Appendix A

Mathematical Tools

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Explain the relationship between exponential functions and natural logarithms.
2. Explain and apply scientific notation.
3. Define a linear relationship, as opposed to a nonlinear relationship.
4. Compute the elasticity at a point on a function.
5. Explain the concept of a derivative and its relationship to the slope of a function.
6. Compute the derivatives of simple functions and provide their interpretations.
7. Describe the relationship between a derivative and a partial derivative.
8. Explain the concept of an integral.
9. Maximize or minimize functions of one or two variables.
10. Use integration to find the area under curves.
11. Explain and evaluate second derivatives.

KEYWORDS

absolute value
antilogarithm
derivatives
e
elasticity
exponential function
exponents
inequalities
integers
integral
intercept
irrational number
linear relationship
logarithm
marginal effect
maximizing a function
minimizing a function
natural logarithm
nonlinear relationship
partial derivative
percentage change
product rule
quadratic function
quotient rule
rational numbers
real numbers
relative change
scientific notation
second derivative
slope
Taylor series

We assume that you have studied basic math. Hopefully you understand the calculus concepts of differentiation and integration, though these tools are *not required* prerequisites for success using this book. In this appendix we review some essential concepts that you may wish to consult from time to time.[1]

[1] Summation signs and operations are covered in the Probability Primer that precedes Chapter 2.

A.1 Some Basics

A.1.1 Numbers

Integers are the whole numbers, 0, ±1, ±2, ±3, ... The positive integers are the counting numbers. **Rational numbers** can be written as a/b, where a and b are integers and $b \neq 0$. The **real numbers** can be represented by points on a line. There are an uncountable number of real numbers, and they are not all rational. Numbers such as $\pi \cong 3.1415927$ and $\sqrt{2}$ are said to be **irrational** since they cannot be expressed as ratios, and have only decimal representations. Numbers like $\sqrt{-2}$ are not real numbers. The **absolute value** of a number is denoted by $|a|$. It is the positive part of the number: $|3| = 3$ and $|-3| = 3$.

Inequalities among numbers obey certain rules. The notation $a < b$, a is less than b, means that a is to the left of b on the number line, and that $b - a > 0$. If a is less than or equal to b, it is written as $a \leq b$. Three basic rules are

$$\text{If } a < b, \text{ then } a + c < b + c$$

$$\text{If } a < b, \text{ then } \begin{cases} ac < bc & \text{if } c > 0 \\ ac > bc & \text{if } c < 0 \end{cases}$$

$$\text{If } a < b \text{ and } b < c, \text{ then } a < c$$

A.1.2 Exponents

Exponents are defined as follows:

$x^n = xx \cdots x$ (n terms) if n is a positive integer

$x^0 = 1$ if $x \neq 0$. 0^0 does not have meaning and is "undefined."

Some common rules for working with exponents, assuming x and y are real, m and n are integers, and a and b are rational, are as follows:

$$x^{-n} = \frac{1}{x^n} \text{ if } x \neq 0. \text{ For example, } x^{-1} = \frac{1}{x}$$

$$x^{1/n} = \sqrt[n]{x}. \text{ For example, } x^{1/2} = \sqrt{x} \text{ and } x^{-1/2} = \frac{1}{\sqrt{x}}$$

$$x^{m/n} = \left(x^{1/n}\right)^m. \text{ For example, } 8^{4/3} = \left(8^{1/3}\right)^4 = 2^4 = 16$$

$$x^a x^b = x^{a+b}, \ \frac{x^a}{x^b} = x^{a-b}$$

$$\left(\frac{x}{y}\right)^a = \frac{x^a}{y^a}, \ (xy)^a = x^a y^a$$

A.1.3 Scientific Notation

Scientific notation is useful for very large or very small numbers. A number in scientific notation is written as a number between 1 and 10 multiplied by a power of 10. So, for example: $5.1 \times 10^5 = 510,000$, and $0.00000034 = 3.4 \times 10^{-7}$. Scientific notation makes handling large numbers much easier, because complex operations can be broken into simpler ones. For example,

$$\begin{aligned} 510{,}000 \times 0.00000034 &= \left(5.1 \times 10^5\right) \times \left(3.4 \times 10^{-7}\right) \\ &= (5.1 \times 3.4) \times \left(10^5 \times 10^{-7}\right) \\ &= 17.34 \times 10^{-2} \\ &= 0.1734 \end{aligned}$$

and

$$\frac{510{,}000}{0.00000034} = \frac{5.1 \times 10^5}{3.4 \times 10^{-7}} = \frac{5.1}{3.4} \times \frac{10^5}{10^{-7}} = 1.5 \times 10^{12}$$

Computer programs sometimes write $5.1 \times 10^5 = 5.1\text{E}5$ or $5.1\text{D}5$ and $3.4 \times 10^{-7} = 3.4\text{E}{-7}$ or $3.4\text{D}{-7}$.

A.1.4 Logarithms and the Number *e*

Logarithms are exponents. If $x = 10^b$, then b is the **logarithm** of x using the base 10. The **irrational number** $\boldsymbol{e} \cong 2.718282$ is used in mathematics and statistics as the base for logarithms. If $x = e^b$, then b is the logarithm of x using the base e. Logarithms using the number e as base are called **natural logarithms**. *All logarithms in this book are natural logarithms*. We express the natural logarithm of x as $\ln(x)$,

For any positive number, $x > 0$,

$$e^{\ln(x)} = \exp\left[\ln(x)\right] = x$$

and

$$\ln\left(e^x\right) = x$$

Note that $\ln(1) = 0$, using the laws of exponents. Table A.1 gives the logarithms of some powers of 10. For example, $e^{2.3025851} = 10$ and $e^{4.6051702} = 100$.

Note that logarithms have a compressed scale compared to the original numbers. Since logarithms are exponents, they follow similar rules:

$$\ln(xy) = \ln(x) + \ln(y)$$
$$\ln(x/y) = \ln(x) - \ln(y)$$
$$\ln\left(x^a\right) = a\ln(x)$$

For example, if $x = 1000$ and $y = 10,000$, then

$$\begin{aligned}\ln(1000 \times 10{,}000) &= \ln(1000) + \ln(10{,}000)\\ &= 6.9077553 + 9.2103404\\ &= 16.118096\end{aligned}$$

What is the advantage of this? The value of xy is a multiplication problem, which by using logarithms we can turn into an addition problem. We need a way to go backward, from the logarithm of a number to the number itself. By definition,

$$x = e^{\ln(x)} = \exp\left[\ln(x)\right]$$

TABLE A.1 Some Natural Logarithms

x	$\ln(x)$
1	0
10	2.3025851
100	4.6051702
1,000	6.9077553
10,000	9.2103404
100,000	11.512925
1,000,000	13.815511

When there is an **exponential function** with a complicated exponent, the notation **exp** is often used, so that $e^{(\cdot)} = \exp(\cdot)$. The exponential function is the **antilogarithm**, because we can recover the value of x using it. Then,

$$1000 \times 10000 = \exp(16.118096) = 10{,}000{,}000$$

You will not be doing many calculations like these, but the knowledge of logarithms and exponents is quite critical in economics and econometrics.

A.1.5 Decimals and Percentages

Suppose the value of a variable y changes from the value $y = y_0$ to $y = y_1$. The difference between these values is often denoted by $\Delta y = y_1 - y_0$, where the notation Δy is read "change in y," or "delta-y." The **relative change in** y is defined to be

$$\text{relative change in } y = \frac{y_1 - y_0}{y_0} = \frac{\Delta y}{y_0} \tag{A.1}$$

For example, if $y_0 = 3$ and $y_1 = 3.02$, then the relative change in y is

$$\frac{y_1 - y_0}{y_0} = \frac{3.02 - 3}{3} = 0.0067$$

Often the relative change in y is written as $\Delta y/y$, omitting the subscript.

A relative change is a decimal. The corresponding **percentage change** in y is 100 times the relative change.

$$\text{percentage change in } y = 100\frac{y_1 - y_0}{y_0} = \%\Delta y \tag{A.2}$$

If $y_0 = 3$ and $y_1 = 3.02$, then the percentage change in y is

$$\%\Delta y = 100\frac{y_1 - y_0}{y_0} = 100\frac{3.02 - 3}{3} = 0.67\%$$

A.1.6 Logarithms and Percentages

A feature of logarithms that helps greatly in their economic interpretation is that they can be approximated very simply. Let y_1 be a positive value of y, and let y_0 be a value of y that is "close" to y_1. A useful approximation rule is

$$100\left[\ln(y_1) - \ln(y_0)\right] \cong \%\Delta y = \text{percentage change in } y \tag{A.3}$$

That is, 100 times the difference in the logarithms is the approximate percentage difference between y_0 and y_1, if y_0 and y_1 are close.

Derivation of the Approximation The result in (A.3) follows from the mathematical tool called a **Taylor series** approximation, which is developed in Example A.3 in Section A.3.1. Using this approximation, the value of $\ln(y_1)$ can be written as

$$\ln(y_1) \cong \ln(y_0) + \frac{1}{y_0}(y_1 - y_0) \tag{A.4}$$

For example, let $y_1 = 1 + x$ and let $y_0 = 1$. Then, as long as x is small,

$$\ln(1 + x) \cong x$$

Subtracting $\ln(y_0)$ from both sides of (A.4), we obtain

$$\ln(y_1) - \ln(y_0) = \Delta\ln(y) \cong \frac{1}{y_0}(y_1 - y_0) = \text{relative change in } y$$

The symbol $\Delta\ln(y)$ represents the "difference" between two logarithms. Using (A.2),

$$\begin{aligned} 100\Delta\ln(y) &= 100\left[\ln(y_1) - \ln(y_0)\right] \\ &\cong 100 \times \frac{(y_1 - y_0)}{y_0} = \%\Delta y \\ &= \text{percentage change in } y \end{aligned}$$

A.2 Linear Relationships

In economics, and in econometrics, we study linear and nonlinear relationships between variables. In this section, we review basic characteristics of **linear relationships**. Let y and x be variables. The standard form for a linear relationship is

$$y = mx + b \tag{A.5}$$

In Figure A.1, the **slope** is m and the y-**intercept** is b. The symbol Δ represents "a change in," so Δx is read as a "change in x." The slope of the line is

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x}$$

For the straight-line relationship in Figure A.1, the slope m is the ratio of the change in vertical distance (rise) to the change in horizontal distance (run) as a point moves along the line in either direction. The slope of a straight line is constant; the rate at which y changes as x changes is constant over the length of the straight line.

The slope m is very meaningful to economists as it is the **marginal effect** of a change in x on y. To see this, solve the **slope** definition $m = \Delta y/\Delta x$ for Δy, obtaining

$$\Delta y = m\Delta x \tag{A.6}$$

If x changes by one unit, $\Delta x = 1$, then the corresponding change in y is $\Delta y = m$. The marginal effect, m, is always the same for a linear relationship like (A.5), because the slope is constant.

The **intercept** parameter indicates where the linear relationship crosses the vertical axis—that is, it is the value of y when x is zero,

$$y = mx + b = m \times 0 + b = b$$

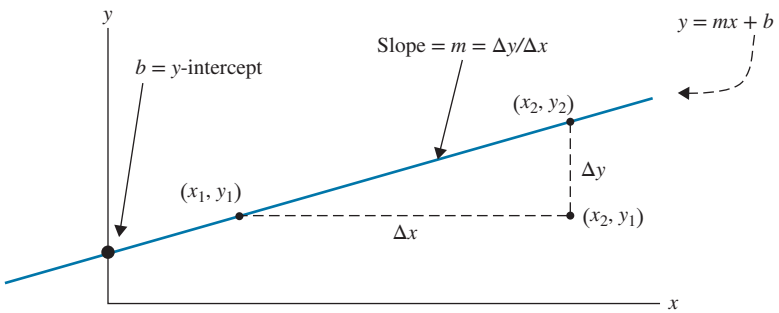


FIGURE A.1 **A linear relationship.**

A.2.1 Slopes and Derivatives

Derivatives have an important role in econometrics. In a relationship between two variables, $y = f(x)$, the **first derivative** measures the slope. The slope of the line $y = f(x) = mx + b$ is denoted as dy/dx. The notation dy/dx is a "stylized" version of $\Delta y/\Delta x$, and for the linear relationship (A.5) the first derivative is

$$dy/dx = m \tag{A.7}$$

In general, the first derivative measures the change in the function value y given an infinitesimal change in x. For the linear function the first derivative is the constant $m = \Delta y/\Delta x$. The "infinitesimal" does not matter in this case, because the rate of change of y with respect to changes in x is a constant.

A.2.2 Elasticity

A favorite tool of the economist is **elasticity**. It is the percentage change in one variable associated with a 1% change in another variable for movements along a specific curve. That is, if we move from one point on a curve to another point on the curve, what are the relative percentage changes? For example, in Figure A.1, what is the percentage change in y relative to the percentage change in x as we move from the point (x_1, y_1) to (x_2, y_2)? For a linear relationship, the elasticity of y with respect to a change in x is

$$\varepsilon_{yx} = \frac{\%\Delta y}{\%\Delta x} = \frac{100(\Delta y/y)}{100(\Delta x/x)} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x} \times \frac{x}{y} = slope \times \frac{x}{y} \tag{A.8}$$

The elasticity is the product of the slope of the relationship and the ratio of an x value to a y value. In a linear relationship, such as Figure A.1, while the slope is constant, $m = \Delta y/\Delta x$, the elasticity changes at every (x, y) point on the line.

Consider, for example, the linear function $y = 1x + 1$. At the point $x = 2$ and $y = 3$, which is on the line, the elasticity is $\varepsilon_{yx} = m(x/y) = 1 \times (2/3) = 0.67$. That is, at the point $(x = 2, y = 3)$ a 1% change in x is associated with a 0.67% change in y. Specifically, at $x = 2$ a 1% (1% = 0.01 in decimal form) change is $\Delta x = 0.01 \times 2 = 0.02$. If x increases to $x = 2.02$, the value of y increases to 3.02. The **relative change** in y is $\Delta y/y = 0.02/3 = 0.0067$. This, however, is not the percentage change in y, but rather the decimal equivalent. To obtain the percentage change in y, which we denote $\%\Delta y$, we multiply the relative change $\Delta y/y$ by 100. The **percentage change** in y is

$$\%\Delta y = 100 \times (\Delta y/y) = 100 \times 0.02/3 = 100 \times 0.0067 = 0.67\%$$

A.3 Nonlinear Relationships

While linear relationships are intuitive and easy to work with, many real-world economic relationships are nonlinear, as illustrated in Figure A.2.

The slope of this curve is not constant. The slope measures the marginal effect of x on y, and for a **nonlinear relationship** like that in Figure A.2, the slope is different at every point on the curve. The changing slope tells us that the relationship is not linear. Since the slope is different at every point, we can only talk about the effect of small changes in x on y. In (A.6) we replace Δ, the symbol for "a change in," with d, which we will take to mean an "infinitesimal change in." In the linear case when we made this replacement, the slope was given by $dy/dx = m$, where m was a constant. See equation (A.7).

However, with nonlinear functions such as that in Figure A.2, the slope (derivative) is not constant, but changes as x changes, and must be determined at each point. Strictly speaking, the slope of a curve is the slope of the **tangent** to the curve at a specific point. To work out the slope at different points on a nonlinear curve, we need some rules for obtaining the derivative dy/dx.

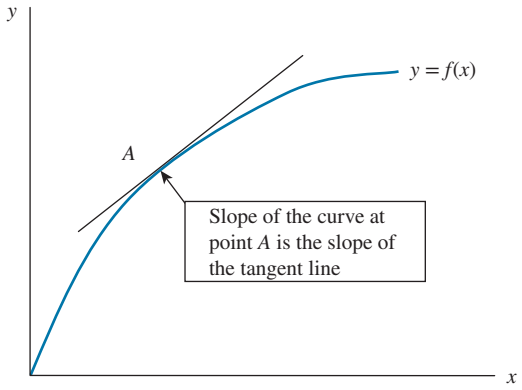


FIGURE A.2 **A nonlinear relationship.**

A.3.1 Rules for Derivatives

Some rules for finding derivatives are the following:

Derivative Rule 1. The derivative of a constant c is zero, that is, if $y = f(x) = c$, then

$$\frac{dy}{dx} = 0$$

Derivative Rule 2. If $y = x^n$, then

$$\frac{dy}{dx} = nx^{n-1}$$

Derivative Rule 3. If $y = cu$ and $u = f(x)$, then

$$\frac{dy}{dx} = c\frac{du}{dx}$$

Constants can be factored out of functions before taking the derivative.

Derivative Rule 4. If $y = cx^n$, using Rules 2 and 3,

$$\frac{dy}{dx} = cnx^{n-1}$$

Derivative Rule 5. If $y = u + v$, where $u = f(x)$ and $v = g(x)$ are functions of x, then

$$\frac{dy}{dx} = \frac{du}{dx} + \frac{dv}{dx}$$

The derivative of the sum (or difference) of two functions is the sum (or difference) of the derivatives. This rule extends to more than two terms in a sum.

Derivative Rule 6. If $y = uv$, where $u = f(x)$ and $v = g(x)$ are functions of x, then

$$\frac{dy}{dx} = \frac{du}{dx}v + u\frac{dv}{dx}$$

This is called the **product rule**. The **quotient rule**, for $y = u/v$, is obtained by inserting v^{-1} for v in the product rule.

Derivative Rule 7. If $y = e^x$, then

$$\frac{dy}{dx} = e^x$$

If $y = \exp(ax + b)$, then

$$\frac{dy}{dx} = \exp(ax + b) \times a$$

In general, the derivative of the exponential function is the exponential function times the derivative of the exponent.

Derivative Rule 8. If $y = \ln(x)$, then

$$\frac{dy}{dx} = \frac{1}{x}, \quad x > 0$$

If $y = \ln(ax + b)$, then

$$\frac{dy}{dx} = \frac{1}{ax + b} \times a$$

Derivative Rule 9. (The Chain Rule of Differentiation). Let $y = f[u(x)]$, so that y depends on u which in turn depends on x. Then

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$$

For example, in Derivative Rule 8, $y = \ln(ax + b)$, or $y = \ln[u(x)]$ where $u = ax + b$. Then

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx} = \frac{1}{u} \times a = \frac{1}{ax + b} \times a$$

EXAMPLE A.1 | Slope of a Linear Function

The derivative of $y = f(x) = 4x + 1$ is

$$\frac{dy}{dx} = \frac{d(4x)}{dx} + \frac{d(1)}{dx} = 4$$

Because this function is the equation of a straight line, $y = mx + b$, its slope is constant and given by the coefficient of x, which in this case is 4.

EXAMPLE A.2 | Slope of a Quadratic Function

Consider the function $y = x^2 - 8x + 16$, shown in Figure A.3. This **quadratic function** is a parabola. Using the rules of derivatives, the slope of a line tangent to the curve is

$$\frac{dy}{dx} = \frac{d(x^2 - 8x + 16)}{dx} = \frac{d(x^2)}{dx} - 8\frac{d(x^1)}{dx} + \frac{d(16)}{dx}$$

$$= 2x^1 - 8x^0 + 0 = 2x - 8$$

This result means that the slope of the tangent line to this curve is $dy/dx = 2x - 8$. The derivative and function values are shown for several values of x in Table A.2.

Note a few things. First, the slope is different at each value of x. The slope is negative for values of $x < 4$, the slope is zero when $x = 4$, and the slope is positive for values of $x > 4$. To interpret these slopes, recall that the derivative of a function at a point is the slope of the tangent at that point. The slope of the tangent is the **rate of change** of the function—how much $y = f(x)$ is changing as x changes. At $x = 0$, the derivative is -8, indicating that y is falling as x increases, and that the rate of change is 8 units in y per unit change in x. At $x = 2$, the rate of change of the function has diminished, and at $x = 4$, the rate of change of the function is $dy/dx = 0$. That is, at $x = 4$, the slope of the tangent to the curve is zero. For values of $x > 4$, the derivative is positive, which indicates that the function $y = f(x)$ is increasing as x increases.

TABLE A.2 **The Function $y = x^2 - 8x + 16$ and Derivative Values**

x	$y = f(x)$	dy/dx
0	16	−8
2	4	−4
4	0	0
6	4	4
8	16	8

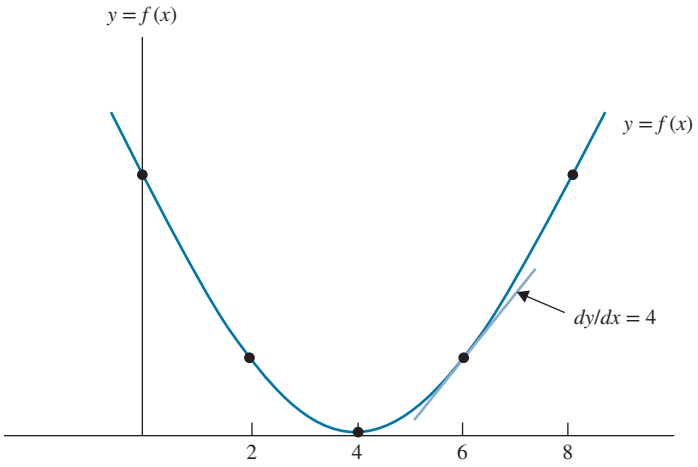


FIGURE A.3 **The function $y = x^2 - 8x + 16$.**

EXAMPLE A.3 | Taylor Series Approximation

The approximation of the logarithm in (A.4) uses a very powerful tool called a Taylor series approximation. For the function $f(y) = \ln(y)$ it is illustrated in Figure A.4. Assume that we know the point A on the function: for $y = y_0$, we know the function value $f(y_0) = \ln(y_0)$. The approximation idea is to draw a line tangent to the curve $f(y) = \ln(y)$ at A, then approximate the point on the curve $f(y_1) = \ln(y_1)$ by the point B on the tangent line. For a smooth curve like $\ln(y)$, this strategy works well, and the approximation error will be small if y_1 is close to y_0. The slope of the tangent line at point A, $(y_0, f(y_0) = \ln(y_0))$, is the derivative of the function $f(y) = \ln(y)$ evaluated at y_0. Using Derivative Rule 8, we have

$$\left.\frac{d\ln(y)}{dy}\right|_{y=y_0} = \left.\frac{1}{y}\right|_{y=y_0} = \frac{1}{y_0}$$

The value of the linear approximation at B is given by geometry. Recall that the slope of the tangent (straight) line is "the

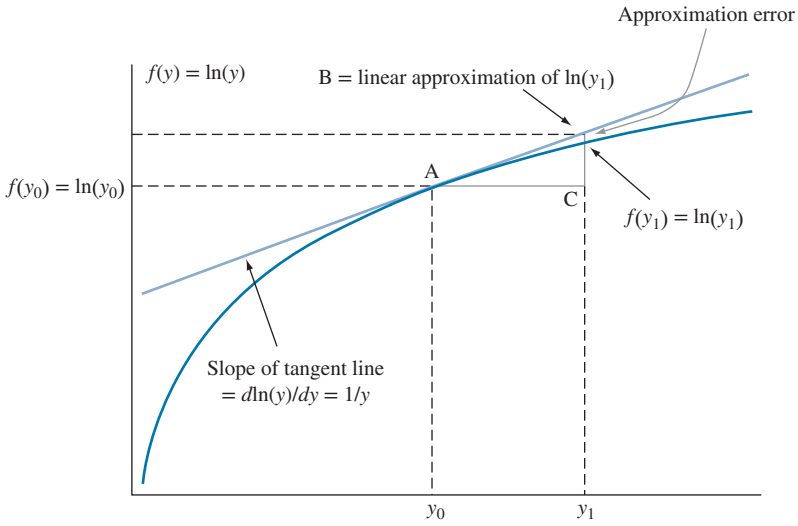


FIGURE A.4 **Taylor series approximation of ln(*y*).**

rise over the run." The "run" is A to C, or $(y_1 - y_0)$, and the corresponding "rise" is C to B. Then

$$\text{tangent slope} = \left.\frac{d\ln(y)}{dy}\right|_{y=y_0} = \frac{1}{y_0} = \frac{\text{rise}}{\text{run}} = \frac{\overline{\text{CB}}}{\overline{\text{AC}}} = \frac{\text{B} - \ln(y_0)}{y_1 - y_0}$$

Solving this equation for B = approximate value of $f(y_1)$, we obtain the expression in (A.4),

$$\text{B} = \ln(y_0) + \left.\frac{d\ln(y)}{dy}\right|_{y=y_0}(y_1 - y_0) = \ln(y_0) + \frac{1}{y_0}(y_1 - y_0)$$

The Taylor series approximation is used in many contexts.

Derivative Rule 10. (Taylor series approximation). If $f(x)$ is a smooth function, then

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x - a) = f(a) + f'(a)(x - a)$$

where $f'(a)$ is a common notation for the first derivative of the function $f(x)$ evaluated at $x = a$. The approximation is good for x close to a. See Exercise A.16 for a **second-order Taylor series approximation**.

A.3.2 Elasticity of a Nonlinear Relationship

Given the slope of a curve, the elasticity of y with respect to changes in x is given by a slightly modified (A.8),

$$\varepsilon_{yx} = \frac{dy/y}{dy/x} = \frac{dy}{dx} \times \frac{x}{y} = slope \times \frac{x}{y}$$

For example, the quadratic function $y = ax^2 + bx + c$ is a parabola. The slope (derivative) is $dy/dy = 2ax + b$. The elasticity is

$$\varepsilon_{yx} = slope \times \frac{x}{y} = (2ax + b)\frac{x}{y}$$

As a numerical example, consider the curve defined by $y = f(x) = x^2 - 8x + 16$. The graph of this quadratic function is shown in Figure A.3. The slope of the curve is $dy/dx = 2x - 8$. When $x = 6$, the slope of the tangent line is $dy/dx = 4$. When $x = 6$, the corresponding value of $y = 4$. So the elasticity at that point is

$$\varepsilon_{xy} = (dy/dx) \times (x/y) = (2x - 8)(x/y) = 4(6/4) = 6$$

A 1% increase in x is associated with a 6% change in y.

A.3.3 Second Derivatives

Since the derivative dy/dx of $f(x)$ is a function of x itself, we can define the derivative of the first derivative of $f(x)$, or **second derivative** of $f(x)$, as

$$\frac{d^2y}{dx^2} = \frac{d(dy/dx)}{dx}$$

The second derivative of a function is interpreted as the rate of change of the first derivative and indicates whether the function is increasing or decreasing at an increasing, constant or decreasing rate.

EXAMPLE A.4 | Second Derivative of a Linear Function

Find the second derivative of $y = 4x + 1$. Using the rules of differentiation

$$\frac{dy}{dx} = \frac{d(4x+1)}{dx} = 4$$

and

$$\frac{d^2y}{dx^2} = \frac{d(dy/dx)}{dx} = \frac{d(4)}{dx} = 0.$$

The function $y = f(x) = 4x + 1$ is a straight line and has a constant first derivative, or slope, 4. The rate of change of the first derivative is zero, and the function increases at a constant rate.

EXAMPLE A.5 | Second Derivative of a Quadratic Function

Find the second derivative of the function $y = x^2 - 8x + 16$ shown in Figure A.3.

$$\frac{dy}{dx} = \frac{d(x^2 - 8x + 16)}{dx} = 2x - 8$$

$$\frac{d^2y}{dx^2} = \frac{d(2x-8)}{dx} = 2$$

The second derivative of $y = f(x)$ is positive and the constant 2, which indicates that the first derivative is increasing for $-\infty < x < \infty$. For $x < 4$ the function is decreasing at a decreasing rate since the negative slope becomes less steep; for $x > 4$ the function increases at an increasing rate. At $x = 4$ the function is at its minimum and the slope is zero.

A.3.4 Maxima and Minima

Using first and second derivatives, we can define relative, or local, maxima and minima of functions, as shown in Figure A.5.

The function $y = f(x)$ has a relative or local maximum at $x = a$ if $f(a)$ is greater than any other value of $f(x)$ in an interval around $x = a$; the function $y = f(x)$ has a relative or local minimum at $x = a$ if $f(a)$ is less than any other value of $f(x)$ in an interval around $x = a$. The conditions for a local maximum or minimum of a function $y = f(x)$ at $x = a$ are as follows:

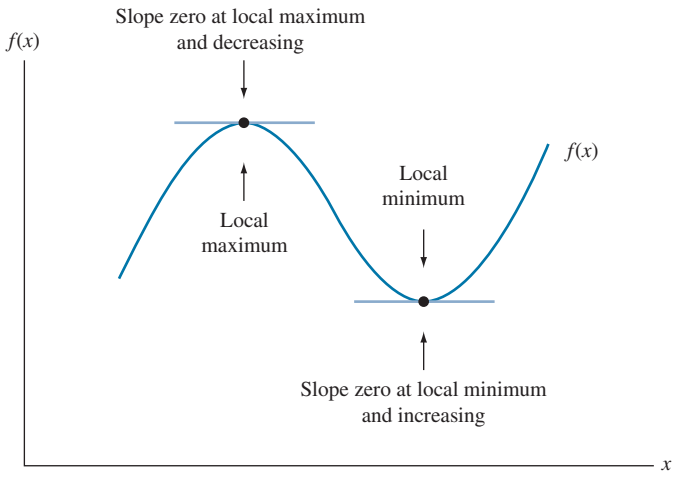


FIGURE A.5 **Local maxima and minima.**

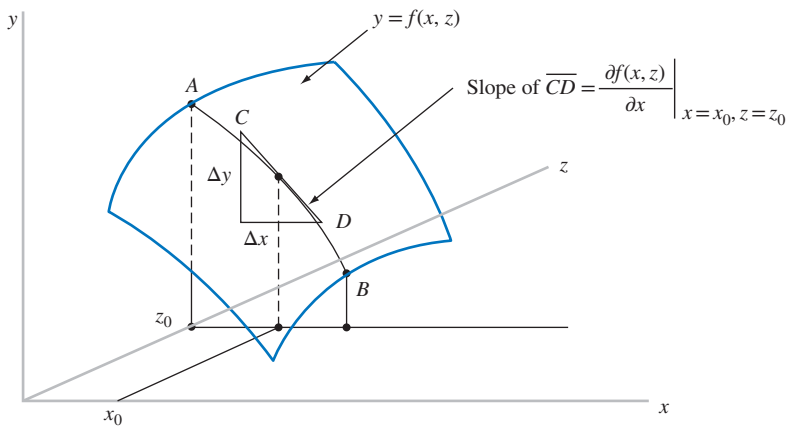


FIGURE A.6 **Three-dimensional diagram of a partial derivative.**

If $y = f(x)$ and dy/dx are nice (continuous) functions at $x = a$, and if $dy/dx = 0$ at $x = a$ then

1. If $d^2y/dx^2 < 0$ at $x = a$ then $f(a)$ is a local maximum.
2. If $d^2y/dx^2 > 0$ at $x = a$ then $f(a)$ is a local minimum.

EXAMPLE A.6 | Finding the Minimum of a Quadratic Function

In Examples A.3 and A.5, we considered the function $y = x^2 - 8x + 16$. To locate possible local minima or maxima, obtain the first derivative, set it to zero, and solve for values of x where $dy/dx = 0$. For this function, $dy/dx = 2x - 8 = 0$ implies that at $x = 4$ we may have a local maximum or a local minimum. Since $d^2y/dx^2 = 2 > 0$, the function is increasing at an increasing rate at $x = 4$ (and everywhere else), and thus $f(4) = 0$ is a local minimum of $y = x^2 - 8x + 16$.

Two notes regarding Example A.6: first, $y = f(x)$ achieves its global or absolute minimum at $x = 4$ as well as its local minimum. Second, if $dy/dx = 0$ at a point $x = a$ where $d^2y/dx^2 = 0$ then the "test" for a local maxima or minima using first and second derivatives does not apply.

A.3.5 Partial Derivatives

When a functional relationship includes several variables, such as $y = f(x, z)$, the slope depends on the values of x and z, and there are slopes in two directions rather than one. In Figure A.6, we illustrate the **partial derivative** of the function with respect to x, holding z constant at the value $z = z_0$.

At the point (x_0, z_0), the value of the function is $y_0 = f(x_0, z_0)$. The slope of the tangent line $\overline{CD}$ is the partial derivative.

$$\text{Slope of } \overline{CD} = \left.\frac{\partial f(x,z)}{\partial x}\right|_{x=x_0, z=z_0}$$

The vertical bar indicates that the partial derivative function is evaluated at the point (x_0, z_0).

To find the partial derivative, we use the already established rules. Consider the function

$$y = f(x, z) = ax^2 + bx + cz + d$$

To find the partial derivative of y with respect to x, treat z as a constant. Then

$$\frac{\partial y}{\partial x} = \frac{d(ax^2)}{dx} + \frac{d(bx)}{dx} + \frac{d(cz)}{dx} + \frac{d(d)}{dx} = 2ax + b$$

Using Derivative Rule 1, the third and fourth terms in the derivative are zero, because cz and d are treated as constants.

A.3.6 Maxima and Minima of Bivariate Functions

Let $y = f(x, z)$ be a continuous function of two variables, or a **bivariate function**, with continuous first derivatives. In order for the point $(x = a, z = b)$ to be a local maximum or minimum three conditions must be met.

1. The two partial derivatives be zero when evaluated at that point:

$$\left.\frac{\partial y}{\partial x}\right|_{x=a,\, z=b} = 0, \qquad \left.\frac{\partial y}{\partial z}\right|_{x=a,\, z=b} = 0$$

   These slope conditions are depicted in Figure A.7.

2. For a local maximum, shown in Figure A.7(a), the second partial derivatives must both be negative at the point $(x = a, z = b)$

$$\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b} < 0, \qquad \left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b} < 0$$

   These two conditions ensure that the function is concave and moving downward in the directions of the x and z axes.

   For a local minimum, shown in Figure A.7(b), the second partial derivatives must both be positive at the point $(x = a, z = b)$ so that the function is convex and the function is moving upward in both the x and z directions

$$\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b} > 0, \qquad \left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b} > 0$$

3. For a local maximum or minimum, the product of the second-order direct partials evaluated at $(x = a, z = b)$ must be larger than the square of the second-order cross-partial derivative at $(x = a, z = b)$, that is,

$$\left(\left.\frac{\partial^2 y}{\partial x^2}\right|_{x=a,\, z=b}\right)\left(\left.\frac{\partial^2 y}{\partial z^2}\right|_{x=a,\, z=b}\right) > \left(\left.\frac{\partial^2 y}{\partial x \partial z}\right|_{x=a,\, z=b}\right)^2$$

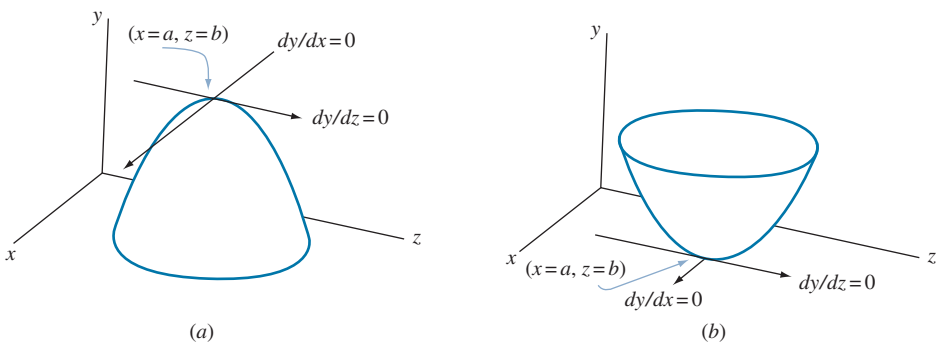


FIGURE A.7 **(*a*) Local maximum and (*b*) local minimum.**

For a local maximum, this condition ensures that the function is moving downward in all directions from $(x = a, z = b)$, not just along the x and z axes. For a local minimum, this condition ensures that the function is moving upward in all directions from $(x = a, z = b)$, not just along the x and z axes.

EXAMPLE A.7 | Maximizing a Profit Function

A firm produces two goods, x and y. The firm's profit function is $\pi = 64x - 2x^2 + 4xy - 4y^2 + 32y - 14$. Find the profit maximizing level of output of x and y. The first partial derivatives are

$$\partial\pi/\partial x = 64 - 4x + 4y, \qquad \partial\pi/\partial y = 4x - 8y + 32$$

The first condition for a maximum or minimum is to set these first derivatives to zero and solve for possible profit maximizing values (x^*, y^*)

$$\left.\begin{aligned} 64 - 4x + 4y &= 0 \\ 4x - 8y + 32 &= 0 \end{aligned}\right\} \Rightarrow x^* = 40, \quad y^* = 24$$

These two values may maximize profit, minimize profit, or neither. We must check the second and third conditions above. The second direct and cross-partial derivatives are

$$\frac{\partial^2\pi}{\partial x^2} = \frac{\partial(64 - 4x + 4y)}{\partial x} = -4$$

$$\frac{\partial^2\pi}{\partial y^2} = \frac{\partial(4x - 8y + 32)}{\partial y} = -8$$

$$\frac{\partial^2\pi}{\partial x\partial y} = \frac{\partial(64 - 4x + 4y)}{\partial y} = 4$$

Both of the second direct partial derivatives are negative, satisfying the second condition for a local maximum. The third condition is that

$$\left(\frac{\partial^2\pi}{\partial x^2}\right)\left(\frac{\partial^2\pi}{\partial y^2}\right) > \left(\frac{\partial^2\pi}{\partial x\partial y}\right)^2$$

This condition is satisfied too, since $(-4)(-8) = 32 > (4)^2 = 16$. Thus, profit is maximized at $x^* = 40$, $y^* = 24$, and the maximum profit is $\pi^* = 1650$.

EXAMPLE A.8 | Minimizing a Sum of Squared Differences

The least squares problem is to find values α and β that minimize the objective function $S(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ where (y_i, x_i), $i = 1, \ldots, n$ are data values. Given three pairs of data values $(y_1, x_1) = (1, 1)$, $(y_2, x_2) = (5, 2)$, and $(y_3, x_3) = (2, 3)$, find the minimizing values of α and β.

To find the minimizing values we first expand

$$\begin{aligned} S(\alpha, \beta) &= \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 \\ &= \sum_{i=1}^{n}(y_i^2 + \alpha^2 + \beta^2 x_i^2 - 2\alpha y_i - 2\beta x_i y_i + 2\alpha\beta x_i) \\ &= \sum_{i=1}^{n} y_i^2 + n\alpha^2 + \beta^2\sum_{i=1}^{n} x_i^2 - 2\alpha\sum_{i=1}^{n} y_i - 2\beta\sum_{i=1}^{n} x_i y_i \\ &\quad + 2\alpha\beta\sum_{i=1}^{n} x_i \end{aligned}$$

For the $n = 3$ given data pairs

$$\sum_{i=1}^{3} y_i^2 = 30, \quad \sum_{i=1}^{3} x_i^2 = 14, \quad \sum_{i=1}^{3} y_i = 8,$$

$$\sum_{i=1}^{3} x_i y_i = 17, \quad \sum_{i=1}^{3} x_i = 6$$

The objective function is then

$$\begin{aligned} S(\alpha, \beta) &= 30 + 3\alpha^2 + \beta^2(14) - 2\alpha(8) - 2\beta(17) + 2\alpha\beta(6) \\ &= 30 + 3\alpha^2 + 14\beta^2 - 16\alpha - 34\beta + 12\alpha\beta \end{aligned}$$

The first direct partial derivatives are

$$\frac{\partial S(\alpha, \beta)}{\partial\alpha} = 6\alpha - 16 + 12\beta, \qquad \frac{\partial S(\alpha, \beta)}{\partial\beta} = 28\beta - 34 + 12\alpha$$

Setting these two equations to zero and solving yields $\alpha^* = 5/3$ and $\beta^* = 1/2$. The second-order partial derivatives are

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha^2} = \frac{\partial(6\alpha - 16 + 12\beta)}{\partial\alpha} = 6$$

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\beta^2} = \frac{\partial(28\beta - 34 + 12\alpha)}{\partial\beta} = 28$$

$$\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha\partial\beta} = \frac{\partial(6\alpha - 16 + 12\beta)}{\partial\beta} = 12$$

Both second direct partial derivatives are positive, and the third condition is satisfied because

$$\begin{aligned} \left(\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha^2}\right)\left(\frac{\partial^2 S(\alpha, \beta)}{\partial\beta^2}\right) &= 6(28) \\ &= 168 > \left(\frac{\partial^2 S(\alpha, \beta)}{\partial\alpha\partial\beta}\right)^2 = 144 \end{aligned}$$

Thus, the values $\alpha^* = 5/3$, $\beta^* = 1/2$ minimize the least squares objective function, which takes the value $S(\alpha^*, \beta^*) \cong 8.167$.

A.4 Integrals

An **integral** is an "antiderivative." If $f(x)$ is a function, we can ask the question, "Of what function $F(x)$ is this the derivative?" The answer is given by the **indefinite integral**

$$\int f(x)\,dx = F(x) + C$$

The function $f(x) + C$, where C is a constant called the **constant of integration**, is an antiderivative of $f(x)$ because

$$\frac{d[F(x) + C]}{dx} = \frac{d[F(x)]}{dx} + \frac{d[C]}{dx} = f(x)$$

Finding $F(x)$ is an application of reversing the rules for derivatives. For example, using the rules of derivatives,

$$\frac{d\left(x^n + C\right)}{dx} = nx^{n-1}$$

Thus, $\int nx^{n-1}dx = x^n + C = F(x) + C$, so in this case $F(x) = x^n$. Many indefinite integrals have been worked out and are tabled in your favorite calculus book and at many websites.

Some handy facts about integrals are as follows:

Integral Rule 1.

$$\int \left[f(x) + g(x)\right]dx = \int f(x)dx + \int g(x)dx$$

An integral of a sum is the sum of the integrals.

Integral Rule 2.

$$\int cf(x)dx = c\int f(x)dx$$

Constants can be factored out of integrals.

These rules can be combined so that

Integral Rule 3.

$$\int \left[c_1 f(x) + c_2 g(x)\right]dx = c_1 \int f(x)\,dx + c_2 \int g(x)dx$$

Integral Rule 4 (power rule).

$$\int x^n dx = \frac{1}{n+1}x^{n+1} + C, \quad \text{where } n \neq -1$$

Integral Rule 5 (power rule $n = -1$).

$$\int x^{-1}dx = \ln(x) + C \text{ for } x > 0$$

Integral Rule 6 (constant function).

$$\int k\,dx = kx + C$$

Integral Rule 7 (exponential function).

$$\int e^{kx}dx = \frac{1}{k}e^{kx} + C$$

A.4.1 Computing the Area Under a Curve

An important use of integrals in econometrics and statistics is to calculate areas under curves. For example, in Figure A.8, what is the shaded area under the curve $f(x)$?

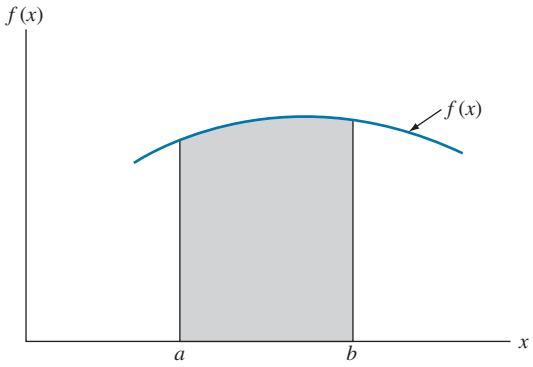


FIGURE A.8 **Area under a curve.**

The area between a curve $f(x)$ and the x-axis, between the limits a and b, is given by the **definite integral**

$$\int_a^b f(x)dx$$

The value of this integral is provided by the **fundamental theorem of calculus**, which says that

$$\int_a^b f(x)dx = F(b) - F(a)$$

EXAMPLE A.9 | Area Under a Curve

Consider the function

$$f(x) = \begin{cases} 2x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{(A.12)}$$

This is the equation of a straight line through the origin, as shown in Figure A.9.

What is the shaded area in Figure A.9, the area under the line between a and b? The answer can be found using the geometry of triangles. The area of a triangle is half the base times the height, $\frac{1}{2} \times base \times height$. Triangles can be identified by their corners. Let $\Delta 0bc$ represent the area of the triangle formed by the points 0 (the origin), b, and c. Similarly $\Delta 0ad$ represents the area of the smaller triangle formed by the points 0, a, and d. The shaded area that represents the area under $f(x) = 2x$ between a and b is the difference between the areas of these two triangles.

$$\begin{aligned} \text{Area} &= \Delta 0bc - \Delta 0ad \\ &= \left(\frac{1}{2}b\right)(2b) - \frac{1}{2}a(2a) \\ &= b^2 - a^2 \end{aligned} \qquad \text{(A.13)}$$

Equation (A.13) gives us an easy formula for calculating the area under $f(x) = 2x$ falling between a and b.

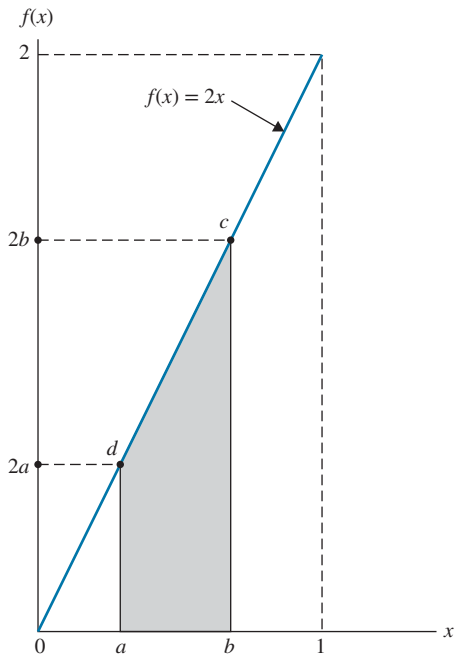


FIGURE A.9 **Area under the curve $f(x) = 2x, 0 \le x \le 1$.**

Using integration, the area under the curve $f(x) = 2x$ and above the x-axis between the limits $x = a$ and $x = b$ is obtained by finding the **definite integral** of $f(x) = 2x$. To use the fundamental theorem of calculus, we need the indefinite integral. Using the power rule, Integral Rule 4, we obtain

$$\int 2xdx = 2\int xdx = 2\left[\frac{1}{2}x^2 + C\right] = x^2 + 2C$$
$$= x^2 + C_1 = F(x) + C_1$$

where $F(x) = x^2$ and the constant of integration is C_1. The area we seek is given by

$$\int_a^b 2xdx = F(b) - F(a) = b^2 - a^2 \qquad \text{(A.14)}$$

This is the same answer we obtained in (A.13) using geometry.

Many times the algebra is abbreviated, because the constant of integration does not affect the definite integral. You will see for definite integrals

$$\int_a^b 2xdx = x^2\Big|_a^b = b^2 - a^2$$

The vertical bar notation means: evaluate the expression first at b and subtract from it the value of the expression at a.

A.5 Exercises

A.1 Each of the following formulas, (1), (2), and (3), represents a supply or demand relation.

(1) $Q = -3 + 2P$ where $P = 10$
(2) $Q = 100 - 20P$ where $P = 4$
(3) $Q = 50P^{-2}$ where $P = 2$

a. Calculate the slope of each function at the given point.
b. Interpret the slope found in (a). Do the slopes change for different values of P and Q? Is it a supply curve (positive relationship) or a demand curve (inverse relationship)?
c. Calculate the elasticity of each function at the given point.
d. Interpret the elasticity found in (c). Do the elasticities change for different values of P and Q?

A.2 The infant mortality rate (*MORTALITY*) for a country is related to the annual per capita income (*INCOME*, U.S. $1000) in that country. Three relationships that may describe this relationship are

(1) $\ln(MORTALITY) = 7.5 - 0.5\ln(INCOME)$
(2) $MORTALITY = 1400 - 100INCOME + 1.67INCOME^2$
(3) $MORTALITY = 1500 - 50INCOME$

a. Sketch each of these relationships between *MORTALITY* and *INCOME* between $INCOME = 0$ and $INCOME = 30$.
b. For each of these relationships, calculate the elasticity of infant mortality with respect to income if (i) $INCOME = 1$, (ii) $INCOME = 3$, and (iii) $INCOME = 25$.

A.3 Suppose the rate of inflation *INF*, the annual percentage increase in the general price level, is related to the annual unemployment rate *UNEMP* by the equation $INF = -3 + 7 \times (1/UNEMP)$.

a. Sketch the curve for values of *UNEMP* between 1 and 10.
b. Where is the impact of a change in the unemployment rate the largest?
c. If the unemployment rate is 5%, what is the marginal effect of an increase in the unemployment rate on the inflation rate?

A.4 Simplify the following expressions:

a. $x^{2/3}x^{2/7}$
b. $x^{2/3} \div x^{2/7}$
c. $\left(x^6y^4\right)^{-1/2}$

A.5 Below are the 2015 *GDP* ($US) figures provided by the World Bank for a few countries.

- **a.** Express each in scientific notation.
  - **i.** Maldives *GDP* $3,142,812,004
  - **ii.** Nicaragua *GDP* $12,692,562,187
  - **iii.** Ecuador *GDP* $100,871,770,000
  - **iv.** New Zealand *GDP* $173,754,075,210
  - **v.** India *GDP* $2,073,542,978,208
  - **vi.** United States *GDP* $17,946,996,000,000
- **b.** Using scientific notation divide the U.S. *GDP* by the *GDP* in (i) Maldives (ii) Ecuador.
- **c.** The population of New Zealand in 2015 was 4.595 million. Use calculations with scientific notation to compute the per capita income in New Zealand. Express the result in scientific notation.
- **d.** The 2015 population of St. Lucia was 184,999 and its *GDP* was $1,436,390,325. Use calculations with scientific notation to compute the per capita income in St. Lucia. Express the result in scientific notation.
- **e.** Using scientific notation, express the sum of the U.S. and New Zealand *GDP* values. [*Hint*: Write each number as $a10^x$ where x is a convenient number for both and a is a numerical value, then simplify.]

A.6 Technology affects agricultural production by increasing yield over time. Let $WHEAT_t$ = average wheat production (tonnes per hectare) for the period 1950–2000 ($t = 1, \ldots, 51$) in Western Australia's Mullewa Shire.

- **a.** Suppose production is defined by $WHEAT_t = 0.58 + 0.14\ln(t)$. Plot this curve. Find the slope and elasticity at the point $t = 49$ (1998).
- **b.** Suppose production is defined by $WHEAT_t = 0.78 + 0.0003\,t^2$. Plot this curve. Find the slope and elasticity at the point $t = 49$ (1998).

A.7 Consider the function $WAGE = f(AGE) = 10 + 200AGE - 2AGE^2$.

- **a.** Sketch the curve for values of *AGE* between $AGE = 20$ and $AGE = 70$.
- **b.** Find the derivative $dWAGE/dAGE$ and evaluate it at $AGE = 30$, $AGE = 50$, and $AGE = 60$. On the curve in part (a), sketch the tangent to the curve at $AGE = 30$.
- **c.** Find the *AGE* at which *WAGE* is maximized.
- **d.** Compute $WAGE_1 = f(29.99)$ and $WAGE_2 = f(30.01)$. Locate these values (approximately) on your sketch from part (a).
- **e.** Evaluate $m = [f(30.01) - f(29.99)]/0.02$. Compare this value to the value of the derivative computed in (b). Explain, geometrically, why the values should be close. The value m is a "numerical derivative," which is useful for approximating derivatives.

A.8 Sketch each of the demand curves below. (i) Indicate the area under the curve between prices $P = 1$ and $P = 2$ on the sketch. (ii) Using integration, calculate the area under the curve between prices $P = 1$ and $P = 2$.

- **a.** $Q = 15 - 5P$
- **b.** $Q = 10P^{-1/2}$
- **c.** $Q = 10/P$

A.9 Consider the function $f(y) = 1/100$ over the interval $0 < y < 100$ and $f(y) = 0$ otherwise.

- **a.** Calculate the area under the curve $f(y)$ for the interval $30 < y < 50$ using a geometric argument.
- **b.** Calculate the area under the curve $f(y)$ for the interval $30 < y < 50$ as an integral.
- **c.** What is a general expression for the area under $f(y)$ over the interval $[a, b]$, where $0 < a < b < 100$?
- **d.** Calculate the integral from $y = 0$ to $y = 100$ of the function $yf(y) = y/100$.

A.10 Consider the function $f(y) = 2e^{-2y}$ for $0 < y < \infty$.

- **a.** Draw a sketch of the function.
- **b.** Compute the integral of $f(y)$ from $y = 1$ to $y = 2$ and illustrate the value on the part (a) sketch.

A.11 Let $y_0 = 1$. For each of the values $y_1 = 1.01, 1.05, 1.10, 1.15, 1.20$, and 1.25 compute

- **a.** The actual percentage change in y using equation (A.2).
- **b.** The approximate percentage change in y using equation (A.3).
- **c.** Comment on how well the approximation in equation (A.3) works as the value of y_1 increases.

A.12 A firm uses labor (L) and capital (K) to produce output (Q). Suppose the production function is $Q = 6L^{1/2}K^{1/3}$. The firm sells its product at price $P = 4$ and pays its labor a wage $W = 12$ with the price of capital being $R = 5$.

- **a.** Find the combination of labor and capital that maximizes profits $\pi = P \times Q - (W \times L) - (R \times K)$ where Q is given by the production function. Check all conditions for a relative maximum.
- **b.** Find the marginal product of labor, $\partial Q/\partial L$, and the marginal product of capital, $\partial Q/\partial K$, at the profit maximizing amounts of labor and capital.

A.13 Use Derivative Rule 10 (Taylor Series approximation) to approximate each of the functions below at $x = 1.5$ and $x = 2$. Let $a = 1$. Calculate the percentage approximation error in each case.

- **a.** $f(x) = 3x^2 - 5x + 1$
- **b.** $f(x) = \ln(2x)$
- **c.** $f(x) = e^{2x}$

A.14 Suppose that a person's earnings (*INCOME*) are determined by their education (*EDUC*) and experience (*EXPER*) according to the relation

$$INCOME = -2EDUC^2 + 78EDUC - 2EXPER^2 + 66EXPER - 2EDUC \times EXPER$$

Find the values of education and experience that maximize the person's income.

A.15 A variable y changes value from $y_0 = 4$ to $y_1 = 4.6$.

- **a.** Compute the relative change in y.
- **b.** Compute the percentage change in y.
- **c.** If the value of y is 4, what is the value of y if it increases by 18%?

A.16 Derivative Rule 10 is a "first-order" Taylor series approximation. A **"second-order" Taylor series approximation** is

$$f(x) \cong f(a) + \left.\frac{df(x)}{dx}\right|_{x=a}(x-a) + \frac{1}{2}\left.\frac{d^2f(x)}{dx^2}\right|_{x=a}(x-a)^2$$

$$= f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

where $f''(a)$ represents the second derivative of the function evaluated at the point $x = a$.

- **a.** Use both the first- and second-order Taylor series approximations to approximate the function $f(x) = e^{2x}$ at $x = 1.5$ and $x = 2$. Let $a = 1$. Calculate the percentage approximation error in each case.
- **b.** Draw a sketch of the function $f(x) = e^{2x}$ for $0 < x < 3$. On the sketch show the tangent line to the function at $a = 1$. On the same graph extrapolate the tangent line to show the location of the first-order approximation when $x = 2$. Show the value of the second-order approximation when $x = 2$.
- **c.** Calculate the percentage approximation error for the first- and second-order Taylor series approximations in part (b). Which is better in this case?

A.17 In 2015, the *GDP* (in nominal U.S. dollars) of Belarus was $GDP_B = \$54,608,962,634.99$ and that of Poland was $GDP_P = \$474,783,393,022.95$.

- **a.** Write GDP_B in scientific notation.
- **b.** Use scientific notation to divide GDP_P by GDP_B. Show your work.
- **c.** Write the natural log of GDP_P.
- **d.** Find $\exp\left[\ln\left(GDP_A\right) - \ln\left(GDP_B\right)\right]$. Write the solution in scientific notation. Show your work.

A.18 Carry out the following:

- **a.** Suppose your wage rate increases from \$17/hr to \$18/hr. What is the percentage increase in your wage?
- **b.** Calculate 100[ln(18) – ln(17)].
- **c.** Suppose your wage rate increases from \$17/hr to \$28/hr. What is the percentage increase in your wage?
- **d.** Calculate 100[ln(28) – ln(17)].
- **e.** Calculate ln(1.02).
- **f.** Calculate ln(1.57).

A.19 Suppose your wage rate is determined by

$$WAGE = -19.68 + 2.52EDUC + 0.55EXPER - 0.007EXPER^2$$

where *EDUC* is years of schooling and *EXPER* is years of work experience. Using calculus, what value of *EXPER* maximizes *WAGE* for a person with 16 years of education? Show your work.

A.20 Suppose wages are determined by the following equation. *EDUC* = years of education, *EXPER* = years of work experience, and *FEMALE* = 1 if person is female, 0 otherwise.

$$WAGE = -23.06 + 2.85EDUC + 0.80EXPER - 0.008EXPER^2 - 9.21FEMALE + 0.34(FEMALE \times EDUC) - 0.015(EDUC \times EXPER)$$

Find $\partial WAGE/\partial EDUC$ for a female with 16 years of schooling and 10 years of experience. Show your work.

Appendix B

Probability Concepts

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Explain the difference between a random variable and its values, and give an example.
2. Explain the difference between discrete and continuous random variables, and give examples of each.
3. State the characteristics of probability density functions (*pdf*) for discrete and continuous random variables, and give examples illustrating these characteristics.
4. Compute probabilities of events, given the probability density function for a discrete or continuous random variable.
5. Show, geometrically and algebraically, using integration, how to compute probabilities given a *pdf* for a continuous random variable.
6. Use the definitions of expected values for discrete and continuous random variables to compute expectations, given a *pdf* $f(x)$ and a function $g(x)$.
7. Define the variance of a random variable, and explain in what sense the values of a random variable are more spread out if the variance is larger.
8. Use a joint *pdf* for two continuous random variables to compute probabilities of joint events, and to find the (marginal) *pdf* of each individual random variable.
9. Find the conditional *pdf* for one random variable given the value of another and their joint *pdf*, and use it to compute conditional probabilities, the conditional mean, and the conditional variance.
10. Define the covariance and correlation between two random variables, and compute these values given a joint probability function.
11. Explain and apply the law of iterated expectations. Explain the variance and covariance decompositions.
12. Find the distribution of a random variable $Y = g(X)$, when $g(X)$ is a strictly increasing or decreasing function, given the probability density function $f(x)$ for the random variable X.
13. Obtain a random number from a probability density function $f(x)$ when its cumulative distribution function $F(x)$ is invertible.
14. Explain in what sense random numbers generated by a computer are random, and in what sense they are not.

KEYWORDS

binary variable
binomial random variable
cdf
change of variable technique
chi-square distribution
conditional *pdf*
conditional probability
continuous random variables
correlation
covariance
covariance decomposition
cumulative distribution function
degrees of freedom
discrete random variable
expected value
experiment
F-distribution
inversion method
iterated expectation
Jacobian
joint probability density function
marginal distributions
mean
median

modulus
normal distribution
pdf
Poisson distribution
probability
probability density function
pseudo-random numbers
random number
random number seed
random variable
standard deviation
standard normal distribution
statistically independent
strictly monotonic
t-distribution
uniform distribution
variance
variance decomposition

We assume that you have had a basic probability and statistics course and that you have read the Probability Primer that precedes Chapter 2. If you have not read the Probability Primer, then do so now.

In this appendix we summarize rules of expected values and variances for **discrete random variables** for easy reference. We then develop similar rules for **continuous random variables** that will require the use of integral concepts introduced in Appendix A.4. We review the properties of some important discrete and continuous random variables, including the *t*-, chi-square, and *F*-distributions. Finally, we introduce concepts related to computer-generated random numbers.

B.1 Discrete Random Variables

In this section we provide a summary of operations with discrete random variables. See the Probability Primer for examples and general background discussion.

A **random variable** is a variable whose value is unknown until it is observed; in other words, it is a variable that is not perfectly predictable. A **discrete random variable** can take only a limited, or countable, number of values. An example of a discrete random variable is the number of late credit card bill payments last year by a *randomly* selected individual. A special case occurs when a random variable can only be one of two possible values. A payment is either late or it is not. Outcomes like this can be characterized by a **binary variable** taking the value one for late payments and zero for those that are on time. Such variables are also called **indicator variables**, or **dummy variables**.

We summarize the probabilities of possible outcomes using a **probability density function** (***pdf***). The ***pdf*** for a discrete random variable indicates the **probability** of each possible value occurring. For a discrete random variable X the value of the probability density function $f(x)$ is the probability that the random variable X takes the value x, $f(x) = P(X = x)$. Because $f(x)$ is a probability, it must be true that $0 \leq f(x) \leq 1$ and, if X takes n possible values $x_1, \ldots, x_n$, then the sum of their probabilities must be one

$$P(X = x_1) + P(X = x_2) + \cdots + P(X = x_n) = f(x_1) + f(x_2) + \cdots + f(x_n) = 1$$

The **cumulative distribution function** (***cdf***) is an alternative way to represent probabilities. The *cdf* of the random variable X, denoted by $F(x)$, gives the probability that X is less than or equal to a specific value x. That is,

$$F(x) = P(X \leq x) \tag{B.1}$$

Two key features of a probability distribution are its center (location) and width (dispersion). A measure of the center is the **mean**, or **expected value**; measures of dispersion are **variance**, and its square root—the **standard deviation**.

B.1.1 Expected Value of a Discrete Random Variable

The **mean** of a random variable is given by its **mathematical expectation**. If X is a discrete random variable taking the values $x_1, \ldots, x_n$ then the mathematical expectation, or **expected value**,

of X is

$$\mu_X = E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \cdots + x_n P(X = x_n) \tag{B.2a}$$

The expected value, or mean, of X is a weighted average of its values, the weights being the probabilities that the values occur. The mean is often symbolized by μ or μ_X. It is the average value of the random variable in all possible experimental outcomes from the underlying **experiment**. Because the probability that the discrete random variable X takes the value x is given by its *pdf* $f(x)$, $P(X = x) = f(x)$, the expected value in (B.2a) can be written equivalently as

$$\begin{aligned}\mu_X = E(X) &= x_1 f(x_1) + x_2 f(x_2) + \cdots + x_n f(x_n) \\ &= \sum_{i=1}^{n} x_i f(x_i) = \sum_x x f(x)\end{aligned} \tag{B.2b}$$

Functions of random variables are also random. Expected values are obtained using calculations similar to those in (B.2). If X is a discrete random variable and $g(X)$ is a function of it, then

$$E[g(X)] = \sum_x g(x) f(x) \tag{B.3}$$

Using (B.3) we can develop some frequently used rules. If a is a constant, then

$$E(aX) = aE(X) \tag{B.4}$$

Similarly, if a and b are constants, then we can show that

$$E(aX + b) = aE(X) + b \tag{B.5}$$

To see how this result is obtained, we apply the definition in (B.3) to the function $g(X) = aX + b$

$$\begin{aligned}E[g(X)] &= \sum g(x) f(x) = \sum (ax + b) f(x) = \sum [ax f(x) + b f(x)] \\ &= \sum [ax f(x)] + \sum [b f(x)] = a \sum x f(x) + b \sum f(x) \\ &= aE(X) + b\end{aligned}$$

In the final step we recognize $E(X)$ from its definition in (B.2), and use the fact that $\sum f(x) = 1$.

If $g_1(X)$, $g_2(X)$, …, $g_M(X)$ are functions of X, then

$$E[g_1(X) + g_2(X) + \cdots + g_M(X)] = E[g_1(X)] + E[g_2(X)] + \cdots + E[g_M(X)] \tag{B.6}$$

This rule extends to any number of functions. **The expected value of a sum is always the sum of the expected values**.

A similar rule does not work, in general, for nonlinear functions. That is, $E[g(X)] \neq g[E(X)]$. For example, $E(X^2) \neq [E(X)]^2$.

B.1.2 Variance of a Discrete Random Variable

The **variance** of a discrete random variable X is the expected value of

$$g(X) = [X - E(X)]^2$$

The variance of a random variable is important in characterizing the scale of measurement and the spread of the probability distribution. We give it the symbol σ^2, which is read "sigma squared," or σ_X^2. Algebraically, letting $E(X) = \mu_X$,

$$\text{var}(X) = \sigma_X^2 = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2 \tag{B.7}$$

The variance of a random variable is the *average* squared difference between the random variable X and its mean value μ. The larger the variance of a random variable, the more "spread out" its values are. The square root of the variance is called the **standard deviation**; it is denoted by σ or σ_X. It measures the spread or dispersion of a distribution and has the advantage of being in the same units of measure as the random variable.

A useful property of variances is the following. Let a and b be constants; then

$$\text{var}(aX + b) = a^2\text{var}(X) \tag{B.8}$$

This result is proven in the Probability Primer, Section P.5.4.

Two other characteristics of a probability distribution are its **skewness** and **kurtosis**. These are defined as

$$skewness = \frac{E\left[\left(X - \mu_X\right)^3\right]}{\sigma_X^3} \tag{B.9}$$

and

$$kurtosis = \frac{E\left[\left(X - \mu_X\right)^4\right]}{\sigma_X^4} \tag{B.10}$$

Skewness measures the lack of symmetry of a distribution. If the distribution is symmetric, then its *skewness* $= 0$. Distributions with long tails to the left are negatively skewed, and *skewness* < 0. Distributions with long tails to the right are positively skewed, and *skewness* > 0. Kurtosis measures the "peakedness" of a distribution. A distribution with large kurtosis has more values concentrated near the mean and a relatively high central peak. A distribution that is relatively flat has a lower kurtosis. The benchmark value for kurtosis is 3, which is the kurtosis of the **normal distribution** that we discuss later in this appendix (Section B.3.5).

B.1.3 Joint, Marginal, and Conditional Distributions

If X and Y are discrete random variables, then the joint probability that $X = a$ and $Y = b$ is given by the joint *pdf* of X and Y, written as $f(x, y)$, and $P[X = a, Y = b] = f(a, b)$. The sum of the joint probabilities is one, $\sum_x \sum_y f(x, y) = 1$. Given a **joint probability density function**, we can obtain the probability distributions of individual random variables, which are also known as **marginal distributions**. If X and Y are two discrete random variables, then

$$f_X(x) = \sum_y f(x, y) \text{ for each value } X \text{ can take} \tag{B.11}$$

For discrete random variables, the probability that the random variable Y takes the value y *given* that $X = x$ is written $P(Y = y|X = x)$. This conditional probability is given by the **conditional *pdf*** $f(y|x)$:

$$f(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x, y)}{f_X(x)} \tag{B.12}$$

Two random variables are **statistically independent** if the conditional probability that $Y = y$ given that $X = x$, is the same as the unconditional probability that $Y = y$ for all x and y values. In this case, knowing the value of X does not alter the probability distribution of Y. If X and Y are independent random variables, then

$$P(Y = y|X = x) = P(Y = y) \tag{B.13}$$

Equivalently, if X and Y are independent, then the conditional *pdf* of Y given $X = x$ is the same as the unconditional, or marginal, *pdf* of Y alone,

$$f(y|x) = \frac{f(x, y)}{f_X(x)} = f_Y(y) \tag{B.14}$$

The converse is also true, so that if (B.13) or (B.14) is true *for every possible pair of x and y values*, then X and Y are statistically independent.

Solving (B.14) for the joint *pdf*, we can also say that X and Y are statistically independent if their joint *pdf* factors into the product of their marginal *pdf*s

$$f(x, y) = f_X(x) f_Y(y) \tag{B.15}$$

If (B.15) is true for each and every pair of x and y values, then X and Y are statistically independent. This result extends to more than two random variables. If X, Y, and Z are statistically independent, then their joint probability density function can be factored and written as $f(x, y, z) = f_X(x) \cdot f_Y(y) \cdot f_Z(z)$.

B.1.4 Expectations Involving Several Random Variables

A rule similar to (B.3) exists for functions of several random variables. Let X and Y be discrete random variables with joint *pdf* $f(x, y)$. If $g(X, Y)$ is a function of X and Y, then

$$E\big[g(X, Y)\big] = \sum_x \sum_y g(x, y) f(x, y) \tag{B.16}$$

Using (B.16) we can show that

$$E(X + Y) = E(X) + E(Y) \tag{B.17}$$

This follows by using the definition (B.16) and letting $g(X, Y) = X + Y$. Then

$$\begin{aligned}
E(X + Y) &= \sum_x \sum_y g(x, y) f(x, y) && \text{[general definition]} \\
&= \sum_x \sum_y (x + y) f(x, y) && \text{[specific function]} \\
&= \sum_x \sum_y x f(x, y) + \sum_x \sum_y y f(x, y) && \text{[separate terms]} \\
&= \sum_x x \sum_y f(x, y) + \sum_y y \sum_x f(x, y) && \text{[factor constants from 2nd sum]} \\
&= \sum_x x f(x) + \sum_y y f(y) && \text{[recognize marginal } pdf\text{]} \\
&= E(X) + E(Y) && \text{[recognize expected values]}
\end{aligned}$$

To go from the fourth to the fifth line, we have used (B.11) to obtain the marginal distributions of X and Y, and the fact that the order of summation does not matter. Using the same logic, we can show that

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{B.18}$$

In general, $E[g(X, Y)] \neq g[E(X), E(Y)]$. For example, in general, $E(XY) \neq E(X)E(Y)$. If, however, X and Y are statistically independent, then using (B.16), we can also show that $E(XY) = E(X)E(Y)$. To see this, recall that if X and Y are independent, then their joint *pdf* factors into the product of the marginal *pdf*s, $f(x, y) = f(x)f(y)$. Letting $g(X, Y) = XY$, we have

$$\begin{aligned}
E(XY) = E\big[g(X, Y)\big] &= \sum_x \sum_y xy f(x, y) = \sum_x \sum_y xy f(x) f(y) \\
&= \sum_x x f(x) \sum_y y f(y) = E(X)E(Y)
\end{aligned}$$

This rule can be extended to more independent random variables.

B.1.5 Covariance and Correlation

One particular application of (B.16) is the derivation of the **covariance** between X and Y. Define a function that is the product of X minus its mean times Y minus its mean,

$$g(X, Y) = \big(X - \mu_X\big)\big(Y - \mu_Y\big) \tag{B.19}$$

The covariance is the expected value of (B.19)

$$\text{cov}(X, Y) = \sigma_{XY} = E\Big[\big(X - \mu_X\big)\big(Y - \mu_Y\big)\Big] = E(XY) - \mu_X\mu_Y \tag{B.20}$$

If the covariance σ_{XY} of the variables is positive, then when x values are greater than their mean, the y values also tend to be greater than their mean, and when x values are below their mean, then the y values also tend to be less than their mean. In this case the random variables X and Y are said to be **positively** or **directly associated**. If $\sigma_{XY} < 0$, then the association is negative, or inverse. If $\sigma_{XY} = 0$, then there is neither a positive nor a negative relationship.

Interpreting the actual value of σ_{XY} is difficult, because X and Y may have different units of measurement. Scaling the covariance by the standard deviations of the variables eliminates the units of measurement, and defines the **correlation** between X and Y:

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{B.21}$$

As with the covariance, the correlation ρ between two random variables measures the degree of *linear* association between them. However, unlike the covariance, the correlation must lie between –1 and 1. The correlation between X and Y is 1 if there is a perfect positive linear relationship between X and Y and –1 if there is a perfect negative, or inverse, association between X and Y. If there is no *linear* association between X and Y, then $\text{cov}(X, Y) = 0$ and $\rho = 0$. For other values of correlation, the magnitude of the absolute value $|\rho|$ indicates the "strength" of the linear association between the values of the random variables.

If X and Y are independent random variables, then the covariance and correlation between them are zero. The converse of this relationship is *not* true. Independent random variables X and Y have zero covariance, indicating that there is no linear association between them. However, just because the covariance or correlation between two random variables is zero *does not* mean that they are necessarily independent. There may be more complicated nonlinear associations such as $X^2 + Y^2 = 1$.

In (B.17) we found the expected value of a sum of random variables. There are similar rules for variances. If a and b are constants, then

$$\text{var}(aX + bY) = a^2\text{var}(X) + b^2\text{var}(Y) + 2ab\,\text{cov}(X, Y) \tag{B.22}$$

To see this, it is convenient to define a new discrete random variable $Z = aX + bY$. This random variable has expected value

$$\mu_Z = E(Z) = E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

The variance of Z is

$$\begin{aligned}
\text{var}(Z) &= E\Big[\big(Z - \mu_Z\big)^2\Big] \\
&= E\Big\{\Big[(aX + bY) - \big(a\mu_X + b\mu_Y\big)\Big]^2\Big\} && \text{[substitute } Z\text{]} \\
&= E\Big\{\Big[\big(aX - a\mu_X\big) + \big(bY - b\mu_Y\big)\Big]^2\Big\} && \text{[combine like terms]}
\end{aligned}$$

$$= E\left\{\left[a(X-\mu_X)+b(Y-\mu_Y)\right]^2\right\} \quad \text{[factor]}$$

$$= E\left[a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y)\right] \quad \text{[expand]}$$

$$= E\left[a^2(X-\mu_X)^2\right]+E\left[b^2(Y-\mu_Y)^2\right]+E\left[2ab(X-\mu_X)(Y-\mu_Y)\right] \quad \text{[group terms]}$$

$$= a^2\text{var}(X)+b^2\text{var}(Y)+2ab\,\text{cov}(X,Y) \quad \text{[factor and recognize]}$$

These rules extend to more random variables. For example, if X, Y, and Z are random variables, then

$$\text{var}(aX+bY+cZ) = a^2\text{var}(X)+b^2\text{var}(Y)+c^2\text{var}(Z)+2ab\,\text{cov}(X,Y) + 2bc\,\text{cov}(Y,Z)+2ac\,\text{cov}(X,Z) \tag{B.23}$$

B.1.6 Conditional Expectations

If X and Y are two random variables with joint probability distribution $f(x, y)$, then the conditional probability distribution of Y given X is $f(y|x)$. We can use this conditional *pdf* to compute the **conditional mean** of Y given a value of X. That is, we can obtain the expected value of Y given that $X = x$. The conditional expectation $E(Y|X = x)$ is the average (or mean) value of Y given that X takes the value x. In the discrete case, it is defined to be

$$E(Y|X=x) = \sum_y yP(Y=y|X=x) = \sum_y yf(y|x) \tag{B.24}$$

Similarly, we can define the **conditional variance** of Y given X. This is the variance of the conditional distribution of Y given X. In the discrete case, it is

$$\text{var}(Y|X=x) = \sum_y \left[y-E(Y|X=x)\right]^2 f(y|x) \tag{B.25}$$

B.1.7 Iterated Expectations

The **law of iterated expectations** says that the expected value of Y is equal to the expected value of the conditional expectation of Y given X. That is,

$$E(Y) = E_X\left[E(Y|X)\right] \tag{B.26}$$

In Probability Primer Section P.6.3, we provide a numerical example of the Law of Iterated Expectations, and give the proof.

B.1.8 Variance Decomposition

Just as we can break up the expected value using the Law of Iterated Expectations, we can decompose the variance of a random variable into two parts.

$$\text{Variance Decomposition: } \text{var}(Y) = \text{var}_X\left[E(Y|X)\right] + E_X\left[\text{var}(Y|X)\right] \tag{B.27}$$

This result says that the variance of the random variable Y equals the sum of the variance of the conditional mean of Y given X and the mean of the conditional variance of Y given X. We discuss the **variance decomposition** for discrete random variables in Section P.6.4 of the Probability Primer. Here we provide the proof and a numerical example.

Proof of the Variance Decomposition We use the relationship between the marginal, conditional, and joint *pdf*s to prove the variance decomposition for discrete random variables. First, write out $\text{var}(Y)$ in an expanded form.

$$
\begin{aligned}
\text{var}(Y) &= \sum_y (y-\mu_y)^2 f(y) \\
&= \sum_y (y-\mu_y)^2 \left\{ \sum_x f(x,y) \right\} && \text{[replace marginal density]} \\
&= \sum_y (y-\mu_y)^2 \left\{ \sum_x f(y|x)\, f(x) \right\} && \text{[replace joint density]} \\
&= \sum_x \sum_y (y-\mu_y)^2 f(y|x)\, f(x) && \text{[change order of summation]} \\
&= \sum_x \sum_y \big(y - E(Y|x) + E(Y|x) - \mu_y\big)^2 f(y|x) f(x) && \text{[subtract and add conditional mean]} \\
&= \sum_x \sum_y \Big(\big[y - E(Y|x)\big] + \big[E(Y|x) - \mu_y\big] \Big)^2 f(y|x) f(x) && \text{[group terms, then square and expand]} \\
&= \sum_x \sum_y \Big\{ \big(y - E(Y|x)\big)^2 + \big(E(Y|x) - \mu_y\big)^2 + 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) \Big\} f(y|x) f(x) \\
&= \sum_x \sum_y \big(y - E(Y|x)\big)^2 f(y|x) f(x) && \textbf{[Term 1]} \\
&\quad + \sum_x \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) f(x) && \textbf{[Term 2]} \\
&\quad + \sum_x \sum_y 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) f(x) && \textbf{[Term 3]}
\end{aligned}
$$

Examine the three terms separately.

Term 3:

$$
\begin{aligned}
\textbf{Term 3} &= \sum_x \sum_y 2\big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) f(x) \\
&= 2\sum_x \left\{ \sum_y \big(y - E(Y|x)\big)\big(E(Y|x) - \mu_y\big) f(y|x) \right\} f(x) && \text{[group inner sum]} \\
&= 2\sum_x \left\{ \big(E(Y|x) - \mu_y\big) \left[\sum_y \big(y - E(Y|x)\big) f(y|x) \right] \right\} f(x) && \text{[factor out constant]} \\
&= 2\sum_x \Big\{ \big(E(Y|x) - \mu_y\big)[0] \Big\} f(x) \\
&= 0
\end{aligned}
$$

In the third line above we recognize that in the summation over the values of y the expression $\big(E(Y|x) - \mu_y\big)$ does not vary, so that it can be factored out. The remaining term in the square brackets is zero because

$$
\begin{aligned}
&\sum_y \big(y - E(Y|x)\big) f(y|x) \\
&\quad = \sum_y y f(y|x) - E(Y|x) \sum_y f(y|x) && \big[\text{factor out the constant } E(Y|X)\big] \\
&\quad = E(Y|x) - E(Y|x) = 0 && \left[\text{definition of conditional expectation \& } \sum_y f(y|x) = 1\right]
\end{aligned}
$$

Term 2:

$$\begin{aligned}
\textbf{Term 2} &= \sum_x \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) f(x) \\
&= \sum_x \left\{ \sum_y \big(E(Y|x) - \mu_y\big)^2 f(y|x) \right\} f(x) \\
&= \sum_x \left\{ \big(E(Y|x) - \mu_y\big)^2 \sum_y f(y|x) \right\} f(x) && \left[\text{factor out } \big(E(Y|x) - \mu_y\big)^2\right] \\
&= \sum_x \left\{ \big(E(Y|x) - \mu_y\big)^2 \right\} f(x) && \left[\sum_y f(y|x) = \sum_y P\big(Y = y|X = x\big) = 1\right] \\
&= \sum_x \big(E(Y|x) - \mu_y\big)^2 f(x) \\
&= \text{var}_X\big[E(Y|X)\big]
\end{aligned}$$

In the final step, we label **Term 2** as $\text{var}_X[E(Y|X)] = \sum_x \big(E(Y|x) - \mu_y\big)^2 f(x)$. The intuition behind the terminology is discussed in Section P.6.3. The key point is that $E(Y|X)$ varies as the value of X varies. One way to recognize this is to say $E(Y|X) = g(X)$. Using first principles $\text{var}\big[g(X)\big] = E\big\{g(X) - E\big[g(X)\big]\big\}^2$. Also $E_X\big[g(X)\big] = E_X[E(Y|X)] = E(Y) = \mu_y$ using the law of iterated expectations. Then

$$\text{var}_X\big[g(X)\big] = E_X\left\{\big[g(X) - \mu_y\big]^2\right\} = E_X\left\{\big[E(Y|X) - \mu_y\big]^2\right\} = \sum_x \big[E(Y|x) - \mu_y\big]^2 f(x)$$

Term 1:

$$\begin{aligned}
\textbf{Term 1} &= \sum_x \sum_y \big(y - E(Y|x)\big)^2 f(y|x) f(x) \\
&= \sum_x \left\{ \sum_y \big(y - E(Y|x)\big)^2 f(y|x) \right\} f(x) \\
&= \sum_x \text{var}(Y|x) f(x) \\
&= E_X\big[\text{var}(Y|X)\big]
\end{aligned}$$

Term 1 is the expectation of the conditional variance of Y given X. A key point here, as in **Term 2**, is that the conditional variance of Y given X is a function of X.

EXAMPLE B.1 | Variance Decomposition: Numerical Example

The calculations illustrating the variance decomposition are somewhat involved. We have broken it up into parts to simplify the logic.

Variance of *Y*

For the population in Table P.1, given in the Probability Primer, the unconditional variance of Y is $\text{var}(Y) = E\big(Y^2\big) - \mu_Y^2$. We have shown that $E(Y) = \mu_Y = 2/5$. Also,

$$E\big(Y^2\big) = \sum_y y^2 f_Y(y) = 0^2 \times (6/10) + 1^2 \times (4/10) = 2/5$$

Then $\text{var}(Y) = E\big(Y^2\big) - \mu_Y^2 = 2/5 - (2/5)^2 = 6/25 = 0.24$.

Variance of the Conditional Expectation of *Y* Given *X*

The first component of the variance decomposition is $\text{var}_X[E(Y|X)]$. As we have noted earlier, $E(Y|X) = g(X)$ is a function of X. We computed these values to be $E(Y|X = 1) = 1$, $E(Y|X = 2) = 1/2$, $E(Y|X = 3) = 1/3$, and $E(Y|X = 4) = 1/4$. What is the variance of these terms, treating X as random? The variance of a function of X, $g(X)$, is

$$\text{var}_X\big[g(X)\big] = \sum_x \Big\{ g(x) - E_X\big[g(x)\big] \Big\}^2 f_X(x)$$

Using the law of iterated expectations

$$E_X\big[g(x)\big] = E_X[E(Y|X=x)] = E(Y).$$

The calculation we need is

$$\begin{aligned}\text{var}_X\big[E(Y|X)\big] &= \sum_x \big[E(Y|X=x) - \mu_Y\big]^2 f_X(x) \\ &= \left[\sum_x E(Y|X=x)^2 f_X(x)\right] - \mu_Y^2\end{aligned}$$

Now

$$\begin{aligned}&\sum_x E(Y|X=x)^2 f_X(x) \\ &= E(Y|X=1)^2 f_X(1) + E(Y|X=2)^2 f_X(2) \\ &\quad + E(Y|X=3)^2 f_X(3) + E(Y|X=4)^2 f_X(4) \\ &= 1^2\left(\frac{1}{10}\right) + \left(\frac{1}{2}\right)^2\left(\frac{2}{10}\right) + \left(\frac{1}{3}\right)^2\left(\frac{3}{10}\right) + \left(\frac{1}{4}\right)^2\left(\frac{4}{10}\right) \\ &= \frac{5}{24}\end{aligned}$$

Then,

$$\begin{aligned}\text{var}_X\big[E(Y|X)\big] &= \left[\sum_x E(Y|X=x)^2 f_X(x)\right] - \mu_Y^2 = \frac{5}{24} - \left(\frac{2}{5}\right)^2 \\ &= \frac{29}{600} = 0.048333\ldots\end{aligned}$$

That is, $E(Y|X)$ exhibits variation as X changes and has variance 0.0483.

Expectation of the Conditional Variance of *Y* Given *X*

The second component of the variance decomposition is $E_X[\text{var}(Y|X)]$. The conditional variance $\text{var}(Y|X=x)$ varies randomly as X varies, if we treat X as random, so that finding its expected value makes sense. For the population in Table P.1, we have already computed the conditional means $E(Y|X=x)$ for each x. The conditional variances are $\text{var}(Y|X=x) = E\big(Y^2|X=x\big) - [E(Y|X=x)]^2$ so we need the terms $E\big(Y^2|X=x\big)$ for each value of X. These are

$$\begin{aligned}&E\big(Y^2|X=1\big) = 1, \quad E\big(Y^2|X=2\big) = 1/2, \\ &E\big(Y^2|X=3\big) = 1/3, \quad E\big(Y^2|X=4\big) = 1/4\end{aligned}$$

Then

$$\begin{aligned}\text{var}(Y|X=1) &= E\big(Y^2|X=1\big) - \big[E(Y|X=1)\big]^2 \\ &= 1 - 1^2 = 0 \\ \text{var}(Y|X=2) &= E\big(Y^2|X=2\big) - \big[E(Y|X=2)\big]^2 \\ &= 1/2 - (1/2)^2 = 1/4 \\ \text{var}(Y|X=3) &= E\big(Y^2|X=3\big) - \big[E(Y|X=3)\big]^2 \\ &= 1/3 - (1/3)^2 = 2/9 \\ \text{var}(Y|X=4) &= E\big(Y^2|X=4\big) - \big[E(Y|X=4)\big]^2 \\ &= 1/4 - (1/4)^2 = 3/16\end{aligned}$$

The expected value of the conditional variance is

$$\begin{aligned}E_X\big[\text{var}(Y|X)\big] &= \sum_x \text{var}(Y|X=x) f_X(x) \\ &= 0(1/10) + (1/4)(2/10) \\ &\quad + (2/9)(3/10) + (3/16)(4/10) \\ &= 23/120 = 0.191666\ldots\end{aligned}$$

The interpretation of this expectation is that if we repeatedly drew a random member from the population in Table P.1, and for each value computed the conditional variance $\text{var}(Y|X=x)$, the average of the conditional variance in many trials would approach 0.19167.

Variance of *Y* Decomposed

We have shown that for the population in Table P.1 $\text{var}_X[E(Y|X)] = 29/600$ and $E_X[\text{var}(Y|X)] = 23/120$. The variance decomposed is

$$\begin{aligned}\text{var}(Y) &= \text{var}_X\big[E\big(Y|X\big)\big] + E_X\big[\text{var}(Y|X)\big] \\ &= \frac{29}{600} + \frac{23}{120} = \frac{144}{600} = \frac{6}{25} = 0.24\end{aligned}$$

This is the same value for $\text{var}(Y)$ that we derived in the first step above.

B.1.9 Covariance Decomposition

Recall that the covariance between two random variables Y and X is $\text{cov}(X,Y) = E\big[\big(X-\mu_X\big)\big(Y-\mu_Y\big)\big]$. For discrete random variables the expectation is

$$\text{cov}(X,Y) = \sum_x \sum_y \big(x-\mu_X\big)\big(y-\mu_Y\big) f(x,y)$$

By using the relationships between marginal, conditional, and joint *pdf*s we can show

$$\text{cov}(X,Y) = \sum_x \big(x-\mu_X\big) E(Y|X=x) f(x)$$

Recall that $E(Y|X) = g(X)$ is a function of X. The covariance between X and Y can be calculated as the expected value of X, minus its mean, times a function of X,

$$\text{cov}(X, Y) = E_X\Big[(X - \mu_X)E(Y|X)\Big] \tag{B.28}$$

A numerical example of this **covariance decomposition** is given in the Probability Primer Section P.6.5.

Proof of the Covariance Decomposition

$$\begin{aligned}\text{cov}(X, Y) &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) \\ &= \sum_x \sum_y (x - \mu_X) y f(x, y) - \mu_Y \sum_x \sum_y (x - \mu_X) f(x, y)\end{aligned}$$

In this expression, the second term is zero, because

$$\begin{aligned}\sum_x \sum_y (x - \mu_X) f(x, y) &= \sum_x (x - \mu_X) \sum_y f(x, y) && \Big[\text{factor out } (x - \mu_X)\Big] \\ &= \sum_x (x - \mu_X) f(x) && \Big[\sum_y f(x, y) = f(x)\Big] \\ &= \sum_x x f(x) - \mu_X \sum_x f(x) \\ &= \mu_X - \mu_X = 0 && \Big[\sum_x f(x) = 1\Big]\end{aligned}$$

Then

$$\begin{aligned}\text{cov}(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y) &= \sum_x \sum_y (x - \mu_X) y f(x, y) \\ &= \sum_x (x - \mu_X) \Big\{ \sum_y y f(y|x) \Big\} f(x) \\ &= \sum_x (x - \mu_X) E(Y|X = x) f(x)\end{aligned}$$

B.2 Working with Continuous Random Variables

Continuous random variables can take any value in at least one interval. In economics, variables like income and market prices are treated as continuous random variables. In Figure P.2 of the Probability Primer, we depict the probability density function for a continuous random variable that ranges between zero and infinity, or $x > 0$. Because continuous random variables can take uncountably many values, the probability that any single value occurs in a random experiment is zero. For example, $P(X = 100) = 0$ or $P(X = 200) = 0$. Probability statements for continuous random variables are meaningful when we ask about outcomes within intervals, or ranges. We can ask, "What is the probability that X takes a value between 100 and 200?" These ideas were introduced in Sections P.1 and P.2 of the Probability Primer. There we noted that probabilities like these are areas under a curve that is the probability density function. It would be a good time to review those sections now if the concepts are not fresh in your minds. What we did not discuss in the Probability Primer was how exactly such probabilities are calculated. We delayed that discussion until now, because tools from integral calculus are required.

In this section, we discuss how to work with continuous random variables. The interpretation of probabilities, expected values, and variances carries over from what you learned about discrete random variables. What changes is the algebra—summation signs turn into integrals,

and this takes a little getting used to. If you have not done so, review the discussion of integrals in Appendix A.4.

B.2.1 Probability Calculations

If X is a continuous random variable with probability density function $f(x)$, then $f(x)$ must obey certain properties:

$$f(x) \geq 0 \tag{B.29}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{B.30}$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx \tag{B.31a}$$

Property (B.29) states that the *pdf* cannot take negative values. Property (B.30) states that the total area under the *pdf*, which is the probability that X falls between $-\infty$ and ∞, is one. Property (B.31a) states that the probability that X falls in the interval $[a, b]$ is the area under the curve $f(x)$ between those values. Because a single point has probability zero, it is also true that

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_{a}^{b} f(x)dx \tag{B.31b}$$

The **cumulative distribution function**, *cdf*, for a continuous random variable is $F(x) = P(X \leq x)$. Using the *cdf* we can compute

$$P(X \leq a) = \int_{-\infty}^{a} f(x)dx = F(a) \tag{B.32a}$$

The *cdf* is obtained by integrating the *pdf*. The integral is an "antiderivative," so that we can obtain the *pdf* $f(x)$ by differentiating the *cdf* $F(x)$. That is,

$$f(x) = \frac{dF(x)}{dx} = F'(x) \tag{B.32b}$$

The concept of a *cdf* is useful in many ways, including working with computer software, which includes the *cdf*s of many random variables so that probabilities can be easily computed.

EXAMPLE B.2 | Probability Calculation Using Geometry

Let X be a continuous random variable with *pdf* $f(x) = 2(1 - x)$ for $0 \leq x \leq 1$. This *pdf* is depicted in Figure B.1.

Property (B.29) holds for x in the interval [0, 1]. Furthermore, property (B.30) holds because

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} 2(1-x)dx = \int_{0}^{1} 2dx - \int_{0}^{1} 2xdx$$

$$= 2x\Big|_0^1 - x^2\Big|_0^1 = 2 - 1 = 1$$

Using Figure B.1, we can compute $P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = \frac{1}{2}$ using geometry. Using integration, we come to the same conclusion:

$$P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = \int_{1/4}^{3/4} f(x)dx = \int_{1/4}^{3/4} 2(1-x)dx$$

$$= \int_{1/4}^{3/4} 2dx - \int_{1/4}^{3/4} 2xdx = 2x\Big|_{1/4}^{3/4} - x^2\Big|_{1/4}^{3/4}$$

$$= 1 - \left(\frac{9}{6} - \frac{1}{16}\right) = \frac{1}{2}$$

The cumulative distribution function is $F(x) = 2x - x^2$ for x in the interval [0, 1], so the probability can also be computed as

$$P\left(\frac{1}{4} \leq X \leq \frac{3}{4}\right) = F\left(\frac{3}{4}\right) - F\left(\frac{1}{4}\right)$$

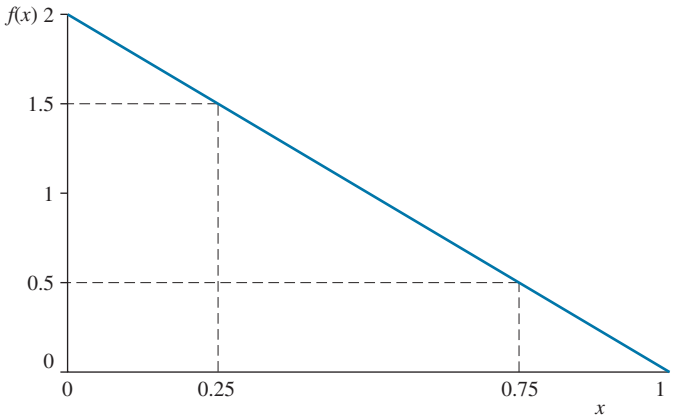


FIGURE B.1 **Probability density function $f(x) = 2(1-x)$.**

EXAMPLE B.3 | Probability Calculation Using Integration

Let X be a continuous random variable with *pdf* $f(x) = 3x^2$ for x in the interval [0,1]. Properties (B.29) and (B.30) hold. Because the *pdf* is a quadratic, we cannot use simple geometry to compute $P\left(\frac{1}{4} \le X \le \frac{3}{4}\right)$. We must use integration, obtaining

$$P\left(\frac{1}{4} \le X \le \frac{3}{4}\right) = \int_{1/4}^{3/4} f(x)\,dx = \int_{1/4}^{3/4} 3x^2 dx = x^3\Big|_{1/4}^{3/4}$$
$$= \frac{9}{64} - \frac{1}{64} = \frac{1}{8}$$

B.2.2 Properties of Continuous Random Variables

If X is a continuous random variable with probability density function $f(x)$, then its expected value is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{B.33}$$

Compare this to the expected value of a discrete random variable in (B.2). An integral has replaced the summation. The interpretation of $E(X)$ is exactly the same as in the discrete case. It is the average value of X that occurs in all possible samples from an underlying experiment.

EXAMPLE B.4 | Expected Value of a Continuous Random Variable

The expected value of the random variable in Example B.2 is

$$\int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x \cdot 2(1-x)\,dx = \int_0^1 \left(2x - 2x^2\right) dx = x^2\Big|_0^1 - \tfrac{2}{3}x^3\Big|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

The variance of a random variable X is defined as $\sigma_X^2 = E\left[\left(X - \mu_X\right)^2\right]$. This definition holds for discrete and continuous random variables. In order to compute the variance we use the analog to the rule in (B.3) for continuous random variables,

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{B.34}$$

Then, letting $g(x) = (X - \mu_X)^2$, we have

$$\begin{aligned}\sigma_X^2 &= E\left[(X-\mu_X)^2\right] = \int_{-\infty}^{\infty}(x-\mu_X)^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty}\left(x^2 + \mu_X^2 - 2x\mu_X\right) f(x)\,dx \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx + \mu_X^2\int_{-\infty}^{\infty} f(x)\,dx - 2\mu_X\int_{-\infty}^{\infty} x f(x)\,dx \\ &= E(X^2) + \mu_X^2 - 2\mu_X^2 \\ &= E(X^2) - \mu_X^2 \end{aligned} \tag{B.35}$$

To go from the third line to the fourth line, we use property (B.30) and the definition of expected value (B.33). The end result is that $\sigma_X^2 = E\left[(X-\mu_X)^2\right] = E(X^2) - \mu_X^2$ as in the discrete case.

EXAMPLE B.5 | Variance of a Continuous Random Variable

To obtain the variance of the random variable described in Example B.2, we first find

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 \cdot 2(1-x)dx = \int_0^1 (2x^2 - 2x^3)dx$$
$$= \tfrac{2}{3}x^3\Big|_0^1 - \tfrac{2}{4}x^4\Big|_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$$

Then,

$$\text{var}(X) = \sigma_X^2 = E(X^2) - \mu_X^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18}$$

B.2.3 Joint, Marginal, and Conditional Probability Distributions

To make simultaneous probability statements about more than one continuous random variable, we need the **joint probability density function** of the random variables. For example, consider the two continuous random variables U = unemployment and P = inflation rate. Suppose that their joint *pdf* is as depicted in Figure B.2.

The joint *pdf* is a surface and probabilities are volumes under the surface. If the two random variables are nonnegative, then we might ask, "What is the probability that inflation is less than 5% and at the same time unemployment is less than 6%?" That is, what is $P(U \le 6, P \le 5)$? Geometrically the answer is that this is the volume under the surface above the rectangle (in the base of the figure) defining the event. Just as an integral is used to obtain the area under a curve, a double integral is used to obtain volumes like that shown in Figure B.2. Given the joint *pdf* $f(u, p)$ we can compute the probability as

$$P(U \le 6, P \le 5) = \int_{u=0}^{6}\int_{p=0}^{5} f(u,p)\,dp\,du$$

If we know the joint *pdf*, can we obtain the marginal *pdf* of one of the random variables? If so, we can answer questions like "What is the probability that unemployment will be between 2% and 5%?" Analogous to (B.11) for discrete random variables, we integrate out the unwanted random variable. That is, the **marginal probability density function** for U is

$$f(u) = \int_{-\infty}^{\infty} f(u,p)\,dp \tag{B.36}$$

Then, for example, $P(2 \le U \le 5) = \int_2^5 f(u)\,du$.

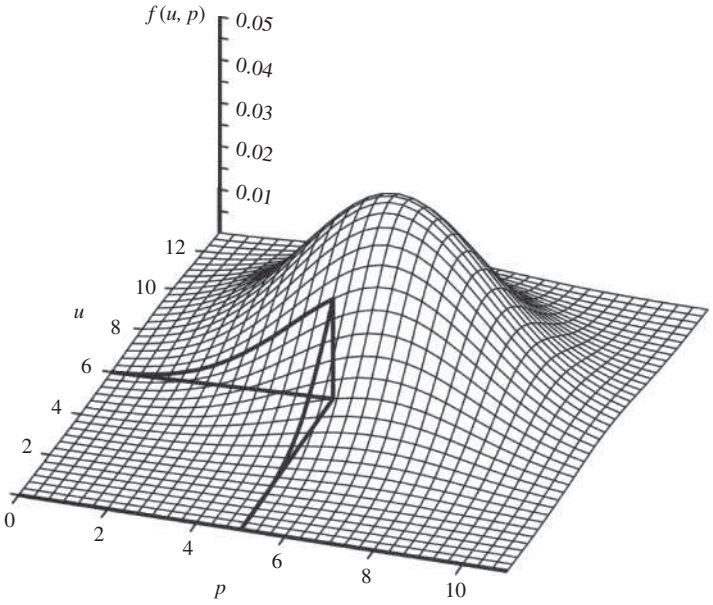


FIGURE B.2 **A joint probability density function.**

We might ask "What is the probability that unemployment will be between 2% and 5% **if** we can use monetary policy to keep the inflation rate at 2%?" This is a question about a **conditional probability**. Given that $P = 2$, what is the probability that $2 \leq U \leq 5$? Or in terms of conditioning notation, what is $P(2 \leq U \leq 5|P = 2)$? To answer such questions for continuous random variables, we need the **conditional probability density function** $f(u|p)$, which is given by

$$f(u|p) = \frac{f(u,p)}{f(p)} \tag{B.37}$$

Unlike the result (B.12) for discrete random variables, we do not obtain the probability from this division, but rather a density function that can be used for probability calculations. Not only can we obtain conditional probabilities using $f(u|p)$, but we can also obtain the **conditional expectation**, or **conditional mean**,

$$E(U|P = p) = \int_{-\infty}^{\infty} u f(u|p)\, du \tag{B.38}$$

Similarly, the **conditional variance** is

$$\text{var}(U|P = p) = \int_{-\infty}^{\infty} \left[u - E(U|P = p)\right]^2 f(u|p)\, du \tag{B.39}$$

The importance of questions involving unemployment and inflation are of great social importance. Economists and econometricians work on these problems, and you will glimpse the issues a few times throughout this book. But it is difficult. So we illustrate the above concepts with a simpler example.

EXAMPLE B.6 | Computing a Joint Probability

Let X and Y be continuous random variables with joint *pdf* $f(x, y) = x + y$ for x in [0, 1] and y in [0, 1]. You might test your geometric skills by creating a three-dimensional graph of this joint density function. Is it a valid density function? It satisfies the more general version of property (B.29), because $f(x, y) \geq 0$ for all points $x \in [0, 1]$ and $y \in [0, 1]$. Also the total amount of probability, the volume under the surface, is

$$\begin{aligned}
\int_{y=0}^{1}\int_{x=0}^{1} f(x,y)dx\,dy &= \int_{y=0}^{1}\int_{x=0}^{1}(x+y)dx\,dy \\
&= \int_{y=0}^{1}\int_{x=0}^{1} xdx\,dy + \int_{y=0}^{1}\int_{x=0}^{1} ydx\,dy \\
&= \int_{y=0}^{1}\left[\int_{x=0}^{1} xdx\right]dy + \int_{x=0}^{1}\left[\int_{y=0}^{1} ydy\right]dx \\
&= \int_{y=0}^{1}\left[\tfrac{1}{2}x^2\Big|_0^1\right]dy + \int_{x=0}^{1}\left[y^2\Big|_0^1\right]dx \\
&= \int_{y=0}^{1}\tfrac{1}{2}dy + \int_{x=0}^{1}\tfrac{1}{2}dx = \frac{1}{2}+\frac{1}{2} = 1
\end{aligned}$$

In the third line, we have used a property of multiple integrals. In the Probability Primer, Section P.4, the rule "Sum 9" states that the order of multiple summations does not matter. Similarly, as long as the limits of integration for one variable do not depend on the value of the other, the order of integration does not matter when we have multiple integrals. However, we must keep the integral symbol with its lower and upper limits paired with the variable of integration, indicated by dx or dy. In the first term in the third line above, we have isolated the integral involving x inside the integral involving y. Multiple integrals are evaluated by working from the "inside out." Solve the inside integral with respect to x, and then solve the outer integral with respect to y.

EXAMPLE B.7 | Another Joint Probability Calculation

For further practice with double integrals find the probability that X is between zero and 1/2 while Y is between 1/4 and 3/4 for the joint *pdf* in Example B.6. This is a joint probability and is computed as follows:

$$\begin{aligned}
&P\left(0 \leq X \leq \tfrac{1}{2}, \tfrac{1}{4} \leq Y \leq \tfrac{3}{4}\right) \\
&\quad= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\int_{x=0}^{\frac{1}{2}} f(x,y)dx\,dy \\
&\quad= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\int_{x=0}^{\frac{1}{2}}(x+y)dx\,dy \\
&\quad= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\left[\int_{x=0}^{\frac{1}{2}} xdx\right]dy + \int_{y=\frac{1}{4}}^{\frac{3}{4}} y\left[\int_{x=0}^{\frac{1}{2}} dx\right]dy \\
&\quad= \int_{y=\frac{1}{4}}^{\frac{3}{4}}\left[\tfrac{1}{2}x^2\Big|_0^{1/2}\right]dy + \int_{y=\frac{1}{4}}^{\frac{3}{4}} y\left[x\Big|_0^{1/2}\right]dy \\
&\quad= \tfrac{1}{8}\int_{y=\frac{1}{4}}^{\frac{3}{4}} dy + \tfrac{1}{2}\int_{y=\frac{1}{4}}^{\frac{3}{4}} ydy \\
&\quad= \tfrac{1}{8}\left(y\Big|_{1/4}^{3/4}\right) + \tfrac{1}{2}\left(\tfrac{1}{2}y^2\Big|_{1/4}^{3/4}\right) \\
&\quad= \frac{1}{8}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{4} = \frac{3}{16}
\end{aligned}$$

In the third step of this example, we did not change the order of integration in the second term. This illustrates another feature of working with multiple integrals. When carrying out the "inside" integration with respect to x the value of y is fixed, and because it is fixed it can be factored out, leaving a simpler inside integral.

EXAMPLE B.8 | Finding and Using a Marginal *pdf*

The marginal *pdf* of X, for $x \in [0, 1]$, is

$$f(x) = \int_{y=0}^{1} f(x,y)\,dy = \int_{y=0}^{1} (x+y)\,dy$$
$$= \int_{y=0}^{1} x\,dy + \int_{y=0}^{1} y\,dy = x{\cdot}y\Big|_0^1 + \tfrac{1}{2}y^2\Big|_0^1 = x + \frac{1}{2}$$

Technically we should also say that $f(x) = 0$ for $x \notin [0, 1]$, but we will generally not explicitly include this extra information. Using similar steps the marginal *pdf* of Y is $f(y) = y + 1/2$ for values of y in the [0, 1] interval. The marginal *pdf* for X can be used to compute probabilities that X falls in intervals in the domain of X, $x \in [0, 1]$. For example,

$$P\left(\tfrac{1}{2} < X < \tfrac{3}{4}\right) = \int_{1/2}^{3/4} \left(x + \tfrac{1}{2}\right) dx = \int_{1/2}^{3/4} x\,dx + \tfrac{1}{2}\int_{1/2}^{3/4} dx$$
$$= \tfrac{1}{2}x^2\Big|_{1/2}^{3/4} + \tfrac{1}{2}x\Big|_{1/2}^{3/4}$$
$$= \frac{1}{2}\left(\frac{9}{16} - \frac{1}{4}\right) + \frac{1}{2}\left(\frac{3}{4} - \frac{1}{2}\right)$$
$$= \frac{1}{2} \times \frac{5}{16} + \frac{1}{2} \times \frac{1}{4} = \frac{9}{32}$$

Using the marginal *pdf* of X, we can find its expected value.

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x)dx = \int_0^1 x\left(x + \tfrac{1}{2}\right) dx$$
$$= \int_0^1 x^2 dx + \int_0^1 \tfrac{1}{2}x dx$$
$$= \tfrac{1}{3}x^3\Big|_0^1 + \tfrac{1}{4}x^2\Big|_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

The limits of integration in the first line change from $(-\infty, \infty)$ to [0, 1], because for $x \notin [0, 1]$, $f(x) = 0$ and the area (probability) under $f(x) = 0$ is zero.

To find the variance of X, we first find

$$E\left(X^2\right) = \int_0^1 x^2 f(x)dx = \int_0^1 x^2\left(x + \tfrac{1}{2}\right)dx$$
$$= \int_0^1 x^3 dx + \int_0^1 \tfrac{1}{2}x^2 dx$$
$$= \tfrac{1}{4}x^4\Big|_0^1 + \tfrac{1}{6}x^3\Big|_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

Then

$$\sigma_X^2 = \text{var}(X) = E\left(X^2\right) - \left[E(X)\right]^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

The conditional *pdf* of Y given that $X = x$ is $f(y|x) = f(x, y)/f(x)$.

EXAMPLE B.9 | Finding and Using a Conditional *pdf*

In Example B.6 the conditional *pdf* is

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{x+y}{x+\frac{1}{2}} \quad \text{for} \quad y \in [0, 1]$$

As a specific example,

$$f\left(y\middle|X = \tfrac{1}{3}\right) = \frac{y + \frac{1}{3}}{\frac{1}{3} + \frac{1}{2}} = \frac{1}{5}(6y + 2) \quad \text{for} \quad y \in [0, 1]$$

The conditional *pdf* can be used to compute probabilities that Y falls in a given interval. Also, we can compute the **conditional mean** of Y given that $X = 1/3$

$$\mu_{Y|X=1/3} = E\left(Y\middle|X = \tfrac{1}{3}\right) = \int_{y=0}^{1} y f\left(y\middle|X = \tfrac{1}{3}\right)dy$$
$$= \int_{y=0}^{1} y\tfrac{1}{5}(6y + 2)dy$$
$$= \int_{y=0}^{1} \tfrac{6}{5}y^2 dy + \int_{y=0}^{1} \tfrac{2}{5}y dy$$
$$= \tfrac{6}{5}\left(\tfrac{1}{3}y^3\Big|_0^1\right) + \tfrac{2}{5}\left(\tfrac{1}{2}y^2\Big|_0^1\right) = \frac{2}{5} + \frac{1}{5} = \frac{3}{5}$$

Note that the conditional expected value is not the same as the **unconditional** expected value $\mu_Y = E(Y) = \frac{7}{12}$. To calculate the **conditional variance**, we first calculate

$$E\left(Y^2 \left| X = \tfrac{1}{3}\right.\right) = \int_{y=0}^{1} y^2 f\left(y \left| X = \tfrac{1}{3}\right.\right) dy$$

$$= \int_{y=0}^{1} y^2 \tfrac{1}{5}(6y + 2) dy = \frac{13}{30}$$

The conditional variance is then

$$\text{var}\left(Y \left| X = \tfrac{1}{3}\right.\right) = E\left(Y^2 \left| X = \tfrac{1}{3}\right.\right) - \left[E\left(Y \left| X = \tfrac{1}{3}\right.\right)\right]^2$$

$$= \frac{11}{150} = 0.07333$$

The unconditional variance is $\sigma_Y^2 = \text{var}(Y) = \frac{11}{144} = 0.07639$.

The **correlation** between X and Y is

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

The covariance between X and Y can be calculated using $\text{cov}(X, Y) = E(XY) - \mu_X \mu_Y$.

EXAMPLE B.10 | Computing a Correlation

To compute the expected value of XY for Example B.6, we calculate the double integral

$$E(XY) = \int_{y=0}^{1} \int_{x=0}^{1} xy f(x, y) dx dy$$

$$= \int_{y=0}^{1} \int_{x=0}^{1} xy(x + y) dx dy$$

$$= \int_{y=0}^{1} \int_{x=0}^{1} x^2 y dx dy + \int_{y=0}^{1} \int_{x=0}^{1} xy^2 dx dy$$

$$= \int_{y=0}^{1} y \left[\int_{x=0}^{1} x^2 dx\right] dy + \int_{y=0}^{1} y^2 \left[\int_{x=0}^{1} x dx\right] dy$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Then

$$\text{cov}(X, Y) = E(XY) - \mu_X \mu_Y = \frac{1}{3} - \left(\frac{7}{12}\right)\left(\frac{7}{12}\right) = \frac{-1}{144}$$

Finally, the correlation between X and Y is

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-1/144}{\sqrt{11/144}\sqrt{11/144}} = \frac{-1}{11} = -0.09091$$

B.2.4 Using Iterated Expectations with Continuous Random Variables

A useful result, proved in Section B.1.7 for the discrete case, is the **law of iterated expectations**. If X and Y are continuous random variables with joint $pdf f(x,y)$, then the expected value of Y can be calculated as

$$E(Y) = E_X\left[E(Y|X)\right]$$

This is the same result as in (B.26) for the discrete case. The exact meaning of this expression is best understood by first deriving it and then carrying through an illustration. To establish that this result is true, we proceed as follows:

$$
\begin{aligned}
E(Y) &= \int_{y=-\infty}^{\infty} y f(y) dy \\
&= \int_{y=-\infty}^{\infty} y \left[\int_{x=-\infty}^{\infty} f(x,y) dx \right] dy && \text{[replacing marginal } pdf\text{]} \\
&= \int_{y} \int_{x} y f(x,y) dx\, dy && \text{[simplifying integral]} \\
&= \int_{y} \int_{x} y \big[f(y|x) f(x) \big] dx\, dy && \text{[replace joint } pdf\text{]} \\
&= \int_{x} \left[\int_{y} y f(y|x) dy \right] f(x) dx && \text{[reverse order of integration]} \\
&= \int_{x} \big[E(Y|X) \big] f(x) dx && \text{[recognize } E(Y|X)\text{]} \\
&= E_X \big[E(Y|X) \big] && \text{[recognize expectation wrt } X\text{]}
\end{aligned}
$$

In the last line of the expression, the notation $E_X[\cdot]$ means that we take the expectation of the term in brackets treating X as random. Note that we also replaced the $(-\infty, \infty)$ integral form with a simpler form in line three indicating "over all values" of the variable of integration.

EXAMPLE B.11 | Using Iterated Expectation

To better understand the **iterated expectation** expression, for Example B.6 find the **conditional expectation** of Y given that $X = x$, where the value x is not specified:

$$
\begin{aligned}
E(Y|X = x) &= \int_{y=0}^{1} y f(y|x)\, dy = \int_{y=0}^{1} y \left[\frac{x+y}{x+\frac{1}{2}} \right] dy \\
&= \frac{2+3x}{3(2x+1)}
\end{aligned}
$$

Note that the integration over the values of Y, treating x as given, leaves us with a function of x. If we now recognize that x can take any value and is thus random, we can find the expected value of the function

$$
g(X) = \frac{2+3X}{3(2X+1)}
$$

The law of iterated expectations says that if we take the expectation of $g(X)$, treating X as random, we should obtain $E(Y)$.

$$
\begin{aligned}
E\big[g(X)\big] &= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} f(x) dx \\
&= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} \left(x + \frac{1}{2} \right) dx \\
&= \int_{x=0}^{1} \frac{2+3x}{3(2x+1)} \frac{1}{2} (2x+1) dx = \int_{x=0}^{1} \tfrac{1}{6} (2+3x) dx \\
&= \int_{x=0}^{1} \tfrac{1}{3} dx + \int_{x=0}^{1} \tfrac{1}{2} x dx = \tfrac{1}{3} x \Big|_0^1 + \tfrac{1}{4} x^2 \Big|_0^1 \\
&= \frac{1}{3} + \frac{1}{4} = \frac{7}{12} = E(Y)
\end{aligned}
$$

There are several important implications of the law of iterated expectations. First, based on $E(Y) = E_X\big[E(Y|X)\big]$, we can see that if $E(Y|X) = 0$, then $E(Y) = E_X[E(Y|X)] = E_X(0) = 0$. If the conditional expectation of Y is zero, then the unconditional expectation of Y is also zero.

Second, if $E(Y|X) = E(Y)$, then $\text{cov}(X, Y) = 0$. To see this, first rewrite $E(XY)$ as

$$\begin{aligned} E(XY) &= \int_x \int_y xyf(x,y)dydx \\ &= \int_x \int_y xyf(y|x)f(x)dydx \\ &= \int_x x\left[\int_y yf(y|x)dy\right]f(x)dx \\ &= \int_x x\big[E(Y|X)\big]f(x)dx \end{aligned} \tag{B.40}$$

If $E(Y|X) = E(Y)$, then the last line of (B.40) becomes

$$\begin{aligned} E(XY) &= \int_x x\big[E(Y)\big]\, f(x)dx = E(Y)\int_x xf(x)dx \\ &= E(Y)E(X) = \mu_Y\mu_X \end{aligned}$$

The covariance between X and Y in this case is

$$\text{cov}(X, Y) = E(XY) - \mu_X\mu_Y = \mu_X\mu_Y - \mu_Y\mu_X = 0$$

An extremely important special case of these two results concerns the consequences of $E(Y|X) = 0$. We have already seen that $E(Y|X) = 0 \Rightarrow E(Y) = 0$. Now we can also see that if $E(Y|X) = E(Y) = 0$, then $\text{cov}(X, Y) = 0$.

B.2.5 Distributions of Functions of Random Variables

As we have noted several times, a function of a random variable is random itself. The question we address in this section is, "What is the probability density function of the new random variable?" For the case of a discrete random variable this problem is not too hard. For example, consider the discrete random variable X that can take the values 1, 2, 3, or 4 with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Let $Y = 2 + 3X = g(X)$. What is the *pdf* for Y? In this case it is clear. The probability that $Y = 5, 8, 11$, or 14 corresponds exactly to the probability that $X = 1, 2, 3$, or 4, respectively, as shown in Table B.1.

What makes this possible is that each value of y corresponds to a unique value of x, and each value of x corresponds to a unique value of y. Another way to say this is that the transformation from X to Y is "one-to-one." This type of relationship is ensured to hold when the function $g(X)$ relating Y to X is either strictly increasing or strictly decreasing. Such functions are said to be *strictly monotonic*. Our function $Y = 2 + 3X = g(X)$ is strictly (monotonically) increasing. This guarantees that if $x_2 > x_1$, then $y_2 = g(x_2) > y_1 = g(x_1)$. Note in particular that we are ruling out the possibility that $y_1 = y_2$.

Determining the distribution of $Y = g(X)$ in the continuous case is a bit more challenging. In the following example, we present the **change-of-variable** technique that applies when the function $g(X)$ is strictly increasing or decreasing.

TABLE B.1 **Change of Variable: Discrete Case**

x	$P(X = x) = P(Y = y)$	y
1	0.1	5
2	0.2	8
3	0.3	11
4	0.4	14

EXAMPLE B.12 | Change of Variable: Continuous Case

Let X be a continuous random variable with *pdf* $f(x) = 2x$ for $0 < x < 1$. Let $Y = g(X) = 2X$ be another random variable. We want to compute probabilities that Y falls in certain intervals. One solution is to compute probabilities for Y based on the probability of the corresponding event for X. For example,

$$P(0 < Y < 1) = P\left(0 < X < \frac{1}{2}\right) = \int_0^{1/2} 2x\,dx = x^2\Big|_0^{1/2} = \frac{1}{4}$$

Although this is reasonable and relatively simple in this case, it will not always be so. It is preferable to determine the *pdf* of Y, say $h(y)$, and use it to compute probabilities for Y. Since $X = Y/2$, we might be tempted to substitute this into the *pdf* $f(x)$ to obtain $h(y) = 2(y/2) = y$ for $0 < y < 2$. This substitution does not work, however, because

$$\int_{-\infty}^{\infty} h(y)\,dy = \int_0^2 y\,dy = \tfrac{1}{2}y^2\Big|_0^2 = 2$$

This violates property (B.30) for a probability density function. Furthermore, using $h(y)$ to compute the probability of Y falling in the interval (0, 1) produces 0.5, which we know is incorrect.

The problem is that we must adjust the height of $h(y)$ to account for the fact that Y can take values in the interval (0, 2) whereas X can take values only in (0, 1). In fact, a change in Y of one unit corresponds to a change in X of half a unit. If we adjust $h(y)$ by this factor, we have

$$h(y) = 2(y/2)\left(\frac{1}{2}\right) = y/2, \quad 0 < y < 2$$

Using this corrected *pdf*, property (B.30) is satisfied:

$$\int_{-\infty}^{\infty} h(y)\,dy = \int_0^2 \tfrac{1}{2}y\,dy = \tfrac{1}{4}y^2\Big|_0^2 = 1$$

Also, we obtain the correct probability that Y falls in the interval (0, 1):

$$P(0 < Y < 1) = \int_0^1 \tfrac{1}{2}y\,dy = \tfrac{1}{4}y^2\Big|_0^1 = \frac{1}{4}$$

Another perspective on the change-of-variable technique is obtained by examining the integral representation for the probability that Y falls in the interval (0, 1):

$$P(0 < Y < 1) = \int_0^1 h(y)\,dy$$

The integral representation of the equivalent X event, showing explicitly the lower and upper limits of the integral, is

$$P(0 < Y < 1) = P\left(0 < X < \frac{1}{2}\right) = \int_{x=0}^{x=1/2} f(x)dx$$
$$= \int_{x=0}^{x=1/2} 2x\,dx$$

Thinking of dx as a small change in X, and noting that $x = y/2$, then $dx = dy/2$. Substituting this into the integral above, we have

$$P(0 < Y < 1) = \int_{y/2=0}^{y/2=1/2} 2\left(\tfrac{1}{2}y\right)\left(\tfrac{1}{2}dy\right) = \int_{y=0}^{y=1} \tfrac{1}{2}y\,dy$$

The adjustment factor 1/2 that we obtained intuitively appears here in the relation of dx to dy. The mathematical name for this adjustment factor is the Jacobian of the transformation (actually its absolute value, as we will soon see). Its purpose is to make the integral expression in terms of x equal to that in terms of y. Now we are ready to describe the change-of-variable technique more precisely.

Let X be a continuous random variable with *pdf* $f(x)$. Let $Y = g(X)$ be a function that is strictly increasing or strictly decreasing. This condition ensures that the function is one-to-one, so that there is exactly one Y value for each X value and exactly one X value for each Y value. The importance of this condition on $g(X)$ is that we can solve $Y = g(X)$ for X. That is, we can find an inverse function $X = w(Y)$. Then the *pdf* for Y is given by

$$h(y) = f[w(y)] \cdot \left|\frac{dw(y)}{dy}\right| \tag{B.41}$$

where $\|$ denotes the absolute value.

Change of Variable Technique to Find the *pdf* of Y: Step by Step

1. Solve $y = g(x)$ for x in terms of y;
2. Substitute this for x in $f(x)$; and
3. Multiply by the absolute value of the derivative $dw(y)/dy$, which is called the Jacobian of the transformation.

The scale factor $|dw(y)/dy|$ is the adjustment factor that makes the probabilities (i.e., the integrals) come out right. In Example B.12 the inverse function is $X = w(Y) = Y/2$. The Jacobian term is $dw(y)/dy = d(y/2)/dy = \frac{1}{2}$, and $\left|dw(y)/dy\right| = \left|\frac{1}{2}\right| = \frac{1}{2}$.

EXAMPLE B.13 | Change of Variable: Continuous Case

Let X be a continuous random variable with *pdf* $f(x) = 2x$ for $0 < x < 1$. Let $Y = g(X) = 8X^3$ be the function of X in which we are interested. The function $Y = g(X) = 8X^3$ is strictly increasing for the set of values that X can take, $0 < x < 1$. The corresponding set of values that Y can take is $0 < y < 8$. Because the function is strictly increasing, we can solve for the inverse function

$$x = w(y) = \left(\frac{1}{8}y\right)^{1/3} = \frac{1}{2}y^{1/3}$$

and

$$\frac{dw(y)}{dy} = \frac{1}{6}y^{-2/3}$$

Applying the change-of-variable formula (B.41), we have

$$\begin{aligned} h(y) &= f\left[w(y)\right] \times \left|\frac{dw(y)}{dy}\right| \\ &= 2\left(\frac{1}{2}y^{1/3}\right) \times \left|\frac{1}{6}y^{-2/3}\right| \\ &= \frac{1}{6}y^{-1/3}, 0 < y < 8 \end{aligned}$$

The change-of-variable technique can be modified for the case of several random variables, X_1, X_2 being transformed into Y_1, Y_2. For a description of the method, which requires matrix algebra, see William Greene (2018) *Econometric Analysis*, 8th edition, Pearson Prentice Hall, pp. 1120–1121.

B.2.6 Truncated Random Variables

A truncated random variable is one whose probability density function is cutoff above or below some specified point. That is suppose that X is a continuous random variable such that $-\infty < x < \infty$ and its *pdf* is $f(x)$. The *pdf* $f(x)$ has the properties (i) $f(x) \geq 0$ and (ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Now suppose that the underlying experiment is such that only x values greater than some value a are possible. What is the probability density function of this random variable? It is not simply $f(x)$ for $x > c$ because the *pdf* would not satisfy condition (ii) above, the area beneath it, which represents probability, would not total one. There is a simple fix-up. The density of a truncated random variable, such that $x > c$, is

$$f(x|x > c) = \frac{f(x)}{P(X > c)}$$

The adjustment makes the area equal to one.

Intuitively, what will happen to the expected value and variance of the truncated random variable, relative to the untruncated one? Thinking about it for a moment you can see that $E(X|x > c) > E(X)$ and $\text{var}(X|x > c) < \text{var}(X)$. Specific examples of truncated random variables will appear in the case of Poisson random variables (Section B.3.3) and normally distributed random variables (Section B.3.5).

B.3 Some Important Probability Distributions

In this section, we give brief descriptions and summarize the properties of the probability distributions used in this book.

B.3.1 The Bernoulli Distribution

Let the random variable X denote an experimental outcome with only two possible outcomes, A or B. Let $X = 1$ if the outcome is A and let $X = 0$ if the outcome is B. Let the probabilities of the outcomes be $P(X = 1) = p$ and $P(X = 0) = 1 - p$ where $0 \leq p \leq 1$. X is said to have a **Bernoulli distribution**. The *pdf* of this Bernoulli random variable is

$$f(x|p) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.42}$$

The expected value of X is $E(X) = p$, and its variance is $\text{var}(X) = p(1 - p)$. This random variable arises in choice models, such as the linear probability model (Chapters 7, 8, and 16) and in binary and multinomial choice models (Chapter 16).

B.3.2 The Binomial Distribution

If $X_1, X_2, \ldots, X_n$ are independent random variables, each having a Bernoulli distribution with parameter p, then $X = X_1 + X_2 + \cdots + X_n$ is a discrete random variable that is the number of successes (i.e., Bernoulli experiments with outcome $X_i = 1$) in n trials of the experiment. The random variable X is said to have a **binomial distribution**. The *pdf* of this random variable is

$$P(X = x|n, p) = f(x|n, p) = \binom{n}{x} p^x(1-p)^{n-x} \quad \text{for} \quad x = 0, 1, \ldots, n \tag{B.43}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

is the number of combinations of n things taken x at a time. This distribution has two parameters, n and p, where n is a positive integer indicating the number of experimental trials and $0 \leq p \leq 1$. These probabilities are tedious to compute by hand, but econometric software has functions to carry out the calculations. The discrete probabilities are illustrated in Figure B.3.

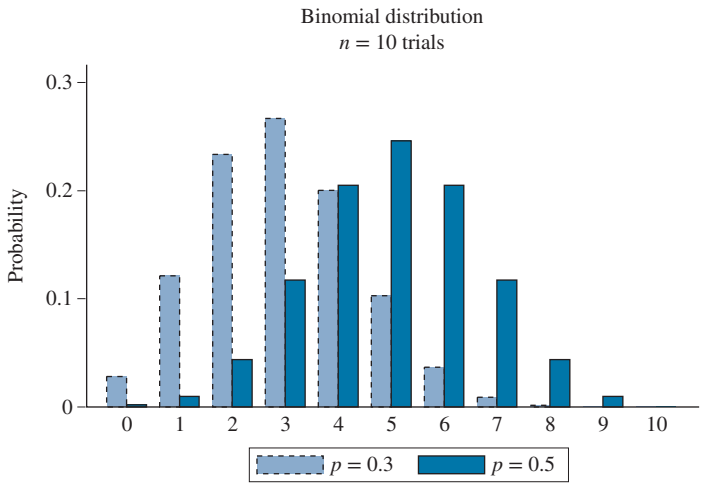


FIGURE B.3 **Binomial distributions for $n = 10$.**

The expected value and variance of X are

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

$$\text{var}(X) = \sum_{i=1}^{n} \text{var}(X_i) = np(1-p)$$

A related random variable is $Y = X/n$, which is the proportion of successes in n trials of an experiment. Its mean and variance are $E(Y) = p$ and $\text{var}(Y) = p(1-p)/n$.

B.3.3 The Poisson Distribution

Whereas a **binomial random variable** is the number of event occurrences in a given number of experimental trials, n, the Poisson random variable is the number of event occurrences in a given interval of time or space. The probability density function for this discrete random variable X is

$$P(X = x|\mu) = f(x|\mu) = \frac{e^{-\mu}\mu^x}{x!} \quad \text{for} \quad x = 0, 1, 2, 3, \ldots \tag{B.44}$$

Probabilities depend on the parameter μ, and $e \cong 2.71828$ is the base of natural logarithms. The expected value and variance of X are $E(X) = \text{var}(X) = \mu$. The **Poisson distribution** is used in models involving count variables (Chapter 16), such as the number of visits a person makes to a physician during a year. Probabilities for $x = 0$ to 10 for distributions with $\mu = 3$ and $\mu = 4$ are shown in Figure B.4.

In applications of count data, we sometimes only observe positive outcomes. For example, suppose we might survey individuals at a shopping mall and ask "How many times have you visited the mall this year?" The answer must be one or more. Using the notion of a truncated random variable introduced in Section B.2.6, the probability function in (B.44) becomes

$$f(x|\mu, x > 0) = \frac{f(x|\mu)}{P(X > 0)}$$

In the case of the Poisson distribution $P(X > 0) = 1 - P(X = 0) = 1 - e^{-\mu}$. Then the **truncated Poisson distribution** is

$$f(x|\mu, x > 0) = \frac{f(x|\mu)}{1 - P(X = 0)} = \frac{(e^{-\mu}\mu^x)/x!}{1 - e^{-\mu}} \quad \text{for} \quad x = 1, 2, 3, \ldots$$

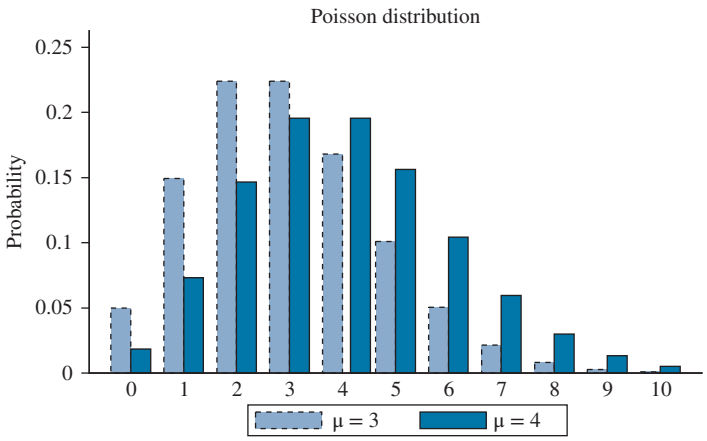


FIGURE B.4 **Poisson distributions.**

B.3.4 The Uniform Distribution

A continuous distribution that is vastly important for theoretical purposes is the **uniform distribution**. The random variable X with values $a \le x \le b$ has a uniform distribution if its *pdf* is given by

$$f(x|a,b) = \frac{1}{b-a} \quad \text{for} \quad a \le x \le b \tag{B.45}$$

The plot of the density function is given in Figure B.5.

The area under $f(x)$ between a and b is one, which is required of any probability density function for a continuous random variable. The expected value of X is the midpoint of the interval $[a, b]$, $E(X) = (a + b)/2$. This can be deduced from the symmetry of the distribution. The variance of X is $\text{var}(X) = E(X^2) - \mu^2 = (b - a)^2/12$.

An interesting special case occurs when $a = 0$ and $b = 1$, so that $f(x) = 1$ for $0 \le x \le 1$. The distribution, shown in Figure B.6, describes one common meaning of "a random number between zero and one."

The uniform distribution has the property that any two intervals of equal width have the same probability of occurring. That is,

$$P(0.1 \le X \le 0.6) = P(0.3 \le X \le 0.8) = P(0.21131 \le X \le 0.71131) = 0.5$$

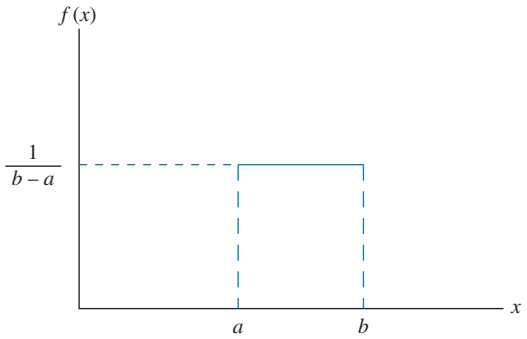


FIGURE B.5 **A uniform distribution.**

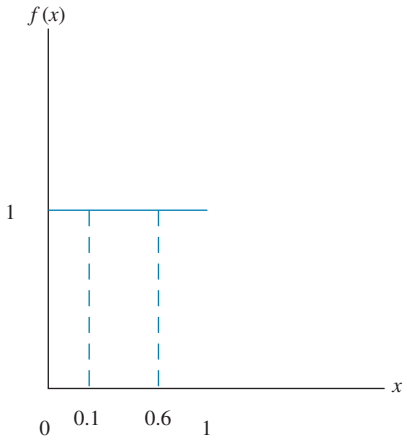


FIGURE B.6 **A uniform distribution on [0, 1] interval.**

Picking a number randomly between zero and one is conceptually complicated by the fact that the interval has an uncountably infinite number of values, and the probability of any one of them occurring is zero. What is more likely meant by such a statement is that each interval of equal width has the same probability of occurring, no matter how narrow. This is exactly the nature of the uniform distribution.

B.3.5 The Normal Distribution

The normal distribution was described in the Probability Primer, Section P.6. A point not stressed at that time was why we must consult tables, like Statistical Table 1 to calculate normal probabilities. For example, we now know that for the continuous and normally distributed random variable X, with mean μ and variance σ^2, the probability that X falls in the interval $[a, b]$ is

$$\int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-(x-\mu)^2/2\sigma^2\right] dx$$

Unfortunately this integral does not have a closed-form algebraic solution. Consequently, we wind up working with tabled values containing numerical approximations to areas under the **standard normal distribution**, or we use computer software functions in a similar manner.

Moments of the Normal Distribution

If X is a random variable, then $E(X^r)$ is called the rth moment of the random variable about the origin. Sometimes they are called *raw* moments. If $X \sim N(\mu,\sigma^2)$, then we have the following useful expressions for the first three moments about the origin:

$$E(X) = \mu$$
$$E(X^2) = \mu^2 + \sigma^2$$
$$E(X^3) = 0$$

For any random variable X, $E(X-\mu)^r$ is the rth moment of the random variable about its mean. Sometimes, these are called *central* moments. For the normal random variable $X \sim N(\mu,\sigma^2)$, these are

$$E(X-\mu) = 0$$
$$E\left[(X-\mu)^2\right] = \sigma^2$$
$$E\left[(X-\mu)^3\right] = 0$$
$$E\left[(X-\mu)^4\right] = 3\sigma^4$$

The second moment about the mean $E\left[(X-\mu)^2\right] = \sigma^2$ is the variance of the random variable. The third moment, $E\left[(X-\mu)^3\right] = 0$, is related to the *skewness* of the probability density function. Because the normal distribution is symmetrical, it is not skewed, its skewness is zero. It is also true that all odd central moments are zero, so that $E\left[(X-\mu)^r\right] = 0$ if r is an odd number. The fourth moment about the mean, $E\left[(X-\mu)^4\right] = 3\sigma^4$, is related to the *kurtosis* of the distribution, which is a measure of the thickness of the tails of the distribution. For the normal distribution, the standardized fourth moment $E\left[(X-\mu)^4/\sigma^4\right] = 3$ is a useful reference point for tail thickness. For more about population moments see Appendix C.4.

The Truncated Normal Distribution

In Section B.2.6, we introduced the notion of a truncated random variable. The truncated normal distribution has been studied quite intensely. Suppose that $X \sim N(\mu,\sigma^2)$ but the distribution is **truncated from below** so that $x > c$. Then

$$f(x|x > c) = \frac{f(x)}{P(X > c)}$$

For the normal distribution

$$P(X > c) = P\left(\frac{X-\mu}{\sigma} > \frac{c-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{c-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

where $\Phi(\alpha)$ is the cumulative distribution function of the standard normal random variable evaluated at $\alpha = (c-\mu)/\sigma$. Then

$$f(x|x > c) = \frac{f(x)}{1 - \Phi(\alpha)}$$

Following Greene (2018, p. 921), define the **Inverse Mill's Ratio** as

$$\lambda(\alpha) = \begin{cases} \dfrac{\phi(\alpha)}{1-\Phi(\alpha)} & \text{if truncation is from below, so that } x > c \\ \dfrac{-\phi(\alpha)}{\Phi(\alpha)} & \text{if truncation is from above, so that } x < c \end{cases}$$

where $\phi(\alpha)$ is the probability density function of the standard normal random variable evaluated at $\alpha = (c-\mu)/\sigma$. Then the expected value of the truncated normal random variable is

$$E(X|\text{truncation}) = \mu + \sigma\lambda(\alpha)$$

Letting $\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$, the variance of the truncated normal random variable is

$$\text{var}(X|\text{truncation}) = \sigma^2[1 - \delta(\alpha)]$$

This is consistent with the intuition about the variance of a truncated variable in Section B.2.6 because $0 < \delta(\alpha) < 1$.

The normal distribution is related to the chi-square, *t*-, and *F*-distributions, which we now discuss.

B.3.6 The Chi-Square Distribution

Chi-square random variables arise when standard normal random variables are squared. If Z_1, Z_2, …, Z_m denote m *independent* $N(0,1)$ random variables, then

$$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2 \sim \chi^2_{(m)} \tag{B.46}$$

The notation $V \sim \chi^2_{(m)}$ is read as: The random variable V has a chi-square distribution with m **degrees of freedom**. The degrees of freedom parameter m indicates the number of *independent* $N(0,1)$ random variables that are squared and summed to form V. The value of m determines the entire shape of the **chi-square distribution**, including its mean and variance as

$$E(V) = E\left[\chi^2_{(m)}\right] = m$$

$$\text{var}(V) = \text{var}\left[\chi^2_{(m)}\right] = 2m \tag{B.47}$$

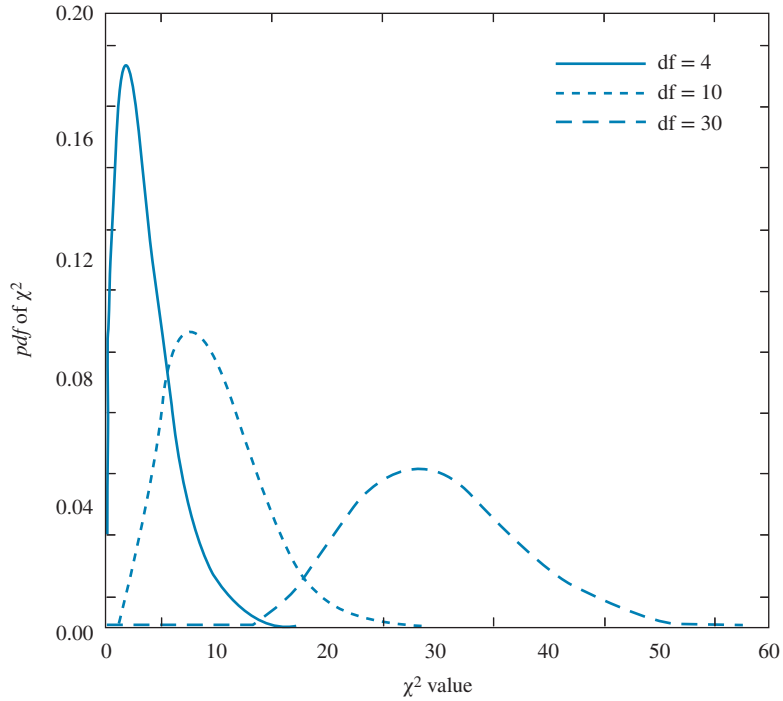


FIGURE B.7(a) **The chi-square distribution.**

In Figure B.7(a) graphs of the chi-square distribution for various degrees of freedom are presented. The values of V must be nonnegative, $v \geq 0$, because V is formed by squaring and summing m standardized normal, $N(0,1)$, random variables. The distribution has a long tail, or is *skewed*, to the right. As the degrees of freedom m gets larger, however, the distribution becomes more symmetric and "bell-shaped." In fact, as m gets larger, the chi-square distribution converges to, and essentially becomes, a normal distribution.

The 90th, 95th, and 99th percentile values of the chi-square distribution for selected values of the degrees of freedom are given in Statistical Table 3. These values are often of interest in hypothesis testing.

In the definition (B.46) of the chi-square random variable the $Z_i,\ i = 1, \ldots, m$ are statistically independent standard normal, $N(0, 1)$, random variables. If, instead, V is equal to the sum of squares of normal random variables $(Z_i + \delta_i)$ that have a non-zero mean δ_i and variance 1, then V has a **non-central chi-square distribution** with m degrees of freedom and **non-centrality parameter** $\delta = \delta_1^2 + \delta_2^2 + \cdots + \delta_m^2$, which is denoted by $\chi^2_{(m,\delta)}$. If all $\delta_i = 0$ then we have the usual **central chi-square** distribution. That is,

$$V = (Z_1 + \delta_1)^2 + (Z_2 + \delta_2)^2 + \cdots + (Z_m + \delta_m)^2 \sim \chi^2_{(m,\delta)}$$

In Figure B.7(b) we plot a few non-central chi-square distributions, all having $m = 10$ degrees of freedom.

The effect of the non-centrality parameter is to shift the chi-square density function to the right, increasing both the mean and the variance, which become $E\left[\chi^2_{(m,\delta)}\right] = m + \delta$ and $\text{var}\left[\chi^2_{(m,\delta)}\right] = 2(m + 2\delta)$.

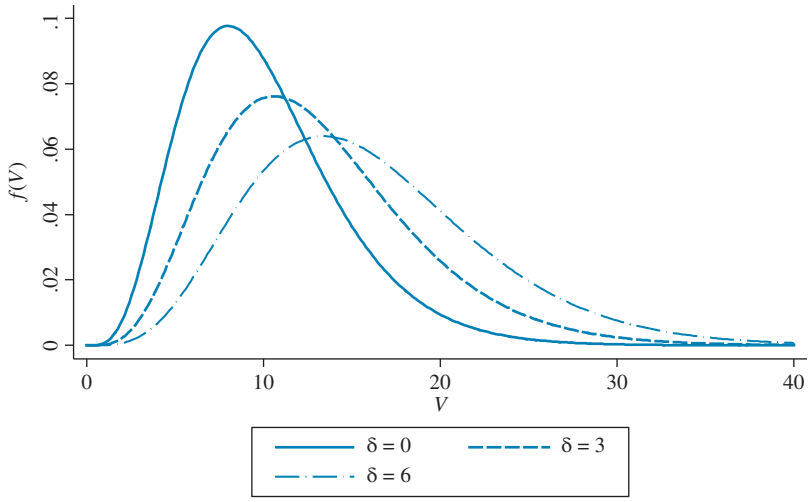


FIGURE B.7(b) **Non-central chi-square distributions, $m = 10$ degrees of freedom and non-centrality $\delta = 0, 3, 6$.**

B.3.7 The *t*-Distribution

A t random variable (no upper case) is formed by dividing a standard normal random variable $Z \sim N(0,1)$ by the square root of an *independent* chi-square random variable, $V \sim \chi^2_{(m)}$, that has been divided by its degrees of freedom m. If $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$, and if Z and V are independent, then

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)} \tag{B.48}$$

The t-distribution's shape is completely determined by the degrees of freedom parameter, m, and the distribution is symbolized by $t_{(m)}$.

Figure B.8(a) shows a graph of the t-distribution with $m = 3$ degrees of freedom relative to the $N(0,1)$. Note that the t-distribution is less "peaked," and more spread out than the $N(0,1)$. The t-distribution is symmetric, with mean $E(t_{(m)}) = 0$ and variance $\text{var}(t_{(m)}) = m/(m-2)$. As the degrees of freedom parameter $m \to \infty$, the $t_{(m)}$ distribution approaches the standard normal $N(0,1)$.

Computer programs have functions for the *cdf* of t-random variables that can be used to calculate probabilities. Since certain probabilities are widely used, Statistical Table 2 contains frequently used percentiles of t-distributions, called **critical values** of the distribution. For example, the 95th percentile of a t-distribution with 20 degrees of freedom is $t_{(0.95,20)} = 1.725$. The t-distribution is symmetric, so Statistical Table 2 shows only the right tail of the distribution.

The statistic formed from a $N(\delta,1)$ random variable and an independent central chi-square random variable with m degrees of freedom is called a **non-central *t*-random variable**,

$$t = \frac{Z + \delta}{\sqrt{V/m}} \sim t_{(m,\delta)}$$

This distribution has two parameters, the degrees of freedom, m, and the **non-centrality parameter** δ. The usual t-random variable in (B.48) has non-centrality parameter $\delta = 0$ and is sometimes called the **central *t*-distribution**. The additive factor in the numerator causes the resulting distribution to be centered at a value other than zero if $\delta \neq 0$. In Figure B.8(b), we plot the $t_{(3,\delta)}$

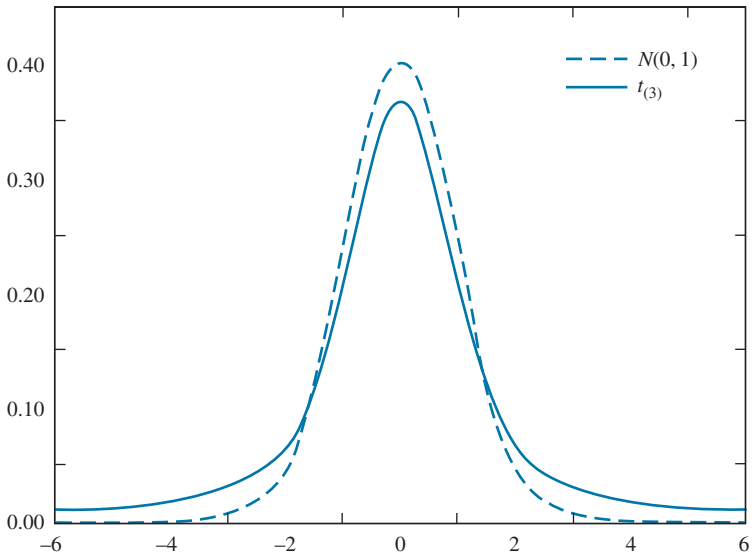


FIGURE B.8(a) **The standard normal and $t_{(3)}$ probability density functions.**

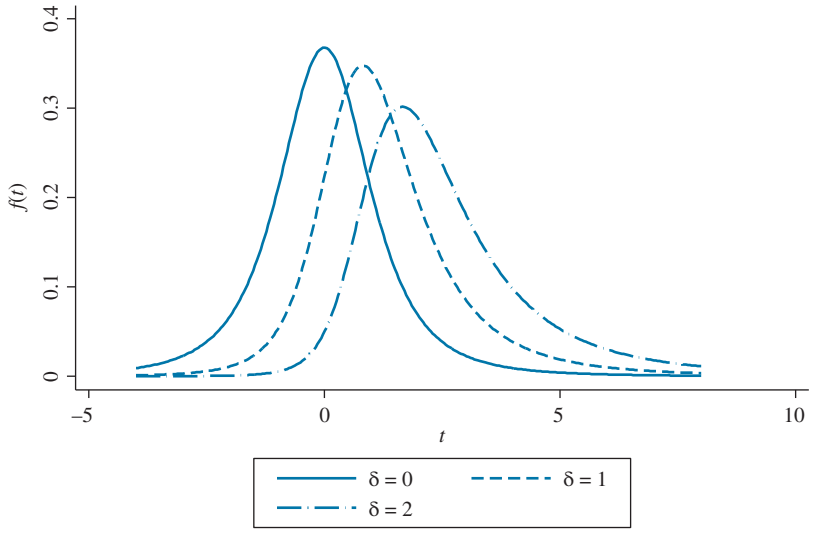


FIGURE B.8(b) **Non-central t-distributions, $m = 3$ degrees of freedom and non-centrality $\delta = 0, 1, 2$.**

density for values of $\delta = 0, 1, 2$. The positive non-centrality parameter shifts the density function rightward.

B.3.8 The F-Distribution

An F-random variable is formed by the ratio of two independent chi-square random variables that have been divided by their degrees of freedom. If $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$, and if V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)} \tag{B.49}$$

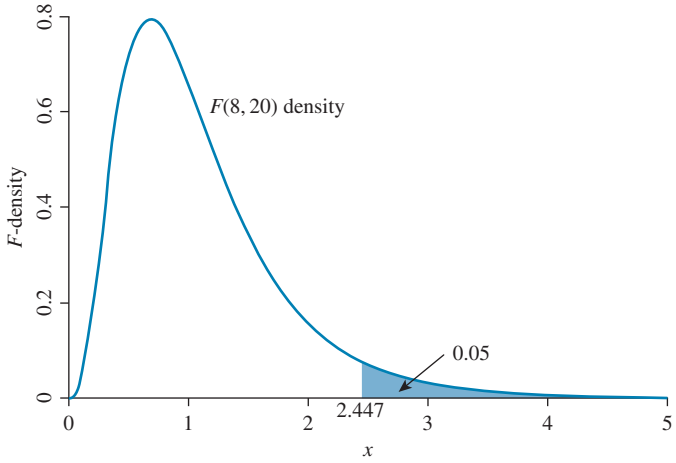


FIGURE B.9(a) **The 95th percentile of an $F_{(8,20)}$-random variable.**

The F-distribution is said to have m_1 ***numerator degrees of freedom*** and m_2 ***denominator degrees of freedom***. The values of m_1 and m_2 determine the shape of the distribution, which in general looks like Figure B.9(a). The range of the random variable is $(0,\infty)$, and it has a long tail to the right. For example, the 95th percentile value for an F-distribution with $m_1 = 8$ numerator degrees of freedom and $m_2 = 20$ denominator degrees of freedom is $F_{(0.95,8,20)} = 2.447$. Critical values (two decimal places) for the ***F*-distribution** are given in Statistical Table 4 (the 95th percentile) and Statistical Table 5 (the 99th percentile).

In the definition (B.49), the numerator chi-square random variable V_1 has a **central chi-square** distribution, with non-centrality parameter $\delta = 0$. The central and non-central chi-square distributions are discussed in Section B.3.6. If the numerator in (B.49) has a non-central chi-square distribution, $V_1 \sim \chi^2_{(m_1,\delta)}$ with m_1 degrees of freedom and non-centrality, δ, then the F-random variable has a **non-central** F-distribution with numerator degrees of freedom m_1, denominator degrees of freedom m_2 and non-centrality parameter δ. This distribution is denoted by $F_{(m_1,m_2,\delta)}$. In Figure B.9(b), we show several density functions for comparison with

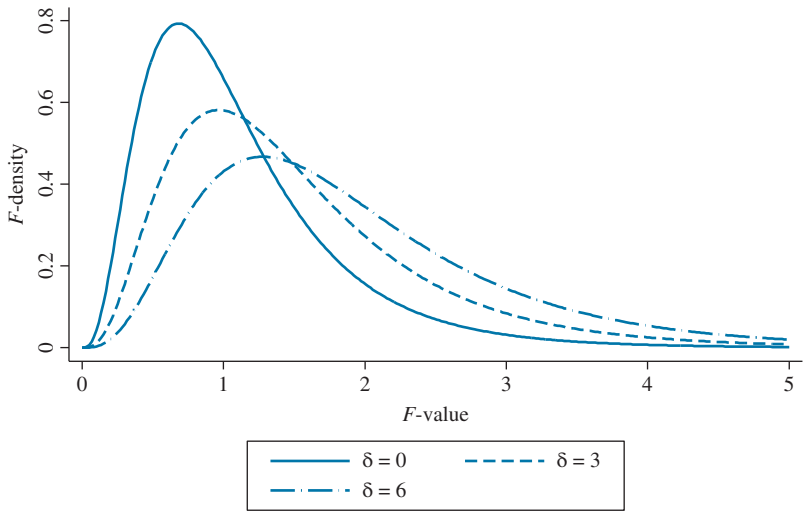


FIGURE B.9(b) **Non-central $F_{(8,\,20,\,\delta)}$-distributions with $\delta = 0,\ 3,\ 6$.**

Figure B.9(a). These have degrees of freedom $m_1 = 8$, $m_2 = 20$, and non-centrality $\delta = 0,\ 3,\ 6$. As the non-centrality parameter increases, the F-density moves to the right, increasing both its mean and variance.

B.3.9 The Log-Normal Distribution

A continuous random variable X is said to have a **log-normal** distribution if

$$\ln(X) \sim N(\mu, \sigma^2), \quad x > 0$$

The probability density function of X is

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

Probabilities are computed using the *cdf* of the standard normal random variable, $\Phi(z)$. That is

$$P(X \le c) = P[\ln(X) \le \ln(c)] = P\left\{[(\ln(X) - \mu)/\sigma] \le [(\ln(c) - \mu)/\sigma]\right\}$$

$$= \Phi[(\ln(c) - \mu)/\sigma]$$

The parameters μ and σ^2 are the mean and variance of $\ln(X)$. The *pdf* of X is not symmetrical. The **median** of X is $m = \exp(\mu)$ and $\mu = \ln(m)$.[1] The expected value of X is

$$E(X) = m\exp(\sigma^2/2) = \exp(\mu)\exp(\sigma^2/2) = \exp(\mu + \sigma^2/2)$$

Using $\omega = \exp(\sigma^2)$, the variance of X is

$$\text{var}(X) = m^2\omega(\omega - 1) = \exp(2\mu)\exp(\sigma^2)[\exp(\sigma^2) - 1] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$$

The mode of the density is m/ω so that $E(X) = \text{mean} > \text{median} > \text{mode}$. In Figure B.10, we plot the log-normal density for several choices of σ with median $m = 1$.

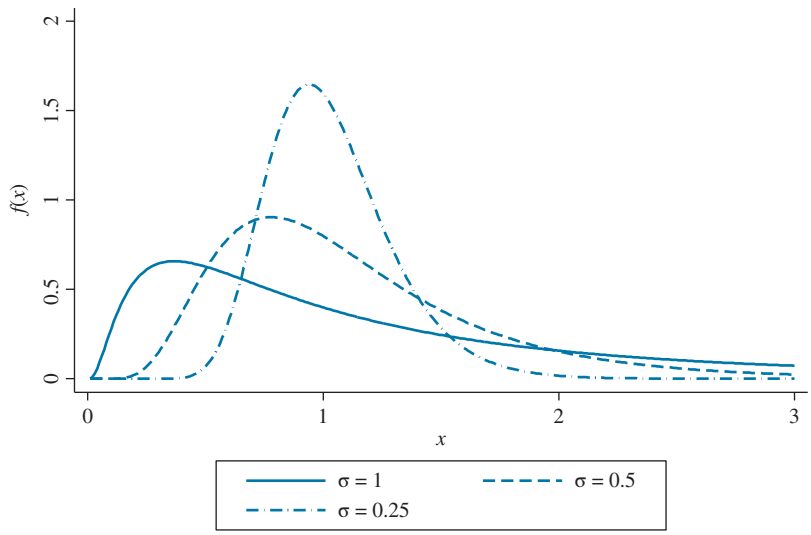


FIGURE B.10 **Log-normal densities. With median $m = 1$ and shape $\sigma = 1, 0.5, 0.25$.**

[1] In the statistics literature σ is sometimes called the **shape** parameter and m the **scale** parameter.

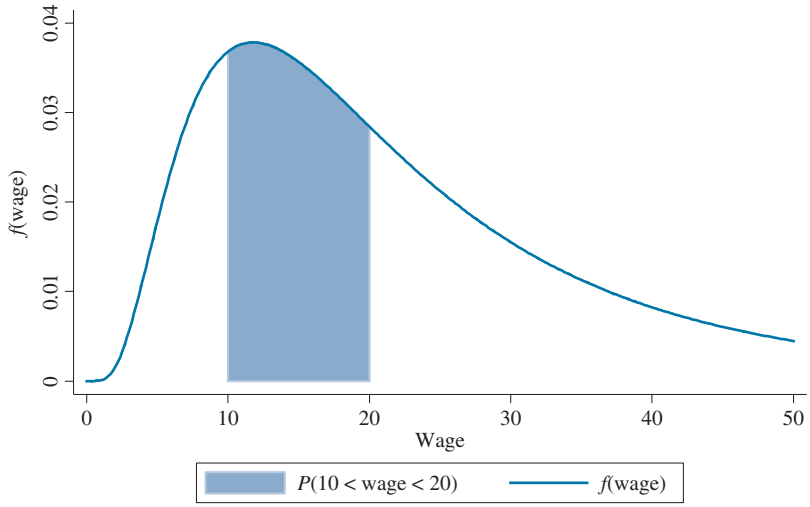


FIGURE B.11 **Hypothetical probability density function for *WAGE*, log-normal with $m = 19.23$ and $\sigma = 0.7$.**

A common use of the log-normal distribution in Economics is for wages, incomes, and house prices. These variables are positive, and the distributions are skewed with a long tail to the right, indicating that a small portion of the population has large values. Using the data file *cps5*, the median wage is \$19.23, and the mean wage is \$23.5. Using the expression for the expected value of a log-normal distribution $E(X) = m\exp\left(\sigma^2/2\right)$, we can calculate the shape parameter $\sigma = \sqrt{2\ln(E(X)/m)}$, which is about 0.7 using the *cps5* data values. Then the implied distribution of *WAGE* is shown in Figure B.11. What is the probability that a randomly chosen worker will have an hourly wage between \$10 and \$20? Graphically, it is the area under the *pdf* between 10 and 20. The calculated probability, using our approximated log-normal distribution is

$$P(10 < WAGE < 20) = \Phi\left[\frac{\ln(20) - \ln(19.23)}{0.7}\right] - \Phi\left[\frac{\ln(10) - \ln(19.23)}{0.7}\right]$$

$$= \Phi(0.05609) - \Phi(-0.93412)$$

$$= 0.52236 - 0.17512 = 0.34724$$

In the *cps5* data, 38.95% of the individuals have a wage between \$10 and \$20, so our rough approximation using the log-normal distribution is not far off.

B.4 Random Numbers

In several chapters we carry out Monte Carlo simulations to illustrate the sampling properties of estimators. See, for example, Chapters 3, 4, 5, 10, 11, and 16. To use Monte Carlo simulations we rely upon the ability to create **random numbers** from specific probability distributions, such

as the uniform and the normal. Using computer simulations is widespread in all sciences. In this section we introduce to you this aspect of computing.[2] You should first realize that the idea of creating random numbers using a computer is paradoxical, because by definition random numbers that are "created" cannot be truly random. The random numbers generated by a computer are **pseudo-random numbers** in that they "behave as if they were random." We present one method for generating pseudo-random numbers called the **inverse transformation** approach, or the **inversion method**. This method assumes that we have the ability to generate pseudo-random numbers from the **uniform distribution** (see Sections B.3.4 and B.4.1) on the (0, 1) interval. The uniformly distributed random variables are then transformed into random variables with other distributions.

EXAMPLE B.14 | An Inverse Transformation

Let U be a random variable with a uniform distribution. It is a continuous random variable with *pdf* $h(u) = 1$ for $u \in (0, 1)$. See Figure B.6 for an illustration. If $Y = U^{1/2}$, then $0 < y < 1$. Furthermore, the square root function is strictly increasing, so that we can apply the change-of-variable technique to find the *pdf* of Y. The inverse function is $U = w(Y) = Y^2$, and the Jacobian of the transformation is $dw(y)/dy = d\left(y^2\right)/dy = 2y$. The *pdf* of Y is then

$$f(y) = h\left[w(y)\right] \times \frac{dw(y)}{dy} = 1 \times 2y = 2y, \quad 0 < y < 1 \tag{B.50}$$

This is a distribution that we have used in Examples B.12 and B.13. The importance of this example is that it shows that we can obtain a random number from the distribution in (B.50) by taking the square root of a random number from a uniform distribution.

Example B.14 leads us toward a general technique, the inversion method, for drawing random numbers from certain distributions. Suppose you wish to obtain a random number from a specific probability distribution, with *pdf* $f(y)$ and *cdf* $F(y)$.

The Inversion Method: Step by Step

1. Obtain a uniform random number u_1 in the (0, 1) interval.
2. Let $u_1 = F\left(y_1\right)$.
3. Solve the equation in step 2 for $y_1 = F^{-1}\left(u_1\right)$.
4. The value y_1 is a random number from the *pdf* $f(y)$.

The inversion method can be used to draw random numbers from any distribution that permits you to carry out step 3. The solution is often denoted $y_1 = F^{-1}\left(u_1\right)$, where F^{-1} is called the **inverse cumulative distribution function**. The *cdf* function F is said to be **invertible**.

[2] A well-written book on the subject is by James E. Gentle (2003) *Random Number Generation and Monte Carlo Methods,* New York: Springer. Also, J. F. Kiviet (2011) Monte Carlo Simulation for Econometricians, *Foundations and Trends® in Econometrics*, vol 5, nos 1–2.

EXAMPLE B.15 | The Inversion Method: An Example

Suppose the target distribution, from which we want a random number, is $f(y) = 2y, 0 < y < 1$. The *cdf* of Y is $P(Y \leq y) = F(y) = y^2, 0 < y < 1$. The two distributions are shown in Figure B.12. Set a uniform random number $u_1 = F(y_1) = y_1^2$ and solve to obtain $y_1 = F^{-1}(u_1) = (u_1)^{1/2}$. The value y_1 is a random value, or a **random draw**, from the probability distribution $f(y) = 2y, 0 < y < 1$. This agrees perfectly with the result in Example B.6, where we showed that the square root of a uniform random variable has this *pdf*.

In Figure B.12(a), suppose the uniform random number value is $u_1 = 0.16$. It falls between 0 and 1, along the vertical axis of the *cdf* function $F(y)$. The value $u_1 = 0.16$ corresponds to the value $y_1 = 0.4 = (u_1)^{1/2} = (0.16)^{1/2}$ on the horizontal axis. In the lower panel we see the connection between the *pdf* and the *cdf*. The area under the *pdf* to the left of $y_1 = 0.4$ is the probability $P(0 < Y < 0.4) = 0.16$. For every randomly drawn uniform random number u_i there is a unique corresponding y_i from the distribution $f(y) = 2y, 0 < y < 1$.

To illustrate, in the data file *uniform1*, we have 1,000 observations on two independent uniform random variables *U1* and *U2*.[3] Figure B.13 shows the histogram of *U1*. There are 10 intervals and approximately 10% of the values fall into each, as we would expect for values from a uniform distribution.

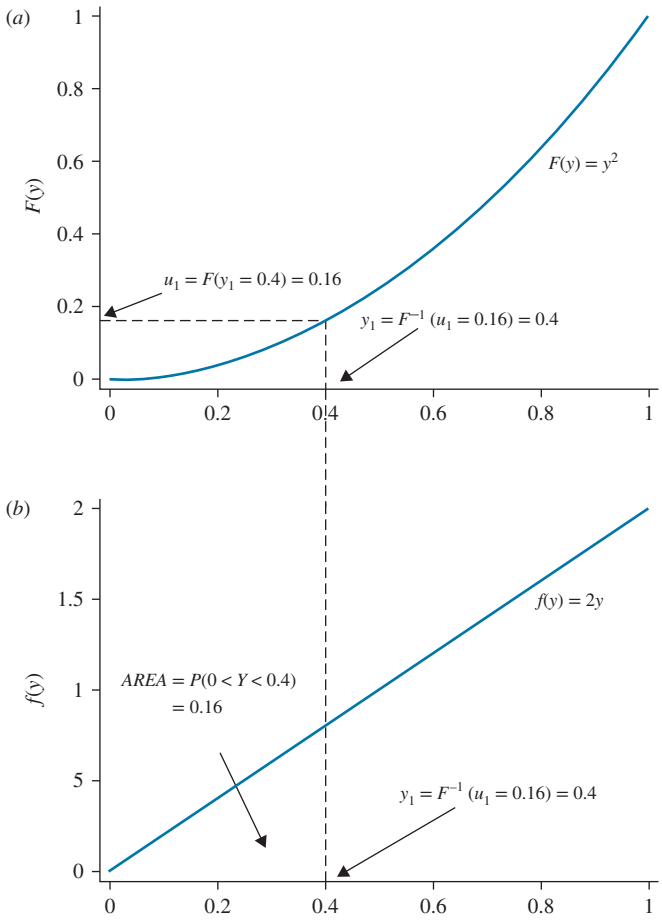


FIGURE B.12 **(*a*) Cumulative distribution function and (*b*) probability density function.**

[3]The data file *uniform2* contains 10,000 observations if you prefer a larger sample.

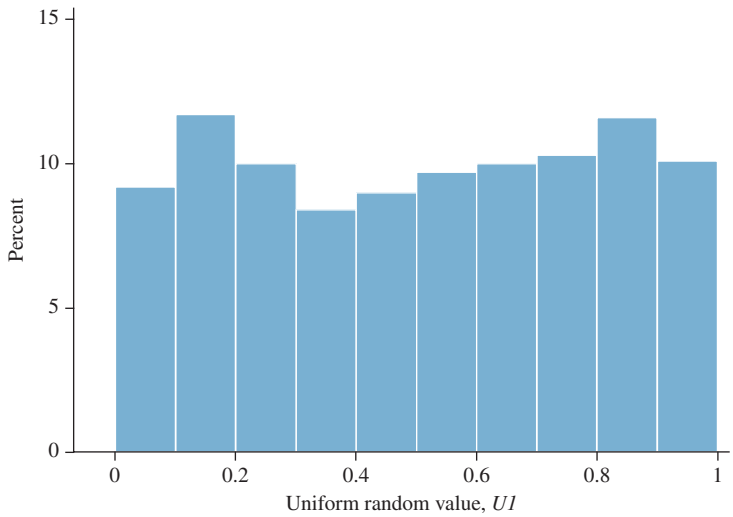


FIGURE B.13 **Histogram of 1,000 uniform random values.**

Let *Y1* be the square root of the *U1* values. The histogram of these values is shown in Figure B.14. It looks like a triangle, doesn't it? Just like the density $f(y) = 2y, 0 < y < 1$.

As a second example, let us consider a slightly more exotic distribution. The **extreme value distribution** is the foundation of logit choice models that are discussed in Chapter 16. It has probability density function $f(v) = \exp(-v)\exp(-\exp(-v))$, depicted in Figure B.15. The extreme value *cdf* is $F(v) = \exp(-\exp(-v))$. Despite its complicated-looking form, we can obtain values from this distribution using $v = F^{-1}(u) = -\ln(-\ln(u))$. Using the 1,000 values *U1* in data file *uniform1*, we obtain the histogram of values from the extreme value distribution shown in Figure B.16.[4] The solid curve superimposed on the histogram looks much like the extreme value density function in Figure B.15.

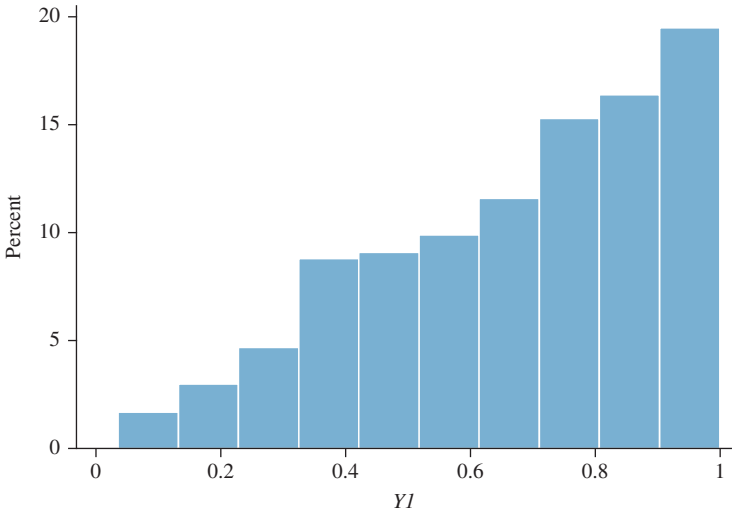


FIGURE B.14 **Histogram of 1,000 square roots of uniform random values.**

[4]The solid curve is a kernel density fitted to the data using a Gaussian kernel. See Appendix C.10 for more on kernel densities.

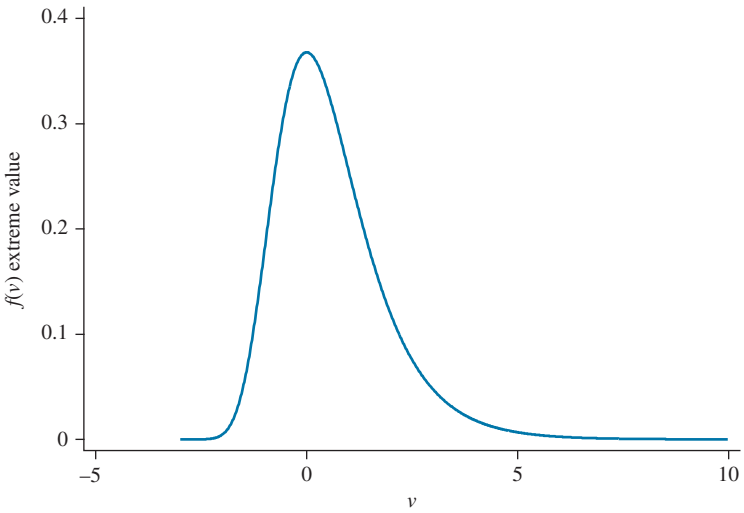


FIGURE B.15 **The extreme value distribution.**

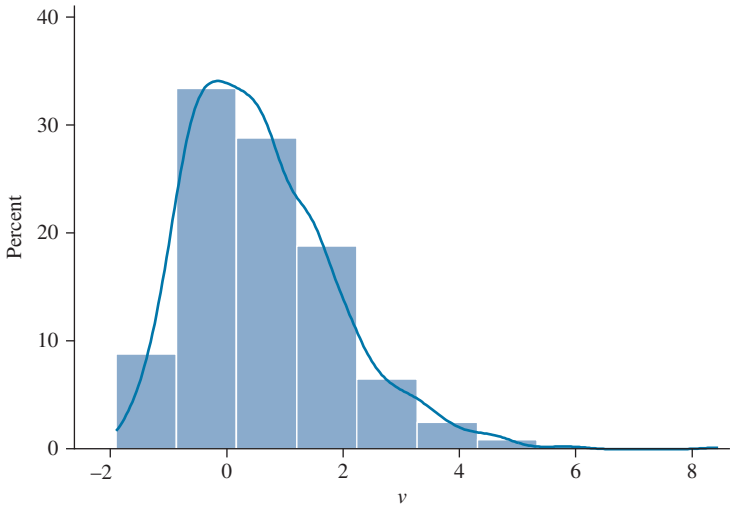


FIGURE B.16 **Histogram of simulated draws from the extreme value distribution.**

To summarize, the **inversion method** for generating random numbers from specific distributions depends upon (1) the ability to obtain uniform random numbers and (2) the distribution having a *cdf* that is invertible. The procedure does not work for joint distributions.

Knowing the inversion method, you can generate random variables from other distributions given a uniform random number generator. Books on statistical distributions[5] have instructions on how to transform uniform random numbers into a wide variety of distributions. A particular method for generating normal random numbers is illustrated in Exercise B.8.

[5]See, for example, Catherine Forbes, Merran Evans, Nicholas Hastings, Brian Peacock (2010) *Statistical Distributions*, 4th ed., John Wiley and Sons, Inc.

B.4.1 Uniform Random Numbers

The inversion method depends upon the ability to obtain random numbers from a uniform distribution. The generation of "random numbers" when used without modifiers usually means uniform random numbers, which is a field of study in and of itself. As noted earlier, the notion of computer-generated random numbers is illogical. Computers use algorithms to do their work; an algorithm is a formula so that the product is not "random," but randomlike. Computers generate **pseudo-random numbers**. Enter that term into your favorite search engine and you will find many, many links.

One bit of notation that appears in citations is for the mathematical **modulus**, denoted "a mod b." This is shorthand for the remainder resulting from dividing a by b. One method for calculating the modulus is[6]

$$n \bmod m = n - m\,\text{ceil}(n/m) + m \tag{B.51}$$

where **ceil** is short for the **ceiling** function that rounds up[7] to the next integer. To see how this works:

$$7 \bmod 3 = 1 = 7 - 3\,\text{ceil}(7/3) + 3 = 7 - 3\,\text{ceil}(2.3333) + 3 = 7 - 3 \cdot 3 + 3 = 1$$

A standard method for creating a uniform random number is the **linear congruential generator**.[8] Consider the recursive relationship

$$X_n = \left(aX_{n-1} + c\right) \bmod m \tag{B.52}$$

where a, c, and m are constants that we choose. It means that X_n takes the value equal to the remainder obtained by dividing $aX_{n-1} + c$ by m. It is a recursive relationship because the nth value depends on the $(n-1)$th. That means we must choose a starting value X_0, which is called the **random number seed**. Everyone using the same seed, and values a, c, and m will generate the same string of numbers. The value m is the divisor in (B.52), and it determines the maximum period of the recursively generated values. The uniform random values falling in the interval (0, 1) are obtained as $U_n = X_n/m$. The value of m is often chosen to be 2^{32} when using computers with 32-bit architecture. The values of a and c are critical to the success of the random number generator. Bad choices result in sequences of numbers that are not random. For example, type RANDU into your search engine. This was a popular random number generator in the 1960s (I used it too!) that was later discovered to be very flawed, failing tests of randomness.[9]

EXAMPLE B.16 | Linear Congruential Generator Example

To illustrate that the process defined in (B.52) can generate apparently random numbers, we choose $X_0 = 1234567$, $a = 1664525$, $b = 1013904223$, and $m = 2^{32}$ and create 10,000 data values, labeled *U1* in the data file *uniform3*.[10] Using a histogram with 20 bins, we would expect 5% of the values in each, and as Figure B.17 illustrates, that is about what we get.

The 10,000 values for *U1* have sample mean 0.4987197 and variance 0.0820758 compared to the true mean and variance for a uniform distribution of 0.5 and 0.08333. The minimum and maximum values are 0.0000327 and 0.9998433, respectively.

[6] www.functions.wolfram.com/IntegerFunctions/Mod/27/01/03/01/0001/.

[7] ceil(x) is the smallest integer not less than x.

[8] A description and link to sources is www.en.wikipedia.org/wiki/Linear_congruential_generator.

[9] George Marsaglia developed a series of tests for randomness that are widely used. They are available at www.stat.fsu.edu/pub/diehard/.

[10] The variable *U2* in this file uses seed 987654321.

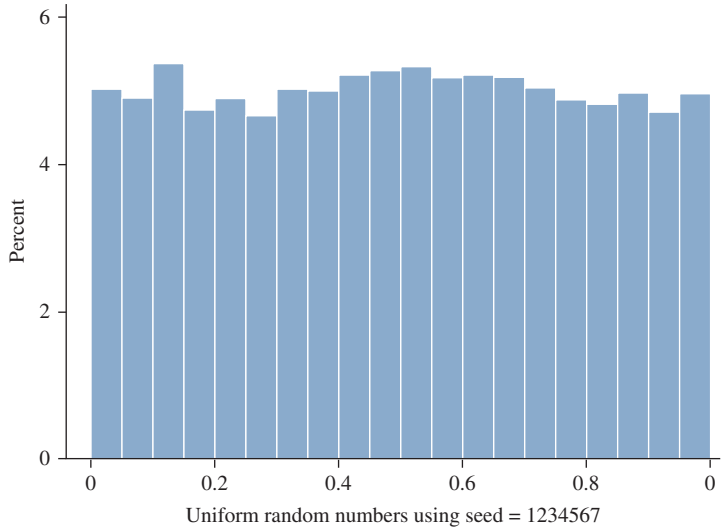


FIGURE B.17 **Histogram of 10,000 generated uniform random values.**

The lessons learned from these exercises are that random numbers are not random, and some random number generators are better than others. Ones that are popularly cited are the Mersenne twister and the KISS+Monster algorithm. New ones continue to be developed, and each software provider uses different algorithms which are predominately kept secret, or difficult to discover at any rate.

The third lesson is that you should probably **not** attempt to write your own random number algorithms. Professor Ken Train, an econometrician who has studied computational methods a great deal, says,[11] "From a practical perspective, my advice is the following: unless one is willing to spend considerable time investigating and resolving (literally, re-solving)…" the issues related to designing pseudo-random number routines "… it is probably better to use available routines rather than write a new one." Our advice is to use your software to generate random numbers, but when documenting your work, cite the software used and the software version, as revisions can change results from one version to another.

B.5 Exercises

B.1 Let $X_1, X_2, \ldots, X_n$ be independent random variables which all have the same probability distribution, with mean μ and variance σ^2. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

a. Use the properties of expected values to show that $E\left(\overline{X}\right) = \mu$.

b. Use the properties of variance to show that $\text{var}\left(\overline{X}\right) = \sigma^2/n$. How have you used the assumption of independence?

[11] *Discrete Choice Methods with Simulation*, 2nd ed., 2009, Cambridge University Press, p. 206.

B.2 Suppose that Y_1, Y_2, Y_3 is a sample of observations from a $N(\mu, \sigma^2)$ population but that Y_1, Y_2, and Y_3 are *not* independent. In fact, suppose that

$$\text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_3) = \text{cov}(Y_1, Y_3) = \frac{\sigma^2}{2}$$

Let $\overline{Y} = (Y_1 + Y_2 + Y_3)/3$.

a. Find $E(\overline{Y})$.

b. Find $\text{var}(\overline{Y})$.

B.3 Let X be a continuous random variable with probability density function given by

$$f(x) = -\frac{1}{2}x + 1, \quad 0 \le x \le 2$$

a. Graph the density function $f(x)$.

b. Find the total area beneath $f(x)$ for $0 \le x \le 2$.

c. Find $P(X \ge 1)$ using both geometry and integration.

d. Find $P(X \le \frac{1}{2})$.

e. Find $P(X = 1\frac{1}{2})$.

f. Find the expected value and variance of X.

g. Find the cumulative distribution function of X.

B.4 Let X be a uniform random variable on the interval (a, b).

a. Use integration techniques to find the mean and variance of X.

b. Find the cumulative distribution function of X.

B.5 Use the recursive relationship in (B.52) with $X_0 = 79, m = 100, a = 263$, and $c = 71$ to generate 40 values $X_1, X_2, \ldots, X_{40}$. Do the resulting numbers appear random? Is this a good random number generator, or not?

B.6 Let X have a normal distribution with mean μ and variance σ^2. Use the change-of-variable technique to find the probability density function of $Y = aX + b$.

B.7 Show that if $E(Y|X) = E(Y)$, then $\text{cov}[Y, g(X)] = 0$ for any function $g(X)$.

B.8 Normal random numbers are useful for Monte Carlo simulations. One way to generate them is using the Box–Muller transformation. The Box–Muller transformation creates two new random variables, *Z1* and *Z2*, that have independent $N(0, 1)$ distributions, using

$$Z1 = \sqrt{-2\ln(U1)}\cos(2\pi U2), \quad Z2 = \sqrt{-2\ln(U1)}\sin(2\pi U2)$$

a. Construct a histogram of *Z1* and *Z2* obtained by using the 1,000 uniform random values *U1* and *U2* in data file *uniform1* (or the 10,000 values in the data file *uniform2*). Is the distribution of values "bell shaped"?

b. Calculate the summary statistics for *Z1* and *Z2*. Are the sample mean and variance close to zero and one, respectively?

c. Construct a scatter diagram for *Z1* and *Z2*. That is, plot *Z1* (vertical axis) and *Z2* (horizontal axis) in the *x*–*y* plane. Is there any evidence of positive or negative correlation?

B.9 Let X be a continuous random variable with *pdf* $f(x) = 3x^2/8$ for $0 < x < 2$. Compute

a. $P(0 < X < \frac{1}{2})$

b. $P(1 < X < 2)$

B.10 A continuous random variable X is said to have an exponential distribution if its *pdf* is $f(x) = e^{-x}$, $x \ge 0$.

a. Plot this density function for $0 \le x \le 10$.

b. The cumulative distribution function for X is $F(x) = 1 - e^{-x}$. Plot this function over the interval $0 \le x \le 10$. Is it strictly increasing or decreasing, or are you unsure?

c. Use the inverse transformation method to draw random values *X1* from this distribution. Use the 1,000 values for *U1* in data file *uniform1* or the 10,000 values for *U1* in data file *uniform2*. Construct a histogram of the values you have created. Does it resemble the plot in (a)?

d. The true mean and variance of X are $\mu = 1$ and $\sigma^2 = 1$. How close are the sample mean and the sample variance to the true values?

B.11 Use the recursive relationship in (B.52) with $X_0 = 1234567$, $m = 2^{32}$, $a = 1103515245$, and $c = 12345$ to generate 1,000 random values called *U1*. Do the resulting numbers appear random? Is this a good random number generator, or not? Choose another seed value and generate another 1,000 values called *U2*. Find the summary statistics and sample correlation for *U1* and *U2*. Do the values behave as you expect them to, or not?

B.12 Suppose that the joint *pdf* of the continuous random variables X and Y is $f(x, y) = 6x^2y$ for $0 \le x \le 1, 0 \le y \le 1$.

a. Does this function satisfy the conditions for a valid *pdf*?
b. Find the marginal *pdf* of X, as well as its mean and variance.
c. Find the marginal *pdf* of Y.
d. Find the conditional *pdf* of X given $Y = \frac{1}{2}$.
e. Find the conditional mean and variance of X given $Y = \frac{1}{2}$.
f. Are X and Y independent? Explain.

B.13 Suppose that X and Y are continuous random variables with joint *pdf* $f(x, y) = \frac{1}{2}$ for $0 \le x \le y \le 2$ and $f(x, y) = 0$ otherwise. Note that the values of X are less than or equal to the values of Y.

a. Verify that the volume under the joint *pdf* is 1.
b. Find the marginal *pdf*s of X and Y.
c. Find $P\left(X < \frac{1}{2}\right)$.
d. Find the *cdf* of Y.
e. Find the conditional probability $P\left(X < \frac{1}{2} \middle| Y = 1.5\right)$. Are X and Y independent?
f. Find the expected value and variance of Y.
g. Use the law of iterated expectations to find $E(X)$.

B.14 Let X and Y be two discrete random variables. X can take the values 1, 2, 3, or 4. Y can take the values 1, 2, 3. Their joint *pdf* is

		X			
		1	**2**	**3**	**4**
Y	1	0.01	0.07	0.09	0.03
	2	0.20	0	0.05	0.25
	3	0.09	0.03	0.06	0.12

a. Find the marginal distributions, the *pdf*s of X and Y.
b. Are these two random variables statistically independent? If not, give an example that disproves independence.
c. Find the conditional *pdf* of X given that $Y = 2$, $f(x|Y = 2)$, for $x = 1, 2, 3$, and 4.
d. Find the expected value of X.
e. Find the expected value of X given that $Y = 2$.

B.15 This exercise uses the random variables X and Y, and their joint *pdf*, from Exercise B.14.

a. Find the variance of X.
b. Find the variance of X given that $Y = 2$, and the variance of X given that $Y = 3$. Are they equal?
c. Find the conditional expectations $E(X|Y = 1)$, $E(X|Y = 2)$, and $E(X|Y = 3)$. Using these values show that $E(X) = \sum_{i=1}^{3} E(X|Y = i)\,P(Y = i)$.
d. Find $E(XY)$.
e. Find $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$.
f. Find the correlation between X and Y.

B.16 Suppose that the two continuous random variables X and Y have joint *pdf* $f(x, y) = \frac{21}{4}x^2y$, if $x^2 \le y \le 1$.

a. Show that the marginal *pdf* of X is $f(x) = \int_{x^2}^{1} \frac{21}{4}x^2y\,dy = \frac{21}{8}x^2\left(1 - x^4\right)$ if $-1 \le x \le 1$.

b. Show that the conditional *pdf* of Y given that $X = \frac{1}{2}$ is $f(y|X = 1/2) = \frac{32}{15}y$ for $0.25 \le y \le 1$.

c. Show that the conditional *pdf* of Y given $X = x$ is $f(y|X = x) = \dfrac{2y}{1 - x^4}$ if $x^2 \le y \le 1$.

d. Using the result in part (c), show that $E(Y|X = x) = \left(\frac{2}{3}\right)\left(\dfrac{1 - x^6}{1 - x^4}\right)$.

e. It can be shown that the *pdf* of Y is $f(y) = \frac{7}{2}y^{5/2}$ if $0 \le y \le 1$. (i) Verify that this is a legitimate *pdf* and (ii) using this result show that $E(Y) = \frac{7}{9}$.

f. Using the results in (d) and (a), use the law of iterated expectations to show that $E(Y) = E_X[E(Y|X = x)] = \frac{7}{9}$.

B.17 Consider the random variable X which is the number of heads occurring in two flips of a fair coin.

a. What values can X take? What are the probabilities of each outcome? What is the probability that $X \le 1.5$?

b. Write down the values of the cumulative distribution function of X. What is the probability that $X \le 1$? What is the probability that $X \le 1.5$?

c. Suppose a wager is proposed in which you will receive winnings of $W = 2X$ dollars. What is the probability distribution of W?

d. What are your expected winnings? Show your work.

e. What is the conditional probability density function of X given that the first flip is a head?

f. What is the conditional expectation of $W = 2X$ given that the first flip is a head?

B.18 Suppose X is a continuous random variable that can take any value between zero and three, $0 < x < 3$. The *pdf* is $f(x) = cx^2$.

a. Find the value of c that makes this a legitimate *pdf*.

b. Using the result in (a) find $P(0 < X < 2)$. Show your work.

c. Find the mathematical equation for the *cdf* $F(x)$. Draw a sketch of the *cdf* for $-\infty < x < \infty$.

d. Use the *cdf* in (c) to compute $P(0.5 < X < 1)$.

e. Find the probability $P(0.5 < X < 1)$ given that $X < 2$.

B.19 The *cdf* of the continuous random variable X is $F(x) = 1 - e^{-2x}$ for $x \ge 0$ and $F(x) = 0$ otherwise.

a. Draw a sketch of the *cdf*.

b. Use the *cdf* to find the probability $P(1 < X < 2)$.

c. Find the *pdf* of X. Sketch the *pdf*.

d. Sketch on the *pdf* the area representing $P(1 < X < 2)$.

B.20 Two discrete random variables X and Y have the joint *pdf* $f(x, y) = c(2x + y)$. The random variable X takes the values $x = 0,\ 1,\ 2$ and the random variable Y takes the values $y = 0,\ 1,\ 2,\ 3$.

a. Find the value c that makes the probabilities sum to 1.

b. Find $P(X \ge 1, Y \le 1)$.

c. Find the marginal *pdf*s of X and Y.

d. Find the probability $P(X \ge 1, Y \le 1)$ given that $Y \le 2$.

e. Find the expected value of X.

f. Find the expected value of X given that $Y \le 2$.

g. Are X and Y statistically independent? Explain.

B.21 This exercise uses the joint *pdf* in Exercise B.14.

a. Find the variance of Y.

b. Find $E(Y|X = 1)$, $E(Y|X = 2)$, $E(Y|X = 3)$, and $E(Y|X = 4)$.

c. Calculate $\sum_{x=1}^{4}[E(Y|X = x) - E(Y)]^2 f(x)$. Which term in equation (B.27) does this represent?

d. Find $\text{var}(Y|X = 1)$, $\text{var}(Y|X = 2)$, $\text{var}(Y|X = 3)$, $\text{var}(Y|X = 4)$.

e. Calculate $\sum_{x=1}^{4}[\text{var}(Y|X = x)]\, f(x)$. Which term in equation (B.27) does this represent?

f. Use the results in parts (c) and (e) to compute $\text{var}(Y)$.

B.22 An econometrics instructor randomly chooses $n = 5$ students and gives each a problem to solve. Let the random variables $X_i = 1$ if the ith student answers correctly and $X_i = 0$ if the student does not answer the question correctly. Suppose that the probability that each student answers correctly is 0.80. Let $X = \sum_{i=1}^{5} X_i$ be the number of students who answer correctly.

a. Use the binomial distribution (B.43) to compute $P(X = 3|n = 5,\ p = 0.80)$.

b. From the first group of five students selected, four answered correctly. From a second randomly selected group of five students, two answered correctly. How does this illustrate the concept of sampling variation?

c. The instructor repeats the experiment of randomly selecting five students many times, recording the value X in each experiment. What will the average number of students answering correctly converge toward, as the number of experiments becomes very large?

d. Draw a sketch of the *pdf* of this random variable, locating $E(X)$ on the graph.

e. Find var(X). How does this value relate to the concept of sampling variation?

B.23 Suppose that for the population of married U.S. women, the average number of extramarital affairs, X, is $\mu = 2$.

a. Use the Poisson density function in (B.44) to find the probability that a randomly chosen married women will have $X = 2$ affairs.

b. Find the probability that a randomly chosen married woman will have two or more extramarital affairs. [*Hint:* First compute $P(X = 0)$ and $P(X = 1)$.]

c. Instead of sampling the entire population of married U.S. women, suppose that we sample the population of women who are known to have had at least one extramarital affair. Find the probability that a randomly chosen married woman will have two or more extramarital affairs given that she will have had at least one. That is, find $P(X \geq 2 | X \geq 1)$.

B.24 **Chebyshev's inequality** is a remarkable statistical result. Suppose X is a discrete or continuous random variable with mean μ and variance σ^2. Let ε be any positive number, then $P(|X - \mu| \geq \varepsilon) \leq \sigma^2/\varepsilon^2$.

a. Let X be a normal random variable with mean $\mu = 1$ and variance $\sigma^2 = 1$. Draw a sketch of the *pdf* of X.

b. Let $\varepsilon = 1$. On the sketch in (a) show $P(|X - 1| \geq 1)$.

c. Using the normal probabilities in Statistical Table 1, or your computer software, compute $P(|X - 1| \geq 1)$. Does the calculated value agree with Chebyshev's inequality?

B.25 Chebyshev's inequality is given in Exercise B.24.

a. If we let $\varepsilon = k\sigma$ what does the inequality become?

b. Let X be a normal random variable with mean $\mu = 1$ and variance $\sigma^2 = 1$. Find the exact probability $P(|X - 1| \geq 2\sigma)$. Does the value you calculate agree with the version of Chebyshev's inequality derived in part (a)?

c. Let U be a uniform random variable, see Section B.3.4, on the interval [0, 1]. Find the exact probability $P(|U - 0.5| > 2\sigma)$. Does this result agree with the revised version of Chebyshev's inequality derived in part (a)?

d. Let Y be a binomial random variable based on $n = 10$ trials each with probability $p = 0.8$. For this binomial distribution, what are the mean μ and standard deviation σ? Using your computer software, compute $P(|Y - \mu| > 2\sigma)$. Does your computed value agree with the revised version of Chebyshev's inequality derived in part (a)?

B.26 Suppose that X is a random variable, and $g(X)$ is a *convex* function of X. Then **Jensen's inequality**, as used in probability theory, says $g[E(X)] \leq E\big[g(X)\big]$. A convex function "curves up" without any inflection points. If a function $g(X)$ has second derivative that is positive over an interval, then it is convex over the interval.

a. Consider the function $g(X) = X^2$ over the interval $X > 0$. Find the second derivative of this function. Is $g(X)$ convex for $X > 0$? Draw a simple sketch of the function.

b. Suppose X is a discrete random variable taking the values $x = 1, 2, 3, 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Find $E(X)$ and $E(X^2)$. Is $[E(X)]^2 \leq E\big(X^2\big)$?

c. The variance of the random variable X is $E\big\{[X - E(X)]^2\big\} = E\big(X^2\big) - [E(X)]^2$. Using Jensen's inequality what can we say about the variance of a random variable?

B.27 Suppose that X is a random variable, and $g(X)$ is a *concave* function of X. Then Jensen's inequality, as used in probability theory, says $g[E(X)] \geq E\big[g(X)\big]$. A concave function has a continuously diminishing slope. If a function $g(X)$ has second derivative that is negative over an interval, then it is concave over the interval.

a. Consider the function $g(X) = \ln(X)$ over the interval $X > 0$. Find the second derivative of this function. Is $g(X)$ concave for $X > 0$? Draw a simple sketch of the function.

b. Suppose X is a discrete random variable taking the values $x = 1, 2, 3, 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Find $E(X)$ and $E[\ln(X)]$. Is $\ln[E(X)] \geq E[\ln(X)]$?

c. Jensen's inequality is also true for sample averages. Suppose $x_1, x_2, \ldots, x_n$ are numbers and $g(x)$ is a concave function. Then $g\left(\sum_{i=1}^{n} x_i/n\right) \geq \sum_{i=1}^{n} g(x_i)/n$. Suppose $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, and $x_4 = 4$. Show that $\ln\left(\sum_{i=1}^{4} x_i/4\right) \geq \sum_{i=1}^{4} \ln(x_i)/4$.

B.28 Let X and Y be random variables. The **Cauchy–Schwarz inequality**, as used in probability theory, is $[E(XY)]^2 \leq E(X^2)\,E(Y^2)$.

a. Using the joint probabilities in Table P.3, in the Probability Primer, Section P.3.2, verify that $[E(XY)]^2 \leq E(X^2)\,E(Y^2)$ holds.

b. Replace the random variables X and Y by $X - E(X) = X - \mu_X$ and $Y - E(Y) = Y - \mu_Y$. Show that the Cauchy–Schwarz inequality implies $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$.

c. Using the joint probabilities in Table P.3, in the Probability Primer, Section P.3.2, verify that $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$.

d. Use the fact that $[\text{cov}(X, Y)]^2 \leq \text{var}(X)\,\text{var}(Y)$ to prove that the correlation ρ_{XY} must fall in the interval $[-1, 1]$.

e. Show that $[\text{cov}(X, Y)]^2 = \text{var}(X)\,\text{var}(Y)$ if $Y = a + bX$, where a and b are constants.

B.29 Let X be a random variable and consider a function $g(X) \geq 0$ for every value of X. Assume $E[g(X)]$ exists. Then **Markov's inequality** is $P(g(X) \geq c) \leq c^{-1}E[g(X)]$.

a. Suppose X is a discrete random variable taking the values $x = 1,\ 2,\ 3,\ 4$ with probabilities 0.1, 0.2, 0.3, and 0.4, respectively. Let $g(X) = X^2$. Find $P[X^2 \geq 5]$. Find $E(X^2)$. Is $P[X^2 \geq 5] \leq E(X^2)/5$?

b. Let $g(X) = (X - \mu_X)^2$, where $\mu_X = E(X)$. Let $c = k^2\sigma_X^2$. Show that Markov's inequality leads to Chebyshev's inequality. [Author's note: Many mathematical inequalities are used in probability and statistics. A good list is in Dale J. Poirier (1995) *Intermediate Statistics and Econometrics: A Comparative Approach*, MIT Press, Chapter 2.8. There Poirier (page 76) also relates a conversation between Nobel Prize winning economist Lawrence Klein and statistician Harold Freeman. *Lawrence Klein* "If the Devil promised you a theorem in return for your immortal soul, would you accept the bargain?" *Harold Freeman* "No. But I would for an inequality."]

B.30 Suppose X is a uniformly distributed variable on the (0, 1) interval. That is, $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. Further, suppose that the conditional *pdf* of Y given $X = x$ is $f(y|x) = 1/x$ for $0 < y < x$ and $f(y|x) = 0$ otherwise. [Adapted from Takeshi Amemiya (1994) *Introduction to Statistics and Econometrics*, Harvard University Press.]

a. Use the law of iterated expectations to show that $E(Y) = E_X E(Y|X) = 1/4$.

b. Show that $f(x, y) = 1/x$ for $0 < x < 1$ and $0 < y < x$, but $f(x, y) = 0$ otherwise. Then show $f(y) = \ln(y)$ for $0 < y < 1$. Then find $E(Y) = \displaystyle\int_0^1 y f(y)\,dy$.

B.31 Suppose X is a uniformly distributed variable on the (0, 1) interval. That is, $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. Suppose the random variable Y takes the values 1 and 0, and the conditional probabilities of these values are $P(Y = 1|X = x) = x$ and $P(Y = 0|X = x) = 1 - x$. [Adapted from Takeshi Amemiya (1994) *Introduction to Statistics and Econometrics*, Harvard University Press.]

a. Use the law of iterated expectations to show $E(Y) = 1/2$.

b. Use the variance decomposition to show that $\text{var}(Y) = 1/4$.

Appendix C

Review of Statistical Inference

LEARNING OBJECTIVES

Based on the material in this appendix, you should be able to

1. Discuss the difference between a population and a sample, and why we use samples of data as a basis for inference about population parameters.
2. Connect the concepts of a population and a random variable, indicating how the probability density function of a random variable, and the expected value and variance of the random variable, inform us about the population.
3. Explain the difference between the population mean and the sample mean.
4. Explain the difference between an estimate and an estimator, and why the latter is a random variable.
5. Explain the terms sampling variation and sampling distribution.
6. Explain the concept of unbiasedness, and use the rules of expected values to show that the sample mean is unbiased.
7. Explain why we prefer unbiased estimators with smaller variances to those with larger variances.
8. Describe the central limit theorem, and its implications for statistical inference.
9. Explain the relation between the population ''standard deviation'' and the standard error of the sample mean.
10. Explain the difference between point and interval estimation, and construct and interpret interval estimates of a population mean given a sample of data.
11. Give, in simple terms, a clarification of what the phrase ''95% level of confidence'' does and does not mean in relation to interval estimation.
12. Explain the purpose of hypothesis testing, and list the elements that must be present when carrying out a test.
13. Discuss the implications of the possible alternative hypotheses when testing the null hypothesis $H_0 : \mu = 7$. Give an economic example in which this hypothesis might be tested against one of the alternatives.
14. Describe the level of significance of a test, and explain the difference between the level of significance and the *p*-value of a test.
15. Define Type I error and its relationship to the level of significance of a test.
16. Explain the difference between one-tail tests and two-tail tests, describing when one is preferred to the other.
17. Explain the difference and implications between the statements ''I accept the null hypothesis'' and ''I do not reject the null hypothesis.''
18. Give an intuitive explanation of maximum likelihood estimation, and describe the properties of the maximum likelihood estimator.
19. List the three types of tests associated with maximum likelihood estimation and comment on their similarities and differences.
20. Distinguish between parametric and nonparametric estimation.
21. Understand how a kernel density estimator fits an empirical distribution.

KEYWORDS

alternative hypothesis
asymptotic distribution
BLUE
central limit theorem
central moments
estimate
estimator
experimental design
information measure
interval estimate
kernel density estimator
Lagrange multiplier test
law of large numbers
level of significance
likelihood function
likelihood ratio test
linear estimator
log-likelihood function
maximum likelihood estimation
nonparametric
null hypothesis
parametric
point estimate
population parameter
p-value
random sample
rejection region
sample mean
sample variance
sampling distribution
sampling variation
standard error
standard error of the estimate
standard error of the mean
statistical inference
test statistic
two-tail tests
Type I error
Type II error
unbiased estimators
Wald test

Economists are interested in relationships between economic variables. For example, how much can we expect the sales of Frozen Delight ice cream to rise if we reduce the price by 5%? How much will household food expenditure rise if household income rises by $100 per month? Questions such as these are the main focus of this book.

However, sometimes questions of interest focus on a single economic variable. For example, an airplane seat designer must consider the average hip size of passengers in order to allow adequate room for each person, while still designing the plane to carry the profit-maximizing number of passengers. What is the average hip size, or more precisely hip width, of U.S. flight passengers? If a seat 18 inches wide is planned, what percent of customers will not be able to fit? Questions like this must be faced by manufacturers of everything from golf carts to women's jeans. How can we answer these questions? We certainly cannot take the measurements of every man, woman, and child in the U.S. population. This is a situation when statistical inference is used. Infer means "to conclude by reasoning from something known or assumed." **Statistical inference** means that we will draw conclusions about a population based on a sample of data.

C.1 A Sample of Data

To carry out statistical inference, we need data. The data should be obtained from the population in which we are interested. For the airplane seat designer this is essentially the entire U.S. population above the age of two, since small children can fly "free" on the laps of their suffering parents. A separate branch of statistics, called **experimental design**, is concerned with the question of how to actually collect a representative sample. How would you proceed if you were asked to obtain 50 measurements of hip size representative of the entire population? This is not such an easy task. Ideally the 50 individuals will be randomly chosen from the population, in such a way that there is no pattern of choices. Suppose we focus on only the population of adult flyers, since usually there are few children on planes. Our experimental design specialist draws a sample that is shown in Table C.1 and stored in the data file *hip*.

TABLE C.1 **Sample Hip Size Data**

14.96	14.76	15.97	15.71	17.77
17.34	17.89	17.19	13.53	17.81
16.40	18.36	16.87	17.89	16.90
19.33	17.59	15.26	17.31	19.26
17.69	16.64	13.90	13.71	16.03
17.50	20.23	16.40	17.92	15.86
15.84	16.98	20.40	14.91	16.56
18.69	16.23	15.94	20.00	16.71
18.63	14.21	19.08	19.22	20.23
18.55	20.33	19.40	16.48	15.54

EXAMPLE C.1 | Histogram of Hip Width Data

A first step when analyzing a sample of data is to examine it visually. Figure C.1 is a histogram of the 50 data points. Based on this figure, the "average" hip size in this sample seems to be between 16 and 18 inches. For our profit-maximizing designer this casual estimate is not sufficiently precise. In the next section we set up an econometric model that will be used as a basis for inference in this problem.

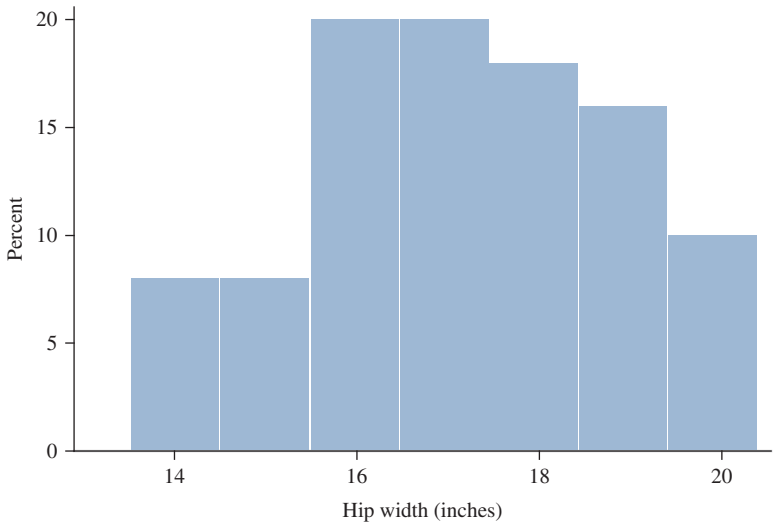


FIGURE C.1 **Histogram of hip sizes.**

C.2 An Econometric Model

The data in Table C.1 were obtained by sampling. Sampling from a population is an experiment. The variable of interest in this experiment is an individual's hip size. Before the experiment is performed we do not know what the values will be, thus the hip size of a randomly chosen person is a random variable. Let us denote this random variable as Y. We choose a sample of $N = 50$ individuals, $Y_1, Y_2, \ldots, Y_N$, where each Y_i represents the hip size of a different person. The data values in

Table C.1 are specific values of the variables, which we denote as $y_1, y_2, \ldots, y_N$. We assume that the population has a center, which we describe by the expected value of the random variable Y,

$$E(Y) = \mu \tag{C.1}$$

We use the Greek letter μ ("mu") to denote the mean of the random variable Y, and also the mean of the population we are studying. Thus if we knew μ we would have the answer to the question "What is the average hip size of adults in the United States?" To indicate its importance to us in describing the population we call μ a **population parameter**, or, more briefly, a parameter. Our objective is to use the sample of data in Table C.1 to make inferences, or judgments, about the unknown population parameter μ.

The other random variable characteristic of interest is its variability, which we measure by its variance,

$$\text{var}(Y) = E\left[Y - E(Y)\right]^2 = E\left[Y - \mu\right]^2 = \sigma^2 \tag{C.2}$$

The variance σ^2 is also an unknown population parameter. As described in the Probability Primer, the variance of a random variable measures the "spread" of a probability distribution about the population mean, with a larger variance meaning a wider spread, as shown in Figure P.3. In the context of the hip data, the variance tells us how much hip sizes can vary from one randomly chosen person to the next. To economize on space, we will denote the mean and variance of a random variable as $Y \sim (\mu, \sigma^2)$ where $\sim$ means "is distributed as." The first element in parentheses is the population mean and the second is the population variance. So far we have not said what kind of probability distribution we think Y has.

The econometric model is not complete. If our sample is drawn randomly, we can assume that $Y_1, Y_2, \ldots, Y_N$ are statistically independent. The hip size of any one individual is independent of the hip size of another randomly drawn individual. Furthermore, we assume that each of the observations we collect is from the population of interest, so each random variable Y_i, has the same mean and variance, or $Y_i \sim (\mu, \sigma^2)$. The Y_i constitute a **random sample**, in the statistical sense, because $Y_1, Y_2, \ldots, Y_N$ are statistically independent with identical probability distributions. It is sometimes reasonable to assume that population values are *normally* distributed, which we represent by $Y \sim N(\mu, \sigma^2)$.

C.3 Estimating the Mean of a Population

How shall we estimate the population mean μ given our sample of data values in Table C.1? The population mean is given by the expected value $E(Y) = \mu$. The expected value of a random variable is its average value in the population. It seems reasonable, by analogy, to use the average value in the sample, or **sample mean**, to estimate the population mean. Denote by $y_1, y_2, \ldots, y_N$ the sample of N observations. Then the sample mean is

$$\bar{y} = \sum y_i / N \tag{C.3}$$

The notation $\bar{y}$ (pronounced "y-bar") is widely used for the sample mean, and you probably encountered it in your statistics courses.

EXAMPLE C.2 | Sample Mean of Hip Width Data

For the hip data in Table C.1 we obtain $\bar{y} = 17.1582$, thus we estimate that the average hip size in the population is 17.1582 inches.

Given the estimate $\bar{y} = 17.1582$ we are inclined to ask, "How good an estimate is 17.1582?" By that we mean how close is 17.1582 to the true population mean, μ? Unfortunately this is an ill-posed question in the sense that it can never be answered. In order to answer it, we would have to know μ, in which case we would not have tried to estimate it in the first place!

Instead of asking about the quality of the *estimate* we will ask about the quality of the *estimation procedure*, or **estimator**. How good is the sample mean as an estimator of the mean of a population? This is a question we can answer. To distinguish between the estimate and the estimator of the population mean μ we will write the estimator as

$$\overline{Y} = \sum_{i=1}^{N} Y_i / N \tag{C.4}$$

In (C.4) we have used Y_i instead of y_i to indicate that this general formula is used whatever the sample values turn out to be. In this context Y_i are random variables, and thus the estimator $\overline{Y}$ is random too. We do not know the value of the estimator $\overline{Y}$ until a data sample is obtained, and different samples will lead to different values.

EXAMPLE C.3 | Sampling Variation of Sample Means of Hip Width Data

To illustrate, we collect 10 more samples of size $N = 50$ and calculate the average hip size, as shown in Table C.2. The estimates differ from sample to sample because $\overline{Y}$ is a random variable. This variation, due to collection of different random samples, is called **sampling variation**. It is an inescapable fact of statistical analysis that the estimator $\overline{Y}$—indeed, all statistical estimation procedures—are subject to sampling variability. Because of this terminology, an estimator's probability density function is called its **sampling distribution**.

TABLE C.2 Sample Means from 10 Samples

Sample	$\bar{y}$
1	17.3544
2	16.8220
3	17.4114
4	17.1654
5	16.9004
6	16.9956
7	16.8368
8	16.7534
9	17.0974
10	16.8770

We can determine how good the estimator $\overline{Y}$ is by examining its expected value, variance, and sampling distribution.

C.3.1 The Expected Value of $\overline{Y}$

Write out formula (C.4) fully as

$$\overline{Y} = \sum_{i=1}^{N} \frac{1}{N} Y_i = \frac{1}{N} Y_1 + \frac{1}{N} Y_2 + \cdots + \frac{1}{N} Y_N \tag{C.5}$$

From (P.16) the expected value of this sum is the sum of expected values

$$\begin{aligned} E\left(\overline{Y}\right) &= E\left(\frac{1}{N} Y_1\right) + E\left(\frac{1}{N} Y_2\right) + \cdots + E\left(\frac{1}{N} Y_N\right) \\ &= \frac{1}{N} E(Y_1) + \frac{1}{N} E(Y_2) + \cdots + \frac{1}{N} E(Y_N) \\ &= \frac{1}{N}\mu + \frac{1}{N}\mu + \cdots + \frac{1}{N}\mu \\ &= \mu \end{aligned}$$

The expected value of the estimator $\overline{Y}$ *is* the population mean μ that we are trying to estimate. What does this mean? The expectation of a random variable is its average value in all possible random samples from the population. If we did obtain many samples of size N, and obtained their average values, like those in Table C.2, then the average of all *those* values would equal the true population mean μ. This property is a good one for estimators to have. Estimators with this property are called **unbiased estimators**. The sample mean $\overline{Y}$ is an unbiased estimator of the population mean μ.

Unfortunately, while unbiasedness is a good property for an estimator to have, it does not tell us anything about whether our estimate $\bar{y} = 17.1582$, based on a single sample of data, is close to the true population mean value μ. To assess how far the estimate might be from μ, we will determine the variance of the estimator.

C.3.2 The Variance of $\overline{Y}$

The variance of $\overline{Y}$ is obtained using the procedure for finding the variance of a sum of uncorrelated (zero covariance) random variables in (P.23). We can apply this rule if our data are obtained by random sampling, because with random sampling the observations are statistically independent, and thus are uncorrelated. Furthermore, we have assumed that $\text{var}(Y_i) = \sigma^2$ for all observations. Carefully note how these assumptions are used in the derivation of the variance of $\overline{Y}$, which we write as $\text{var}(\overline{Y})$:

$$\begin{aligned}
\text{var}(\overline{Y}) &= \text{var}\left(\frac{1}{N}Y_1 + \frac{1}{N}Y_2 + \cdots + \frac{1}{N}Y_N\right) \\
&= \frac{1}{N^2}\text{var}(Y_1) + \frac{1}{N^2}\text{var}(Y_2) + \cdots + \frac{1}{N^2}\text{var}(Y_N) \\
&= \frac{1}{N^2}\sigma^2 + \frac{1}{N^2}\sigma^2 + \cdots + \frac{1}{N^2}\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned} \tag{C.6}$$

This result tells us that (i) the variance of $\overline{Y}$ is *smaller* than the population variance, because the sample size $N \geq 2$, and (ii) the larger the sample size, the smaller the sampling variation of $\overline{Y}$ as measured by its variance.

C.3.3 The Sampling Distribution of $\overline{Y}$

If the population data are normally distributed, then we say that the random variable Y_i follows a normal distribution. In this case the estimator $\overline{Y}$ also follows a normal distribution. In (P.36) it is noted that weighted averages of normal random variables are normal themselves. From (C.5) we know that $\overline{Y}$ is a weighted average of Y_i. If $Y_i \sim N(\mu, \sigma^2)$, then $\overline{Y}$ is also normally distributed, or $\overline{Y} \sim N(\mu, \sigma^2/N)$.

We can gain some intuition about the meaning and usefulness of the finding that $\overline{Y} \sim N(\mu, \sigma^2/N)$ if we examine Figure C.2. Each of the normal distributions in this figure is a sampling distribution of $\overline{Y}$. The differences among them are the sample sizes used in estimation. The sample size $N_3 > N_2 > N_1$. Increasing the sample size decreases the variance of the estimator $\overline{Y}$, $\text{var}(\overline{Y}) = \sigma^2/N$, and this increases the probability that the sample mean will be "close" to the true population parameter μ. When examining Figure C.2, recall that an area under a probability density function (*pdf*) measures the probability of an event. If ε represents a positive number, the probability that $\overline{Y}$ falls in the interval between $\mu - \varepsilon$ and $\mu + \varepsilon$ is greater for larger samples.

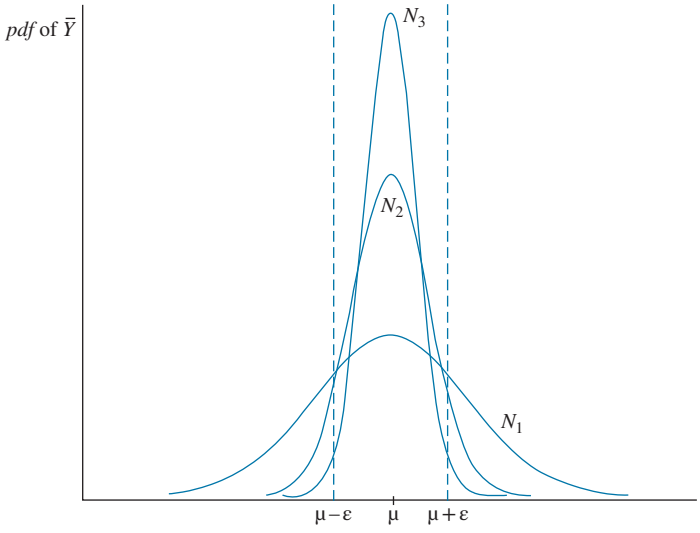


FIGURE C.2 **Increasing sample size and sampling distributions of $\bar{Y}$.**

The lesson here is that having more data is better than having less data, because having a larger sample increases the probability of obtaining an estimate "close" or "within ε" of the true population parameter μ.

EXAMPLE C.4 | The Effect of Sample Size on Sample Mean Precision

In our numerical example, suppose we want our estimate of μ to be within 1 inch of the true value. Let us compute the probability of getting an estimate within $\varepsilon = 1$ inch of μ—that is, within the interval $[\mu - 1, \mu + 1]$. For the purpose of illustration assume that the population is normal, $\sigma^2 = 10$ and $N = 40$. Then $\bar{Y} \sim N(\mu, \sigma^2/N = 10/40 = 0.25)$. We can compute the probability that $\bar{Y}$ is within 1 inch of μ by calculating $P[\mu - 1 \le \bar{Y} \le \mu + 1]$. To do so we standardize $\bar{Y}$ by subtracting its mean μ and dividing by its standard deviation $\sigma/\sqrt{N}$, and then use the standard normal distribution and Statistical Table 1:

$$P\left[\mu - 1 \le \bar{Y} \le \mu + 1\right] = P\left[\frac{-1}{\sigma/\sqrt{N}} \le \frac{\bar{Y} - \mu}{\sigma/\sqrt{N}} \le \frac{1}{\sigma/\sqrt{N}}\right]$$
$$= P\left[\frac{-1}{\sqrt{0.25}} \le Z \le \frac{1}{\sqrt{0.25}}\right]$$
$$= P[-2 \le Z \le 2] = 0.9544$$

Thus, if we draw a random sample of size $N = 40$ from a normal population with variance 10, using the sample mean as an estimator will provide an estimate within 1 inch of the true value about 95% of the time. If $N = 80$, the probability that $\bar{Y}$ is within 1 inch of μ increases to 0.995.

C.3.4 The Central Limit Theorem

We were able to carry out the above analysis because we assumed that the population we are considering, hip width of U.S. adults, has a normal distribution. This implies that $Y_i \sim N(\mu, \sigma^2)$, and $\bar{Y} \sim N(\mu, \sigma^2/N)$. A question we need to ask is "If the population is not normal, then what is the sampling distribution of the sample mean?" The **central limit theorem** provides an answer to this question.

Central Limit Theorem:

If $Y_1, \ldots, Y_N$ are independent and identically distributed random variables with mean μ and variance σ^2, and $\overline{Y} = \sum Y_i/N$, then

$$Z_N = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}}$$

has a probability distribution that converges to the standard normal $N(0, 1)$ as $N \to \infty$.

This theorem says that the sample average of N independent random variables from *any* probability distribution will have an approximate standard normal distribution after standardizing (i.e., subtracting the mean and dividing by the standard deviation), if the sample is sufficiently large. A shorthand notation is $\overline{Y} \overset{a}{\sim} N(\mu, \sigma^2/N)$, where the symbol $\overset{a}{\sim}$ means *asymptotically distributed*. The word **asymptotic** implies that the approximate normality of $\overline{Y}$ depends on having a large sample. Thus even if the population is not normal, if we have a sufficiently large sample, we can carry out calculations like those in the previous section. How large does the sample have to be? In general, it depends on the complexity of the problem, but in the simple case of estimating a population mean, if $N \geq 30$ then you can feel pretty comfortable in assuming that the sample mean is approximately normally distributed, $\overline{Y} \overset{a}{\sim} N(\mu, \sigma^2/N)$, as indicated by the central limit theorem.

EXAMPLE C.5 | Illustrating the Central Limit Theorem

To illustrate how well the central limit theorem actually works, we carry out a simulation experiment. Let the continuous random variable Y have a triangular distribution, with probability density function

$$f(y) = \begin{cases} 2y & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

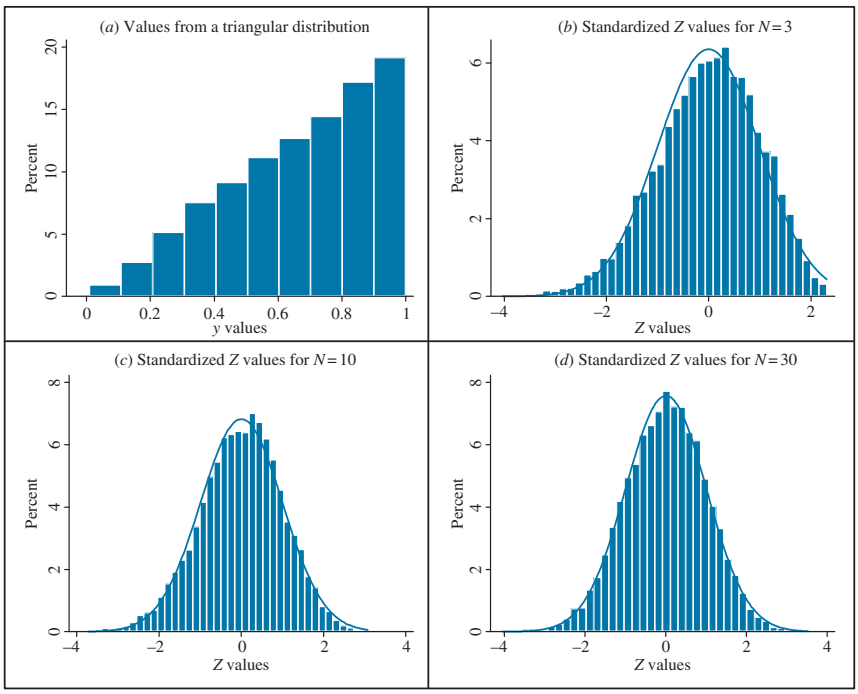


FIGURE C.3 **Central limit theorem.**

Draw a sketch of the triangular *pdf* to understand its name. The expected value of Y is $\mu = E(Y) = 2/3$, and its variance is $\sigma^2 = \text{var}(Y) = 1/18$. The central limit theorem says that if $Y_1, \ldots, Y_N$ are independent and identically distributed with density $f(y)$ then

$$Z_N = \frac{\overline{Y} - 2/3}{\sqrt{\frac{1/18}{N}}}$$

has a probability distribution that approaches the standard normal distribution as N approaches infinity.

We use a random number generator to create random values from the triangular *pdf*. Plotting 10,000 values gives the histogram in Figure C.3(a). We generate 10,000 samples of sizes $N = 3$, 10, and 30, compute the sample means of each sample, and create Z_N. Their histograms are shown in Figures C.3(b)–(d). You see the amazing convergence of the standardized sample mean's distribution to a distribution that is bell shaped, centered at zero, symmetric, with almost all values between -3 and 3, just like a standard normal distribution, with a sample size as small as $N = 10$.

C.3.5 Best Linear Unbiased Estimation

Another powerful finding about the estimator $\overline{Y}$ of the population mean is that it is the best of all possible estimators that are both *linear* and *unbiased*. A **linear estimator** is simply one that is a weighted average of Y_i's, such as $\overline{Y} = \sum a_i Y_i$, where a_i are constants. The sample mean $\overline{Y}$, given in (C.4), is a linear estimator with $a_i = 1/N$. The fact that $\overline{Y}$ is the "best" linear unbiased estimator (**BLUE**) accounts for its wide use. "Best" means that it is the linear unbiased estimator with the smallest possible variance. In the previous section we demonstrated that it is better to have an estimator with a smaller variance rather than a larger one—because it increases the chances of getting an estimate close to the true population mean μ. This important result about the estimator $\overline{Y}$ is true *if* the sample values $Y_i \sim (\mu, \sigma^2)$ are uncorrelated and identically distributed. It does not depend on the population being normally distributed. A proof of this result is in Section C.9.2.

C.4 Estimating the Population Variance and Other Moments

The sample mean $\overline{Y}$ is an estimate of the population mean μ. The population mean is often called the "first moment" since it is the expected value of Y to the first power. Higher moments are obtained by taking expected values of higher powers of the random variable, so the second moment of Y is $E(Y^2)$, the third moment is $E(Y^3)$, and so on. When the random variable has its population mean subtracted, it is said to be *centered*. Expected values of powers of centered random variables are called **central moments**, and they are often denoted as μ_r, so that the rth central moment of Y is

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

The value of the first central moment is zero since $\mu_1 = E(Y - \mu)^1 = E(Y) - \mu = 0$. It is the higher central moments of Y that are interesting:

$$\mu_2 = E[(Y - \mu)^2] = \sigma^2$$
$$\mu_3 = E[(Y - \mu)^3]$$
$$\mu_4 = E[(Y - \mu)^4]$$

You recognize that the second central moment of Y is its variance, and the third and fourth moments appear in the definitions of skewness and kurtosis introduced in Appendix B.1.2. The question we address in this section is, now that we have an excellent estimator of the mean of a population, how do we estimate these higher moments? We will first consider estimation of the population variance, and then address the problem of estimating the third and fourth moments.

C.4.1 Estimating the Population Variance

The population variance is $\text{var}(Y) = \sigma^2 = E[Y - \mu]^2$. An expected value is an "average" of sorts, so if we knew μ we could estimate the variance by using the sample analog $\tilde{\sigma}^2 = \sum(Y_i - \mu)^2/N$. We do not know μ, so replace it by its estimator $\overline{Y}$, giving

$$\tilde{\sigma}^2 = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{N}$$

This estimator is not a bad one. It has a logical appeal, and it can be shown to converge to the true value of σ^2 as the sample size $N \to \infty$, but it is biased. To make it unbiased, we divide by $N - 1$ instead of N. This correction is needed since the population mean μ has to be estimated before the variance can be estimated. This change does not matter much in samples of at least 30 observations, but it does make a difference in smaller samples. The unbiased estimator of the population variance σ^2 is

$$\hat{\sigma}^2 = \frac{\sum\left(Y_i - \overline{Y}\right)^2}{N - 1} \tag{C.7}$$

You may remember this estimator from a prior statistics course as the "sample variance." Using the **sample variance** we can estimate the variance of the estimator $\overline{Y}$ as

$$\widehat{\text{var}}\left(\overline{Y}\right) = \hat{\sigma}^2/N \tag{C.8}$$

In (C.8) note that we have put a "hat" ($\hat{}$) over this variance to indicate that it is an estimated variance. The square root of the estimated variance is called the **standard error** of $\overline{Y}$ and is also known as the **standard error of the mean** and the **standard error of the estimate**,

$$\text{se}\left(\overline{Y}\right) = \sqrt{\widehat{\text{var}}\left(\overline{Y}\right)} = \hat{\sigma}/\sqrt{N} \tag{C.9}$$

C.4.2 Estimating Higher Moments

Recall that central moments are expected values, $\mu_r = E\left[(Y - \mu)^r\right]$, and thus are averages in the population. In statistics the **law of large numbers** says that sample means converge to population averages (expected values) as the sample size $N \to \infty$. We can estimate the higher moments by finding the sample analog and replacing the population mean μ by its estimate $\overline{Y}$, so that

$$\tilde{\mu}_2 = \sum\left(Y_i - \overline{Y}\right)^2/N = \tilde{\sigma}^2$$

$$\tilde{\mu}_3 = \sum\left(Y_i - \overline{Y}\right)^3/N$$

$$\tilde{\mu}_4 = \sum\left(Y_i - \overline{Y}\right)^4/N$$

Note that in these calculations we divide by N and not by $N - 1$, since we are using the law of large numbers (i.e., large samples) as justification, and in large samples the correction has little effect. Using these sample estimates of the central moments we can obtain estimates of the skewness coefficient (S) and kurtosis coefficient (K) as

$$\widehat{skewness} = S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{kurtosis} = K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

EXAMPLE C.6 | Sample Moments of the Hip Data

The sample variance for the hip data is

$$\hat{\sigma}^2 = \frac{\sum(y_i - \bar{y})^2}{N-1} = \frac{\sum(y_i - 17.1582)^2}{49} = \frac{159.9995}{49} = 3.2653$$

This means that the estimated variance of the sample mean is

$$\widehat{\text{var}}(\bar{Y}) = \frac{\hat{\sigma}^2}{N} = \frac{3.2653}{50} = 0.0653$$

and the standard error of the mean is

$$\text{se}(\bar{Y}) = \hat{\sigma}/\sqrt{N} = 0.2556$$

The estimated skewness is $S = -0.0138$ and the estimated kurtosis is $K = 2.3315$ using

$$\tilde{\sigma} = \sqrt{\sum\left(Y_i - \bar{Y}\right)^2/N} = \sqrt{159.9995/50} = 1.7889$$

$$\tilde{\mu}_3 = \sum\left(Y_i - \bar{Y}\right)^3/N = -0.0791$$

$$\tilde{\mu}_4 = \sum\left(Y_i - \bar{Y}\right)^4/N = 23.8748$$

Thus, the hip data is slightly negatively skewed and is slightly less peaked than would be expected for a normal distribution. Nevertheless, as we will see in Section C.7.4, we cannot conclude that the hip data follow a non-normal distribution.

EXAMPLE C.7 | Using the Hip Data Estimates

How can we summarize what we have learned? Our estimates suggest that the hip size of U.S. adults is normally distributed with mean 17.158 inches and with a variance of 3.265; $Y \sim N(17.158, 3.265)$. Based on this information, if an airplane seat is 18 inches wide, what percentage of customers will not be able to fit? We can recast this question as asking what the probability is that a randomly drawn person will have hips larger than 18 inches,

$$P(Y > 18) = P\left(\frac{Y-\mu}{\sigma} > \frac{18-\mu}{\sigma}\right)$$

We can give an approximate answer to this question by replacing the unknown parameters by their estimates,

$$\widehat{P(Y > 18)} \cong P\left(\frac{Y-\bar{y}}{\hat{\sigma}} > \frac{18-17.158}{1.8070}\right) = P(Z > 0.4659) = 0.3207$$

Based on our estimates, 32% of the population would not be able to fit into a seat that is 18 inches wide.

How large would a seat have to be to fit 95% of the population? If we let y^* denote the required seat size, then

$$\widehat{P(Y \le y^*)} \cong P\left(\frac{Y-\bar{y}}{\hat{\sigma}} \le \frac{y^* - 17.1582}{1.8070}\right) = P\left(Z \le \frac{y^* - 17.1582}{1.8070}\right) = 0.95$$

Using your computer software, or the table of normal probabilities, the value of Z such that $P(Z \le z^*) = 0.95$ is $z^* = 1.645$. Then

$$\frac{y^* - 17.1582}{1.8070} = 1.645 \Rightarrow y^* = 20.1305$$

Thus, to accommodate 95% of U.S. adult passengers, we estimate that the seats should be slightly greater than 20 inches wide.

C.5 Interval Estimation

In contrast to a **point estimate** of the population mean μ, like $\bar{y} = 17.158$, a confidence interval, or **interval estimate**, is a range of values that may contain the true population mean. A confidence interval contains information not only about the location of the population mean, but also about the precision with which we estimate it.

C.5.1 Interval Estimation: σ^2 Known

Let Y be a normally distributed random variable, $Y \sim N(\mu, \sigma^2)$. Assume that we have a random sample of size N from this population, $Y_1, Y_2, \ldots, Y_N$. The estimator of the population mean

is $\overline{Y} = \sum_{i=1}^{N} Y_i/N$. Because we have assumed that Y is normally distributed, it is also true that $\overline{Y} \sim N\left(\mu, \sigma^2/N\right)$.

For the present, let us assume that the population variance σ^2 is known. This assumption is not likely to be true, but making it allows us to introduce the notion of confidence intervals with few complications. In the next section we introduce methods for the case when σ^2 is unknown. Create a standard normal random variable

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/N}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1) \tag{C.10}$$

Cumulative probabilities for the standard normal are given by its cumulative distribution function (see the Probability Primer, Section P.7)

$$P(Z \leq z) = \Phi(z)$$

These values are given in Statistical Table 1. Let z_c be a "critical value" for the standard normal distribution, such that $\alpha = 0.05$ of the probability is in the tails of the distribution, with $\alpha/2 = 0.025$ of the probability in the tail to the right of z_c and $\alpha/2 = 0.025$ of the probability in the tail to the left of $-z_c$. The critical value is the 97.5 percentile of the standard normal distribution, $z_c = 1.96$, with $\Phi(1.96) = 0.975$. It is shown in Figure C.4. Thus, $P(Z \geq 1.96) = P(Z \leq -1.96) = 0.025$ and

$$P(-1.96 \leq Z \leq 1.96) = 1 - 0.05 = 0.95 \tag{C.11}$$

Substitute (C.10) into (C.11) and rearrange to obtain

$$P\left(\overline{Y} - 1.96\sigma/\sqrt{N} \leq \mu \leq \overline{Y} + 1.96\sigma/\sqrt{N}\right) = 0.95$$

In general,

$$P\left(\overline{Y} - z_c\frac{\sigma}{\sqrt{N}} \leq \mu \leq \overline{Y} + z_c\frac{\sigma}{\sqrt{N}}\right) = 1 - \alpha \tag{C.12}$$

where z_c is the appropriate critical value for a given value of tail probability α such that $\Phi\left(z_c\right) = 1 - \alpha/2$. In (C.12) we have defined the **interval estimator**

$$\overline{Y} \pm z_c\frac{\sigma}{\sqrt{N}} \tag{C.13}$$

Our choice of the phrase interval *estimator* is a careful one. Intervals constructed using (C.13), in repeated sampling from the population, have a $100(1 - \alpha)\%$ chance of containing the population mean μ.

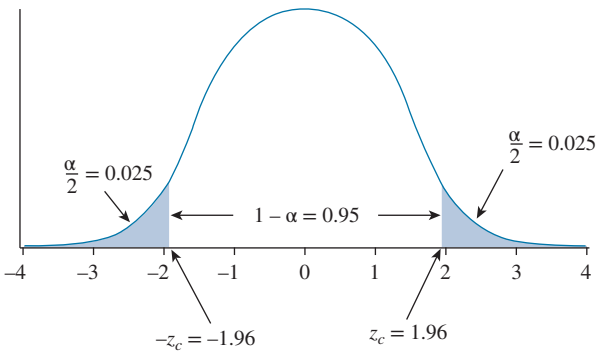


FIGURE C.4 **$\alpha = 0.05$ Critical values for the $N(0, 1)$ distribution.**

EXAMPLE C.8 | Simulating the Hip Data: Interval Estimates

In order to use the interval estimator in (C.13) we must have data from a normal population with a known variance. To illustrate the computation, and the meaning of interval estimation, we will create a sample of data using a computer simulation. Statistical software programs contain random number generators. These are routines that create values from a given probability distribution. Table C.3 (data file *table_c3*) contains 30 random values from a normal population with mean $\mu = 10$ and variance $\sigma^2 = 10$.

TABLE C.3 30 Values from $N(10, 10)$

11.939	11.407	13.809
10.706	12.157	7.443
6.644	10.829	8.855
13.187	12.368	9.461
8.433	10.052	2.439
9.210	5.036	5.527
7.961	14.799	9.921
14.921	10.478	11.814
6.223	13.859	13.403
10.123	12.355	10.819

The sample mean of these values is $\bar{y} = 10.206$ and the corresponding interval estimate for μ, obtained by applying the interval estimator in (C.13) with a 0.95 probability content, is $10.206 \pm 1.96 \times \sqrt{10/30} = [9.074, 11.338]$. To appreciate how the sampling variability of an interval estimator arises, consider Table C.4, which contains the interval estimate for the sample in Table C.3, as well as the sample means and interval estimates from another 9 samples of size 30, like that in Table C.3. The whole 10 samples are stored in the data file *table_c4*.

TABLE C.4 Confidence Interval Estimates from 10 Samples of Data

Sample	$\bar{y}$	Lower Bound	Upper Bound
1	10.206	9.074	11.338
2	9.828	8.696	10.959
3	11.194	10.063	12.326
4	8.822	7.690	9.953
5	10.434	9.303	11.566
6	8.855	7.723	9.986
7	10.511	9.380	11.643
8	9.212	8.080	10.343
9	10.464	9.333	11.596
10	10.142	9.010	11.273

Table C.4 illustrates the sampling variation of the estimator $\overline{Y}$. The sample mean varies from sample to sample. In this simulation, or Monte Carlo experiment, we know that the true population mean, $\mu = 10$, and the estimates $\overline{Y}$ are centered at that value. The half-width of the interval estimates is $1.96\sigma/\sqrt{N}$. Note that while the point estimates $\overline{Y}$ in Table C.4 fall near the true value $\mu = 10$, not all of the interval estimates contain the true value. Intervals from samples 3, 4, and 6 do not contain the true value $\mu = 10$. However, in 10,000 simulated samples the average value of $\bar{y} = 10.004$ and 94.86% of intervals constructed using (C.13) contain the true parameter value $\mu = 10$.

These numbers in Example C.8 reveal what is, and what is not, true about interval estimates.

- Any one interval estimate may or may not contain the true population parameter value.
- If *many* samples of size N are obtained, and intervals are constructed using (C.13) with $(1 - \alpha) = 0.95$, then 95% of them will contain the true parameter value.
- A 95% level of "confidence" is the probability that the interval estimator will provide an interval containing the true parameter value. Our confidence is in the procedure, not in any one interval estimate.

Since 95% of intervals constructed using (C.13) will contain the true parameter $\mu = 10$, we will be surprised if an interval estimate based on one sample does not contain the true parameter. Indeed, the fact that 3 of the 10 intervals in Table C.4 do not contain $\mu = 10$ is surprising, since out of 10 we would assume that only one 95% interval estimate might not contain the true parameter. This just goes to show that what happens in any one sample, or just a few samples, is not what sampling properties tell us. Sampling properties tell us what happens in many repeated experimental trials, or in all possible samples from a population.

C.5.2 Interval Estimation: σ^2 Unknown

The standardization in (C.10) assumes that the population variance σ^2 is known. When σ^2 is unknown, it is natural to replace it with its estimator $\hat{\sigma}^2$ given in (C.7)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2}{N-1}$$

When we do so, the resulting standardized random variable has a t-distribution (see Appendix B.3.7) with $(N-1)$ degrees of freedom,

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)} \tag{C.14}$$

The notation $t_{(N-1)}$ denotes a t-distribution with $N-1$ "degrees of freedom." Let the critical value t_c be the $100(1-\alpha/2)$ -percentile value $t_{(1-\alpha/2,N-1)}$. This critical value has the property that $P\left[t_{(N-1)} \le t_{(1-\alpha/2,N-1)}\right] = 1-\alpha/2$. Critical values for the t-distribution are contained in Statistical Table 2. If t_c is a critical value from the t-distribution, then

$$P\left(-t_c \le \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \le t_c\right) = 1-\alpha$$

Rearranging, we obtain

$$P\left(\overline{Y} - t_c\frac{\hat{\sigma}}{\sqrt{N}} \le \mu \le \overline{Y} + t_c\frac{\hat{\sigma}}{\sqrt{N}}\right) = 1-\alpha$$

The $100(1-\alpha)\%$ interval estimator for μ is

$$\overline{Y} \pm t_c\frac{\hat{\sigma}}{\sqrt{N}} \quad \text{or} \quad \overline{Y} \pm t_c \text{se}\left(\overline{Y}\right) \tag{C.15}$$

Unlike the interval estimator for the known σ^2 case in (C.13), the interval in (C.15) has center *and width* that vary from sample to sample.

Remark

The confidence interval (C.15) is based upon the assumption that the population is normally distributed, so that $\overline{Y}$ is normally distributed. If the population is not normal, then we invoke the central limit theorem, and say that $\overline{Y}$ is approximately normal in "large" samples, which from Figure C.3 you can see might be as few as 30 observations. In this case, we can use (C.15), recognizing that there is an approximation error introduced in smaller samples.

EXAMPLE C.9 | Simulating the Hip Data: Continued

Table C.5 contains estimated values of σ^2 and interval estimates using (C.15) for the same 10 samples used for Table C.4. For the sample size $N = 30$ and the 95% confidence level, the t-distribution critical value $t_c = t_{(0.975,29)} = 2.045$. The estimates $\overline{Y}$ are the same as in Table C.4. The estimates $\hat{\sigma}^2$ vary about the true value $\sigma^2 = 10$. Of these 10 intervals, those for samples 4 and 6 do not contain the true parameter $\mu = 10$. Nevertheless, in 10,000 simulated samples 94.82% of them contain the true population mean $\mu = 10$.

TABLE C.5 **Interval Estimates Using (C.15) from 10 Samples**

Sample	$\bar{y}$	$\hat{\sigma}^2$	Lower Bound	Upper Bound
1	10.206	9.199	9.073	11.338
2	9.828	6.876	8.849	10.807
3	11.194	10.330	9.994	12.394
4	8.822	9.867	7.649	9.995
5	10.434	7.985	9.379	11.489
6	8.855	6.230	7.923	9.787
7	10.511	7.333	9.500	11.523
8	9.212	14.687	7.781	10.643
9	10.464	10.414	9.259	11.669
10	10.142	17.689	8.571	11.712

EXAMPLE C.10 | Interval Estimation Using the Hip Data

We have introduced the empirical problem faced by an airplane seat design engineer. Given a random sample of size $N = 50$ we estimated the mean U.S. hip width to be $\bar{y} = 17.158$ inches. Furthermore we estimated the population variance to be $\hat{\sigma}^2 = 3.265$; thus the estimated standard deviation is $\hat{\sigma} = 1.807$. The standard error of the mean is $\hat{\sigma}/\sqrt{N} = 1.807/\sqrt{50} = 0.2556$. The critical value for interval estimation comes from a t-distribution with $N - 1 = 49$ degrees of freedom. While this value is not in Statistical Table 2, the correct value using our software is $t_c = t_{(0.975,49)} = 2.0095752$, which we round to $t_c = 2.01$. To construct a 95% interval estimate we use (C.15), replacing estimates for the estimators, to give

$$\bar{y} \pm t_c \frac{\hat{\sigma}}{\sqrt{N}} = 17.1582 \pm 2.01 \frac{1.807}{\sqrt{50}}$$
$$= [16.6447, 17.6717]$$

We estimate that the population mean hip size falls between 16.645 and 17.672 inches. Although we do not know if this interval contains the true population mean hip size for sure, we know that the procedure used to create the interval "works" 95% of the time; thus we would be surprised if the interval did not contain the true population value μ.

C.6 Hypothesis Tests About a Population Mean

Hypothesis testing procedures compare a conjecture, or a hypothesis, that we have about a population to the information contained in a sample of data. The conjectures we test here concern the mean of a normal population. In the context of the problem faced by the airplane seat designer, suppose that airplanes since 1970 have been designed assuming that the mean population hip width is 16.5 inches. Is that figure still valid today?

C.6.1 Components of Hypothesis Tests

Hypothesis tests use sample information about a parameter—namely, its point estimate and its standard error—to draw a conclusion about the hypothesis. In every hypothesis test, five ingredients must be present:

Components of Hypothesis Tests

A *null* hypothesis, H_0
An *alternative* hypothesis, H_1
A test *statistic*
A *rejection* region
A conclusion

The Null Hypothesis

The "null" hypothesis, which is denoted by H_0 (*H-naught*), specifies a value c for a parameter. We write the **null hypothesis** as $H_0: \mu = c$. A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true, in which case we *reject* the null hypothesis.

The Alternative Hypothesis

Paired with every null hypothesis is a logical alternative hypothesis, H_1, that we will accept if the null hypothesis is rejected. The **alternative hypothesis** is flexible and depends to some extent on the problem at hand. For the null hypothesis $H_0: \mu = c$ three possible alternative hypotheses are

- $H_1: \mu > c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ is greater than c.
- $H_1: \mu < c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ is less than c.
- $H_1: \mu \neq c$. If we reject the null hypothesis that $\mu = c$, we accept the alternative that μ takes a value other than (not equal to) c.

The Test Statistic

The sample information about the null hypothesis is embodied in the sample value of a **test statistic**. Based on the value of a test statistic, we decide either to reject the null hypothesis or not to reject it. A test statistic has a very special characteristic: its probability distribution is completely known when the null hypothesis is true, and it has some other distribution if the null hypothesis is not true.

Consider the null hypothesis $H_0: \mu = c$. If the sample data come from a normal population with mean μ and variance σ^2, then

$$t = \frac{\overline{Y} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)}$$

If the null hypothesis $H_0: \mu = c$ is true, then

$$t = \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \sim t_{(N-1)} \tag{C.16}$$

If the null hypothesis is not true, then the t-statistic in (C.16) does not have the usual t-distribution.

Remark

The test statistic distribution in (C.16) is based on an assumption that the population is normally distributed. If the population is not normal, then we invoke the central limit theorem, and say that $\overline{Y}$ is approximately normal in "large" samples. We can use (C.16), recognizing that there is an approximation error introduced if our sample is small.

The Rejection Region The **rejection region** depends on the form of the alternative. It is the range of values of the test statistic that leads to rejection of the null hypothesis. They are values that are *unlikely* and have low probability of occurring when the null hypothesis is true. The chain of logic is "If a value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus it is unlikely that the null hypothesis is true." If the alternative hypothesis is true, then values of the test statistic will tend to be unusually large or unusually small. The terms "large" and "small" are determined by choosing a probability α, called the **level of significance** of the test, which provides a meaning for "an *unlikely* event." The level of significance of the test α is usually chosen to be 0.01, 0.05, or 0.10.

Conclusion When you have completed a hypothesis test, you should state your conclusion, whether you reject the null hypothesis. However, we urge you to make it standard practice to say what the conclusion means in the economic context of the problem you are working on—that is, interpret the results in a meaningful way. This should be a point of emphasis in all statistical work that you do.

We will now discuss the mechanics of carrying out alternative versions of hypothesis tests.

C.6.2 One-Tail Tests with Alternative "Greater Than" (>)

If the alternative hypothesis $H_1: \mu > c$ is true, then the value of the t-statistic (C.16) tends to become larger than usual for the t-distribution. Let the critical value t_c be the $100(1-\alpha)$-percentile $t_{(1-\alpha, N-1)}$ from a t-distribution with $N-1$ degrees of freedom. Then $P(t \le t_c) = 1-\alpha$, where α is the level of significance of the test. If the t-statistic is greater than or equal to t_c, then we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu > c$, as shown in Figure C.5.

If the null hypothesis $H_0: \mu = c$ is *true*, then the test statistic (C.16) has a t-distribution, and its values would tend to fall in the center of the distribution, where most of the probability is contained. If $t < t_c$, then there is no evidence against the null hypothesis, and we do not reject it.

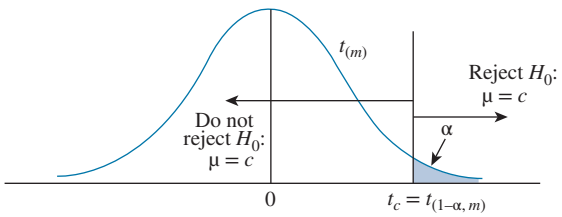


FIGURE C.5 **The rejection region for the one-tail test of $H_0: \mu = c$ against $H_1: \mu > c$.**

C.6.3 One-Tail Tests with Alternative "Less Than" (<)

If the alternative hypothesis $H_1: \mu < c$ is true, then the value of the t-statistic (C.16) tends to become smaller than usual for the t-distribution. The critical value $-t_c$ is the 100α-percentile $t_{(\alpha, N-1)}$ from a t-distribution with $N-1$ degrees of freedom. Then $P(t \le -t_c) = \alpha$, where α is the level of significance of the test as shown in Figure C.6. If $t \le -t_c$, then we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu < c$. If $t > -t_c$, then we do not reject $H_0: \mu = c$.

Memory Trick

The rejection region for a one-tail test is in the direction of the arrow in the alternative. If alternative is ">", then reject in right tail. If alternative is "<", reject in left tail.

C.6.4 Two-Tail Tests with Alternative "Not Equal To" (≠)

If the alternative hypothesis $H_1: \mu \neq c$ is true, then values of the test statistic may be unusually "large" or unusually "small." The rejection region consists of the two "tails" of the t-distribution, and this is called a **two-tail test**. In Figure C.7, the critical values for testing $H_0: \mu = c$ against $H_1: \mu \neq c$ are depicted. The critical value is the $100(1-\alpha/2)$-percentile from a t-distribution with $N-1$ degrees of freedom, $t_c = t_{(1-\alpha/2, N-1)}$, so that $P(t \ge t_c) = P(t \le -t_c) = \alpha/2$.

If the value of the test statistic t falls in the rejection region, either tail of the $t_{(N-1)}$ distribution, then we reject the null hypothesis $H_0: \mu = c$ and accept the alternative $H_1: \mu \neq c$. If the value of the test statistic t falls in the nonrejection region, between the critical values $-t_c$ and t_c, then we do not reject the null hypothesis $H_0: \mu = c$.

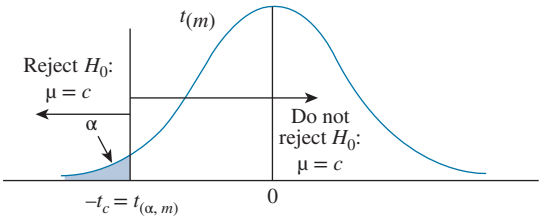


FIGURE C.6 **Critical value for one-tail test $H_0: \mu = c$ versus $H_1: \mu < c$.**

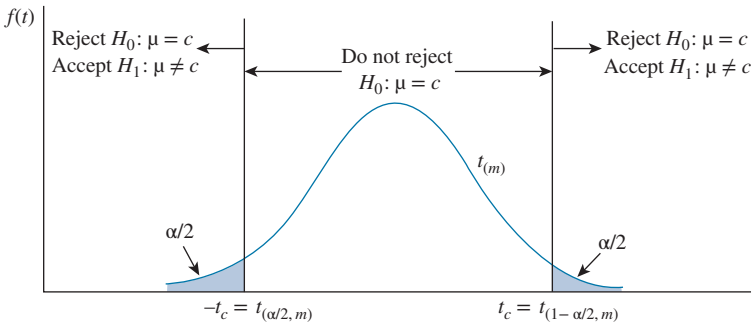


FIGURE C.7 **Rejection region for a test of $H_0: \mu = c$ against $H_1: \mu \neq c$.**

EXAMPLE C.11 | One-tail Test Using the Hip Data

Let us illustrate by testing the null hypothesis that the population hip size is 16.5 inches, against the alternative that it is *greater* than 16.5 inches. The following five-step format is recommended.

1. The null hypothesis is $H_0: \mu = 16.5$. The alternative hypothesis is $H_1: \mu > 16.5$.
2. The test statistic $t = \left(\overline{Y} - 16.5\right) \Big/ \left(\hat{\sigma}/\sqrt{N}\right) \sim t_{(N-1)}$ *if the null hypothesis is true*.
3. Let us select the level of significance $\alpha = 0.05$. The critical value $t_c = t_{(0.95,49)} = 1.6766$ for a t-distribution with $N - 1 = 49$ degrees of freedom. Thus we will reject the null hypothesis in favor of the alternative if $t \geq 1.68$.
4. Using the hip data, the estimate of μ is $\overline{y} = 17.1582$, with estimated variance $\hat{\sigma}^2 = 3.2653$, so $\hat{\sigma} = 1.807$. The value of the test statistic is
$$t = \frac{17.1582 - 16.5}{1.807/\sqrt{50}} = 2.5756$$
5. *Conclusion*: Since $t = 2.5756 > 1.68$, we *reject* the null hypothesis. The sample information we have is *incompatible* with the hypothesis that $\mu = 16.5$. We accept the alternative that the population mean hip size is greater than 16.5 inches, at the $\alpha = 0.05$ level of significance.

EXAMPLE C.12 | Two-tail Test Using the Hip Data

Let us test the null hypothesis that the population hip size is 17 inches, against the alternative that it is *not equal to* 17 inches. The steps of the test are

1. The null hypothesis is $H_0: \mu = 17$. The alternative hypothesis is $H_1: \mu \neq 17$.
2. The test statistic $t = \left(\overline{Y} - 17\right) \Big/ \left(\hat{\sigma}/\sqrt{N}\right) \sim t_{(N-1)}$ *if the null hypothesis is true*.
3. Let us select the level of significance $\alpha = 0.05$. In a two-tail test $\alpha/2 = 0.025$ of probability is allocated to each tail of the distribution. The critical value is the 97.5 percentile of the t-distribution, which leaves 2.5% of the probability in the upper tail, $t_c = t_{(0.975,49)} = 2.01$ for a t-distribution with $N - 1 = 49$ degrees of freedom. Thus, we will reject the null hypothesis in favor of the alternative if $t \geq 2.01$ or if $t \leq -2.01$.
4. Using the hip data, the estimate of μ is $\overline{y} = 17.1582$, with estimated variance $\hat{\sigma}^2 = 3.2653$, so $\hat{\sigma} = 1.807$. The value of the test statistic is
$$t = (17.1582 - 17) \Big/ \left(1.807/\sqrt{50}\right) = 0.6191.$$
5. *Conclusion*: Since $-2.01 < t = 0.6191 < 2.01$ we *do not reject* the null hypothesis. The sample information we have is *compatible* with the hypothesis that the population mean hip size $\mu = 17$.

Warning

Care must be taken when interpreting the outcome of a statistical test. One of the basic precepts of hypothesis testing is that finding a sample value of the test statistic in the nonrejection region does not make the null hypothesis true! Suppose another null hypothesis is $H_0: \mu = c^*$, where c^* is "close" to c. If we fail to reject the hypothesis $\mu = c$, then we will likely fail to reject the hypothesis that $\mu = c^*$. In the example above, at the $\alpha = 0.05$ level, we fail to reject the hypothesis that μ is 17, 16.8, 17.2, or 17.3. In fact, in any problem there are many hypotheses that we would fail to reject, but that does not make any of them true. The weaker statements "we do not reject the null hypothesis" or "we fail to reject the null hypothesis" do not send a misleading message.

C.6.5 The *p*-Value

When reporting the outcome of statistical hypothesis tests it has become common practice to report the ***p*-value** of the test. If we have the *p*-value of a test, p, we can determine the outcome of the test by comparing the *p*-value to the chosen level of significance, α, *without* looking up or calculating the critical values ourselves. The rule is

> ***p*-Value Rule**
>
> Reject the null hypothesis when the *p*-value is less than, or equal to, the level of significance α. That is, if $p \leq \alpha$ then reject H_0. If $p > \alpha$, then do not reject H_0.

If you have chosen the level of significance to be $\alpha = 0.01, 0.05, 0.10$, or any other value, you can compare it to the *p*-value of a test and then reject, or not reject, without checking the critical value t_c.

How the *p*-value is computed depends on the alternative. If t is the calculated value (not the critical value t_c) of the t-statistic with $N - 1$ degrees of freedom, then

- if $H_1 : \mu > c$, p = probability to the right of t
- if $H_1 : \mu < c$, p = probability to the left of t
- $H_1 : \mu \neq c$, p = *sum* of probabilities to the right of $|t|$ and to the left of $-|t|$

The direction of the alternative indicates the tail(s) of the distribution in which the *p*-value falls.

EXAMPLE C.13 | One-tail Test *p*-value: The Hip Data

In Example C.11 we used the hip data to test $H_0 : \mu = 16.5$ against $H_1 : \mu > 16.5$. The calculated t-statistic value was $t = 2.5756$. In this case, since the alternative is "greater than" (>), the *p*-value of this test is the probability that a t-random variable with $N - 1 = 49$ degrees of freedom is greater than 2.5756. This probability value cannot be found in the usual t-table of critical values, but it is easily found using the computer. Statistical software packages, and spreadsheets such as Excel, have simple commands to evaluate the *cumulative distribution function* (*cdf*) (see the Probability Primer, Section P.2) for a variety of probability distributions. If $F_X(x)$ is the *cdf* for a random variable X, then for any value $x = c$, $P[X \leq c] = F_X(c)$. Given such a function for the t-distribution, we compute the desired *p*-value as

$$p = P\left(t_{(49)} \geq 2.576\right) = 1 - P\left(t_{(49)} \leq 2.576\right) = 0.0065$$

Given the *p*-value, we can immediately conclude that at $\alpha = 0.01$ or 0.05 we reject the null hypothesis in favor of the alternative, but if $\alpha = 0.001$ we would not reject the null hypothesis.

The logic of the *p*-value rule is shown in Figure C.8. If 0.0065 of the probability lies to the right of $t = 2.5756$, then the critical value t_c that leaves a probability of $\alpha = 0.01\left(t_{(0.99,49)}\right)$ or $\alpha = 0.05\left(t_{(0.95,49)}\right)$ in the tail must be to the left of 2.5756. In this case, when the *p*-value $\leq \alpha$, it must be true that $t \geq t_c$, and we should reject the null hypothesis for either of these levels of significance. On the other hand, it must be true that the critical value for $\alpha = 0.001$ must fall to the right of 2.5756, meaning that we should not reject the null hypothesis at this level of significance.

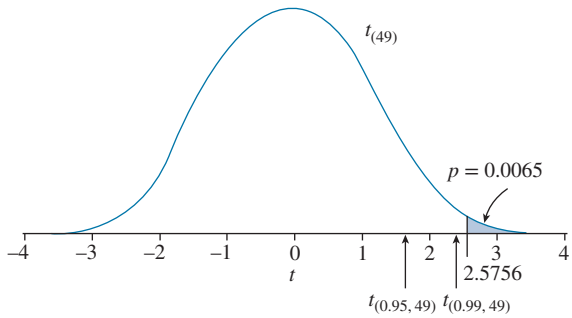


FIGURE C.8 ***p*-value for a right-tail test.**

EXAMPLE C.14 | Two-Tail Test *p*-Value: The Hip Data

For a two-tail test, the rejection region is in the two tails of the *t*-distribution, and the *p*-value must similarly be calculated in the two tails of the distribution. For the hip data, we tested the null hypothesis $H_0: \mu = 17$ against $H_1: \mu \neq 17$, yielding the test statistic value $t = 0.6191$. The *p*-value is

$$p = P\left[t_{(49)} \geq 0.6191\right] + P\left[t_{(49)} \leq -0.6191\right]$$
$$= 2 \times 0.2694 = 0.5387$$

Since the *p*-value $= 0.5387 > \alpha = 0.05$, we do not reject the null hypothesis $H_0: \mu = 17$ at $\alpha = 0.05$ or any other common level of significance. The two-tail *p*-value is shown in Figure C.9.

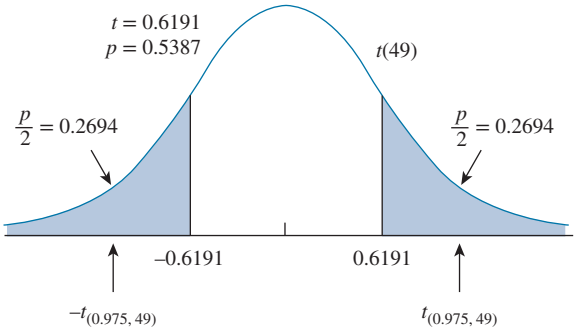


FIGURE C.9 **The *p*-value for a two-tail test.**

C.6.6 A Comment on Stating Null and Alternative Hypotheses

A statistical test procedure cannot prove the truth of a null hypothesis. When we fail to reject a null hypothesis, all the hypothesis test can establish is that the information in a sample of data is *compatible* with the null hypothesis. On the other hand, a statistical test can lead us to *reject* the null hypothesis, with only a small probability, α, of rejecting the null hypothesis when it is actually true. Thus rejecting a null hypothesis is a stronger conclusion than failing to reject it.

The null hypothesis is usually stated in such a way that if our theory is correct, then we will reject the null hypothesis. For example, our airplane seat designer has been operating under the assumption (the maintained or null hypothesis) that the population mean hip width is 16.5 inches. Casual observation suggests that people are getting larger all the time. If we are larger, and if the airline wants to continue to accommodate the same percentage of the population, then the seat widths must be increased. This costly change should be undertaken only if there is statistical evidence that the population hip size is indeed larger. When using a hypothesis test we would like to find out whether there is statistical evidence against our current "theory," or whether the data are compatible with it. With this goal, we set up the null hypothesis that the population mean is 16.5 inches, $H_0: \mu = 16.5$, against the alternative that it is greater than 16.5 inches, $H_1: \mu > 16.5$. In this case, if we reject the null hypothesis, we have shown that there has been a "statistically significant" increase in hip width.

You may view the null hypothesis to be too limited in this case, since it is feasible that the population mean hip width is now smaller than 16.5 inches. The hypothesis test of the null hypothesis $H_0: \mu \leq 16.5$ against the alternative hypothesis $H_1: \mu > 16.5$ is exactly the same as the test for $H_0: \mu = 16.5$ against the alternative hypothesis $H_1: \mu > 16.5$. The test statistic and rejection region are exactly the same. For a one-tail test you can form the null hypothesis in either of these ways.

Finally, it is important to set up the null and alternative hypotheses before you analyze or even collect the sample of data. Failing to do so can lead to errors in formulating the alternative hypothesis. Suppose that we wish to test whether $\mu > 16.5$ and the sample mean is $\bar{y} = 15.5$. Does that mean we should set up the alternative $\mu < 16.5$, to be consistent with the estimate? The answer is no. The alternative is formed to state the conjecture that we wish to establish, $\mu > 16.5$.

C.6.7 Type I and Type II Errors

Whenever we reject—or do not reject—a null hypothesis, there is a chance that we may be making a mistake. This is unavoidable. In any hypothesis testing situation, there are two ways that we can make a correct decision and two ways that we can make an incorrect decision.

Correct Decisions

The null hypothesis is *false* and we decide to *reject* it.

The null hypothesis is *true* and we decide *not* to reject it.

Incorrect Decisions

The null hypothesis is *true* and we decide to *reject* it (a Type I error).

The null hypothesis is *false* and we decide *not* to reject it (a Type II error).

When we reject the null hypothesis we risk what is called a **Type I error**. The probability of a Type I error is α, the level of significance of the test. When the null hypothesis is true, the t-statistic falls in the rejection region with probability α. Thus hypothesis tests will *reject* a true hypothesis $100\alpha\%$ of the time. The good news here is that we can control the probability of a Type I error by choosing the level of significance of the test, α.

We risk a **Type II error** when we do not reject the null hypothesis. Hypothesis tests will lead us to fail to reject null hypotheses that are false with a certain probability. The magnitude of the probability of a Type II error is not under our control and cannot be computed, because it depends on the true value of μ, which is unknown. However, we do know that

- The probability of a Type II error varies inversely with the level of significance of the test, α, which is the probability of a Type I error. If you choose to make α smaller, the probability of a Type II error increases.
- If the null hypothesis is $\mu = c$, and if the true (unknown) value of μ is *close* to c, then the probability of a Type II error is high.
- The larger the sample size N, the lower the probability of a Type II error, given a level of Type I error α.

An easy to remember example of the difference between Type I and Type II errors is from the U.S. legal system. In a trial, a person is presumed innocent. This is the "null" hypothesis, the alternative hypothesis being that the person is guilty. If we convict an innocent person, then we have rejected a null hypothesis that is true, committing a Type I error. If we fail to convict a guilty person, failing to reject the false null hypothesis, then we commit a Type II error. Which is the more costly error in this context? Is it better to send an innocent person to jail, or to let a guilty person go free? It is better in this case to make the probability of a Type I error very small.

C.6.8 A Relationship Between Hypothesis Testing and Confidence Intervals

There is an algebraic relationship between two-tail hypothesis tests and confidence interval estimates that is sometimes useful. Suppose that we are testing the null hypothesis $H_0 : \mu = c$ against the alternative $H_1 : \mu \neq c$. If we fail to reject the null hypothesis at the α level of significance,

then the value c will fall within a $100(1-\alpha)\%$ confidence interval estimate of μ. Conversely, if we reject the null hypothesis, then c will fall outside the $100(1-\alpha)\%$ confidence interval estimate of μ. This algebraic relationship is true because we fail to reject the null hypothesis when $-t_c \le t \le t_c$, or when

$$-t_c \le \frac{\overline{Y} - c}{\hat{\sigma}/\sqrt{N}} \le t_c$$

which when rearranged becomes

$$\overline{Y} - t_c \frac{\hat{\sigma}}{\sqrt{N}} \le c \le \overline{Y} + t_c \frac{\hat{\sigma}}{\sqrt{N}}$$

The endpoints of this interval are the same as the endpoints of a $100(1-\alpha)\%$ confidence interval estimate of μ. Thus for any value of c within the confidence interval, we do not reject $H_0: \mu = c$ against the alternative $H_1: \mu \ne c$. For any value of c outside the confidence interval, we reject $H_0: \mu = c$ and accept the alternative $H_1: \mu \ne c$.

This relationship can be handy if you are given only a confidence interval and want to determine what the outcome of a two-tail test would be.

C.7 Some Other Useful Tests

In this section we very briefly summarize some additional tests. These tests are not only useful in and of themselves, but also illustrate the use of test statistics with chi-square and F-distributions. These distributions were introduced in Appendix B.3.

C.7.1 Testing the Population Variance

Let Y be a normally distributed random variable, $Y \sim N(\mu, \sigma^2)$. Assume that we have a random sample of size N from this population, $Y_1, Y_2, \ldots, Y_N$. The estimator of the population mean is $\overline{Y} = \sum Y_i/N$, and the unbiased estimator of the population variance is $\hat{\sigma}^2 = \sum\left(Y_i - \overline{Y}\right)^2/(N-1)$. To test the null hypothesis $H_0: \sigma^2 = \sigma_0^2$, we use the test statistic

$$V = \frac{(N-1)\,\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{(N-1)}$$

If the null hypothesis is true, then the test statistic has the indicated chi-square distribution with $N-1$ degrees of freedom. If the alternative hypothesis is $H_1: \sigma^2 > \sigma_0^2$, then we carry out a one-tail test. If we choose the level of significance $\alpha = 0.05$, then the null hypothesis is rejected if $V \ge \chi^2_{(0.95,N-1)}$, where $\chi^2_{(0.95,N-1)}$ is the 95th percentile of the chi-square distribution with $N-1$ degrees of freedom. These values can be found in Statistical Table 3, or computed using statistical software. If the alternative hypothesis is $H_1: \sigma^2 \ne \sigma_0^2$, then we carry out a two-tail test, and the null hypothesis is rejected if $V \ge \chi^2_{(0.975,N-1)}$ or if $V \le \chi^2_{(0.025,N-1)}$. The chi-square distribution is skewed, with a long tail to the right, so we cannot use the properties of symmetry when determining the left- and right-tail critical values.

C.7.2 Testing the Equality of Two Population Means

Let two normal populations be denoted by $N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$. In order to estimate and test the difference between means, $\mu_1 - \mu_2$, we must have random samples of data from each of the two populations. We draw a sample of size N_1 from the first population, and a sample of size N_2 from the second population. Using the first sample we obtain the sample mean $\overline{Y}_1$ and sample

variance $\hat{\sigma}_1^2$; from the second sample we obtain $\overline{Y}_2$ and $\hat{\sigma}_2^2$. How the null hypothesis $H_0: \mu_1 - \mu_2 = c$ is tested depends on whether the two population variances are equal or not.

Case 1: *Population variances are equal* If the population variances are equal, so that $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$, then we use information in both samples to estimate the common value σ_p^2. This "pooled variance estimator" is

$$\hat{\sigma}_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}$$

If the null hypothesis $H_0: \mu_1 - \mu_2 = c$ is true, then

$$t = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - c}{\sqrt{\hat{\sigma}_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim t_{(N_1+N_2-2)}$$

As usual, we can construct a one-sided alternative, such as $H_1: \mu_1 - \mu_2 > c$, or the two-sided alternative $H_1: \mu_1 - \mu_2 \neq c$.

Case 2: *Population variances are unequal* If the population variances are not equal, then we cannot use the pooled variance estimate. Instead, we use

$$t^* = \frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - c}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

The exact distribution of this test statistic is neither normal nor the usual t-distribution. The distribution of t^* can be approximated by a t-distribution with degrees of freedom

$$df = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\frac{\left(\hat{\sigma}_1^2/N_1\right)^2}{N_1 - 1} + \frac{\left(\hat{\sigma}_2^2/N_2\right)^2}{N_2 - 1}\right)}$$

This is one of several approximations that appear in the statistics literature, and your software may well use a different one.

C.7.3 Testing the Ratio of Two Population Variances

Given two normal populations, denoted by $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, we can test the null hypothesis $H_0: \sigma_1^2/\sigma_2^2 = 1$. If the null hypothesis is true, then the population variances are equal. The test statistic is derived from the results that $(N_1 - 1)\hat{\sigma}_1^2/\sigma_1^2 \sim \chi^2_{(N_1-1)}$ and $(N_2 - 1)\hat{\sigma}_2^2/\sigma_2^2 \sim \chi^2_{(N_2-1)}$. In Appendix B.3.8 we define an F-random variable, which is formed by taking the ratio of two independent chi-square random variables that have been divided by their degrees of freedom. In this case, the relevant ratio is

$$F = \frac{\frac{(N_1 - 1)\,\hat{\sigma}_1^2/\sigma_1^2}{(N_1 - 1)}}{\frac{(N_2 - 1)\,\hat{\sigma}_2^2/\sigma_2^2}{(N_2 - 1)}} = \frac{\hat{\sigma}_1^2/\sigma_1^2}{\hat{\sigma}_2^2/\sigma_2^2} \sim F_{(N_1-1, N_2-1)}$$

If the null hypothesis $H_0: \sigma_1^2/\sigma_2^2 = 1$ is true then the test statistic is $F = \hat{\sigma}_1^2/\hat{\sigma}_2^2$, which has an F-distribution with $N_1 - 1$ numerator and $N_2 - 1$ denominator degrees of freedom. If the

alternative hypothesis is $H_1: \sigma_1^2/\sigma_2^2 \neq 1$, then we carry out a two-tail test. If we choose level of significance $\alpha = 0.05$, then we reject the null hypothesis if $F \geq F_{(0.975, N_1-1, N_2-1)}$ or if $F \leq F_{(0.025, N_1-1, N_2-1)}$, where $F_{(\alpha, N_1-1, N_2-1)}$ denotes the 100α-percentile of the F-distribution with the specified degrees of freedom. If the alternative is one sided, $H_1: \sigma_1^2/\sigma_2^2 > 1$, then we reject the null hypothesis if $F \geq F_{(0.95, N_1-1, N_2-1)}$.

C.7.4 Testing the Normality of a Population

The tests for means and variances we have developed began with the assumption that the populations were normally distributed. Two questions immediately arise. How well do the tests work when the population is not normal? Can we test for the normality of a population? The answer to the first question is that the tests work pretty well even if the population is not normal, so long as samples are sufficiently large. How large must the samples be? There is no easy answer, since it depends on how "nonnormal" the populations are. The answer to the second question is yes, we can test for normality. Statisticians have been vitally interested in this question for a long time, and a variety of tests have been developed, but the tests and underlying theory are very complicated and far outside the scope of this book.

However, we can present a test that is slightly less ambitious. The normal distribution is symmetric and has a bell shape with a peakedness and tail thickness leading to a kurtosis of three. Thus we can test for departures from normality by checking the skewness and kurtosis from a sample of data. If skewness is not close to zero, or if kurtosis is not close to three, then we reject the normality of the population. In Section C.4.2 we developed sample measures of skewness and kurtosis as

$$\widehat{skewness} = S = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3}$$

$$\widehat{kurtosis} = K = \frac{\tilde{\mu}_4}{\tilde{\sigma}^4}$$

The **Jarque–Bera** test statistic allows a joint test of these two characteristics,

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

If the true distribution is symmetric and has kurtosis three, which includes the normal distribution, then the JB test statistic has a chi-square distribution with two degrees of freedom if the sample size is sufficiently large. If $\alpha = 0.05$, then the critical value of the $\chi^2_{(2)}$ distribution is 5.99. We reject the null hypothesis and conclude that the data are nonnormal if $JB \geq 5.99$. If we reject the null hypothesis, then we know that the data have nonnormal characteristics, but we do not know what distribution the population might have.

EXAMPLE C.15 | Testing the Normality of the Hip Data

For the hip data, skewness and kurtosis measures were estimated in Example C.6. Plugging these values into the JB test statistic formula we obtain

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$
$$= \frac{50}{6}\left((-0.0138)^2 + \frac{(2.3315-3)^2}{4}\right) = 0.9325$$

Since $JB = 0.9325$ is less than the critical value 5.99, we conclude that we cannot reject the normality of the hip data. The p-value for this test is the tail area of a $\chi^2_{(2)}$ -distribution to the right of 0.9325,

$$p = P\left[\chi^2_{(2)} \geq 0.9325\right] = 0.6273$$

C.8 Introduction to Maximum Likelihood Estimation[1]

Maximum likelihood estimation is a powerful procedure that can be used when the population distribution is known. In this section we introduce the concept with a very simple but revealing example.

EXAMPLE C.16 | The "Wheel of Fortune" Game: $p = 1/4$ or $3/4$

Consider the following "Wheel of Fortune" game. You are a contestant faced with two wheels, each of which is partly shaded and partly nonshaded (see Figure C.10). Suppose that after spinning a wheel, you win if a pointer is in the shaded area, and you lose if the pointer is in the nonshaded area. On wheel A 25% of the area is shaded so that the probability of winning is 1/4. On wheel B 75% of the area is shaded so that the probability of winning is 3/4. The game that you must play is this. One of the wheels is chosen and spun three times, with outcomes WIN, WIN, LOSS. You *do not* know which wheel was chosen, and must pick which wheel was spun. Which would you select?

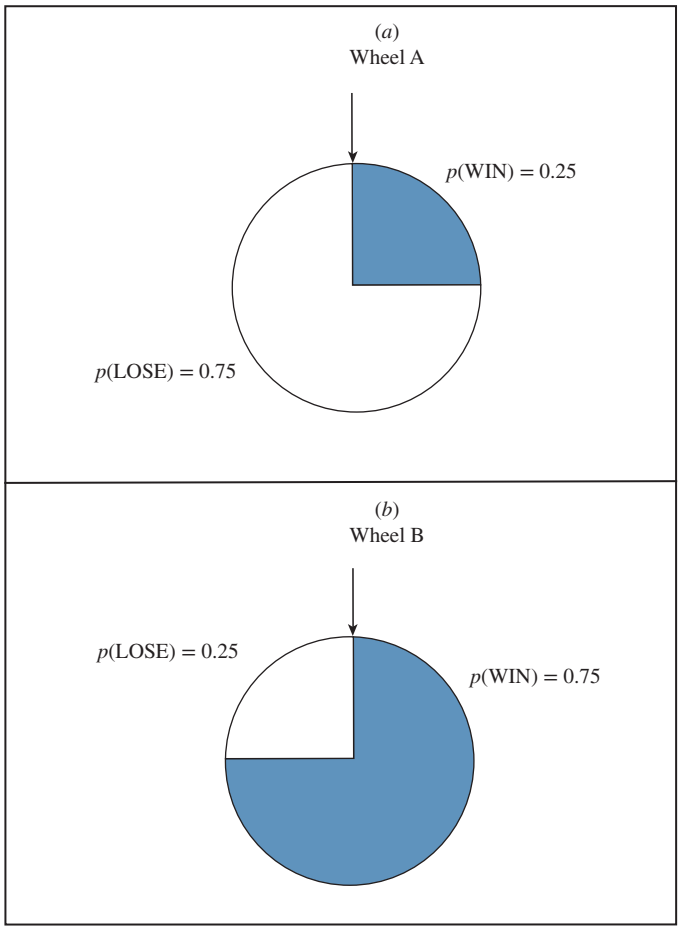


FIGURE C.10 **Wheel of fortune game.**

[1]This section contains some advanced material.

One intuitive approach is the following: let p denote the probability of winning on one spin of a wheel. Choosing between wheels A and B means choosing between $p = 1/4$ and $p = 3/4$. You are estimating p, but there are only two possible estimates, and you must choose based on the observed data. Let us compute the probability of each sequence of outcomes for each of the wheels.

For wheel A, with $p = 1/4$, the probability of observing WIN, WIN, LOSS is

$$\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64} = 0.0469$$

That is, the probability, or **likelihood**, of observing the sequence WIN, WIN, LOSS when $p = 1/4$ is 0.0469.

For wheel B, with $p = 3/4$, the probability of observing WIN, WIN, LOSS is

$$\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{9}{64} = 0.1406$$

The probability, or likelihood, of observing the sequence WIN, WIN, LOSS when $p = 3/4$ is 0.1406.

If we had to choose wheel A or B based on the available data, we would choose wheel B because it has a higher probability of having produced the observed data. It is more *likely* that wheel B was spun than wheel A, and $\hat{p} = 3/4$ is called the **maximum likelihood estimate** of p. The **maximum likelihood principle** seeks the parameter values that maximize the probability, or likelihood, of observing the outcomes actually obtained.

EXAMPLE C.17 | The "Wheel of Fortune" Game: $0 < p < 1$

Now suppose p can be any probability between zero and one, not just 1/4 or 3/4. We have one wheel with a proportion of it shaded, which is the probability of WIN, but we do not know the proportion. In three spins we observe WIN, WIN, LOSS. What is the most likely value of p? The probability of observing WIN, WIN, LOSS is the likelihood L and is

$$L(p) = p \times p \times (1 - p) = p^2 - p^3 \qquad \text{(C.17)}$$

The likelihood L depends on the unknown probability p of a WIN, which is why we have given it the notation $L(p)$, indicating a functional relationship. We would like to find the value of p that maximizes the likelihood of observing the outcomes actually obtained. The graph of the likelihood function (C.17) and the choice of p that maximizes this function is shown in Figure C.11. The maximizing value is denoted as $\hat{p}$ and is called the maximum likelihood estimate of p. To find this value of p we can use calculus. Differentiate $L(p)$ with respect to p,

$$\frac{dL(p)}{dp} = 2p - 3p^2$$

Set this derivative to zero:

$$2p - 3p^2 = 0 \Rightarrow p(2 - 3p) = 0$$

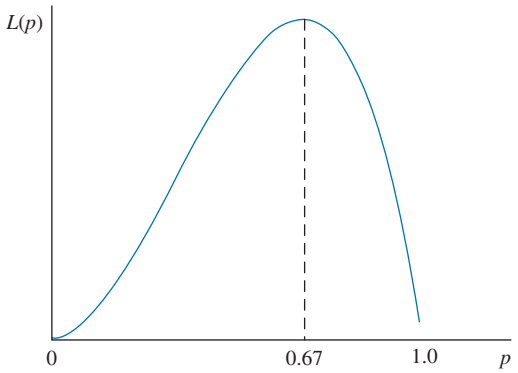


FIGURE C.11 **A likelihood function.**

There are two solutions to this equation, $p = 0$ or $p = 2/3$. The value that maximizes $L(p)$ is $\hat{p} = 2/3$, which is the maximum likelihood estimate. That is, of all possible values of p, between zero and one, the value that maximizes the probability of observing two wins and one loss (the order does not matter) is $\hat{p} = 2/3$.

Can we derive a more general formula that can be used for any observed data? In Appendix B.3.1 we introduced the Bernoulli distribution. Let us define the random variable X that takes the values $x = 1$ (WIN) and $x = 0$ (LOSS) with probabilities p and $1 - p$. The probability function for this random variable can be written in mathematical form as

$$P(X = x) = f(x|p) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

If we spin the "wheel" N times we observe N sample values $x_1, x_2, \ldots, x_N$. Assuming that the spins are independent, we can form the joint probability function

$$\begin{aligned} f(x_1, \ldots, x_N|p) &= f(x_1|p) \times \cdots \times f(x_N|p) \\ &= p^{\Sigma x_i}(1-p)^{N-\Sigma x_i} \\ &= L(p|x_1, \ldots, x_N) \end{aligned} \tag{C.18}$$

The joint probability function gives the probability of observing a specific set of outcomes, and it is a generalization of (C.17). In the last line we have indicated that the joint probability function is algebraically equivalent to the **likelihood function** $L(p|x_1, \ldots, x_N)$. The notation emphasizes that the likelihood function depends upon the unknown probability p *given* the sample outcomes, which we observe. For notational simplicity we will continue to denote the likelihood function as $L(p)$.

EXAMPLE C.18 | The "Wheel of Fortune" Game: Maximizing the Log-likelihood

In the "Wheel of Fortune" game, the maximum likelihood estimate is that value of p that maximizes $L(p)$. To find this estimate using calculus we use a trick to simplify the algebra. The value of p that maximizes $L(p) = p^2(1-p)$ is the same value of p that maximizes the **log-likelihood function** $\ln L(p) = 2\ln(p) + \ln(1-p)$, where "ln" is the natural logarithm. The plot of the log-likelihood function is shown in Figure C.12. Compare Figures C.11 and C.12. The maximum of the likelihood function is $L(\hat{p}) = 0.1481$. The maximum of the log-likelihood function is $\ln L(\hat{p}) = -1.9095$. Both of these maximum values occur at $\hat{p} = 2/3 = 0.6667$.

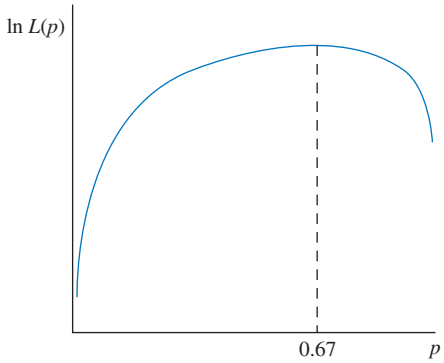


FIGURE C.12 **A log-likelihood function.**

The trick in Example C.18 works for all likelihood and log-likelihood functions and their parameters, so when you see maximum likelihood estimation being discussed it will always be in terms of maximizing the log-likelihood function. For the general problem we are considering, the log-likelihood function is the logarithm of (C.18)

$$\begin{aligned} \ln L(p) &= \sum_{i=1}^{N} \ln\left[f(x_i|p)\right] \\ &= \left(\sum_{i=1}^{N} x_i\right)\ln(p) + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1-p) \end{aligned} \tag{C.19}$$

The first derivative is

$$\frac{d\ln L(p)}{dp} = \frac{\sum x_i}{p} - \frac{N - \sum x_i}{1-p}$$

Setting this to zero and replacing p by $\hat{p}$ to denote the value that maximizes $\ln L(p)$ yields

$$\frac{\sum x_i}{\hat{p}} - \frac{N - \sum x_i}{1-\hat{p}} = 0$$

To solve this equation, multiply both sides by $\hat{p}(1-\hat{p})$. This gives

$$(1-\hat{p})\sum x_i - \hat{p}\left(N - \sum x_i\right) = 0$$

Finally, solving for $\hat{p}$ yields

$$\hat{p} = \frac{\sum x_i}{N} = \bar{x} \tag{C.20}$$

The estimator $\hat{p}$ is the **sample proportion**; $\sum x_i$ is the total number of 1s (wins) out of N spins. As you can see, $\hat{p}$ is also the sample mean of x_i. This result is completely general. Any time we have two outcomes that can occur with probabilities p and $1 - p$, then the maximum likelihood estimate based on a sample of N observations is the sample proportion (C.20).

EXAMPLE C.19 | Estimating a Population Proportion

This estimation strategy can be used if you are a pollster trying to estimate the proportion of the population intending to vote for candidate A rather than candidate B, a medical researcher who wishes to estimate the proportion of the population having a particular defective gene, or a marketing researcher trying to discover whether the population of customers prefers a blue box or a green box for their morning cereal. Suppose in this latter case that you select 200 cereal consumers at random and ask whether they prefer blue boxes or green. If 75 prefer a blue box, then we would estimate that the population proportion preferring blue is $\hat{p} = \sum x_i/N = 75/200 = 0.375$. Thus, we estimate that 37.5% of the population prefers a blue box.

C.8.1 Inference with Maximum Likelihood Estimators

If we use maximum likelihood estimation, how do we perform hypothesis tests and construct confidence intervals? The answers to these questions are found in some remarkable properties of estimators obtained using maximum likelihood methods. Let us consider a general problem. Let X be a random variable (either discrete or continuous) with a probability density function $f(x|\theta)$, where θ is an unknown parameter. The log-likelihood function, based on a random sample $x_1, \ldots, x_N$ of size N, is

$$\ln L(\theta) = \sum_{i=1}^{N} \ln\left[f\left(x_i|\theta\right)\right]$$

If the probability density function of the random variable involved is relatively smooth, and if certain other technical conditions hold, then in large samples the maximum likelihood estimator $\hat{\theta}$ of a parameter θ has a probability distribution that is approximately normal, with expected value θ and a variance $V = \text{var}\left(\hat{\theta}\right)$ that we will discuss in a moment. That is, we can say

$$\hat{\theta} \overset{a}{\sim} N(\theta, V) \tag{C.21}$$

where the symbol $\overset{a}{\sim}$ denotes "asymptotically distributed." The word "asymptotic" refers to estimator properties when the sample size N becomes large, or as $N \to \infty$. To say that an estimator is asymptotically normal means that its probability distribution, which may be unknown when samples are small, becomes approximately normal in large samples. This is analogous to the central limit theorem we discussed in Section C.3.4.

Based on the normality result in (C.21) it will not surprise you that we can immediately construct a t-statistic and obtain both a confidence interval and a test statistic from it. Specifically, if we wish to test the null hypothesis $H_0: \theta = c$ against a one-tail or two-tail alternative hypothesis, then we can use the test statistic

$$t = \frac{\hat{\theta} - c}{\text{se}\left(\hat{\theta}\right)} \overset{a}{\sim} t_{(N-1)} \tag{C.22}$$

If the null hypothesis is true, then this t-statistic has a distribution that can be approximated by a t-distribution with $N - 1$ degrees of freedom in large samples. The mechanics of carrying out the hypothesis test are exactly those in Section C.6.

If t_c denotes the 100(1 – α/2)-percentile $t_{(1-\alpha/2,N-1)}$, then a 100(1 − α)% confidence interval for θ is

$$\hat{\theta} \pm t_c \text{se}\left(\hat{\theta}\right)$$

This confidence interval is interpreted just like those in Section C.5.

Remark

These asymptotic results in (C.21) and (C.22) hold only in large samples. We have indicated that the distribution of the test statistic can be approximated by a t-distribution with $N - 1$ degrees of freedom. If N is truly large, then the $t_{(N-1)}$-distribution converges to the standard normal distribution $N(0, 1)$ and the 100(1 − α/2)-percentile value $t_{(1-\alpha/2,N-1)}$ converges to the corresponding percentile from the standard normal distribution. Asymptotic results are used, rightly or wrongly, when the sample size N may not be large. We prefer using the t-distribution critical values, which are adjusted for small samples by the degrees of freedom correction, when obtaining interval estimates and carrying out hypothesis tests.

C.8.2 The Variance of the Maximum Likelihood Estimator

A key ingredient in both the test statistic and confidence interval expressions is the standard error $\text{se}\left(\hat{\theta}\right)$. Where does this come from? Standard errors are square roots of estimated variances. The part we have delayed discussing until now is how we find the variance of the maximum likelihood estimator, $V = \text{var}\left(\hat{\theta}\right)$. The variance V is given by the inverse of the negative expectation of the second derivative of the log-likelihood function,

$$V = \text{var}\left(\hat{\theta}\right) = \left[-E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)\right]^{-1} \tag{C.23}$$

This looks quite intimidating, and you can see why we put it off. What does this mean? First of all, the second derivative measures the curvature of the log-likelihood function. A second derivative is literally the derivative of the derivative see Appendix A.3.3. A single derivative, the first, measures the slope of a function or the rate of change of the function. The second derivative measures the rate of change of the slope. To obtain a maximum of the log-likelihood function, it must be an "inverted bowl" shape, like those shown in Figure C.13.

At any point to the left of the maximum point, the slope of the log-likelihood function is positive. At any point to the right of the maximum, the slope is negative. As we progress from left to right the slope is *decreasing* (becoming less positive or more negative), so that the second derivative must be negative. A larger absolute magnitude of the second derivative implies a

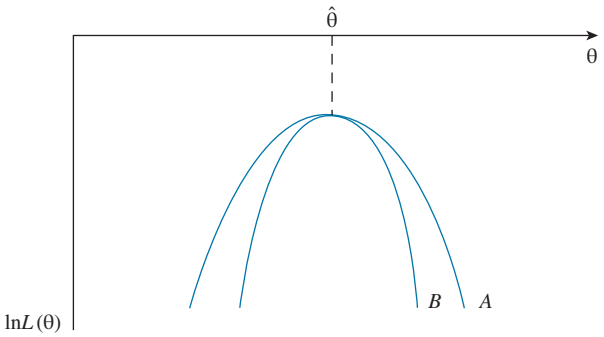


FIGURE C.13 **The log-likelihood functions.**

more rapidly changing slope, indicating a more sharply curved log-likelihood. This is important. In Figure C.13 the two log-likelihood functions A and B have the same maximizing value $\hat{\theta}$. Imagine yourself a climber who is trekking up one of these mountains. For which mountain is the summit most clearly defined? For log-likelihood B, the summit is a sharp peak, and its maximum is more easily located than that for log-likelihood A. The sharper peak has less "wiggle room" at the summit. The smaller amount of wiggle room means that there is less uncertainty as to the location of the maximizing value $\hat{\theta}$; in estimation terminology, less uncertainty means greater precision, and a smaller variance. The more sharply curved log-likelihood function, the one whose second derivative is larger in absolute magnitude, leads to more precise maximum likelihood estimation, and to a maximum likelihood estimator with smaller variance. Thus the variance V of the maximum likelihood estimator is inversely related to the (negative) second derivative. The expected value "E" must be present because this quantity depends on the data and is thus random, so we average over all possible data outcomes.

C.8.3 The Distribution of the Sample Proportion

It is time for an example. At the beginning of Section C.8 we introduced a random variable X that takes the values $x = 1$ and $x = 0$ with probabilities p and $1 - p$. It has log-likelihood given in (C.19). In this problem the parameter θ that we are estimating is the population proportion p, the proportion of $x = 1$ values in the population. We already know that the maximum likelihood estimator of p is the sample proportion $\hat{p} = \sum x_i / N$. The second derivative of the log-likelihood function (C.19) is

$$\frac{d^2 \ln L(p)}{dp^2} = -\frac{\sum x_i}{p^2} - \frac{N - \sum x_i}{(1-p)^2} \tag{C.24}$$

To calculate the variance of the maximum likelihood estimator we need the "expected value" of expression (C.24). In the expectation we treat the x_i values as random because these values vary from sample to sample. The expected value of this discrete random variable is obtained using (P.9) in the probability primer:

$$E(x_i) = 1 \times P(x_i = 1) + 0 \times P(x_i = 0) = 1 \times p + 0 \times (1-p) = p$$

Then, using a generalization of (P.16) (the expected value of a sum is the sum of the expected values and constants can be factored out of expectations) we find the expected value of the second derivative as

$$\begin{aligned} E\left(\frac{d^2 \ln L(p)}{dp^2}\right) &= -\frac{\sum E(x_i)}{p^2} - \frac{N - \sum E(x_i)}{(1-p)^2} \\ &= -\frac{Np}{p^2} - \frac{N - Np}{(1-p)^2} \\ &= -\frac{N}{p(1-p)} \end{aligned}$$

The variance of the sample proportion, which is the maximum likelihood estimator of p, is then

$$V = \text{var}(\hat{p}) = \left[-E\left(\frac{d^2 \ln L(p)}{dp^2}\right)\right]^{-1} = \frac{p(1-p)}{N}$$

The **asymptotic distribution** of the sample proportion, which is valid in large samples, is

$$\hat{p} \overset{a}{\sim} N\left(p, \frac{p(1-p)}{N}\right)$$

To estimate the variance V we must replace the true population proportion by its estimate,

$$\hat{V} = \frac{\hat{p}(1-\hat{p})}{N}$$

The standard error that we need for hypothesis testing and confidence interval estimation is the square root of this estimated variance:

$$\text{se}(\hat{p}) = \sqrt{\hat{V}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

EXAMPLE C.20 | Testing a Population Proportion

As a numerical example, suppose a cereal company CEO conjectures that 40% of the population prefers a blue box. To test this hypothesis, we construct the null hypothesis $H_0: p = 0.4$ and use the two-tail alternative $H_1: p \neq 0.4$. If the null hypothesis is true, then the test statistic $t = (\hat{p} - 0.4)/\text{se}(\hat{p}) \overset{a}{\sim} t_{(N-1)}$. For a sample of size $N = 200$ the critical value from the t-distribution is $t_c = t_{(0.975,199)} = 1.97$. Therefore we reject the null hypothesis if the calculated value of $t \geq 1.97$ or $t \leq -1.97$. If 75 of the respondents prefer a blue box, then the sample proportion is $\hat{p} = 75/200 = 0.375$. The standard error of this estimate is

$$\text{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \sqrt{\frac{0.375 \times 0.625}{200}} = 0.0342$$

The value of the test statistic is

$$t = \frac{\hat{p} - 0.4}{\text{se}(\hat{p})} = \frac{0.375 - 0.4}{0.0342} = -0.7303$$

This value is in the nonrejection region, $-1.97 < t = -0.7303 < 1.97$, so we do not reject the null hypothesis that $p = 0.4$. The sample data are compatible with the conjecture that 40% of the population prefer a blue box.

The 95% interval estimate of the population proportion p who prefer a blue box is

$$\hat{p} \pm 1.97\text{se}(\hat{p}) = 0.375 \pm 1.97(0.0342) = [0.3075, 0.4424]$$

We estimate that between 30.8% and 44.3% of the population prefer a blue box.

C.8.4 Asymptotic Test Procedures

When using maximum likelihood estimation, there are three test procedures that can be used, with the choice depending on which one is most convenient in a given case. The tests are *asymptotically equivalent* and will give the same result in large samples. Suppose that we are testing the null hypothesis $H_0: \theta = c$ against the alternative hypothesis $H_1: \theta \neq c$. In (C.22) we have the t-statistic for carrying out the test. How does this test really work? Basically it is measuring the distance $\hat{\theta} - c$ between the estimate of θ and the hypothesized value c. This distance is normalized by the standard error of $\hat{\theta}$ to adjust for how precisely we have estimated θ. If the distance between the estimate $\hat{\theta}$ and the hypothesized value c is large, then that is taken as evidence against the null hypothesis, and if the distance is large enough, we conclude that the null hypothesis is not true.

There are other ways to measure the distance between $\hat{\theta}$ and c that can be used to construct test statistics. Each of the three testing principles takes a different approach to measuring the distance between $\hat{\theta}$ and the hypothesized value.

The Likelihood Ratio (LR) Test Consider Figure C.14. A log-likelihood function is shown, along with the maximum likelihood estimate $\hat{\theta}$ and the hypothesized value c. Note that the distance between $\hat{\theta}$ and c is also reflected by the distance between the log-likelihood function value evaluated at the maximum likelihood estimate $\ln L\left(\hat{\theta}\right)$ and the log-likelihood function value evaluated at the hypothesized value $\ln L(c)$. We have labeled the difference between these two log-likelihood values (1/2)LR for a reason that will become clear. If the estimate $\hat{\theta}$ is close to c, then the difference between the log-likelihood values will be small. If $\hat{\theta}$ is far from c, then

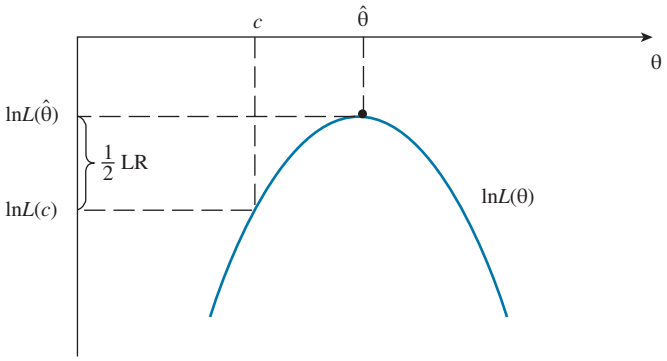


FIGURE C.14 **The likelihood ratio test.**

the difference between the log-likelihood values will be large. This observation leads us to the **likelihood ratio statistic**, which is twice the difference between $\ln L\left(\hat{\theta}\right)$ and $\ln L(c)$,

$$\text{LR} = 2\left[\ln L\left(\hat{\theta}\right) - \ln L(c)\right] \tag{C.25}$$

Based on some advanced statistical theory, it can be shown that if the null hypothesis is true, then the LR test statistic has a chi-square distribution (see Appendix B.3.6) with $J = 1$ degree of freedom. In more general contexts J is the number of hypotheses being tested and it can be greater than 1. If the null hypothesis is not true, then the LR test statistic becomes large. We reject the null hypothesis at the α level of significance if $\text{LR} \geq \chi^2_{(1-\alpha,J)}$, where $\chi^2_{(1-\alpha,J)}$ is the $100(1-\alpha)$-percentile of a chi-square distribution with J degrees of freedom, as shown in Figure C.15. The 90th, 95th, and 99th percentile values of the chi-square distribution for various degrees of freedom are given in Statistical Table 3.

When estimating a population proportion p the log-likelihood function is given by (C.19). The value of p that maximizes this function is $\hat{p} = \sum x_i/N$. Thus, the maximum value of the log-likelihood function is

$$\begin{aligned}\ln L(\hat{p}) &= \left(\sum_{i=1}^{N} x_i\right)\ln\hat{p} + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1-\hat{p}) \\ &= N\hat{p}\ln\hat{p} + (N - N\hat{p})\ln(1-\hat{p}) \\ &= N\left[\hat{p}\ln\hat{p} + (1-\hat{p})\ln(1-\hat{p})\right]\end{aligned}$$

where we have used the fact that $\sum x_i = N\hat{p}$.

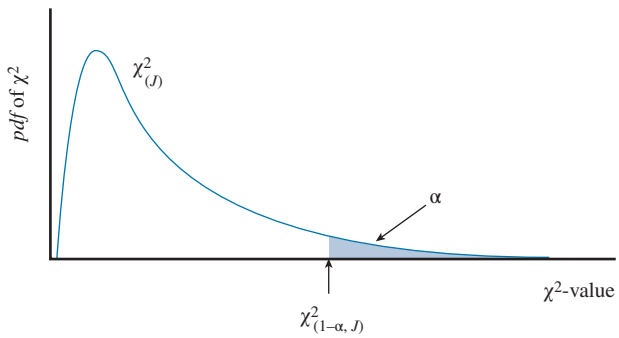


FIGURE C.15 **Critical value from a chi-square distribution.**

EXAMPLE C.21 | Likelihood Ratio Test of the Population Proportion

For our cereal box problem, $\hat{p} = 0.375$ and $N = 200$, so we have

$$\ln L(\hat{p}) = 200\left[0.375 \times \ln(0.375) + (1 - 0.375)\ln(1 - 0.375)\right]$$
$$= -132.3126$$

The value of the log-likelihood function assuming $H_0: p = 0.4$ is true is

$$\ln L(0.4) = \left(\sum_{i=1}^{N} x_i\right)\ln(0.4) + \left(N - \sum_{i=1}^{N} x_i\right)\ln(1 - 0.4)$$
$$= 75 \times \ln(0.4) + (200 - 75) \times \ln(0.6)$$
$$= -132.5750$$

The problem is to assess whether -132.3126 is significantly different from -132.5750. The LR test statistic (C.25) is

$$LR = 2\left[\ln L(\hat{p}) - \ln L(0.4)\right]$$
$$= 2 \times (-132.3126 - (-132.575)) = 0.5247$$

If the null hypothesis $p = 0.4$ is true, then the LR test statistic has a $\chi^2_{(1)}$-distribution. If we choose $\alpha = 0.05$, then the test critical value is $\chi^2_{(0.95,1)} = 3.84$, the 95th percentile from the $\chi^2_{(1)}$-distribution. Since $0.5247 < 3.84$ we do not reject the null hypothesis.

The Wald Test In Figure C.14 it is clear that the distance (1/2)LR will depend on the curvature of the log-likelihood function. In Figure C.16 we show two log-likelihood functions with the hypothesized value c and the distances (1/2)LR for each of the log-likelihoods. The log-likelihoods have the same maximum value $\ln L\left(\hat{\theta}\right)$, but the values of the log-likelihood evaluated at the hypothesized value c are different.

The distance $\hat{\theta} - c$ translates into a larger value of (1/2)LR for the more highly curved log-likelihood, B, so it seems reasonable to construct a test measure by weighting the distance $\hat{\theta} - c$ by the magnitude of the log-likelihood's curvature, which we measure by the negative of its second derivative. This is exactly what the Wald statistic does:

$$W = \left(\hat{\theta} - c\right)^2 \left[-\frac{d^2 \ln L(\theta)}{d\theta^2}\right] \tag{C.26}$$

The value of the Wald statistic is larger for log-likelihood function B (more curved) than log-likelihood function A (less curved).

If the null hypothesis is true, then the Wald statistic (C.26) has a $\chi^2_{(1)}$ -distribution, and we reject the null hypothesis if $W \geq \chi^2_{(1-\alpha,1)}$. In more general situations we may test $J > 1$ hypotheses jointly, in which case we work with a chi-square distribution with J degrees of freedom, as shown in Figure C.15.

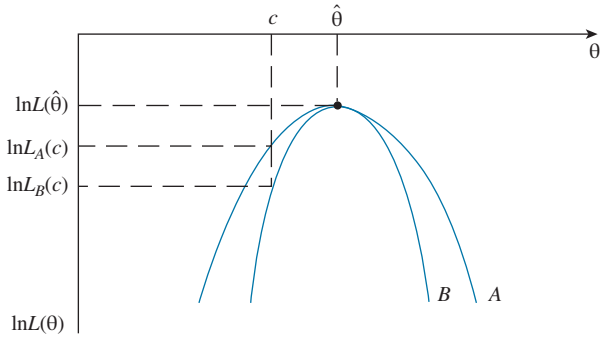


FIGURE C.16 **The Wald statistic.**

There is a linkage between the curvature of the log-likelihood function and the precision of maximum likelihood estimation. The greater the curvature of the log-likelihood function, the smaller the variance V in (C.23) and the more precise maximum likelihood estimation becomes, meaning that we have more **information** about the unknown parameter θ. Conversely, the more information we have about θ, the smaller the variance of the maximum likelihood estimator. Using this idea we define an **information measure** to be the reciprocal of the variance V,

$$I(\theta) = -E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right] = V^{-1} \tag{C.27}$$

As the notation indicates the information measure $I(\theta)$ is a function of the parameter θ. Substitute the information measure for the second derivative in the Wald statistic in (C.26) to obtain

$$W = \left(\hat{\theta} - c\right)^2 I(\theta) \tag{C.28}$$

In large samples the two versions of the Wald statistic are the same. An interesting connection here is obtained by rewriting (C.28) as

$$W = \left(\hat{\theta} - c\right)^2 V^{-1} = \left(\hat{\theta} - c\right)^2 \Big/ V \tag{C.29}$$

To implement the **Wald test**, we use the estimated variance

$$\hat{V} = \left[I\left(\hat{\theta}\right)\right]^{-1} \tag{C.30}$$

Then, taking the square root, we obtain the t-statistic in (C.22),

$$\sqrt{W} = \frac{\hat{\theta} - c}{\sqrt{\hat{V}}} = \frac{\hat{\theta} - c}{\text{se}\left(\hat{\theta}\right)} = t$$

That is, the t-test is also a Wald test.

EXAMPLE C.22 | Wald Test of the Population Proportion

In our blue box–green box example, we know that the maximum likelihood estimate $\hat{p} = 0.375$. To implement the Wald test we calculate

$$I(\hat{p}) = \hat{V}^{-1} = \frac{N}{\hat{p}(1-\hat{p})} = \frac{200}{0.375(1-0.375)} = 853.3333$$

where $V = p(1-p)/N$ and $\hat{V}$ were obtained in Section C.7.3. Then the calculated value of the Wald statistic is

$$W = (\hat{p} - c)^2 I(\hat{p}) = (0.375 - 0.4)^2 \times 853.3333 = 0.5333$$

In this case the value of the Wald statistic is close in magnitude to the LR statistic and the test conclusion is the same. Also, when testing one hypothesis, the Wald statistic is the square of the t-statistic, $W = t^2 = (-0.7303)^2 = 0.5333$.

The Lagrange Multiplier (LM) Test

The third testing procedure that comes from maximum likelihood theory is the Lagrange multiplier (LM) test. Figure C.17 illustrates another way to measure the distance between $\hat{\theta}$ and c. The slope of the log-likelihood function, which is sometimes called the *score*, is

$$s(\theta) = \frac{d \ln L(\theta)}{d\theta} \tag{C.31}$$

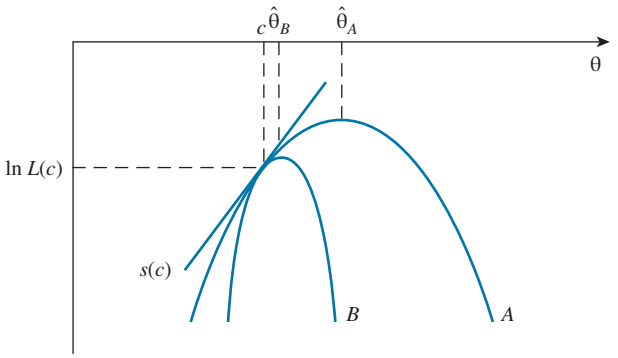


FIGURE C.17 **Motivating the Lagrange multiplier test.**

The slope of the log-likelihood function depends on the value of θ, as our function notation $s(\theta)$ indicates. The slope of the log-likelihood function at the maximizing value is zero, $s\left(\hat{\theta}\right) = 0$. The LM test examines the slope of the log-likelihood function at the point c. The logic of the test is that if $\hat{\theta}$ is close to c then the slope $s(c)$ of the log-likelihood function evaluated at c should be close to zero. In fact testing the null hypothesis $\theta = c$ is equivalent to testing $s(c) = 0$.

The difference between c and the maximum likelihood estimate $\hat{\theta}_B$ (maximizing ln L_B) is smaller than the difference between c and $\hat{\theta}_A$. In contrast to the Wald test, more curvature in the log-likelihood function implies a smaller difference between the maximum likelihood estimate and c. If we use the information measure $I(\theta)$ as our measure of curvature (more curvature means more information), the **Lagrange multiplier test** statistic can be written as

$$\text{LM} = \frac{\left[s(c)\right]^2}{I(\theta)} = \left[s(c)\right]^2\left[I(\theta)\right]^{-1} \tag{C.32}$$

The LM statistic for log-likelihood function A (less curved) is greater than the LM statistic for log-likelihood function B (more curved). If the null hypothesis is true, LM test statistic (C.32) has a $\chi^2_{(1)}$-distribution, and the rejection region is the same as for the LR and Wald tests. The LM, LR, and Wald tests are asymptotically equivalent and will lead to the same conclusion in sufficiently large samples.

In order to implement the LM test we can evaluate the information measure at the point $\theta = c$, so that it becomes

$$\text{LM} = \left[s(c)\right]^2\left[I(c)\right]^{-1}$$

In cases in which the maximum likelihood estimate is difficult to obtain (which it can be in more complex problems) the LM test has an advantage because $\hat{\theta}$ is not required. On the other hand, the Wald test in (C.28) uses the information measure evaluated at the maximum likelihood estimate $\hat{\theta}$,

$$W = \left(\hat{\theta} - c\right)^2 I\left(\hat{\theta}\right)$$

It is preferred when the maximum likelihood estimate and its variance are easily obtained. The likelihood ratio test statistic (C.25) requires calculation of the log-likelihood function at both the maximum likelihood estimate and the hypothesized value c. As noted, the three tests are asymptotically equivalent, and the choice of which to use is often made on the basis of convenience. In complex situations, however, the rule of convenience may not be a good one. The **likelihood ratio test** is relatively reliable in most circumstances, so if you are in doubt, it is a safe one to use.

EXAMPLE C.23 | Lagrange Multiplier Test of the Population Proportion

In the blue box–green box example, the value of the score, based on the first derivative shown just below (C.19), evaluated at the hypothesized value $c = 0.4$ is

$$s(0.4) = \frac{\sum x_i}{c} - \frac{N - \sum x_i}{1 - c} = \frac{75}{0.4} - \frac{200 - 75}{1 - 0.4} = -20.8333$$

The calculated information measure is

$$I(0.4) = \frac{N}{c(1 - c)} = \frac{200}{0.4(1 - 0.4)} = 833.3333$$

The value of the LM test statistic is

$$\text{LM} = \left[s(0.4)\right]^2 \left[I(0.4)\right]^{-1} = [-20.8333]^2 [833.3333]^{-1} = 0.5208$$

Thus, in our example, the values of the LR, Wald, and LM test statistics are very similar and give the same conclusion. This was to be expected, since the sample size $N = 200$ is large, and the problem is a simple one.

C.9 Algebraic Supplements

C.9.1 Derivation of Least Squares Estimator

In this section we illustrate how to use the least squares principle to obtain the sample mean as an estimator of the population mean. Represent a sample of data as $y_1, y_2, \ldots, y_N$. The population mean is $E(Y) = \mu$. The least squares principle says to find the value of μ that minimizes

$$S = \sum_{i=1}^{N} (y_i - \mu)^2$$

where S is the sum of squared deviations of the data values from μ.

The motivation for this approach can be deduced from the following example. Suppose you are going shopping at a number of shops along a certain street. Your plan is to shop at one store and return to your car to deposit your purchases. Then you visit a second store and return again to your car, and so on. After visiting each shop you return to your car. Where would you park to minimize the total amount of walking between your car and the shops you visit? You want to minimize the *distance* traveled. Think of the street along which you shop as a number line. The Euclidean distance between a shop located at y_i and your car at point μ is

$$d_i = \sqrt{(y_i - \mu)^2}$$

The squared distance, which is mathematically more convenient to work with, is

$$d_i^2 = (y_i - \mu)^2$$

To minimize the total squared distance between your parking spot μ and all the shops located at $y_1, y_2, \ldots, y_N$ you would minimize

$$S(\mu) = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (y_i - \mu)^2$$

which is the sum of squares function. Thus the least squares principle is really the least *squared distance* principle.

Since the values of y_i are known given the sample, the sum of squares function $S(\mu)$ is a function of the unknown parameter μ. Multiplying out the sum of squares, we have

$$S(\mu) = \sum_{i=1}^{N} y_i^2 - 2\mu \sum_{i=1}^{N} y_i + N\mu^2 = a_0 - 2a_1\mu + a_2\mu^2$$

EXAMPLE C.24 | Hip Data: Minimizing the Sum of Squares Function

For the data in Table C.1 we have

$$a_0 = \sum y_i^2 = 14880.1909,\ a_1 = \sum y_i = 857.9100,$$
$$a_2 = N = 50$$

The plot of the sum of squares parabola is shown in Figure C.18. The minimizing value appears to be a bit larger than 17 in the figure. Now we will determine the minimizing value exactly.

The value of μ that minimizes $S(\mu)$ is the "least squares estimate." From calculus, we know that the minimum of the function occurs where its slope is zero. See Appendix A.3.4. The function's derivative gives its slope, so by equating the first derivative of $S(\mu)$ to zero and solving, we can obtain the minimizing value exactly. The derivative of $S(\mu)$ is

$$\frac{dS(\mu)}{d\mu} = -2a_1 + 2a_2\mu$$

Setting the derivative to zero determines the least squares estimate of μ, which we denote as $\hat{\mu}$. Setting the derivative to zero,

$$-2a_1 + 2a_2\hat{\mu} = 0$$

Solving for $\hat{\mu}$ yields the formula for the least squares estimate,

$$\hat{\mu} = \frac{a_1}{a_2} = \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}$$

Thus, the least squares estimate of the population mean is the sample mean, $\bar{y}$. This formula can be used in general, for any sample values that might be obtained, meaning that the least squares estimator is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} Y_i}{N} = \overline{Y}$$

For the hip data in Table C.1

$$\hat{\mu} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{857.9100}{50} = 17.1582$$

Thus, we estimate that the average hip size in the population is 17.1582 inches.

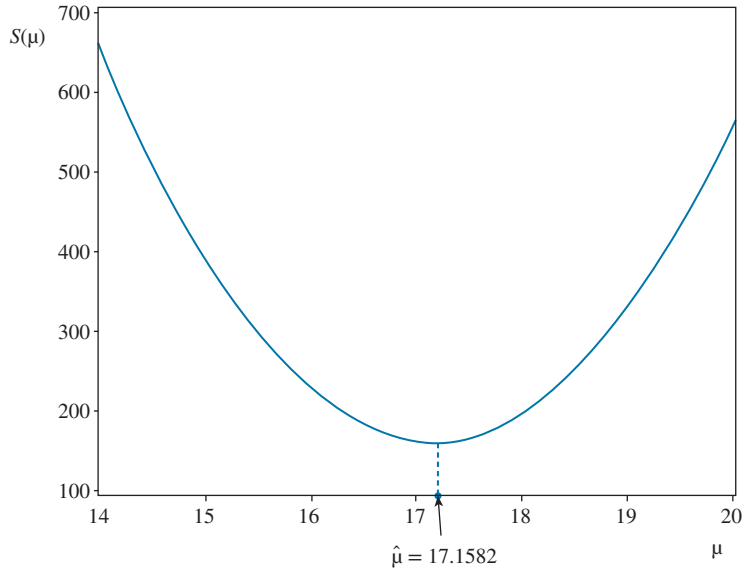


FIGURE C.18 **The sum of squares parabola for the hip data.**

C.9.2 Best Linear Unbiased Estimation

One of the powerful findings about the sample mean (which is also the least squares estimator) is that it is the best of all possible estimators that are both *linear* and *unbiased*. The fact that $\overline{Y}$ is the best linear unbiased estimator (BLUE) accounts for its wide use. In this context we mean by best that it is the estimator with the smallest variance of all linear and unbiased estimators. It is better to have an estimator with a smaller variance than one with a larger variance; it increases the

chances of getting an estimate close to the true population mean μ. This important result about the least squares estimator is true *if* the sample values $Y_i \sim (\mu, \sigma^2)$ are uncorrelated and identically distributed. It does not depend on the population being normally distributed. The fact that $\overline{Y}$ is BLUE is so important that we will prove it.

The sample mean is a weighted average of the sample values,

$$\begin{aligned}\overline{Y} &= \sum_{i=1}^{N} Y_i/N = \frac{1}{N}Y_1 + \frac{1}{N}Y_2 + \cdots + \frac{1}{N}Y_N \\ &= a_1Y_1 + a_2Y_2 + \cdots + a_NY_N \\ &= \sum_{i=1}^{N} a_iY\end{aligned}$$

where the weights $a_i = 1/N$. Weighted averages are also called linear combinations, so we call the sample mean a **linear estimator**. In fact, any estimator that can be written as $\sum_{i=1}^{N} a_iY_i$ is a linear estimator. For example, suppose the weights a_i^* are constants different from $a_i = 1/N$. Then we can define another linear estimator of μ as

$$\tilde{Y} = \sum_{i=1}^{N} a_i^*Y_i$$

To ensure that $\tilde{Y}$ is different from $\overline{Y}$, let us define

$$a_i^* = a_i + c_i = \frac{1}{N} + c_i$$

where c_i are constants that are not all zero. Thus,

$$\begin{aligned}\tilde{Y} &= \sum_{i=1}^{N} a_i^*Y_i = \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)Y_i \\ &= \sum_{i=1}^{N}\frac{1}{N}Y_i + \sum_{i=1}^{N} c_iY \\ &= \overline{Y} + \sum_{i=1}^{N} c_iY\end{aligned}$$

The expected value of the new estimator $\tilde{Y}$ is

$$\begin{aligned}E[\tilde{Y}] &= E\left[\overline{Y} + \sum_{i=1}^{N} c_iY_i\right] = \mu + \sum_{i=1}^{N} c_iE[Y_i] \\ &= \mu + \mu\sum_{i=1}^{N} c_i\end{aligned}$$

The estimator $\tilde{Y}$ is not unbiased unless $\sum c_i = 0$. We want to compare the sample mean to other linear and unbiased estimators, so we will assume that $\sum c_i = 0$ holds. Now we find the variance of $\tilde{Y}$. The linear unbiased estimator with the smaller variance will be best.

$$\begin{aligned}\text{var}(\tilde{Y}) &= \text{var}\left(\sum_{i=1}^{N} a_i^*Y_i\right) = \text{var}\left(\sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)Y_i\right) = \sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2\text{var}(Y_i) \\ &= \sigma^2\sum_{i=1}^{N}\left(\frac{1}{N} + c_i\right)^2 = \sigma^2\sum_{i=1}^{N}\left(\frac{1}{N^2} + \frac{2}{N}c_i + c_i^2\right) = \sigma^2\left(\frac{1}{N} + \frac{2}{N}\sum_{i=1}^{N} c_i + \sum_{i=1}^{N} c_i^2\right) \\ &= \sigma^2/N + \sigma^2\sum_{i=1}^{N} c_i^2 \quad \left(\text{since}\sum_{i=1}^{N} c_i = 0\right) \\ &= \text{var}(\overline{Y}) + \sigma^2\sum_{i=1}^{N} c_i^2\end{aligned}$$

It follows that the variance of $\tilde{Y}$ must be greater than the variance of $\overline{Y}$, unless all the c_i values are zero, in which case $\tilde{Y} = \overline{Y}$.

C.10 Kernel Density Estimator

As econometricians, we work with data that are drawings from unknown distributions. For example, Figure C.19 shows the empirical distributions of two datasets, presented here as histograms. The variables X and Y are in the data file *kernel*. The problem before us is to estimate the density functions that yielded the observations. Knowledge about the distributions is important for statistical inference.

There are two main ways to estimate the distribution. We can use a parametric density estimator, or we can use a nonparametric **kernel density estimator**. In the **parametric approach**, we rely on density functions with well-defined functional forms characterized by parameters. For example, the normal probability density distribution $f(\cdot)$ has a specific functional form defined by two parameters—the mean μ and the standard deviation σ:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Once we have estimates of the mean and the standard deviation, $\hat{\mu}$ and $\hat{\sigma}$, we plug these into the normal density function formula to obtain

$$\widehat{f(x)} = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right)$$

Figure C.20 shows our application of this approach; the generated normal density functions are superimposed onto the histograms of the data. We have applied this parametric approach in the discussion about the Central Limit Theorem (C.3.4) and in discussion about ARCH models (Chapter 14).

The histogram of the variable X, on the left in Figure C.20, is unimodal, and the normal distribution appears to fit the shape of the data well. In contrast, the histogram of the variable Y on the right in Figure C.20 is bimodal, and the normal distribution is a poor representation of the underlying density function. We could try fitting the data with other parametric distributional forms, but rather than do that, let us adopt a nonparametric kernel density estimator to capture the shape of the data in a smooth continuous form.

Nonparametric methods do not require specific functional forms (e.g., the normal distribution formula) to generate the distribution. Instead, smoothing functions, called **kernels**, are used to "fit" the shape of the distribution of the data.

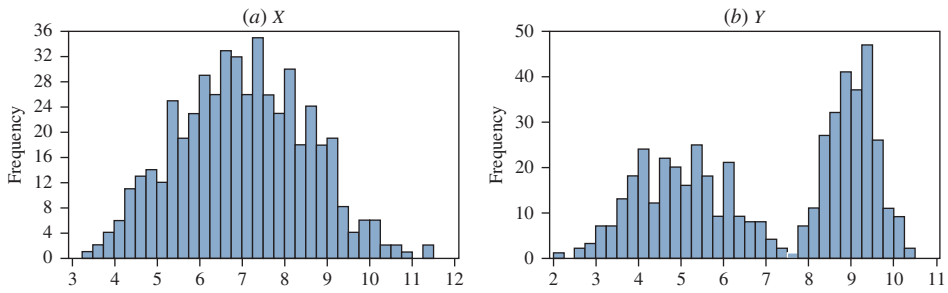


FIGURE C.19 **Histograms of variables (*a*) unimodal variable *X* and (*b*) bimodal variable *Y*.**

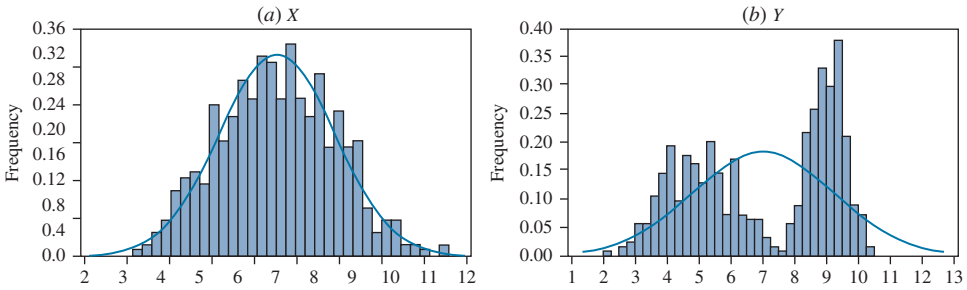


FIGURE C.20 **Parametric density estimator (*a*) unimodal variable *X* and (*b*) bimodal variable *Y*.**

The logic of the nonparametric approach can be grasped intuitively by thinking about how we set up histograms. Figure C.21 shows two histograms for the dataset Y. The one on the left has nine bins (i.e., the rectangles in the histogram) with bin width = 1 whereas the one on the right has many bins each with bin width = 0.1. The histogram with less bins has the higher frequency per bin as more observations fall into the larger bin width. More specifically, if x_k is the midpoint of the kth bin and h is the bin width, the range of values in the bin is $x_k \pm h/2$, and the frequency count n_k is the number of observations which falls in that range. The sum of all frequencies equals the sample size n, while the sum of the areas equals nh, since each area is $n_k h$ and $\sum_k n_k = n$. Note, too, that the shapes of the histograms are similar, but that the one with the larger bin width is "smoother" (fewer spikes and dips).

We can think of the histogram as a density function estimator $\widehat{f(x)}$, where x takes values over the domain of x and

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} 1\left(A_i\right)$$

The expression $1(A_i)$ is an **indicator function** taking on the value of 1 if A_i is true; A_i is the condition that x_i is in the same bin as x. For example, suppose we wish to find $\widehat{f(x)}$ for an x that lies in the kth bin. Then, A_i is true for all x_i such that $x_k - h/2 < x_i < x_k + h/2$. Thus, in the kth bin, $\sum_{i=1}^{n} 1\left(A_i\right) = n_k$, and the histogram density estimator for all x in the kth bin is $\widehat{f(x)} = n_k/nh$. The divisor nh ensures that the bin areas sum to one.

Now consider another density estimator where, instead of having a number of predetermined bins with midpoints x_k, we consider a bin with midpoint x and count the number of observations in the range $x \pm h/2$. If we repeat this process for all values of x, we can picture it as creating an infinite number of overlapping bins along the domain of x. In this case the density estimator is given by

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} 1\left(x - \frac{h}{2} < x_i < x + \frac{h}{2}\right) = \frac{1}{nh}\sum_{i=1}^{n} 1\left(-\frac{1}{2} < \frac{x_i - x}{h} < \frac{1}{2}\right)$$

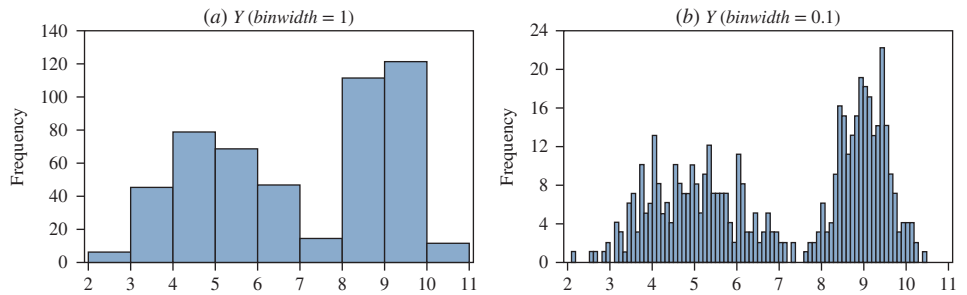


FIGURE C.21 **Histograms with different bin widths (*a*) width = 1 (*b*) width = 0.1.**

In practice, as you sum over the observations, the indicator function ensures that you only "count" the relevant observations. However, this density function will not be smooth, because each observation is given a weight of either zero or one—that is, it is either in or out, according to the condition specified in the indicator function.

Suppose we now replace this simple counting rule with a more sophisticated weighting function known as a **kernel**:

$$\widehat{f(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

where K is a kernel, h is a smoothing parameter called the **bandwidth**, and x is any value over the domain of possible values. There are many kernel functions; one of them is Gaussian and is described as follows:

$$K\left(\frac{x_i - x}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - x}{h}\right)^2\right)$$

Figure C.22 shows the application of this kernel estimator to variable Y in data file *kernel* with four different bandwidths. Note how the shape of the density function is controlled by the bandwidth. The smaller the bandwidth, the better the fit, but there is a tradeoff between the number of "humps" captured and the smoothness of the fit. Intuitively, decreasing the bandwidth is like decreasing the bin width in the histogram, and the kernel is like a "counter"—but one which puts less weight on observations that are further away from the point being evaluated. (Imagine moving from the histogram on the right in Figure C.21 to the one on the left as you increase the bandwidth, and then imagine the use of the kernel to smooth the bars.) The kernel (Gaussian) density function with bandwidth equal to 0.4 appears to have captured the bimodality in the data.

There is a vast literature about the optimal choice of bandwidth as well as extensions of the nonparametric methods to regression analysis. Useful references include Pagan, A. and Ullah, A., *Nonparametric Econometrics*, Cambridge University Press, 1999; and Li, Q. and Racine, J.S. *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, 2007.

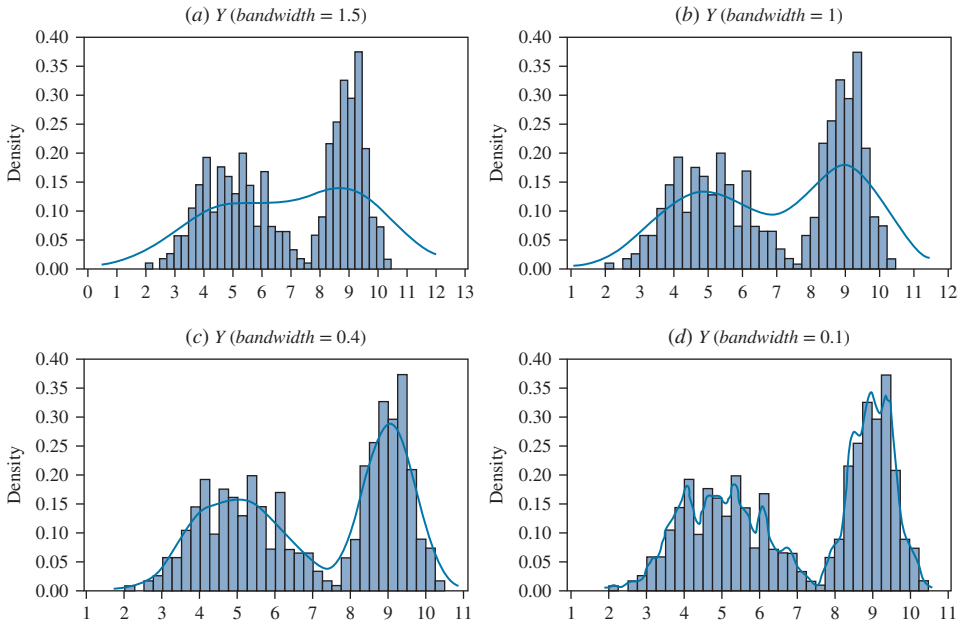


FIGURE C.22 **Fitting with a nonparametric density estimator (*a*) bandwidth = 1.5, (*b*) bandwidth = 1, (*c*) bandwidth = 0.4, and (*d*) bandwidth 0.1.**

C.11 Exercises

C.11.1 Problems

C.1 Suppose $Y_1, Y_2, \ldots, Y_N$ is a random sample from a population with mean μ and variance σ^2. Rather than using all N observations, consider an easy estimator of μ that uses only the first two observations

$$Y^* = \frac{Y_1 + Y_2}{2}$$

a. Show that Y^* is a linear estimator.
b. Show that Y^* is an unbiased estimator.
c. Find the variance of Y^*.
d. Explain why the sample mean of all N observations is a better estimator than Y^*.

C.2 Suppose that Y_1, Y_2, Y_3 is a random sample from a $N(\mu, \sigma^2)$ population. To estimate μ, consider the weighted estimator

$$\tilde{Y} = \frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3$$

a. Show that $\tilde{Y}$ is a linear estimator.
b. Show that $\tilde{Y}$ is an unbiased estimator.
c. Find the variance of $\tilde{Y}$ and compare it to the variance of the sample mean $\overline{Y}$.
d. Is $\tilde{Y}$ as good an estimator as $\overline{Y}$?
e. If $\sigma^2 = 9$, calculate the probability that each estimator is within one unit on either side of μ.

C.3 The hourly sales of fried chicken at Louisiana Fried Chicken are normally distributed with mean 2,000 pieces and standard deviation 500 pieces. What is the probability that in a 9-hour day more than 20,000 pieces will be sold?

C.4 Starting salaries for economics majors have a mean of \$47,000 and a standard deviation of \$8,000. What is the probability that a random sample of 40 economics majors will have an average salary of more than \$50,000?

C.5 A store manager designs a new accounting system that will be cost-effective if the mean monthly charge account balance is more than \$170. A sample of 400 accounts is randomly selected. The sample mean balance is \$178 and the sample standard deviation is \$65. Can the manager conclude that the new system will be cost-effective?

a. Carry out a hypothesis test to answer this question. Use the $\alpha = 0.05$ level of significance.
b. Compute the p-value of the test.

C.6 An econometric professor's rule of thumb is that students should expect to spend 2 hours outside of class on coursework for each hour in class. For a three-hour-per-week class, this means that students are expected to do 6 hours of work outside class. The professor randomly selects eight students from a class, and asks how many hours they studied econometrics during the past week. The sample values are 1, 3, 4, 4, 6, 6, 8, 12.

a. Assuming that the population is normally distributed, can the professor conclude at the 0.05 level of significance that the students are studying on average more than 6 hours per week?
b. Construct a 90% confidence interval for the population mean number of hours studied per week.

C.7 Modern labor practices attempt to keep labor costs low by hiring and laying off workers to meet demand. Newly hired workers are not as productive as experienced ones. Assume that assembly line workers with experience handle 500 pieces per day. A manager concludes that it is cost-effective to maintain the current practice if new hires, with a week of training, can process more than 450 pieces per day. A random sample of $N = 50$ trainees is observed. Let Y_i denote the number of pieces each handles on a randomly selected day. The sample mean is $\bar{y} = 460$, and the estimated sample standard deviation is $\hat{\sigma} = 38$.

a. Carry out a test of whether or not there is evidence to support the conjecture that current hiring procedures are effective, at the 5% level of significance. Pay careful attention when formulating the null and alternative hypotheses.

b. What exactly would a Type I error be in this example? Would it be a costly one to make?
c. Compute the p-value for this test.

C.8 To evaluate alternative retirement benefit packages for its employees, a large corporation must determine the mean age of its workforce. Assume that the age of its employees is normally distributed. Since the corporation has thousands of workers, a sample is to be taken. If the standard deviation of ages is known to be $\sigma = 21$ years, how large should the sample be to ensure that a 95% interval estimate of mean age is no more than four years wide?

C.9 Consider the discrete random variable Y that takes the values $y = 1, 2, 3,$ and 4 with probabilities 0.1, 0.2, 0.3, and 0.4, respectively.

a. Sketch this *pdf*.
b. Find the expected value of Y.
c. Find the variance of Y.
d. If we take a random sample of size $N = 3$ from this distribution, what are the mean and variance of the sample mean, $\overline{Y} = \left(y_1 + y_2 + y_3\right)/3$?

C.10 The sample proportion $\hat{p}$ is an estimator of the population proportion p. The variance of the estimator $\hat{p}$ is $\text{var}(\hat{p}) = p(1-p)/N$, where N is the sample size. Suppose we sample $N = 100$ voters. Of the 100 people sampled, 54 preferred candidate Hillary to candidate Donald.

a. Construct a 95% interval estimate of the population proportion using the approximately correct critical value 1.96 and the estimated variance $\widehat{\text{var}}(\hat{p}) = \hat{p}(1-\hat{p})/N$.
b. Calculate the alternative variance estimate, $\widetilde{\text{var}}(\hat{p}) = 0.5(1-0.5)/N$. Is this variance estimate larger or smaller than the one in part (a)? Will using the alternative variance make for a more conservative, wider, interval estimate or a less conservative, narrower, one?
c. Repeat the calculation of the interval estimate using the alternative variance estimate from part (b) and using the easier to work with critical value 2.0. Is it correct to say that this interval estimate has "a margin of error approximately equal to plus or minus 10 percent?"
d. Define the rough and conservative "margin of error" for the sample proportion interval to be $2\left[0.5(1-0.5)/N\right]^{1/2}$. Calculate the sample size required so that the margin of error is 0.07. What sample sizes are required for 0.05, 0.03, and 0.01 margins of error?
e. A February, 2017, Gallup poll on NAFTA (North American Free Trade Agreement) resulted in 48% saying it "has been a good thing." The poll was based on telephone interviews conducted Feb. 1-5, 2017, with a random sample of 1,035 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. Construct a conservative interval estimate of the true proportion of the 18 or older population thinking NAFTA has been a good thing. A news report based on the poll said "U.S. voters are deeply divided" on NAFTA. Do you think that is a fair statement? [One disheartening comment on the article by a reader said "I don't trust Poles."]

C.11 Let X denote the birthweight of a child, measured in hundreds of grams, whose mother did not smoke. Using a sample of $N = 968$ newly born children, we find the sample mean birthweight to be $\overline{X} = 34.2514$ hundred grams. Also $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 = 33296.003$, $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^3 = -137910.04$, $\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^4 = 6392783.3$

a. Use these values to compute the sample variance, as shown in (C.7) and the sample standard deviation, as shown in (C.9).
b. Use these values to compute $\tilde{\mu}_2$, $\tilde{\mu}_3$, $\tilde{\mu}_4$, as shown in Section C.4.2.
c. Calculate the skewness (S) and kurtosis (K) coefficients given in Section C.4.2. Are the values compatible with the normal distribution?
d. Test the normality of the data using the Jarque–Bera test in Section C.7.4.

C.12 Let Y denote the number of doctor visits in one month by a randomly chosen person. Assume that this count variable has a Poisson distribution with $E(Y) = \text{var}(Y) = \lambda$.

a. Calculate the probabilities $P(Y = 0)$, $P(Y = 1)$, and $P(Y = 2)$ assuming $\lambda = 1$.
b. We choose a random sample of $N = 3$ individuals and observe that the first and second people had two doctor visits, and the third person had one. Calculate the joint probability $P\left(Y_1 = 2, Y_2 = 2, Y_3 = 1\right)$ given that $\lambda = 1$.
c. Show that in general $P\left(Y_1 = 2, Y_2 = 2, Y_3 = 1 \middle| \lambda\right) = 0.25\lambda^5 e^{-3\lambda}$.

d. The likelihood function is $L(\lambda|Y_1 = 2, Y_2 = 2, Y_3 = 1) = 0.25\lambda^5 e^{-3\lambda}$. Write down the algebraic form of the log-likelihood, $\ln L(\lambda|Y_1 = 2, Y_2 = 2, Y_3 = 1)$.

e. Find the first derivative of the log-likelihood, set it to zero, and solve for the solution value $\tilde{\lambda}$.

f. Find the second derivative of the log-likelihood. Determine the sign of this derivative.

g. Can we claim that $\tilde{\lambda}$ is the maximum likelihood estimate of $E(Y) = \text{var}(Y) = \lambda$?

C.13 This exercise extends Exercise C.12 to the general case with a random sample of N observations, $Y_1, \ldots, Y_N$ from a population. Each outcome is assumed to have a Poisson distribution with $E(Y) = \text{var}(Y) = \lambda$.

a. Show that the log-likelihood function is $\ln L(\lambda|y_1, \ldots, y_N) = (\ln \lambda) \sum_{i=1}^N y_i - N\lambda - \sum_{i=1}^N \ln(y_i!)$.

b. Show that the maximum likelihood estimate is $\tilde{\lambda} = \sum_{i=1}^N y_i / N$.

c. Show that the second derivative of the log-likelihood function is $-\left(\sum_{i=1}^N y_i\right)/\lambda^2$. What is the sign of the second derivative?

d. The maximum likelihood estimator is $\tilde{\lambda} = \sum_{i=1}^N Y_i / N$. Assuming we have a random sample from a population with $E(Y) = \text{var}(Y) = \lambda$, find $E(\tilde{\lambda})$ and $\text{var}(\tilde{\lambda})$.

e. The information measure $I(\lambda) = -\left\{E\left[\frac{d^2 \ln L(\lambda)}{d\lambda^2}\right]\right\}$, where $\left[\frac{d^2 \ln L(\lambda)}{d\lambda^2}\right] = -\left(\sum_{i=1}^N Y_i\right)/\lambda^2$. Show that the information measure in this case is $I(\lambda) = N/\lambda$.

C.14 Let X denote the birthweight of a child, measured in hundreds of grams. Consider children whose mothers smoked ($SMOKE = 1$) and children whose mothers did not smoke ($SMOKE = 0$). Summary statistics for the birthweights for these two groups are in Table C.6.

TABLE C.6 Summary Statistics for Birthweights

SMOKE	*N*	Mean	Variance	Std. Dev.	Skewness	Kurtosis
0	968	34.25	34.43	5.87	−0.71	5.58
1	232	31.37	34.42	5.87	−1.26	7.66

a. Use the Jarque–Bera test to test the normality of each of these populations. Do we reject the null hypothesis of normality or fail to reject normality?

b. Construct a 95% interval estimate for the population mean birthweight born to mothers who did not smoke, μ_0. Construct a 95% interval estimate for the population mean birthweight born to mothers who did smoke, μ_1. Select any value c in the 95% interval estimate for μ_0. What is the outcome of a two-tail test of the hypothesis $\mu_1 = c$ using the 5% level of significance?

c. Test the null hypothesis that the population mean birthweight is the same for the two populations, $H_0: \mu_0 = \mu_1$ against the alternative $H_1: \mu_0 \neq \mu_1$. Explain your choice to use a pooled variance estimator or to assume that the pooled variance is inappropriate. Use the 5% level of significance.

d. Repeat the test in part (c) for the null and alternative hypotheses $H_0: \mu_0 \leq \mu_1$ and $H_1: \mu_0 > \mu_1$.

C.15 In this exercise we use the data from Exercise C.12 and the results in Exercise C.13 to carry out a hypothesis test concerning the parameter λ in the Poisson distribution.

a. Using the maximum likelihood estimate from Exercise C.12, compute the information measure $I(\tilde{\lambda})$, given in Exercise C.13 (e).

b. Carry out a likelihood ratio test of the null hypothesis $H_0: \lambda = 1$, using the test statistic in equation (C.25), versus the alternative $H_1: \lambda \neq 1$ at the 5% level of significance.

c. Use the Wald statistic in equation (C.26) to carry out the test from part (b).

d. An alternative version of the Wald statistic replaces the second derivative term, $-d^2 \ln L(\lambda)/d\lambda^2$, with $I(\tilde{\lambda})$, as shown in equation (C.28). Carry out the test from part (b) using the modified Wald test.

e. Evaluate the score function, shown in equation (C.31), assuming the null hypothesis is true.

f. Evaluate the information measure $I(\lambda)$ assuming the null hypothesis is true.

g. Using the results in parts (e) and (f), carry out the LM test of the null hypothesis in part (b).

C.16 Two independent food scientists are researching the shelf-life (Y) of "Bill's Big Red" spaghetti sauce. The first collects a random sample of 25 jars and finds their average shelf life to be $\overline{Y} = 48$ months. The second researcher collects a random sample of 100 jars and finds their average shelf life to be $\overline{Y}_2 = 40$ months.

a. Find the ratio of the standard error of $\overline{Y}_1$ relative to the standard error of $\overline{Y}_2$.

b. A combined estimate can be obtained by finding the weighted average $\tilde{Y} = c\overline{Y}_1 + (1-c)\overline{Y}_2$. Is there any value of c that makes this estimator of μ unbiased?

c. What value of c yields the combined estimate with the smallest standard error? Explain the intuition behind your solution, and why weighting the two means equally, with $c = 0.5$, is not the best choice.

C.17 Suppose school children are subjected to a standardized math test each spring. In the population of comparable children, the test score Y is normally distributed with mean 500 and standard deviation 100, $Y \sim N(\mu = 500, \sigma^2 = 100^2)$. It is claimed that reducing class sample size will increase test scores.

a. How can we tell if reducing class size actually does increase test scores? Would you be convinced if a sample of $N = 25$ students from the smaller classes had an average test score of 510? Calculate the probability of obtaining a sample mean of $\overline{Y} = 510$, or more, even if smaller classes actually have no effect on test performance.

b. Show that a class average of 533 will be reached by chance only 5% of the time, if the smaller class sizes have no effect. Is the following statement correct or incorrect? "We can conclude that smaller classes raise average test scores if a class of 25 students has an average test score of 533 or better, with this result being due to sampling error with probability 5%."

c. Suppose that smaller classes actually do improve the average mean population test score to 550. What is the probability of observing a class of 25 with an average score of 533 or better? If our objective is to determine whether smaller classes increase test scores, is it better for this number to be larger or smaller?

d. If smaller classes increase average test score to 550, what is the probability of having a small class average of less than or equal to 533?

e. Draw a figure showing two normal distributions, one with mean 500 and standard deviation 100, and the other with mean 550 and standard deviation 100. On the figure locate the value 533. In part (b) we showed that if the change in class size has no effect on test scores, we would still obtain a class average of 533 or more by chance 5% of the time; we would incorrectly conclude that the smaller classes helped test scores, which is a Type I error. In part (d) we derived the probability that we would obtain a class average test score of less than 533, making us unable to conclude that smaller classes help, even though smaller classes did help. This is a Type II error. If we push the threshold to the right, say 540, what happens to Type I and Type II errors? If we push the threshold to the left, say 530, what happens to the probability of Type I and Type II errors?

C.11.2 Computer Exercises

C.18 Does being in a small class help primary school students learning, and performance on achievement tests? Use the sample data file *star5_small* to explore this question.

a. Consider students in regular-sized classes, with *REGULAR* = 1. Construct a histogram of *MATHSCORE*. Carry out the Jarque–Bera test for normality at the 5% level of significance. What do you conclude about the normality of the data?

b. Calculate the sample mean, standard deviation and standard error of the mean for *MATHSCORE* in regular-sized classes. Use the t-statistic in equation (C.16) to test the null hypothesis that the population mean (the population of students who are enrolled in regular-sized classes) μ_R is 490 versus the alternative that it isn't. Use the 5% level of significance. What is your conclusion?

c. Given the result of the normality test in (a), do you think the test in part (b) is justifiable? Explain your reasoning.

d. Construct a 95% interval estimate for the mean μ_R.

e. Repeat the test in (b) for the population of students in small classes, *SMALL* = 1. Denote the population mean for these students as μ_S. Use the 5% level of significance. What is your conclusion?

f. Let μ_R and μ_S denote the population mean test scores on the math achievement test, *MATHSCORE*. Using the appropriate test, outlined in Section C.7.2, test the null hypothesis $H_0: \mu_S - \mu_R \leq 0$ against the alternative $H_1: \mu_S - \mu_R > 0$. Use the 1% level of significance. Does it appear that being in a small class increases the expected math test score, or not?

C.19 Does having a household member with an advanced degree increase household income relative to a household that includes a member having only a college degree? Use the sample data file *cex5_small* to explore this question.

- **a.** Construct a histogram of incomes for households that include a member with an advanced degree. Construct another histogram of incomes for households that include a member with a college degree. What do you observe about the shape and location of these two histograms?
- **b.** In the sample that includes a member with an advanced degree, what percentage of households have household incomes greater than $10,000 per month? What is the percentage for households that include a member having a college degree?
- **c.** Test the null hypothesis that the population mean income for households including a member with an advanced degree, μ_{ADV}, is less than, or equal to, $9,000 per month against the alternative that it is greater than $9,000 per month. Use the 5% level of significance.
- **d.** Test the null hypothesis that the population mean income for households including a member with a college degree, μ_{COLL}, is less than, or equal to, $9,000 per month against the alternative that it is greater than $9,000 per month. Use the 5% level of significance.
- **e.** Construct 95% interval estimates for μ_{ADV} and μ_{COLL}.
- **f.** Test the null hypothesis $\mu_{ADV} \leq \mu_{COLL}$ against the alternative $\mu_{ADV} > \mu_{COLL}$. Use the 5% level of significance. What is your conclusion?

C.20 How much variation is there in household incomes in households including a member with an advanced degree? Use the sample data file *cex5_small* to explore this question. Let σ^2_{ADV} denote the population variance.

- **a.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} > 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region. What is the ***p*-value** for this test?
- **b.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} < 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region. What is the *p*-value for this test?
- **c.** Test the null hypothesis $\sigma^2_{ADV} = 2500$ against the alternative $\sigma^2_{ADV} \neq 2500$. Use the 5% level of significance. Clearly state the test statistic and the rejection region.

C.21 School officials consider performance on a standardized math test acceptable if 40% of the population of students score at least 500 points. Use the sample data file *star5_small* to explore this topic.

- **a.** Compute the sample proportion of students enrolled in regular-sized classes who score 500 points or more. Calculate a 95% interval estimate of the population proportion. Based on this interval can we reject the null hypothesis that the population proportion of students in regular-sized classes who score 500 points or better is $p = 0.4$?
- **b.** Test the null hypothesis that the population proportion p of students in a regular-sized class who score 500 points or more is less than or equal to 0.4 against the alternative that the true proportion is greater than 0.4. Use the 5% level of significance.
- **c.** Test the null hypothesis that the population proportion p of students in a regular-sized class who score 500 points or more is equal to 0.4 against the alternative that the true proportion is less than 0.4. Use the 5% level of significance.
- **d.** Repeat parts (a)–(c) for students in small classes.

C.22 Consider two populations of Chinese chemical firms: those who export their products and those who do not. Let us consider the sales revenue for these two types of firms. Use the data file *chemical_small* for this exercise. It contains data on 1200 firms in 2006.

- **a.** The variable *LSALES* is ln(*SALES*). Construct a histogram for this variable and test whether the data are normally distributed using the Jarque–Bera test with 10% level of significance.
- **b.** Create the variable $SALES = \exp(LSALES)$. Construct a histogram for this variable and test whether the data are normally distributed using the Jarque–Bera test with 10% level of significance.
- **c.** Consider two populations of firms: those who export ($EXPORT = 1$) and those who do not ($EXPORT = 0$). Let μ_1 be the population mean of *LSALES* for firms that export, and let μ_0 be the population mean of *LSALES* for firms that do not export. Estimate the difference in means $\mu_1 - \mu_0$ and interpret this value. [*Hint*: Use the properties of differences in log-variables.]
- **d.** Test the hypothesis that the means of these two populations are equal. Use the test that assumes the population variances are unequal. What do you conclude?

C.23 Does additional education have as large a payoff for females as males? Use the data file *cps5* to explore this question. If your software does not permit using this larger sample use *cps5_small*.

- **a.** Calculate the sample mean wage of females who have 12 years of education. Calculate the sample mean wage of females with 16 years of education. What did you discover?
- **b.** Calculate a 95% interval estimate for the population mean wage of females with 12 years of education. Repeat the calculation for the wages of females with 16 years of education. Do the intervals overlap?
- **c.** Calculate the sample mean wage of males who have 12 years of education. Calculate the sample mean wage of males with 16 years of education. What did you discover? How does the difference in wages for males compare to the difference of wages for females in part (a)?
- **d.** Calculate a 95% interval estimate for the population mean wage of males with 12 years of education. Repeat the calculation for the wages of males with 16 years of education. Does the interval for males with 12 years of education overlap with the comparable interval for females? Does the interval for males with 16 years of education overlap with the comparable interval for females?
- **e.** Denote the population means of interest by μ_{F16}, μ_{F12}, μ_{M16}, μ_{M12} where F and M denote female and male, and 12 and 16 denote years of education. Estimate the parameter $\theta = (\mu_{F16} - \mu_{F12}) - (\mu_{M16} - \mu_{M12})$ by replacing population means by sample means.
- **f.** Calculate a 95% interval estimate of θ. Based on the interval estimate, what can you say about the benefits of the addition of four years of education for males versus females? Use the 97.5 percentile from the standard normal, 1.96, when calculating the interval estimate.

C.24 How much does the variation in wages change when individuals receive more education? Is the variation different for males and females? Use the data file *cps5* to explore this question. If your software does not permit using this larger sample use *cps5_small*.

- **a.** Calculate the sample variance of wages of females who have 12 years of education. Calculate the sample variance of wages of females who have 18 years of education. What did you discover?
- **b.** Carry out a two-tail test, using a 5% level of significance, of the hypothesis that the variance of wage is the same for females with 12 years of education and females with 18 years of education.
- **c.** Calculate the sample variance of wages of males who have 12 years of education. Calculate the sample variance of wages of males who have 18 years of education. What did you discover?
- **d.** Carry out a two-tail test, using a 5% level of significance, of the hypothesis that the variance of wage is the same for males with 12 years of education and males with 18 years of education.
- **e.** Carry out a two-tail test of the null hypothesis that the mean wage for males with 18 years of education is the same as the mean wage of females with 18 years of education. Use the 1% level of significance.

C.25 What happens to the household budget share of necessity items, like food, when total household expenditures increase? Use data file *malwai_small* for this exercise.

- **a.** Obtain the summary statistics, including the median and 90th percentile, of total household expenditures.
- **b.** Construct a 95% interval estimate for the proportion of income spent on food by households with total expenditures less than or equal to the median.
- **c.** Construct a 95% interval estimate for the proportion of income spent on food by households with total expenditures more than or equal to the 90th percentile.
- **d.** Summarize your findings from parts (b) and (c).
- **e.** Test the null hypothesis that the population mean proportion of income spent on food by households is 0.4. Use a two-tail test and the 5% level of significance. Carry out the test separately using the complete sample, and using the samples of households with total expenditures less than or equal to the median, and again for households whose total expenditures are in the top 10%.

C.26 At the famous Fulton Fish Market in New York City sales of Whiting (a type of fish) vary from day to day. Over a period of several months, daily quantities sold (in pounds) were observed. These data are in the data file *fultonfish*.

- **a.** Using the data for Monday sales, test the null hypothesis that the mean quantity sold is greater than or equal to 10,000 pounds a day, against the alternative that the mean quantity sold is less than 10,000 pounds. Use the $\alpha = 0.05$ level of significance. Be sure to (i) state the null and alternative hypotheses, (ii) give the test statistic and its distribution, (iii) indicate the rejection region, including

a sketch, (iv) state your conclusion, and (v) calculate the *p*-value for the test. Include a sketch showing the *p*-value.

b. Assume that daily sales on Tuesday (X_2) and Wednesday (X_3) are normally distributed with means μ_2 and μ_3, and variances σ_2^2 and σ_3^2, respectively. Assume that sales on Tuesday and Wednesday are independent of each other. Test the hypothesis that the variances σ_2^2 and σ_3^2 are equal against the alternative that the variance on Tuesday is larger. Use the $\alpha = 0.05$ level of significance. Be sure to (i) state the null and alternative hypotheses, (ii) give the test statistic and its distribution, (iii) indicate the rejection region, including a sketch, (iv) state your conclusion, and (v) calculate the *p*-value for the test. Include a sketch showing the *p*-value.

c. We wish to test the hypothesis that mean daily sales on Tuesday and Wednesday are equal against the alternative that they are not equal. Using the result in part (b) as a guide to the appropriate version of the test (Appendix C.7), carry out this hypothesis test using the 5% level of significance.

d. Let the daily sales for Monday, Tuesday, Wednesday, Thursday, and Friday be denoted as X_1, X_2, X_3, X_4, and X_5, respectively. Assume that $X_i \sim NID(\mu_i, \sigma_i^2)$. Define total weekly sales as $W = X_1 + X_2 + X_3 + X_4 + X_5$. Derive the expected value and variance of W, using appropriate theorems about normal distributions. Be sure to show your work and justify your answer.

e. Referring to part (d), let $E(W) = \mu$. Assume that we estimate μ using

$$\hat{\mu} = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \overline{X}_4 + \overline{X}_5$$

where $\overline{X}_i$ is the sample mean for the *i*th day. Derive the probability distribution of $\hat{\mu}$ and construct an approximate (valid in large samples) 95% interval estimate for μ. Justify the validity of your interval estimator.

C.27 A **credit score** is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of the person. A credit score is primarily based on a credit report information typically sourced from credit bureaus. Use the data file *lasvegas* for this exercise.

a. Construct a histogram for the variable *CREDIT*. Does the histogram look symmetrical and "bell-shaped?" Test the normality of the variable *CREDIT* using the Jarque–Bera test and level of significance 5%.

b. Let two populations of *CREDIT* be defined by those who were delinquent (*DELINQUENT* = 1) and those who were not delinquent (*DELINQUENT* = 0). Using the test described in Section C.7.3, carry out a test of the hypothesis that the variances in these two populations are equal against the alternative that they are not equal. Use the 5% level of significance.

c. Use the appropriate one-tail test in Section C.7.2, based on your answer in part (b), to test the equality of *CREDIT* means for the two populations.

d. Using the test in Section C.7.1, test the null hypothesis that the variance of the population who was not delinquent is 3600 versus the alternative that it is not 3600.

C.28 Is it true that more capable individuals ultimately attain more years of schooling? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979.

a. In the data the variable *SCORE* is an index based on 10 aptitude/IQ tests given in 1980. We can loosely use this variable as some measure of ability. Construct a histogram of *SCORE*. What are the sample mean and the standard deviation?

b. The variable *EDUC* is the individual's years of schooling completed by 1993. What percentage of the men had completed at least 12 years of education by 1993?

c. Calculate the sample mean number of years of schooling completed by men with *SCORE* greater than or equal to zero. Calculate the sample mean number of years of schooling completed by men with *SCORE* less than zero. Test the null hypothesis that the population of men with $SCORE \geq 0$ have mean years of education, μ_1, that is greater than the mean number of years of education, μ_0, for those with lower scores. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

d. Some of the men came from broken homes, as indicated by the variable *BROKEN*. Test the null hypothesis that the population of men from broken homes have mean years of education, μ_1, that is less than the mean number of years of education, μ_0, for those who were not from broken homes. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

C.29 Do more highly educated parents tend to have more educated children? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979.

a. The variable *EDUC* is the individual's years of schooling completed by 1993. What percentage of the men had completed at least 16 years of education by 1993? What percentage of the men's mothers had at least 16 years of education? What percentage of fathers had at least 16 years of education?

b. Calculate the sample mean number of years of schooling completed by men with fathers who had 16 or more years of education. Calculate the sample mean number of years of schooling completed by men with fathers who had less than 16 years of education. Test the null hypothesis that the population of men with more educated fathers have mean years of education, μ_1, that is greater than the mean number of years of education, μ_0, for those with less educated fathers. State the null and alternative hypothesis, give the test statistic, and your conclusion using a 5% level of significance.

c. Investigate the question of whether more highly educated men, those with more than 12 years of schooling, tend to marry more highly educated women, those with more than 12 years of schooling. State the null and alternative hypotheses, give the test statistic, and your conclusion using a 5% level of significance.

C.30 Do households with more children tend to result in more broken homes? Use the data file *koop_tobias_87* to study this question. The data file includes 1987 information on males who were between 14 and 22 years of age in 1979. It includes the number of siblings the man had as well as whether he came from a broken home.

a. Create the variable $KIDS = SIBS + 1$. To simplify the following arithmetic, let $KIDS = 3$ if the number of household children is equal to 3 or more. The variable *KIDS* takes the values 1, 2, and 3. Calculate the number of men who came from families with $KIDS = 1$, and $KIDS = 2$, and $KIDS = 3$.

b. Calculate the number of households that were broken having 1, 2, or 3 children. Calculate the number of households that were not broken with $KIDS = 1$, $KIDS = 2$, and $KIDS = 3$.

c. The famous statistician Karl Pearson developed a test for the null hypothesis that two characteristics are unrelated versus the alternative that they are related. If the number of children and broken homes are unrelated, we should expect 176.167 of 1057 households with each of the six possible outcomes. Pearson's chi-square test is calculated as

$$PEARSON = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_6 - E_6)^2}{E_6}$$

where E_i is the "expected" number of outcomes and O_i is the "observed" number of outcomes for each of six outcomes. If there is no relation between the variables the test statistic has a $\chi^2_{(m)}$ distribution, with $m = (c_1 - 1) \times (c_2 - 1)$ degrees of freedom, where c_1 is the number of categories for variable 1 and c_2 is the number of categories for variable 2. The null hypothesis that the variables are unrelated is rejected if the value of *PEARSON* is greater than the $100(1-\alpha)$-percentile from the chi-square distribution. Carry out Pearson's test for the existence of a relationship between *BROKEN* and *KIDS* at the 5% level.

d. Explore your software. Does it have a command to automatically create two-way tables of frequencies? Does it have a command to calculate Pearson's chi-square statistic? If so, carry out the test in part (c) *without* modifying the variable *KIDS* to have only three outcomes. Report the two-way table and the test result.

Appendix D

Statistical Tables

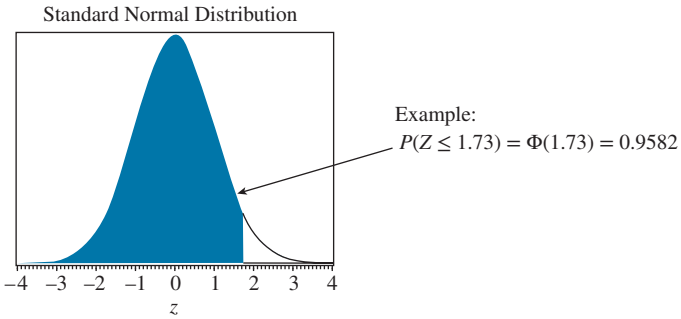


TABLE D.1 **Cumulative Probabilities for the Standard Normal Distribution $\Phi(z) = P(Z \leq z)$**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Source: This table was generated using the SAS® function PROBNORM.

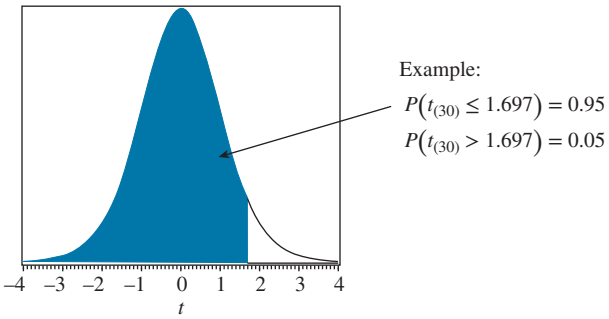


TABLE D.2 **Percentiles of the *t*-distribution**

df	$t_{(0.90,\,df)}$	$t_{(0.95,\,df)}$	$t_{(0.975,\,df)}$	$t_{(0.99,\,df)}$	$t_{(0.995,\,df)}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
36	1.306	1.688	2.028	2.434	2.719
37	1.305	1.687	2.026	2.431	2.715
38	1.304	1.686	2.024	2.429	2.712
39	1.304	1.685	2.023	2.426	2.708
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
∞	1.282	1.645	1.960	2.326	2.576

Source: This table was generated using the SAS® function TINV.

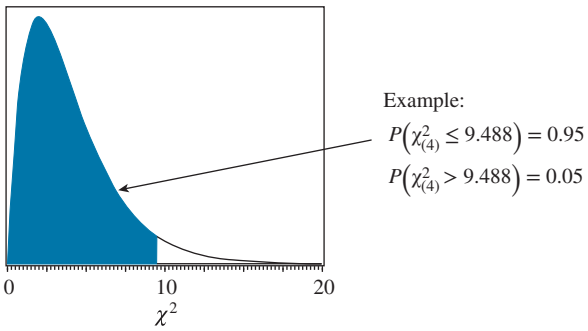
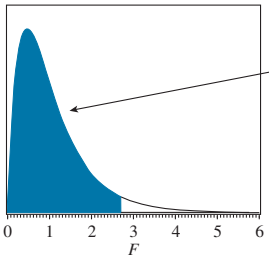


TABLE D.3 **Percentiles of the Chi-square Distribution**

df	$\chi^2_{(0.90,\,df)}$	$\chi^2_{(0.95,\,df)}$	$\chi^2_{(0.975,\,df)}$	$\chi^2_{(0.99,\,df)}$	$\chi^2_{(0.995,\,df)}$
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672
35	46.059	49.802	53.203	57.342	60.275
40	51.805	55.758	59.342	63.691	66.766
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.379	91.952
70	85.527	90.531	95.023	100.425	104.215
80	96.578	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169
110	129.385	135.480	140.917	147.414	151.948
120	140.233	146.567	152.211	158.950	163.648

Source: This table was generated using the SAS® function CINV.



Example:

$P(F_{(4,30)} \leq 2.69) = 0.95$

$P(F_{(4,30)} > 2.69) = 0.05$

TABLE D.4 **95th Percentile for the *F*-distribution**

v_2/v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.10	252.20	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.48	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.62	8.57	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.75	5.69	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.50	4.43	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.81	3.74	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.38	3.30	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.08	3.01	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.86	2.79	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.70	2.62	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.25	2.16	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.04	1.95	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.92	1.82	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.84	1.74	1.62
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.04	1.96	1.88	1.79	1.68	1.56
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.74	1.64	1.51
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	1.97	1.89	1.81	1.71	1.60	1.47
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.87	1.78	1.69	1.58	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.65	1.53	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.55	1.43	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.46	1.32	1.00

Source: This table was generated using the SAS® function FINV.

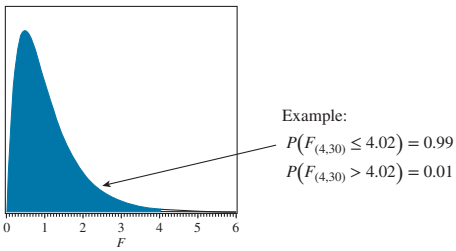


TABLE D.5 **99th Percentile for the *F*-distribution**

ν_2/ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6106.32	6157.28	6208.73	6260.65	6313.03	6365.87
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.47	99.48	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.50	26.32	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.84	13.65	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.38	9.20	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.23	7.06	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.99	5.82	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.20	5.03	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.65	4.48	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.25	4.08	3.91
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.21	3.05	2.87
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.78	2.61	2.42
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.54	2.36	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.39	2.21	2.01
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88	2.74	2.60	2.44	2.28	2.10	1.89
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.20	2.02	1.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74	2.61	2.46	2.31	2.14	1.96	1.74
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.56	2.42	2.27	2.10	1.91	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.03	1.84	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.86	1.66	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.70	1.47	1.00

Source: This table was generated using the SAS® function FINV.

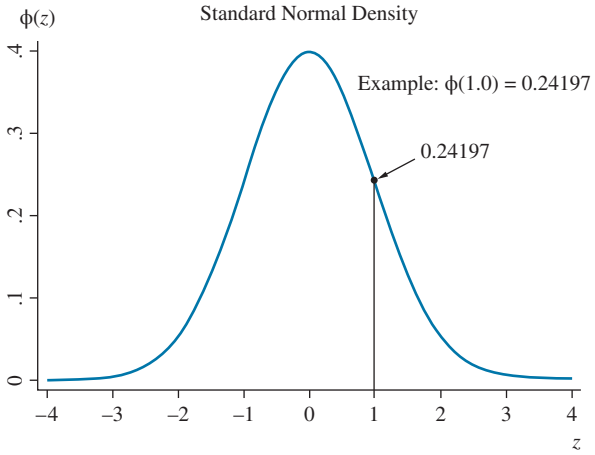


TABLE D.6 **Standard Normal pdf Values $\phi(z)$**

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.39894	0.39892	0.39886	0.39876	0.39862	0.39844	0.39822	0.39797	0.39767	0.39733
0.1	0.39695	0.39654	0.39608	0.39559	0.39505	0.39448	0.39387	0.39322	0.39253	0.39181
0.2	0.39104	0.39024	0.38940	0.38853	0.38762	0.38667	0.38568	0.38466	0.38361	0.38251
0.3	0.38139	0.38023	0.37903	0.37780	0.37654	0.37524	0.37391	0.37255	0.37115	0.36973
0.4	0.36827	0.36678	0.36526	0.36371	0.36213	0.36053	0.35889	0.35723	0.35553	0.35381
0.5	0.35207	0.35029	0.34849	0.34667	0.34482	0.34294	0.34105	0.33912	0.33718	0.33521
0.6	0.33322	0.33121	0.32918	0.32713	0.32506	0.32297	0.32086	0.31874	0.31659	0.31443
0.7	0.31225	0.31006	0.30785	0.30563	0.30339	0.30114	0.29887	0.29659	0.29431	0.29200
0.8	0.28969	0.28737	0.28504	0.28269	0.28034	0.27798	0.27562	0.27324	0.27086	0.26848
0.9	0.26609	0.26369	0.26129	0.25888	0.25647	0.25406	0.25164	0.24923	0.24681	0.24439
1.0	0.24197	0.23955	0.23713	0.23471	0.23230	0.22988	0.22747	0.22506	0.22265	0.22025
1.1	0.21785	0.21546	0.21307	0.21069	0.20831	0.20594	0.20357	0.20121	0.19886	0.19652
1.2	0.19419	0.19186	0.18954	0.18724	0.18494	0.18265	0.18037	0.17810	0.17585	0.17360
1.3	0.17137	0.16915	0.16694	0.16474	0.16256	0.16038	0.15822	0.15608	0.15395	0.15183
1.4	0.14973	0.14764	0.14556	0.14350	0.14146	0.13943	0.13742	0.13542	0.13344	0.13147
1.5	0.12952	0.12758	0.12566	0.12376	0.12188	0.12001	0.11816	0.11632	0.11450	0.11270
1.6	0.11092	0.10915	0.10741	0.10567	0.10396	0.10226	0.10059	0.09893	0.09728	0.09566
1.7	0.09405	0.09246	0.09089	0.08933	0.08780	0.08628	0.08478	0.08329	0.08183	0.08038
1.8	0.07895	0.07754	0.07614	0.07477	0.07341	0.07206	0.07074	0.06943	0.06814	0.06687
1.9	0.06562	0.06438	0.06316	0.06195	0.06077	0.05959	0.05844	0.05730	0.05618	0.05508
2.0	0.05399	0.05292	0.05186	0.05082	0.04980	0.04879	0.04780	0.04682	0.04586	0.04491
2.1	0.04398	0.04307	0.04217	0.04128	0.04041	0.03955	0.03871	0.03788	0.03706	0.03626
2.2	0.03547	0.03470	0.03394	0.03319	0.03246	0.03174	0.03103	0.03034	0.02965	0.02898
2.3	0.02833	0.02768	0.02705	0.02643	0.02582	0.02522	0.02463	0.02406	0.02349	0.02294
2.4	0.02239	0.02186	0.02134	0.02083	0.02033	0.01984	0.01936	0.01888	0.01842	0.01797
2.5	0.01753	0.01709	0.01667	0.01625	0.01585	0.01545	0.01506	0.01468	0.01431	0.01394
2.6	0.01358	0.01323	0.01289	0.01256	0.01223	0.01191	0.01160	0.01130	0.01100	0.01071
2.7	0.01042	0.01014	0.00987	0.00961	0.00935	0.00909	0.00885	0.00861	0.00837	0.00814
2.8	0.00792	0.00770	0.00748	0.00727	0.00707	0.00687	0.00668	0.00649	0.00631	0.00613
2.9	0.00595	0.00578	0.00562	0.00545	0.00530	0.00514	0.00499	0.00485	0.00470	0.00457
3.0	0.00443	0.00430	0.00417	0.00405	0.00393	0.00381	0.00370	0.00358	0.00348	0.00337

Source: This table was generated using the SAS® function PDF("normal," z).

Index

D

E

F

M

Q

R

S

T

U

V

W

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.