

$$\text{Var}[a] = 0. \tag{B-17}$$

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

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

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.

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

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

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

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

and

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

(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:

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

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

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

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 $E[g(x)] \leq g(E[x])$. For example, $E[\log(x)] \leq \log(E[x])$.

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 DISTRIBUTION

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

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

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

[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 (1993). Johnson et al. (1974, 1993, 1994, 1995, 1997) is an encyclopedic reference on the subject of statistical distributions.

is its preservation under linear transformation.

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

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

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

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

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

FIG B.1

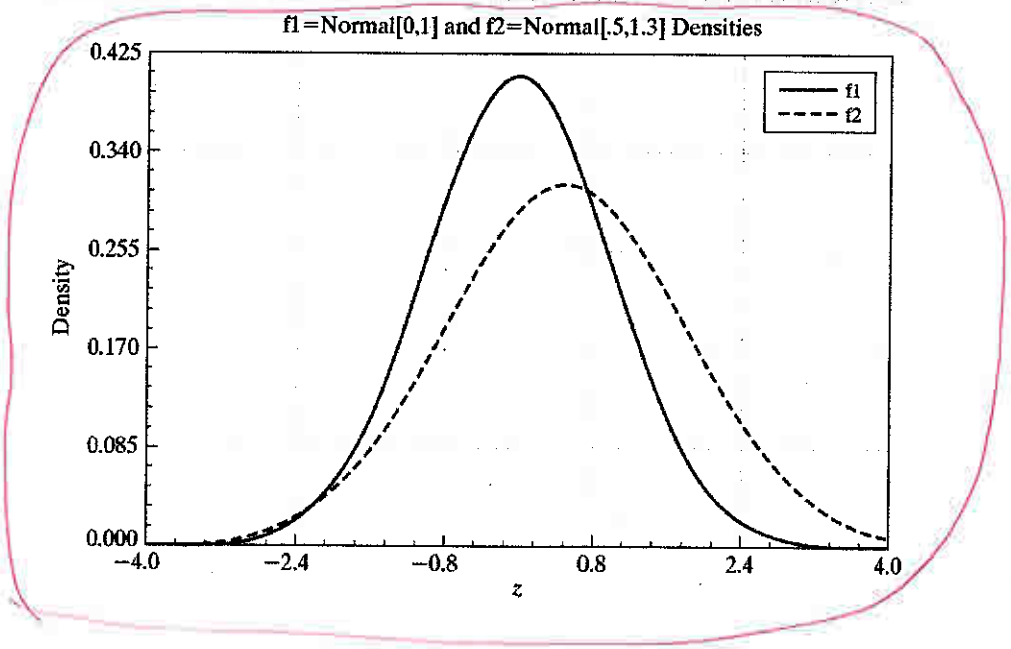
Figure B.1 shows the densities of the standard normal distribution and the normal distribution with mean 0.5, which shifts the distribution to the right, and standard deviation 1.3, which, it can be seen, scales the density so that it is shorter but wider. (The graph is a bit deceiving unless you look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

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

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.



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.

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

$$z^2 \sim \chi^2[1]. \tag{B-30}$$

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

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

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

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

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

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

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

$$x_1 + x_2 \sim \chi^2[n_1 + n_2]. \tag{B-34}$$

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

FIG B.2

Figure B.2 shows the chi-squared density for three 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. Table G.3 in Appendix G of this book gives lower (left) tail areas for a number of values.

- If x_1 and x_2 are two *independent* chi-squared variables with degrees of freedom parameters n_1 and n_2, respectively, then the ratio

$$F[n_1, n_2] = \frac{x_1/n_1}{x_2/n_2} \tag{B-35}$$

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

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.

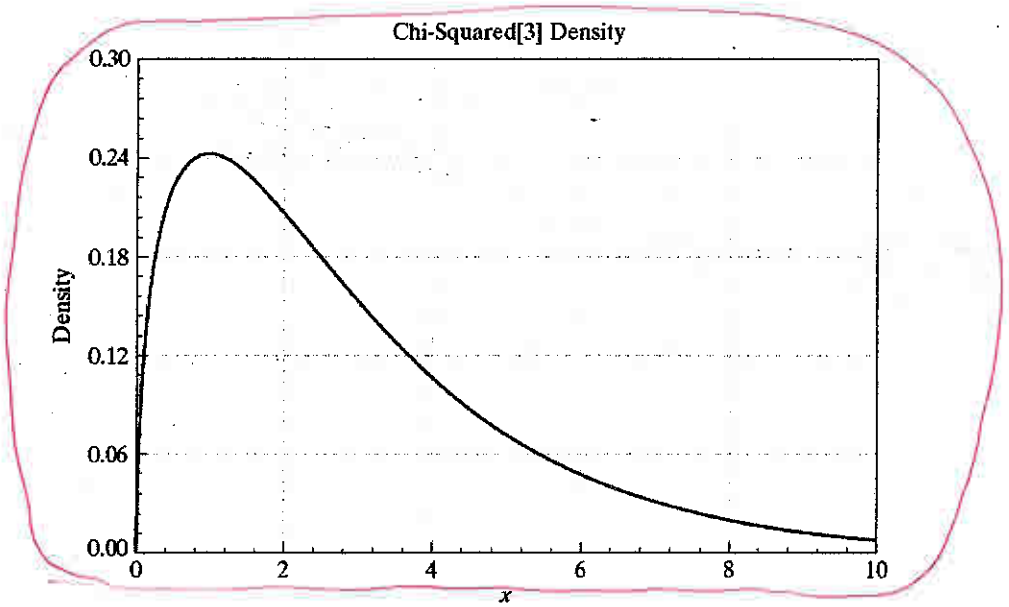


FIGURE B.2 The Chi-Squared [3] Distribution.

- 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.3 illustrates the t distributions with three 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 distributions arise in the F tests of linear restrictions [see (5-6)] 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 $\chi_*^2[J, \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/2]$.
   b. If $\mathbf{z} \sim N[\boldsymbol{\mu}, \mathbf{I}]$ and $\mathbf{M}$ is an idempotent matrix with rank J, then $\mathbf{z}'\mathbf{M}\mathbf{z} \sim \chi_*^2[J, \boldsymbol{\mu}'\mathbf{M}\boldsymbol{\mu}/2]$.

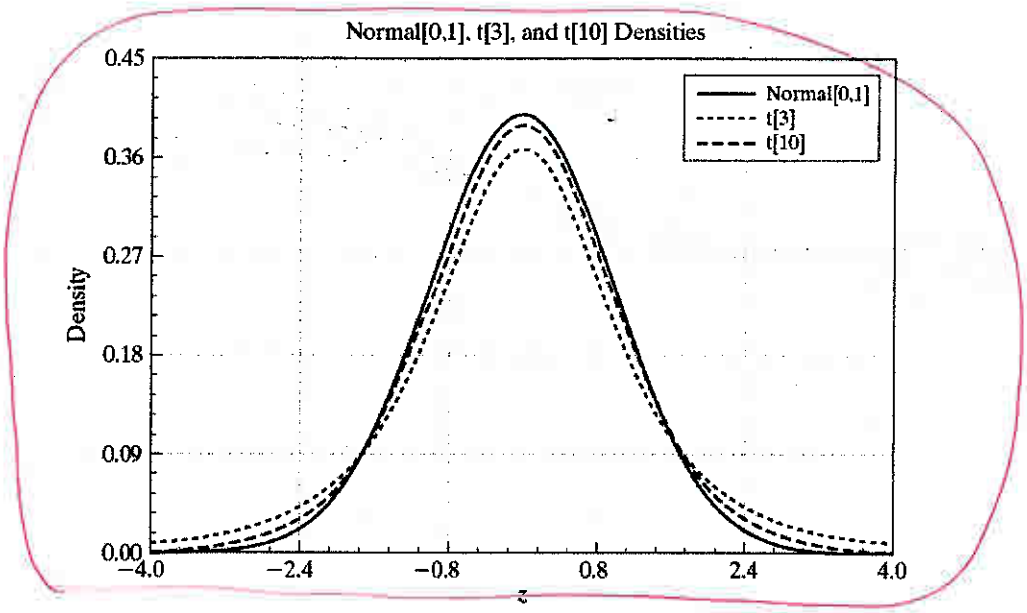


FIGURE B.3 The Standard Normal, t[3], and t[10] Distributions.

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.[2] Note that 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 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,

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

[2] The denominator chi-squared could also be noncentral, but we shall not use any statistics with doubly noncentral distributions.

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.[3] 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.[4]

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

A lognormal variable x has

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

and

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

The relation between the normal and lognormal distributions is

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

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

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

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[5] and production functions.[6] 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

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

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

[5] Salem and Mount (1974).

[6] Greene (1980a).

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

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

This functional form is extremely flexible in the shapes it will accommodate. It is symmetric if $\alpha = \beta$, asymmetric otherwise, and 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.[7]

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 $\pi^2/3$.

B.4.8 THE WISHART DISTRIBUTION

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

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

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

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

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

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

[7] Heckman and Willis (1976).

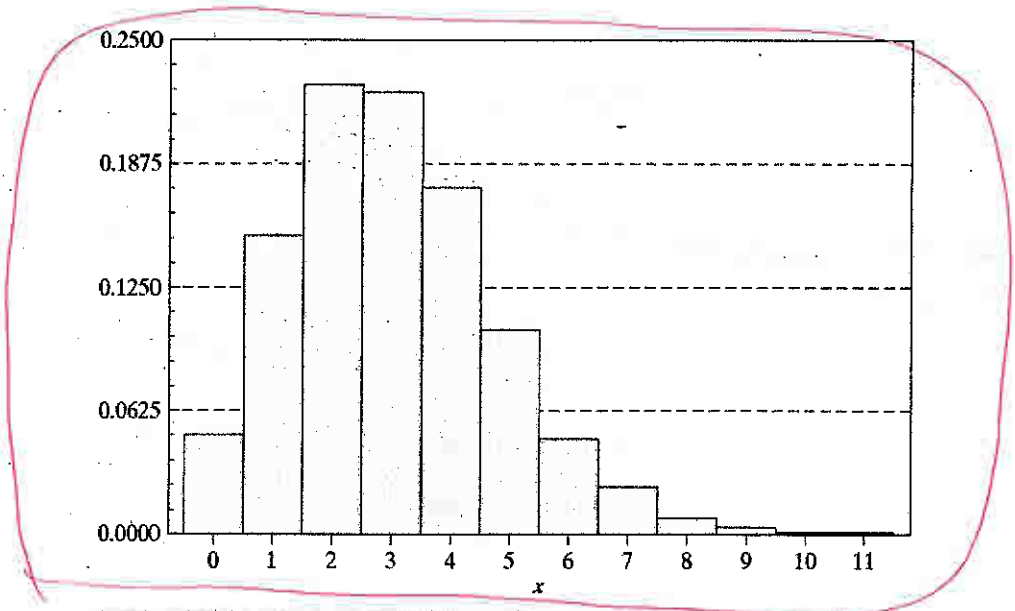


FIGURE B.4 The Poisson [3] Distribution.

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 25.) An example is shown in Figure B.4.

FIG B.4

AU: Confirm x-ref to Chap 25

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.

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 $\text{Prob}(Y = y)$ 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.5.

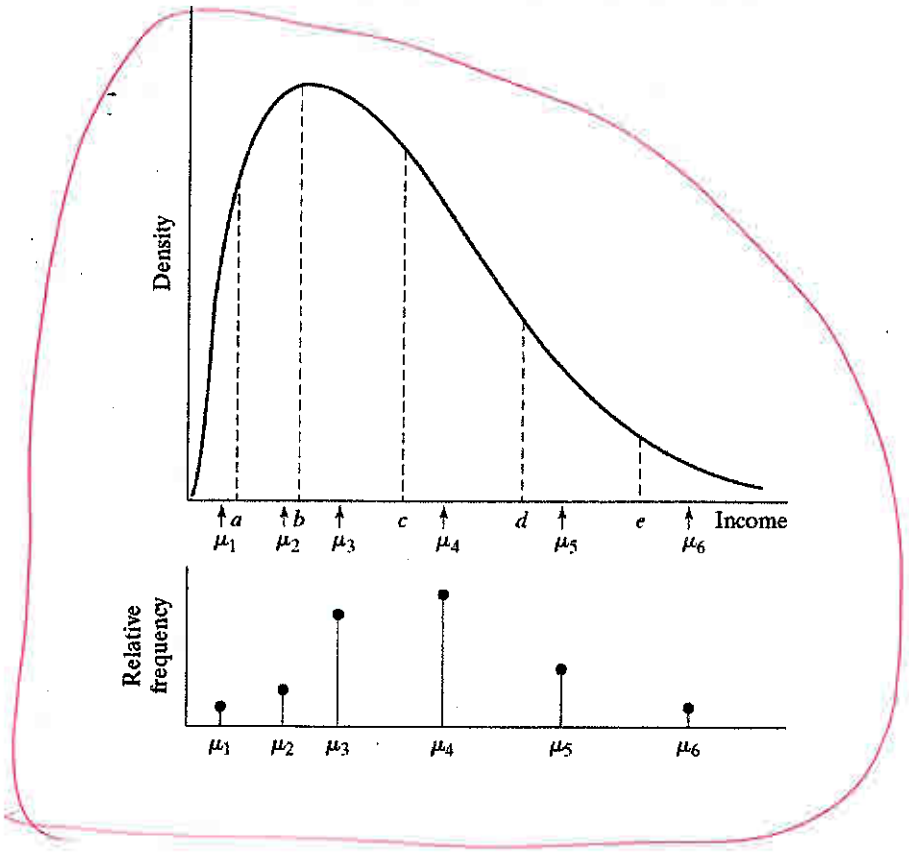
Data are often reported categorically, as shown in the lower part of the figure. Thus, the random variable corresponding to observed income is a discrete transformation of the actual underlying continuous random variable. Suppose, for example, that the transformed variable y is the mean income in the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

$$\text{Prob}(Y = \mu_1) = P(-\infty < X \leq a),$$
$$\text{Prob}(Y = \mu_2) = P(a < X \leq b),$$
$$\text{Prob}(Y = \mu_3) = P(b < X \leq c),$$

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

FIGURE B.5 Censored Distribution.



the cdf of y:

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

This equation can now be written as

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

Hence,

$$f_y(y) = f_x(g^{-1}(y))|g^{-1\prime}(y)|. \tag{B-41}$$

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 $|g^{-1\prime}(y)|$ 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

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

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:

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

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.** 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 (MGF) 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

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

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

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

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:

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

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:

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

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

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

or

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

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,

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

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

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

Variances are computed in the same manner:

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

B.7.3 COVARIANCE AND CORRELATION

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

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

The covariance of x and y is a special case:

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

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:

$$r[x, y] = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{B-53}$$

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

$$E[ax + by + c] = aE[x] + bE[y] + c, \tag{B-54}$$

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

and

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

If X and Y are uncorrelated, then

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

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

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

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,

$$J = \begin{bmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{bmatrix} = \begin{bmatrix} \partial \mathbf{x} \\ \overline{\partial \mathbf{y}'} \end{bmatrix}.$$

The joint distribution of y_1 and y_2 is

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

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

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

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

$$\mathbf{y} = \mathbf{a} + \mathbf{B}\mathbf{x}.$$

The inverse transformation is

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

so the absolute value of the determinant of the Jacobian is

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

The joint distribution of $\mathbf{x}$ 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

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

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

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

The density of y_1 would then be

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

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

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

and

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

It follows from (B-46) that:

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

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

$$\begin{aligned} f(x, y) &= f(y\,|\,x) f_x(x) \\ &= f(x\,|\,y) f_y(y). \end{aligned} \tag{B-61}$$

B.8.1 REGRESSION: THE CONDITIONAL MEAN

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

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

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

A random variable may always be written as

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

B.8.2 CONDITIONAL VARIANCE

A conditional variance is the variance of the conditional distribution:

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

or

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

The computation can be simplified by using

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

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 $\text{Var}[y \mid x]$ equals $\text{Var}[y]$, 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

$$E[y] = E_x[E[y \mid x]]. \tag{B-66}$$

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,

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

(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 $E[y|x] = \alpha + \beta x$, then

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

and

$$\beta = \frac{\text{Cov}[x, y]}{\text{Var}[x]}. \tag{B-68}$$

The proof follows from (B-66).

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,

$$\text{Var}[y] = \text{Var}_x[E[y|x]] + E_x[\text{Var}[y|x]]. \tag{B-69}$$

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,

$$E_x[\text{Var}[y|x]] = \text{Var}[y] - \text{Var}_x[E[y|x]]. \tag{B-70}$$

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 $E[y \mid x] = \alpha + \beta x$ *and if* $\text{Var}[y \mid x]$ *is a constant, then*

$$\text{Var}[y \mid x] = \text{Var}[y]\left(1 - \text{Corr}^2[y, x]\right) = \sigma_y^2\left(1 - \rho_{xy}^2\right). \tag{B-71}$$

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 $E[y \mid x]$ varies with x:

$$\textbf{regression variance} = \text{Var}_x[E[y \mid x]]. \tag{B-72}$$

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

$$\textbf{residual variance} = E_x[\text{Var}[y \mid x]]. \tag{B-73}$$

Thus,

$$\text{Var}[y] = \text{regression variance} + \text{residual variance}. \tag{B-74}$$

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

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

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

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

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.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

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-1/2[(\varepsilon_x^2 + \varepsilon_y^2 - 2\rho\varepsilon_x\varepsilon_y)/(1-\rho^2)]},$$
$$\varepsilon_x = \frac{x - \mu_x}{\sigma_x}, \quad \varepsilon_y = \frac{y - \mu_y}{\sigma_y}. \tag{B-77}$$

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

$$\sigma_{xy} = \rho\sigma_x\sigma_y. \tag{B-78}$$

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

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

then

- The marginal distributions are normal:

$$\begin{aligned} f_x(x) &= N\left[\mu_x, \sigma_x^2\right], \\ f_y(y) &= N\left[\mu_y, \sigma_y^2\right]. \end{aligned} \tag{B-79}$$

- The conditional distributions are normal:

$$\begin{aligned} f(y \mid x) &= N\left[\alpha + \beta x, \sigma_y^2(1-\rho^2)\right], \\ \alpha &= \mu_y - \beta\mu_x, \quad \beta = \frac{\sigma_{xy}}{\sigma_x^2}, \end{aligned} \tag{B-80}$$

  and likewise for $f(x \mid y)$.
- 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

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

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

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

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,

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

which is the **covariance matrix** of the random vector $\mathbf{x}$. 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,

$$\begin{aligned} E[a_1x_1 + a_2x_2 + \cdots + a_nx_n] &= E[\mathbf{a}'\mathbf{x}] \\ &= a_1E[x_1] + a_2E[x_2] + \cdots + a_nE[x_n] \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \\ &= \mathbf{a}'\boldsymbol{\mu}. \end{aligned} \tag{B-84}$$

For the variance,

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

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,

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

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

$$E[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\mu}. \tag{B-86}$$

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 $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$,

$$\text{Var}[\mathbf{Ax}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \tag{B-87}$$

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

B.10.3 NONLINEAR FUNCTIONS

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 ith 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 $\varepsilon_i = (x_i - \mu_i)/\sigma_i$.[8]

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

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

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

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

Finally, if $\sigma = 1$,

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

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

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

[8] 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 $(x_i - \mu_i)/\sigma_i$.

THEOREM B.7 (Continued)
and

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

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

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

where

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

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

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,

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

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:

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

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 $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\sigma}_{\mathbf{xy}}$, where $\boldsymbol{\sigma}_{\mathbf{xy}}$ 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

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

where $\boldsymbol{\beta}$ is given earlier, $\alpha = \mu_y - \boldsymbol{\beta}'\boldsymbol{\mu}_x$, 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,

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

If $\mathbf{A}$ does not have full rank, then $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ 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 $\boldsymbol{\Sigma}$. 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

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

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 $\mathbf{A}$:

$$q = \mathbf{x}'\mathbf{A}\mathbf{x}. \tag{B-105}$$

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

$$q = \mathbf{x}'\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'\mathbf{x}. \tag{B-106}$$

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,

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

If λ_i is always one or zero, then

$$q = \sum_{j=1}^{J} y_j^2, \tag{B-108}$$

which has a chi-squared distribution. The sum is taken over the $j = 1, \ldots, J$ 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.

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

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

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* $\mathbf{A}\mathbf{B} = \mathbf{0}$. (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

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

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

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

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,[9] 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 $\mathbf{A}\mathbf{B} = \mathbf{0}$. 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,

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

We are interested in the distribution of

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

[9] 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.

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

$$q = \mathbf{z}'\boldsymbol{\Sigma}^{-1}\mathbf{z} = \mathbf{z}'\left(\text{Var}[\mathbf{z}]\right)^{-1}\mathbf{z}. \tag{B-117}$$

where $\mathbf{z}$ is normally distributed with mean $\mathbf{0}$. This equation is a quadratic form, but not necessarily in an idempotent matrix.[10] 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 $\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, *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 $\mathbf{x}$ Is Normal
If $\mathbf{x} \sim N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, *then* $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2[n]$.

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.

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