Appendix A

A Review of Some Statistical Concepts

This appendix provides a very sketchy introduction to some of the statistical concepts encountered in the text. The discussion is nonrigorous, and no proofs are given because several excellent books on statistics do that job very well. Some of these books are listed at the end of this appendix.

A.1 Summation and Product Operators

The Greek capital letter $\sum$ (sigma) is used to indicate summation. Thus,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Some of the important properties of the summation operator $\sum$ are

1. $\displaystyle\sum_{i=1}^{n} k = nk$, where k is constant. Thus, $\sum_{i=1}^{4} 3 = 4 \cdot 3 = 12$.
2. $\sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i$, where k is a constant.
3. $\sum_{i=1}^{n}(a + bx_i) = na + b\sum_{i=1}^{n} x_i$, where a and b are constants and where use is made of properties 1 and 2 above.
4. $\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$.

The summation operator can also be extended to multiple sums. Thus, $\sum\sum$, the double summation operator, is defined as

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} &= \sum_{i=1}^{n}(x_{i1} + x_{i2} + \cdots + x_{im})\\ &= (x_{11} + x_{21} + \cdots + x_{n1}) + (x_{12} + x_{22} + \cdots + x_{n2})\\ &\quad + \cdots + (x_{1m} + x_{2m} + \cdots + x_{nm})\end{aligned}$$

Some of the properties of $\sum\sum$ are

1. $\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}$; that is, the order in which the double summation is performed is interchangeable.
2. $\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} y_j$.

3. $\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij}+y_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}$.
4. $\left[\sum_{i=1}^{n}x_i\right]^2 = \sum_{i=1}^{n}x_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}x_ix_j = \sum_{i=1}^{n}x_i^2 + 2\sum_{i<j}x_ix_j$.

The product operator Π is defined as

$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdots x_n$$

Thus,

$$\prod_{i=1}^{3} x_i = x_1 \cdot x_2 \cdot x_3$$

A.2 Sample Space, Sample Points, and Events

The set of all possible outcomes of a random, or chance, experiment is called the **population,** or **sample space,** and each member of this sample space is called a **sample point.** Thus, in the experiment of tossing two coins, the sample space consists of these four possible outcomes: *HH*, *HT*, *TH*, and *TT*, where *HH* means a head on the first toss and also a head on the second toss, *HT* means a head on the first toss and a tail on the second toss, and so on. Each of the preceding occurrences constitutes a sample point.

An **event** is a subset of the sample space. Thus, if we let *A* denote the occurrence of one head and one tail, then, of the preceding possible outcomes, only two belong to *A*, namely *HT* and *TH*. In this case *A* constitutes an event. Similarly, the occurrence of two heads in a toss of two coins is an event. Events are said to be **mutually exclusive** if the occurrence of one event precludes the occurrence of another event. If in the preceding example *HH* occurs, the occurrence of the event *HT* at the same time is not possible. Events are said to be (collectively) **exhaustive** if they exhaust all the possible outcomes of an experiment. Thus, in the example, the events (a) two heads, (b) two tails, and (c) one tail, one head exhaust all the outcomes; hence they are (collectively) exhaustive events.

A.3 Probability and Random Variables

Probability

Let *A* be an event in a sample space. By *P*(*A*), the probability of the event *A*, we mean the proportion of times the event *A* will occur in repeated trials of an experiment. Alternatively, in a total of *n* possible equally likely outcomes of an experiment, if *m* of them are favorable to the occurrence of the event *A*, we define the ratio *m*/*n* as the **relative frequency** of *A*. For large values of *n*, this relative frequency will provide a very good approximation of the probability of *A*.

Properties of Probability

P(*A*) is a real-valued function[1] and has these properties:

1. $0 \leq P(A) \leq 1$ for every A.
2. If $A, B, C, \ldots$ constitute an exhaustive set of events, then $P(A + B + C + \cdots) = 1$, where $A + B + C$ means A or B or C, and so forth.
3. If $A, B, C, \ldots$ are mutually exclusive events, then

$$P(A + B + C + \cdots) = P(A) + P(B) + P(C) + \cdots$$

[1] A function whose domain and range are subsets of real numbers is commonly referred to as a real-valued function. For details, see Alpha C. Chiang, *Fundamental Methods of Mathematical Economics,* 3d ed., McGraw-Hill, 1984, Chapter 2.

EXAMPLE 1

Consider the experiment of throwing a die numbered 1 through 6. The sample space consists of the outcomes 1, 2, 3, 4, 5, and 6. These six events therefore exhaust the entire sample space. The probability of any one of these numbers showing up is 1/6 since there are six equally likely outcomes and any one of them has an equal chance of showing up. Since 1, 2, 3, 4, 5, and 6 form an exhaustive set of events, $P(1+2+3+4+5+6)=1$ where $1, 2, 3, \ldots$ means the probability of number 1 or number 2 or number 3, etc. And since $1, 2, \ldots, 6$ are mutually exclusive events in that two numbers cannot occur simultaneously, $P(1+2+3+4+5+6) = P(1) + P(2) + \cdots + P(6) = 1$.

Random Variables

A variable whose value is determined by the outcome of a chance experiment is called a **random variable** (rv). Random variables are usually denoted by the capital letters X, Y, Z, and so on, and the values taken by them are denoted by small letters x, y, z, and so on.

A random variable may be either **discrete** or **continuous.** A discrete rv takes on only a finite (or countably infinite) number of values.[2] For example, in throwing two dice, each numbered 1 to 6, if we define the random variable X as the sum of the numbers showing on the dice, then X will take one of these values: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12. Hence it is a discrete random variable. A continuous rv, on the other hand, is one that can take on any value in some interval of values. Thus, the height of an individual is a continuous variable—in the range, say, 60 to 65 inches it can take any value, depending on the precision of measurement.

A.4 Probability Density Function (PDF)

Probability Density Function of a Discrete Random Variable

Let X be a discrete rv taking distinct values $x_1, x_2, \ldots, x_n, \ldots$. Then the function

$$\begin{aligned} f(x) &= P(X = x_i) \qquad && \text{for } i = 1, 2, \ldots, n, \ldots \\ &= 0 && \text{for } x \neq x_i \end{aligned}$$

is called the **discrete probability density function** (PDF) of X, where $P(X = x_i)$ means the probability that the discrete rv X takes the value of x_i.

EXAMPLE 2

In a throw of two dice, the random variable X, the sum of the numbers shown on two dice, can take one of the 11 values shown. The PDF of this variable can be shown to be as follows (see also Figure A.1):

$$\begin{array}{lccccccccccc} x = & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ f(x) = & \left(\frac{1}{36}\right) & \left(\frac{2}{36}\right) & \left(\frac{3}{36}\right) & \left(\frac{4}{36}\right) & \left(\frac{5}{36}\right) & \left(\frac{6}{36}\right) & \left(\frac{5}{36}\right) & \left(\frac{4}{36}\right) & \left(\frac{3}{36}\right) & \left(\frac{2}{36}\right) & \left(\frac{1}{36}\right) \end{array}$$

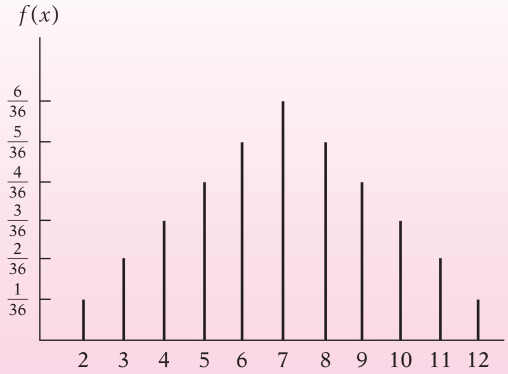
These probabilities can be easily verified. In all there are 36 possible outcomes, of which one is favorable to number 2, two are favorable to number 3 (since the sum 3 can occur either as 1 on the first die and 2 on the second die or 2 on the first die and 1 on the second die), and so on.

(*Continued*)

[2] For a simple discussion of the notion of countably infinite sets, see R. G. D. Allen, *Basic Mathematics,* Macmillan, London, 1964, p. 104.

EXAMPLE 2
(*Continued*)

FIGURE A.1 Density function of the discrete random variable of Example 2.



Probability Density Function of a Continuous Random Variable

Let X be a continuous rv. Then $f(x)$ is said to be the PDF of X if the following conditions are satisfied:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{a}^{b} f(x)\,dx = P(a \leq x \leq b)$$

where $f(x)\,dx$ is known as the *probability element* (the probability associated with a small interval of a continuous variable) and where $P(a \leq x \leq b)$ means the probability that X lies in the interval a to b. Geometrically, we have Figure A.2.

For a continuous rv, in contrast with a discrete rv, the probability that X takes a specific value is zero;[3] probability for such a variable is measurable only over a given range or interval, such as (a, b) shown in Figure A.2.

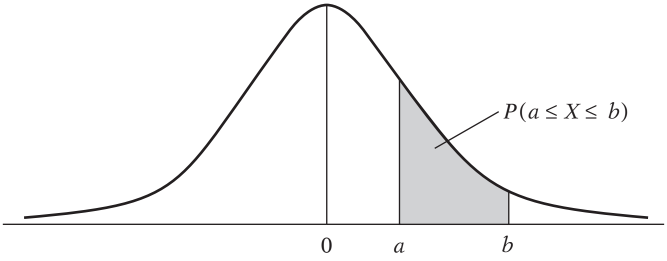
EXAMPLE 3

Consider the following density function:

$$f(x) = \frac{1}{9}x^2 \qquad 0 \leq x \leq 3$$

It can be readily verified that $f(x) \geq 0$ for all x in the range 0 to 3 and that $\int_0^3 \frac{1}{9}x^2 dx = 1$. (*Note:* The integral is $(\frac{1}{27}x^3 \,|_0^3) = 1$.) If we want to evaluate the above PDF between, say, 0 and 1, we obtain $\int_0^1 \frac{1}{9}x^2 dx = (\frac{1}{27}x^3 \,|_0^1) = \frac{1}{27}$; that is, the probability that x lies between 0 and 1 is 1/27.

FIGURE A.2
Density function of a continuous random variable.



[3] *Note:* $\int_a^a f(x)\,dx = 0$.

Joint Probability Density Functions

Discrete Joint PDF

Let X and Y be two discrete random variables. Then the function

$$\begin{aligned} f(x, y) &= P(X = x \text{ and } Y = y) \\ &= 0 \qquad \text{when } X \neq x \text{ and } Y \neq y \end{aligned}$$

is known as the **discrete joint probability density function** and gives the (joint) probability that X takes the value of x and Y takes the value of y.

EXAMPLE 4 The following table gives the joint PDF of the discrete variables X and Y.

		X			
		−2	**0**	**2**	**3**
Y	**3**	0.27	0.08	0.16	0
	6	0	0.04	0.10	0.35

This table tells us that the probability that X takes the value of −2 while Y simultaneously takes the value of 3 is 0.27 and that the probability that X takes the value of 3 while Y takes the value of 6 is 0.35, and so on.

Marginal Probability Density Function

In relation to $f(x, y)$, $f(x)$ and $f(y)$ are called **individual,** or **marginal,** probability density functions. These marginal PDFs are derived as follows:

$$f(x) = \sum_y f(x, y) \qquad \text{marginal PDF of } X$$

$$f(y) = \sum_x f(x, y) \qquad \text{marginal PDF of } Y$$

where, for example, $\sum_y$ means the sum over all values of Y and $\sum_x$ means the sum over all values of X.

EXAMPLE 5 Consider the data given in Example 4. The marginal PDF of X is obtained as follows:

$$f(x = -2) = \sum_y f(x, y) = 0.27 + 0 = 0.27$$

$$f(x = 0) = \sum_y f(x, y) = 0.08 + 0.04 = 0.12$$

$$f(x = 2) = \sum_y f(x, y) = 0.16 + 0.10 = 0.26$$

$$f(x = 3) = \sum_y f(x, y) = 0 + 0.35 = 0.35$$

Likewise, the marginal PDF of Y is obtained as

$$f(y = 3) = \sum_x f(x, y) = 0.27 + 0.08 + 0.16 + 0 = 0.51$$

$$f(y = 6) = \sum_x f(x, y) = 0 + 0.04 + 0.10 + 0.35 = 0.49$$

As this example shows, to obtain the marginal PDF of X we add the column numbers, and to obtain the marginal PDF of Y we add the row numbers. Notice that $\sum_x f(x)$ over all values of X is 1, as is $\sum_y f(y)$ over all values of Y (why?).

Conditional PDF

As noted in Chapter 2, in regression analysis we are often interested in studying the behavior of one variable conditional upon the value(s) of another variable(s). This can be done by considering the conditional PDF. The function

$$f(x \mid y) = P(X = x \mid Y = y)$$

is known as the **conditional PDF** of X; it gives the probability that X takes on the value of x given that Y has assumed the value y. Similarly,

$$f(y \mid x) = P(Y = y \mid X = x)$$

which gives the *conditional PDF of Y.*

The conditional PDFs may be obtained as follows:

$$f(x \mid y) = \frac{f(x, y)}{f(y)} \qquad \text{conditional PDF of } X$$

$$f(y \mid x) = \frac{f(x, y)}{f(x)} \qquad \text{conditional PDF of } Y$$

As the preceding expressions show, the conditional PDF of one variable can be expressed as the ratio of the joint PDF to the marginal PDF of another (conditioning) variable.

EXAMPLE 6

Continuing with Examples 4 and 5, let us compute the following conditional probabilities:

$$f(X = -2 \mid Y = 3) = \frac{f(X = -2, Y = 3)}{f(Y = 3)} = 0.27/0.51 = 0.53$$

Notice that the unconditional probability $f(X = -2)$ is 0.27, but if Y has assumed the value of 3, the probability that X takes the value of -2 is 0.53.

$$f(X = 2 \mid Y = 6) = \frac{f(X = 2, Y = 6)}{f(Y = 6)} = 0.10/0.49 = 0.20$$

Again note that the unconditional probability that X takes the value of 2 is 0.26, which is different from 0.20, which is its value if Y assumes the value of 6.

Statistical Independence

Two random variables X and Y are statistically independent if and only if

$$f(x, y) = f(x)f(y)$$

that is, if the joint PDF can be expressed as the product of the marginal PDFs.

EXAMPLE 7

A bag contains three balls numbered 1, 2, and 3. Two balls are drawn at random, with replacement, from the bag (i.e., the first ball drawn is replaced before the second is drawn). Let X denote the number of the first ball drawn and Y the number of the second ball drawn. The following table gives the joint PDF of X and Y.

EXAMPLE 7 (*Continued*)

		X		
		1	**2**	**3**
Y	**1**	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	3	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

Now $f(X=1, Y=1) = \frac{1}{9}$, $f(X=1) = \frac{1}{3}$ (obtained by summing the first column), and $f(y=1) = \frac{1}{3}$ (obtained by summing the first row). Since $f(X, Y) = f(X)f(Y)$ in this example we can say that the two variables are statistically independent. It can be easily checked that for any other combination of X and Y values given in the above table the joint PDF factors into individual PDFs.

It can be shown that the X and Y variables given in Example 4 are not statistically independent since the product of the two marginal PDFs is not equal to the joint PDF. (*Note:* $f(X, Y) = f(X)f(Y)$ must be true for all combinations of X and Y if the two variables are to be statistically independent.)

Continuous Joint PDF

The PDF $f(x, y)$ of two continuous variables X and Y is such that

$$f(x, y) \geq 0$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$$

$$\int_{c}^{d}\int_{a}^{b} f(x, y)\,dx\,dy = P(a \leq x \leq b, c \leq y \leq d)$$

EXAMPLE 8

Consider the following PDF

$$f(x, y) = 2 - x - y \qquad 0 \leq x \leq 1; 0 \leq y \leq 1$$

It is obvious that $f(x, y) \geq 0$. Moreover[4]

$$\int_{0}^{1}\int_{0}^{1}(2 - x - y)\,dx\,dy = 1$$

The marginal PDF of X and Y can be obtained as

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \qquad \text{marginal PDF of } X$$

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \qquad \text{marginal PDF of } Y$$

[4]

$$\begin{aligned}\int_{0}^{1}\left[\int_{0}^{1}(2 - x - y)\,dx\right]dy &= \int_{0}^{1}\left[\left(2x - \frac{x^2}{2} - xy\right)\Big|_{0}^{1}\right]dy\\ &= \int_{0}^{1}\left(\frac{3}{2} - y\right)dy\\ &= \left(\frac{3}{2}y - \frac{y^2}{2}\right)\Big|_{0}^{1} = 1\end{aligned}$$

Note: The expression $(\frac{3}{2}y - y^2/2)|_0^1$ means the expression in the parentheses is to be evaluated at the upper limit value of 1 and the lower limit value of 0; the latter value is subtracted from the former to obtain the value of the integral. Thus, in the preceding example the limits are $(\frac{3}{2} - \frac{1}{2})$ at $y = 1$ and 0 at $y = 0$, giving the value of the integral as 1.

EXAMPLE 9 The two marginal PDFs of the joint PDF given in Example 8 are as follows:

$$f(x) = \int_0^1 f(x, y)dy = \int_0^1 (2 - x - y)dy$$

$$\left(2y - xy - \frac{y^2}{2}\right)\Bigg|_0^1 = \frac{3}{2} - x \qquad 0 \le x \le 1$$

$$f(y) = \int_0^1 (2 - x - y)dx$$

$$\left(2x - xy - \frac{x^2}{2}\right)\Bigg|_0^1 = \frac{3}{2} - y \qquad 0 \le y \le 1$$

To see if the two variables of Example 8 are statistically independent, we need to find out if $f(x, y) = f(x)f(y)$. Since $(2 - x - y) \neq (\frac{3}{2} - x)(\frac{3}{2} - y)$, we can say that the two variables are not statistically independent.

A.5 Characteristics of Probability Distributions

A probability distribution can often be summarized in terms of a few of its characteristics, known as the **moments** of the distribution. Two of the most widely used moments are the **mean,** or **expected value,** and the **variance.**

Expected Value

The expected value of a discrete rv X, denoted by $E(X)$, is defined as follows:

$$E(X) = \sum_x x f(x)$$

where $\sum_x$ means the sum over all values of X and where $f(x)$ is the (discrete) PDF of X.

EXAMPLE 10 Consider the probability distribution of the sum of two numbers in the throw of two dice given in Example 2. (See Figure A.1.) Multiplying the various X values given there by their probabilities and summing over all the observations, we obtain:

$$\begin{aligned} E(X) &= 2\left(\tfrac{1}{36}\right) + 3\left(\tfrac{2}{26}\right) + 4\left(\tfrac{3}{36}\right) + \cdots + 12\left(\tfrac{1}{36}\right) \\ &= 7 \end{aligned}$$

which is the average value of the sum of numbers observed in a throw of two dice.

EXAMPLE 11 Estimate $E(X)$ and $E(Y)$ for the data given in Example 4. We have seen that

x	-2	0	2	3
$f(x)$	0.27	0.12	0.26	0.35

Therefore,

$$\begin{aligned} E(X) &= \sum_x x f(x) \\ &= (-2)(0.27) + (0)(0.12) + (2)(0.26) + (3)(0.35) \\ &= 1.03 \end{aligned}$$

EXAMPLE 11
(Continued)

Similarly,

y	3	6
$f(y)$	0.51	0.49

$$
\begin{aligned}
E(Y) &= \sum_{y} y f(y) \\
&= (3)(0.51) + (6)(0.49) \\
&= 4.47
\end{aligned}
$$

The expected value of a continuous rv is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

The only difference between this case and the expected value of a discrete rv is that we replace the summation symbol by the integral symbol.

EXAMPLE 12

Let us find out the expected value of the continuous PDF given in Example 3.

$$
\begin{aligned}
E(X) &= \int_{0}^{3} x \left(\frac{x^2}{9}\right) dx \\
&= \frac{1}{9}\left[\left(\frac{x^4}{4}\right)\right]_0^3 \\
&= \frac{9}{4} \\
&= 2.25
\end{aligned}
$$

Properties of Expected Values

1. The expected value of a constant is the constant itself. Thus, if b is a constant, $E(b) = b$.
2. If a and b are constants,

$$E(aX + b) = aE(X) + b$$

This can be generalized. If $X_1, X_2, \ldots, X_N$ are N random variables and $a_1, a_2, \ldots, a_N$ and b are constants, then

$$E(a_1X_1 + a_2X_2 + \cdots + a_NX_N + b) = a_1E(X_1) + a_2E(X_2) + \cdots + a_NE(X_N) + b$$

3. If X and Y are *independent* random variables, then

$$E(XY) = E(X)E(Y)$$

That is, the expectation of the product XY is the product of the (individual) expectations of X and Y.

However, note that

$$E\left(\frac{X}{Y}\right) \neq \frac{E(X)}{E(Y)}$$

even if X and Y are independent.

4. If X is a random variable with PDF $f(x)$ and if $g(X)$ is any function of X, then

$$\begin{aligned} E[g(X)] &= \sum_x g(X)f(x) && \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} g(X)f(x)\,dx && \text{if } X \text{ is continuous} \end{aligned}$$

Thus, if $g(X) = X^2$,

$$\begin{aligned} E(X^2) &= \sum_x x^2 f(X) && \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} x^2 f(X)\,dx && \text{if } X \text{ is continuous} \end{aligned}$$

EXAMPLE 13 Consider the following PDF:

x	-2	1	2
$f(x)$	$\frac{5}{8}$	$\frac{1}{8}$	$\frac{2}{8}$

Then

$$\begin{aligned} E(X) &= -2\left(\tfrac{5}{8}\right) + 1\left(\tfrac{1}{8}\right) + 2\left(\tfrac{2}{8}\right) \\ &= -\tfrac{5}{8} \end{aligned}$$

and

$$\begin{aligned} E(X^2) &= 4\left(\tfrac{5}{8}\right) + 1\left(\tfrac{1}{8}\right) + 4\left(\tfrac{2}{8}\right) \\ &= \tfrac{29}{8} \end{aligned}$$

Variance

Let X be a random variable and let $E(X) = \mu$. The distribution, or spread, of the X values around the expected value can be measured by the variance, which is defined as

$$\operatorname{var}(X) = \sigma_X^2 = E(X - \mu)^2$$

The positive square root of σ_X^2, σ_X, is defined as the **standard deviation** of X. The variance or standard deviation gives an indication of how closely or widely the individual X values are spread around their mean value.

The variance defined previously is computed as follows:

$$\begin{aligned} \operatorname{var}(X) &= \sum_x (X - \mu)^2 f(x) && \text{if } X \text{ is a discrete rv} \\ &= \int_{-\infty}^{\infty} (X - \mu)^2 f(x)\,dx && \text{if } X \text{ is a continuous rv} \end{aligned}$$

For computational convenience, the variance formula given above can also be expressed as

$$\begin{aligned} \operatorname{var}(X) = \sigma_x^2 &= E(X - \mu)^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

Applying this formula, it can be seen that the variance of the random variable given in Example 13 is $\frac{29}{8} - (-\frac{5}{8})^2 = \frac{207}{64} = 3.23$.

EXAMPLE 14 Let us find the variance of the random variable given in Example 3.

$$\text{var}(X) = E(X^2) - [E(X)]^2$$

Now

$$\begin{aligned} E(X^2) &= \int_0^3 x^2\left(\frac{x^2}{9}\right)dx \\ &= \int_0^3 \frac{x^4}{9}dx \\ &= \frac{1}{9}\left[\frac{x^5}{5}\right]_0^3 \\ &= 243/45 \\ &= 27/5 \end{aligned}$$

Since $E(X) = \frac{9}{4}$ (see Example 12), we finally have

$$\begin{aligned} \text{var}(X) &= 243/45 - \left(\tfrac{9}{4}\right)^2 \\ &= 243/720 = 0.34 \end{aligned}$$

Properties of Variance

1. $E(X-\mu)^2 = E(X^2) - \mu^2$, as noted before.
2. The variance of a constant is zero.
3. If a and b are constants, then

$$\text{var}(aX+b) = a^2\,\text{var}(X)$$

4. If X and Y are *independent* random variables, then

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$$

$$\text{var}(X-Y) = \text{var}(X) + \text{var}(Y)$$

This can be generalized to more than two independent variables.

5. If X and Y are *independent* rv's and a and b are constants, then

$$\text{var}(aX+bY) = a^2\,\text{var}(X) + b^2\,\text{var}(Y)$$

Covariance

Let X and Y be two rv's with means μ_x and μ_y, respectively. Then the **covariance** between the two variables is defined as

$$\text{cov}(X,Y) = E\{(X-\mu_x)(Y-\mu_y)\} = E(XY) - \mu_x\mu_y$$

It can be readily seen that the variance of a variable is the covariance of that variable with itself.

The covariance is computed as follows:

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

if X and Y are discrete random variables, and

$$\begin{aligned}\operatorname{cov}(X, Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(X-\mu_x)(Y-\mu_y)f(x, y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} XYf(x, y)\,dx\,dy - \mu_x\,\mu_y\end{aligned}$$

if X and Y are continuous random variables.

Properties of Covariance

1. If X and Y are independent, their covariance is zero, for

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY) - \mu_x\mu_y \\ &= \mu_x\mu_y - \mu_x\mu_y \qquad \text{since } E(XY) = E(X)E(Y) = \mu_x\mu_y \text{ when } X \text{ and } Y \text{ are independent} \\ &= 0\end{aligned}$$

2.

$$\operatorname{cov}(a + bX, c + dY) = bd\operatorname{cov}(X, Y)$$

where a, b, c, and d are constants.

EXAMPLE 15 Let us find out the covariance between discrete random variables X and Y whose joint PDF is as shown in Example 4. From Example 11 we already know that $\mu_x = E(X) = 1.03$ and $\mu_y = E(Y) = 4.47$.

$$\begin{aligned}E(XY) &= \sum_y\sum_x XYf(x, y) \\ &= (-2)(3)(0.27) + (0)(3)(0.08) + (2)(3)(0.16) + (3)(3)(0) \\ &\quad + (-2)(6)(0) + (0)(6)(0.04) + (2)(6)(0.10) + (3)(6)(0.35) \\ &= 6.84\end{aligned}$$

Therefore,

$$\begin{aligned}\operatorname{cov}(X, Y) &= E(XY) - \mu_x\mu_y \\ &= 6.84 - (1.03)(4.47) \\ &= 2.24\end{aligned}$$

Correlation Coefficient

The (population) correlation coefficient ρ (rho) is defined as

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sqrt{\{\operatorname{var}(X)\operatorname{var}(Y)\}}} = \frac{\operatorname{cov}(X, Y)}{\sigma_x\sigma_y}$$

Thus defined, ρ is a measure of *linear* association between two variables and lies between -1 and $+1$, -1 indicating perfect negative association and $+1$ indicating perfect positive association.

From the preceding formula, it can be seen that

$$\operatorname{cov}(X, Y) = \rho\sigma_x\sigma_y$$

EXAMPLE 16 Estimate the coefficient of correlation for the data of Example 4.

From the PDFs given in Example 11 it can be easily shown that $\sigma_x = 2.05$ and $\sigma_y = 1.50$. We have already shown that $\text{cov}(X, Y) = 2.24$. Therefore, applying the preceding formula we estimate ρ as $2.24/(2.05)(1.50) = 0.73$.

Variances of Correlated Variables

Let X and Y be two rv's. Then

$$\begin{aligned}\text{var}(X+Y) &= \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(X, Y)\\ &= \text{var}(X) + \text{var}(Y) + 2\rho\sigma_x\sigma_y\\ \text{var}(X-Y) &= \text{var}(X) + \text{var}(Y) - 2\,\text{cov}(X, Y)\\ &= \text{var}(X) + \text{var}(Y) - 2\rho\sigma_x\sigma_y\end{aligned}$$

If, however, X and Y are independent, $\text{cov}(X, Y)$ is zero, in which case the $\text{var}(X+Y)$ and $\text{var}(X-Y)$ are both equal to $\text{var}(X) + \text{var}(Y)$, as noted previously.

The preceding results can be generalized as follows. Let $\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$, then the variance of the linear combination $\sum X_i$ is

$$\begin{aligned}\text{var}\left(\sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{n} \text{var}\ X_i + 2\sum_{i<j}\sum \text{cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} \text{var}\ X_i + 2\sum_{i<j}\sum \rho_{ij}\sigma_i\sigma_j\end{aligned}$$

where ρ_{ij} is the correlation coefficient between X_i and X_j and where σ_i and σ_j are the standard deviations of X_i and X_j.

Thus,

$$\begin{aligned}\text{var}(X_1 + X_2 + X_3) &= \text{var}\ X_1 + \text{var}\ X_2 + \text{var}\ X_3 + 2\,\text{cov}(X_1, X_2)\\ &\quad + 2\,\text{cov}(X_1, X_3) + 2\,\text{cov}(X_2, X_3)\\ &= \text{var}\ X_1 + \text{var}\ X_2 + \text{var}\ X_3 + 2\rho_{12}\sigma_1\sigma_2\\ &\quad + 2\rho_{13}\sigma_1\sigma_3 + 2\rho_{23}\sigma_2\sigma_3\end{aligned}$$

where σ_1, σ_2, and σ_3 are, respectively, the standard deviations of X_1, X_2, and X_3 and where ρ_{12} is the correlation coefficient between X_1 and X_2, ρ_{13} that between X_1 and X_3, and ρ_{23} that between X_2 and X_3.

Conditional Expectation and Conditional Variance

Let $f(x, y)$ be the joint PDF of random variables X and Y. The conditional expectation of X, given $Y = y$, is defined as

$$\begin{aligned}E(X \mid Y = y) &= \sum_x x f(x \mid Y = y) && \text{if } X \text{ is discrete}\\ &= \int_{-\infty}^{\infty} x f(x \mid Y = y)\, dx && \text{if } X \text{ is continuous}\end{aligned}$$

where $E(X \mid Y = y)$ means the conditional expectation of X given $Y = y$ and where $f(x \mid Y = y)$ is the conditional PDF of X. The conditional expectation of Y, $E(Y \mid X = x)$, is defined similarly.

Conditional Expectation

Note that $E(X \mid Y)$ is a random variable because it is a function of the conditioning variable Y. However, $E(X \big| Y = y)$, where y is a specific value of Y, is a constant.

Conditional Variance

The conditional variance of X given $Y = y$ is defined as

$$\begin{aligned}\operatorname{var}(X \mid Y = y) &= E\{[X - E(X \mid Y = y)]^2 \mid Y = y\} \\ &= \sum_x [X - E(X \mid Y = y)]^2 f(x \mid Y = y) \qquad \text{if } X \text{ is discrete} \\ &= \int_{-\infty}^{\infty} [X - E(X \mid Y = y)]^2 f(x \mid Y = y)\, dx \qquad \text{if } X \text{ is continuous}\end{aligned}$$

EXAMPLE 17 Compute $E(Y \mid X = 2)$ and $\operatorname{var}(Y \mid X = 2)$ for the data given in Example 4.

$$\begin{aligned}E(Y \mid X = 2) &= \sum_y y f(Y = y \mid X = 2) \\ &= 3f(Y = 3 \mid X = 2) + 6f(Y = 6 \mid X = 2) \\ &= 3(0.16/0.26) + 6(0.10/0.26) \\ &= 4.15\end{aligned}$$

Note: $f(Y = 3 \mid X = 2) = f(Y = 3, X = 2)/f(X = 2) = 0.16/0.26$, and $f(Y = 6 \mid X = 2) = f(Y = 6, X = 2)/f(X = 2) = 0.10/0.26$, so

$$\begin{aligned}\operatorname{var}(Y \mid X = 2) &= \sum_y [Y - E(Y \mid X = 2)]^2 f(Y \mid X = 2) \\ &= (3 - 4.15)^2(0.16/0.26) + (6 - 4.15)^2(0.10/0.26) \\ &= 2.13\end{aligned}$$

Properties of Conditional Expectation and Conditional Variance

1. If $f(X)$ is a function of X, then $E(f(X) \mid X) = f(X)$, that is, the function of X behaves as a constant in computation of its expectation conditional on X. Thus, $[E(X^3 \mid X)] = E(X^3)$; this is because, if X is known, X^3 is also known.
2. If $f(X)$ and $g(X)$ are functions of X, then

$$E[f(X)Y + g(X) \mid X] = f(X)E(Y \mid X) + g(X)$$

   For example, $E[XY + cX^2 \mid X] = XE(Y \mid X) + cX^2$, where c is a constant.
3. If X and Y are independent, $E(Y \mid X) = E(Y)$. That is, if X and Y are independent random variables, then the conditional expectation of Y, given X, is the same as the unconditional expectation of Y.

4. **The law of iterated expectations.** It is interesting to note the following relation between the unconditional expectation of a random variable Y, $E(Y)$, and its conditional expectation based on another random variable X, $E(Y \mid X)$:

$$E(Y) = E_X[E(Y \mid X)]$$

   This is known as the law of iterated expectations, which in the present context states that the marginal, or unconditional, expectation of Y is equal to the expectation of its conditional expectation, the symbol E_X denoting that the expectation is taken over the values of X. Put simply, this law states that if we first obtain $E(Y \mid X)$ as a function of X and take its expected value over the distribution of X values, you wind up with $E(Y)$, the unconditional expectation of Y. The reader can verify this using the data given in Example 4.

   An implication of the law of iterated expectations is that if the conditional mean of Y given X (i.e., $E[Y|X]$) is zero, then the (unconditional) mean of Y is also zero. This follows immediately because in that case

$$E[E(Y|X)] = E[0] = 0$$

5. If X and Y are independent, then $\text{var}(Y \mid X) = \text{var}(Y)$.
6. $\text{var}(Y) = E[\text{var}(Y \mid X)] + \text{var}[E(Y \mid X)]$; that is, the (unconditional) variance of Y is equal to expectation of the conditional variance of Y plus the variance of the conditional expectation of Y.

Higher Moments of Probability Distributions

Although mean, variance, and covariance are the most frequently used summary measures of univariate and multivariate PDFs, we occasionally need to consider higher moments of the PDFs, such as the third and the fourth moments. The third and fourth moments of a univariate PDF $f(x)$ around its mean value (μ) are defined as

$$\text{Third moment:} \qquad E(X-\mu)^3$$

$$\text{Fourth moment:} \qquad E(X-\mu)^4$$

In general, the rth moment about the mean is defined as

$$r\text{th moment:} \qquad E(X-\mu)^r$$

The third and fourth moments of a distribution are often used in studying the "shape" of a probability distribution, in particular, its **skewness,** S (i.e., lack of symmetry) and **kurtosis,** K (i.e., tallness or flatness), as shown in Figure A.3.

One measure of skewness is defined as

$$S = \frac{E(X-\mu)^3}{\sigma^3}$$

$$= \frac{\text{third moment about the mean}}{\text{cube of the standard deviation}}$$

A commonly used measure of kurtosis is given by

$$K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2}$$

$$= \frac{\text{fourth moment about the mean}}{\text{square of the second moment}}$$

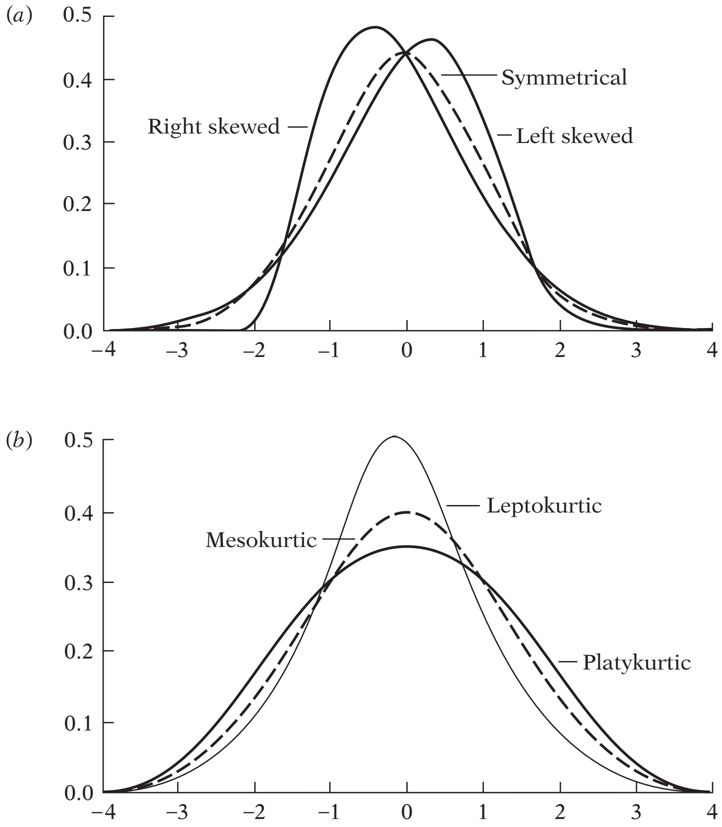


FIGURE A.3
(*a*) Skewness;
(*b*) kurtosis.

PDFs with values of K less than 3 are called **platykurtic** (fat or short-tailed), and those with values greater than 3 are called **leptokurtic** (slim or long-tailed). See Figure A.3. A PDF with a kurtosis value of 3 is known as **mesokurtic,** of which the normal distribution is the prime example. (See the discussion of the normal distribution in Section A.6.)

We will show shortly how the measures of skewness and kurtosis can be combined to determine whether a random variable follows a normal distribution. Recall that our hypothesis-testing procedure, as in the t and F tests, is based on the assumption (at least in small or finite samples) that the underlying distribution of the variable (or sample statistic) is normal. It is therefore very important to find out in concrete applications whether this assumption is fulfilled.

A.6 Some Important Theoretical Probability Distributions

In the text extensive use is made of the following probability distributions.

Normal Distribution

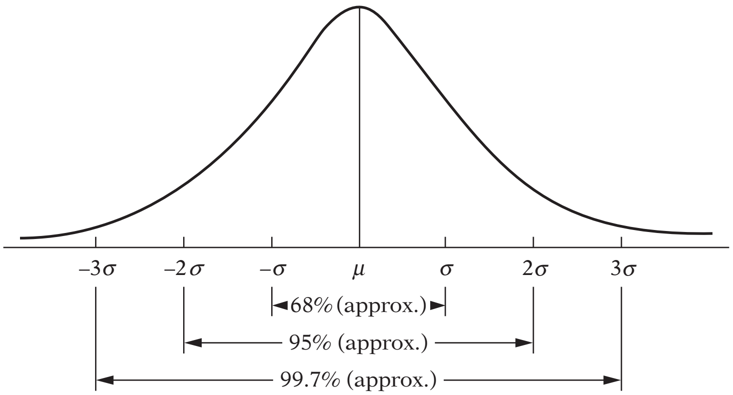
The best known of all the theoretical probability distributions is the normal distribution, whose bell-shaped picture is familiar to anyone with a modicum of statistical knowledge.

A (continuous) random variable X is said to be normally distributed if its PDF has the following form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \qquad -\infty < x < \infty$$

FIGURE A.4

Areas under the normal curve.



where μ and σ^2, known as the *parameters of the distribution,* are, respectively, the mean and the variance of the distribution. The properties of this distribution are as follows:

1. It is symmetrical around its mean value.
2. Approximately 68 percent of the area under the normal curve lies between the values of $\mu \pm \sigma$, about 95 percent of the area lies between $\mu \pm 2\sigma$, and about 99.7 percent of the area lies between $\mu \pm 3\sigma$, as shown in Figure A.4.
3. The normal distribution depends on the two parameters μ and σ^2, so once these are specified, one can find the probability that X will lie within a certain interval by using the PDF of the normal distribution. But this task can be lightened considerably by referring to Table D.1 of **Appendix D.** To use this table, we convert the given normally distributed variable X with mean μ and σ^2 into a **standardized normal variable** Z by the following transformation:

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

An important property of any standardized variable is that its mean value is zero and its variance is unity. Thus Z has zero mean and unit variance. Substituting z into the normal PDF given previously, we obtain

$$f(Z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right)$$

which is the PDF of the standardized normal variable. The probabilities given in **Appendix D,** Table D.1, are based on this standardized normal variable.

By convention, we denote a normally distributed variable as

$$X \sim N(\mu, \sigma^2)$$

where ~ means "distributed as," N stands for the normal distribution, and the quantities in the parentheses are the two parameters of the normal distribution, namely, the mean and the variance. Following this convention,

$$X \sim N(0, 1)$$

means X is a normally distributed variable with zero mean and unit variance. In other words, it is a standardized normal variable Z.

EXAMPLE 18 Assume that $X \sim N(8, 4)$. What is the probability that X will assume a value between $X_1 = 4$ and $X_2 = 12$? To compute the required probability, we compute the Z values as

$$Z_1 = \frac{X_1 - \mu}{\sigma} = \frac{4 - 8}{2} = -2$$

$$Z_2 = \frac{X_2 - \mu}{\sigma} = \frac{12 - 8}{2} = +2$$

Now from Table D.1 we observe that $\Pr(0 \leq Z \leq 2) = 0.4772$. Then, by symmetry, we have $\Pr(-2 \leq Z \leq 0) = 0.4772$. Therefore, the required probability is $0.4772 + 0.4772 = 0.9544$. (See Figure A.4.)

EXAMPLE 19 What is the probability that in the preceding example X exceeds 12?

The probability that X exceeds 12 is the same as the probability that Z exceeds 2. From Table D.1 it is obvious that this probability is (0.5 − 0.4772) or 0.0228.

4. Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and assume that they are independent. Now consider the linear combination

$$Y = aX_1 + bX_2$$

where a and b are constants. Then it can be shown that

$$Y \sim N\big[(a\mu_1 + b\mu_2), \big(a^2\sigma_1^2 + b^2\sigma_2^2\big)\big]$$

This result, which states that *a linear combination of normally distributed variables is itself normally distributed,* can be easily generalized to a linear combination of more than two normally distributed variables.

5. **Central limit theorem.** Let $X_1, X_2, \ldots, X_n$ denote n independent random variables, all of which have the same PDF with mean $= \mu$ and variance $= \sigma^2$. Let $\bar{X} = \sum X_i/n$ (i.e., the sample mean). Then as n increases indefinitely (i.e., $n \to \infty$),

$$\underset{n\to\infty}{\bar{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

That is, $\bar{X}$ approaches the normal distribution with mean μ and variance σ^2/n. Notice that this result holds true regardless of the form of the PDF. As a result, it follows that

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - u)}{\sigma} \sim N(0, 1)$$

That is, Z is a standardized normal variable.

6. The third and fourth moments of the normal distribution around the mean value are as follows:

Third moment: $E(X - \mu)^3 = 0$

Fourth moment: $E(X - \mu)^4 = 3\sigma^4$

Note: All odd-powered moments about the mean value of a normally distributed variable are zero.

7. As a result, and following the measures of skewness and kurtosis discussed earlier, for a normal PDF skewness = 0 and kurtosis = 3; that is, a normal distribution is symmetric

and mesokurtic. Therefore, a simple test of normality is to find out whether the computed values of skewness and kurtosis depart from the norms of 0 and 3. This is in fact the logic underlying the **Jarque–Bera (JB) test of normality** discussed in the text:

$$\text{JB} = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \qquad \textbf{(5.12.1)}$$

where S stands for skewness and K for kurtosis. Under the null hypothesis of normality, JB is distributed as a **chi-square** statistic with 2 df.

8. The mean and the variance of a normally distributed random variable are independent in that one is not a function of the other.
9. If X and Y are jointly normally distributed, then they are independent if, and only if, the covariance between them [i.e., cov (X, Y)] is zero. (See Exercise 4.1.)

The χ^2 (Chi-Square) Distribution

Let $Z_1, Z_2, \ldots, Z_k$ be *independent* standardized normal variables (i.e., normal variables with zero mean and unit variance). Then the quantity

$$Z = \sum_{i=1}^{k} Z_i^2$$

is said to possess the χ^2 distribution with k degrees of freedom (df), where the term df means the number of independent quantities in the previous sum. A chi-square-distributed variable is denoted by χ_k^2, where the subscript k indicates the df. Geometrically, the chi-square distribution appears in Figure A.5.

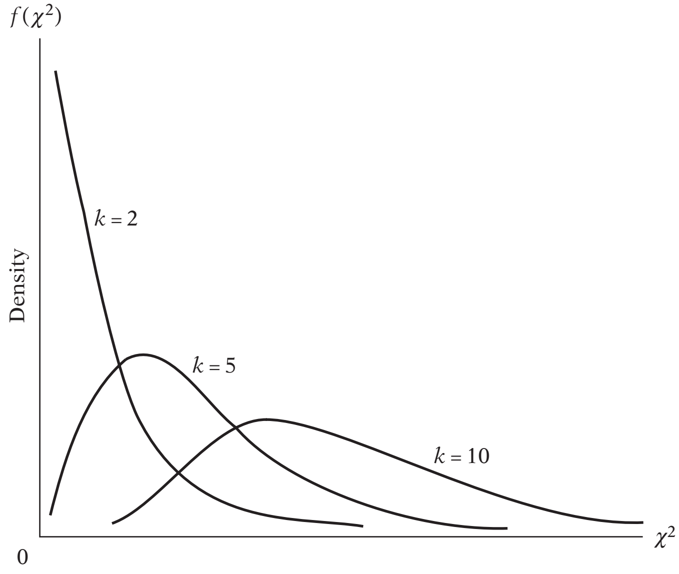
Properties of the χ^2 distribution are as follows:

1. As Figure A.5 shows, the χ^2 distribution is a skewed distribution, the degree of the skewness depending on the df. For comparatively few df, the distribution is highly skewed to the right; but as the number of df increases, the distribution becomes increasingly symmetrical. As a matter of fact, for df in excess of 100, the variable

$$\sqrt{2\chi^2} - \sqrt{(2k-1)}$$

can be treated as a standardized normal variable, where k is the df.

FIGURE A.5
Density function of the χ^2 variable.



2. The mean of the chi-square distribution is k, and its variance is $2k$, where k is the df.
3. If Z_1 and Z_2 are two independent chi-square variables with k_1 and k_2 df, then the sum $Z_1 + Z_2$ is also a chi-square variable with df $= k_1 + k_2$.

EXAMPLE 20

What is the probability of obtaining a χ^2 value of 40 or greater, given the df of 20?

As Table D.4 shows, the probability of obtaining a χ^2 value of 39.9968 or greater (20 df) is 0.005. Therefore, the probability of obtaining a χ^2 value of 40 or greater is less than 0.005, a rather small probability.

Student's *t* Distribution

If Z_1 is a standardized normal variable [that is, $Z_1 \sim N(0, 1)$] and another variable Z_2 follows the chi-square distribution with k df and is distributed independently of Z_1, then the variable defined as

$$t = \frac{Z_1}{\sqrt{(Z_2/k)}}$$

$$= \frac{Z_1\sqrt{k}}{\sqrt{Z_2}}$$

follows Student's t distribution with k df. A t-distributed variable is often designated as t_k, where the subscript k denotes the df. Geometrically, the t distribution is shown in Figure A.6.

Properties of the Student's t distribution are as follows:

1. As Figure A.6 shows, the t distribution, like the normal distribution, is symmetrical, but it is flatter than the normal distribution. But as the df increase, the t distribution approximates the normal distribution.
2. The mean of the t distribution is zero, and its variance is $k/(k-2)$.

The t distribution is tabulated in Table D.2.

EXAMPLE 21

Given df $= 13$, what is the probability of obtaining a t value (*a*) of about 3 or greater, (*b*) of about -3 or smaller, and (*c*) of $|t|$ of about 3 or greater, where $|t|$ means the absolute value (i.e., disregarding the sign) of t?

From Table D.2, the answers are (*a*) about 0.005, (*b*) about 0.005 because of the symmetry of the t distribution, and (*c*) about $0.01 = 2(0.005)$.

FIGURE A.6
Student's t distribution for selected degrees of freedom.

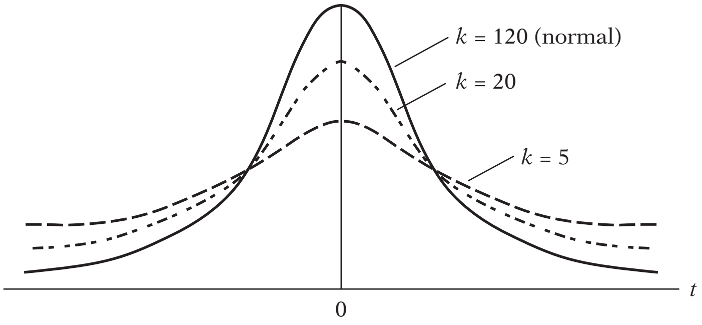
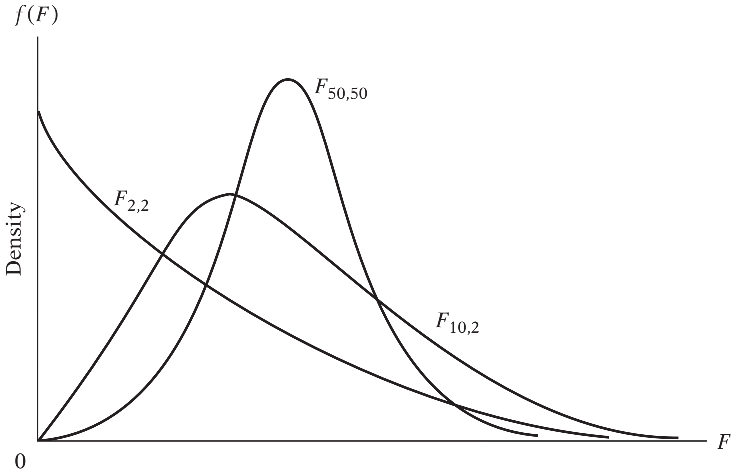


FIGURE A.7
F distribution for various degrees of freedom.



The *F* Distribution

If Z_1 and Z_2 are independently distributed chi-square variables with k_1 and k_2 df, respectively, the variable

$$F = \frac{Z_1/k_1}{Z_2/k_2}$$

follows (Fisher's) *F* distribution with k_1 and k_2 df. An *F*-distributed variable is denoted by F_{k_1,k_2} where the subscripts indicate the df associated with the two *Z* variables, k_1 being called the *numerator df* and k_2 the *denominator df.* Geometrically, the *F* distribution is shown in Figure A.7.

The *F* distribution has the following properties:

1. Like the chi-square distribution, the *F* distribution is skewed to the right. But it can be shown that as k_1 and k_2 become large, the *F* distribution approaches the normal distribution.
2. The mean value of an *F*-distributed variable is $k_2/(k_2 - 2)$, which is defined for $k_2 > 2$, and its variance is

$$\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

which is defined for $k_2 > 4$.

3. The square of a *t*-distributed random variable with *k* df has an *F* distribution with 1 and *k* df. Symbolically,

$$t_k^2 = F_{1,k}$$

EXAMPLE 22

Given $k_1 = 10$ and $k_2 = 8$, what is the probability of obtaining an *F* value (*a*) of 3.4 or greater and (*b*) of 5.8 or greater?

As Table D.3 shows, these probabilities are (*a*) approximately 0.05 and (*b*) approximately 0.01.

4. If the denominator df, k_2, is fairly large, the following relationship holds between the F and the chi-square distributions:

$$k_1 F \sim \chi^2_{k1}$$

That is, for fairly large denominator df, the numerator df times the F value is approximately the same as a chi-square value with numerator df.

EXAMPLE 23 Let $k_1 = 20$ and $k_2 = 120$. The 5 percent critical F value for these df is 1.48. Therefore, $k_1 F = (20)(1.48) = 29.6$. From the chi-square distribution for 20 df, the 5 percent critical chi-square value is about 31.41.

In passing, note that since for large df the t, chi-square, and F distributions approach the normal distribution, these three distributions are known as the *distributions related to the normal distribution.*

The Bernoulli Binomial Distribution

A random variable X is said to follow a distribution named after Bernoulli (a Swiss mathematician) if its probability density (or mass) function (PDF) is:

$$P(X = 0) = 1 - p$$

$$P(X = 1) = p$$

where p, $0 \leq p \leq 1$, is the probability that some event is a "success," such as the probability of obtaining a head in a toss of a coin. For such a variable,

$$E(X) = [1 \times p(X = 1) + 0 \times p(X = 0)] = p$$

$$\text{var}\,(X) = pq$$

where $q = (1 - p)$, that is, the probability of a "failure."

Binomial Distribution

The binomial distribution is the generalization of the Bernoulli distribution. Let n denote the number of independent trials, each of which results in a "success" with probability p and a "failure" with a probability $q = (1 - p)$. If X represents the number of successes in the n trials, then X is said to follow the binomial distribution whose PDF is:

$$f(X) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where x represents the number of successes in n trials and where

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

where $n!$, read as n factorial, means $n(n - 1)(n - 2) \cdots 1$.

The binomial is a two-parameter distribution, n and p. For this distribution,

$$E(X) = np$$

$$\text{var}\,(X) = np(1 - p) = npq$$

For example, if you toss a coin 100 times and want to find out the probability of obtaining 60 heads, you put $p = 0.5$, $n = 100$ and $x = 60$ in the above formula. Computer routines exist to evaluate such probabilities.

You can see how the binomial distribution is a generalization of the Bernoulli distribution.

The Poisson Distribution

A random X variable is said to have the Poisson distribution if its PDF is:

$$f(X) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots, \lambda > 0$$

The Poisson distribution depends on a single parameter, λ. A distinguishing feature of the Poisson distribution is that its variance is equal to its expected value, which is λ. That is,

$$E(X) = \text{var}(X) = \lambda$$

The Poisson model, as we saw in the chapter on nonlinear regression models, is used to model rare or infrequent phenomena, such as the number of phone calls received in a span of, say, 5 minutes, or the number of speeding tickets received in a span of an hour, or the number of patents received by a firm, say, in a year.

A.7 Statistical Inference: Estimation

In Section A.6 we considered several theoretical probability distributions. Very often we know or are willing to assume that a random variable X follows a particular probability distribution but do not know the value(s) of the parameter(s) of the distribution. For example, if X follows the normal distribution, we may want to know the value of its two parameters, namely, the mean and the variance. To estimate the unknowns, the usual procedure is to assume that we have a **random sample** of size n from the known probability distribution and use the sample data to estimate the unknown parameters.[5] This is known as the **problem of estimation.** In this section, we take a closer look at this problem. The problem of estimation can be broken down into two categories: point estimation and interval estimation.

Point Estimation

To fix the ideas, let X be a random variable with PDF $f(x;\theta)$, where θ is the parameter of the distribution (for simplicity of discussion only, we are assuming that there is only one unknown parameter; our discussion can be readily generalized). Assume that we know the functional form—that is, we know the theoretical PDF, such as the t distribution—but do not know the value of θ. Therefore, we draw a random sample of size n from this known PDF and then develop a function of the sample values such that

$$\hat{\theta} = f(x_1, x_2, \ldots, x_n)$$

provides us an estimate of the true θ. $\hat{\theta}$ is known as a **statistic,** or an **estimator,** and a particular numerical value taken by the estimator is known as an **estimate.** Note that $\hat{\theta}$ can be

[5] Let $X_1, X_2, \ldots, X_n$ be n random variables with joint PDF $f(x_1, x_2, \ldots, x_n)$. If we can write

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \cdots f(x_n)$$

where $f(x)$ is the common PDF of each X, then $x_1, x_2, \ldots, x_n$ are said to constitute a random sample of size n from a population with PDF $f(x_n)$.

treated as a random variable because it is a function of the sample data. $\hat{\theta}$ provides us with a rule, or formula, that tells us how we may estimate the true θ. Thus, if we let

$$\hat{\theta} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \bar{X}$$

where $\bar{X}$ is the sample mean, then $\bar{X}$ is an estimator of the true mean value, say, μ. If in a specific case $\bar{X} = 50$, this provides an *estimate of* μ. The estimator $\hat{\theta}$ obtained previously is known as a **point estimator** because it provides only a single (point) estimate of θ.

Interval Estimation

Instead of obtaining only a single estimate of θ, suppose we obtain two estimates of θ by constructing two estimators $\hat{\theta}_1(x_1, x_2, \ldots, x_n)$ and $\hat{\theta}_2(x_1, x_2, \ldots, x_n)$, and say with some confidence (i.e., probability) that the interval between $\hat{\theta}_1$ and $\hat{\theta}_2$ includes the true θ. Thus, in interval estimation, in contrast with point estimation, we provide a range of possible values within which the true θ may lie.

The key concept underlying interval estimation is the notion of the **sampling,** or **probability distribution, of an estimator.** For example, it can be shown that if a variable X is normally distributed, then the sample mean $\bar{X}$ is also normally distributed with mean $= \mu$ (the true mean) and variance $= \sigma^2/n$, where n is the sample size. In other words, the sampling, or probability, distribution of the estimator $\bar{X}$ is $\bar{X} \sim N(\mu, \sigma^2/n)$. As a result, if we construct the interval

$$\bar{X} \pm 2\frac{\sigma}{\sqrt{n}}$$

and say that the probability is approximately 0.95, or 95 percent, that intervals like it will include the true μ, we are in fact constructing an interval estimator for μ. Note that the interval given previously is random since it is based on $\bar{X}$, which will vary from sample to sample.

More generally, in interval estimation we construct two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, both functions of the sample X values, such that

$$\Pr(\hat{\theta}_1 \le \theta \le \hat{\theta}_2) = 1 - \alpha \qquad 0 < \alpha < 1$$

That is, we can state that the probability is $1 - \alpha$ that the interval from $\hat{\theta}_1$ to $\hat{\theta}_2$ contains the true θ. This interval is known as a **confidence interval** of size $1 - \alpha$ for θ, $1 - \alpha$ being known as the **confidence coefficient.** If $\alpha = 0.05$, then $1 - \alpha = 0.95$, meaning that if we construct a confidence interval with a confidence coefficient of 0.95, then in repeated such constructions resulting from repeated sampling we shall be right in 95 out of 100 cases if we maintain that the interval contains the true θ. When the confidence coefficient is 0.95, we often say that we have a 95 percent confidence interval. In general, if the confidence coefficient is $1 - \alpha$, we say that we have a $100(1 - \alpha)\%$ confidence interval. Note that α is known as the **level of significance,** or the probability of committing a Type I error. This topic is discussed in Section A.8.

EXAMPLE 24

Suppose that the distribution of height of men in a population is normally distributed with mean $= \mu$ inches and $\sigma = 2.5$ inches. A sample of 100 men drawn randomly from this population had an average height of 67 inches. Establish a 95 percent confidence interval for the mean height $(= \mu)$ in the population as a whole.

As noted, $\bar{X} \sim N(\mu, \sigma^2/n)$, which in this case becomes $\bar{X} \sim N(\mu, 2.5^2/100)$. From Table D.1 one can see that

$$\bar{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

EXAMPLE 24
(*Continued*)

covers 95 percent of the area under the normal curve. Therefore, the preceding interval provides a 95 percent confidence interval for μ. Plugging in the given values of $\bar{X}$, σ, and n, we obtain the 95 percent confidence interval as

$$66.51 \leq \mu \leq 67.49$$

In repeated such measurements, intervals thus established will include the true μ with 95 percent confidence. A technical point may be noted here. Although we can say that the probability that the random interval $[\bar{X} \pm 1.96(\sigma/\sqrt{n})]$ includes μ is 95 percent, we *cannot* say that the probability is 95 percent that the particular interval (66.51, 67.49) includes μ. Once this interval is fixed, the probability that it will include μ is either 0 or 1. What we can say is that if we construct 100 such intervals, 95 out of the 100 intervals will include the true μ; we cannot guarantee that one particular interval will necessarily include μ.

Methods of Estimation

Broadly speaking, there are three methods of parameter estimation: (1) least squares (LS), (2) maximum likelihood (ML), and (3) method of moments (MOM) and its extension, the generalized method of moments (GMM). We have devoted considerable time to illustrate the LS method. In Chapter 4 we introduced the ML method in the regression context. But the method is of much broader application.

The key idea behind the ML is the **likelihood function.** To illustrate this, suppose the random variable X has PDF $f(X,\theta)$ which depends on a single parameter θ. We know the PDF (e.g., Bernoulli or binomial) but do not know the parameter value. Suppose we obtain a random sample of nX values. The joint PDF of these n values is:

$$g(x_1, x_2, \ldots, x_n; \theta)$$

Because it is a random sample, we can write the preceding joint PDF as a product of the individual PDF as

$$g(x_1, x_2, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$

The joint PDF has a dual interpretation. If θ is known, we interpret it as the joint probability of observing the given sample values. On the other hand, we can treat it as a function of θ for given values of $x_1, x_2, \ldots, x_n$. On the latter interpretation, we call the joint PDF the **likelihood function (LF)** and write it as

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta)$$

Observe the role reversal of θ in the joint probability density function and the likelihood function.

The ML estimator of θ is that value of θ that maximizes the (sample) likelihood function, L. For mathematical convenience, we often take the log of the likelihood, called the **log-likelihood function (log *L*).** Following the calculus rules of maximization, we differentiate the log-likelihood function with respect to the unknown and equate the resulting derivative to zero. The resulting value of the estimator is called the **maximum-likelihood estimator.** One can apply the second-order condition of maximization to assure that the value we have obtained is in fact the maximum value.

In case there is more than one unknown parameter, we differentiate the log-likelihood function with respect to each unknown, set the resulting expressions to zero, and solve them simultaneously to obtain the values of the unknown parameters. We have already shown this for the multiple regression model (see Chapter 4, Appendix 4A.1).

EXAMPLE 25 Assume that the random variable X follows the Poisson distribution with the mean value of λ. Suppose $x_1, x_2, \ldots, x_n$ are independent Poisson random variables each with mean λ. Suppose we want to find out the ML estimator of λ. The likelihood function here is:

$$\begin{aligned} L(x_1, x_2, \ldots, x_n; \lambda) &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!}\cdots\frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{x_1!x_2!\cdots x_n!} \end{aligned}$$

This is a rather unwieldy expression, but if we take its log, it becomes

$$\log(x_1, x_2, \ldots, x_n; \lambda) = -n\lambda + \sum x_i \log\lambda - \log c$$

where $\log c = \prod x_i!$. Differentiating the preceding expression with respect to λ, we obtain $(-n + (\sum x_i)/\lambda)$. By setting this last expression to zero, we obtain $\lambda_{ml} = (\sum x_i)/n = \bar{X}$, which is the ML estimator of the unknown λ.

The Method of Moments

We have given a glimpse of MOM in Exercise 3.4 in the so-called **analogy principle** in which the sample moments try to duplicate the properties of their population counterparts. The generalized method of moments (GMM), which is a generalization of MOM, is now becoming more popular, but not at the introductory level. Hence we will not pursue it here.

The desirable statistical properties fall into two categories: small-sample, or finite-sample, properties and large-sample, or asymptotic, properties. Underlying both of these sets of properties is the notion that an estimator has a sampling, or probability, distribution.

Small-Sample Properties

Unbiasedness

An estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to the true θ; that is,

$$E(\hat{\theta}) = \theta$$

or

$$E(\hat{\theta}) - \theta = 0$$

If this equality does not hold, then the estimator is said to be biased, and the bias is calculated as

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Of course, if $E(\hat{\theta}) = \theta$—that is, $\hat{\theta}$ is an unbiased estimator—the bias is zero.

Geometrically, the situation is as depicted in Figure A.8. In passing, note that unbiasedness is a property of repeated sampling, not of any given sample: Keeping the sample size fixed, we draw several samples, each time obtaining an estimate of the unknown parameter. The average value of these estimates is expected to be equal to the true value if the estimator is unbiased.

Minimum Variance

$\hat{\theta}_1$ is said to be a minimum-variance estimator of θ if the variance of $\hat{\theta}_1$ is smaller than or at most equal to the variance of $\hat{\theta}_2$, which is any other estimator of θ. Geometrically, we have

FIGURE A.8
Biased and unbiased estimators.

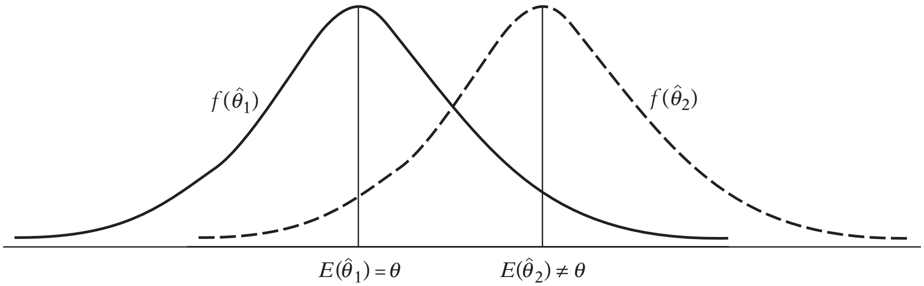


FIGURE A.9
Distribution of three estimators of θ.

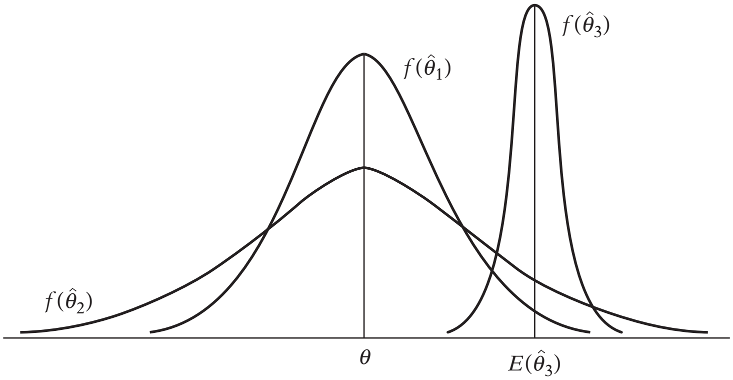


Figure A.9, which shows three estimators of θ, namely $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$, and their probability distributions. As shown, the variance of $\hat{\theta}_3$ is smaller than that of either $\hat{\theta}_1$ or $\hat{\theta}_2$. Hence, assuming only the three possible estimators, in this case $\hat{\theta}_3$ is a minimum-variance estimator. But note that $\hat{\theta}_3$ is a biased estimator (why?).

Best Unbiased, or Efficient, Estimator

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two *unbiased* estimators of θ, and the variance of $\hat{\theta}_1$ is smaller than or at most equal to the variance of $\hat{\theta}_2$, then $\hat{\theta}_1$ is a **minimum-variance unbiased,** or **best unbiased,** or **efficient, estimator.** Thus, in Figure A.9, of the two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, $\hat{\theta}_1$ is best unbiased, or efficient.

Linearity

An estimator $\hat{\theta}$ is said to be a linear estimator of θ if it is a linear function of the sample observations. Thus, the sample mean defined as

$$\bar{X} = \frac{1}{n}\sum X_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

is a linear estimator because it is a linear function of the X values.

Best Linear Unbiased Estimator (BLUE)

If $\hat{\theta}$ is linear, is unbiased, and has minimum variance in the class of all linear unbiased estimators of θ, then it is called a **best linear unbiased estimator,** or **BLUE** for short.

Minimum Mean-Square-Error (MSE) Estimator

The MSE of an estimator $\hat{\theta}$ is defined as

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

This is in contrast with the variance of $\hat{\theta}$, which is defined as

$$\operatorname{var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

The difference between the two is that var $(\hat{\theta})$ measures the dispersion of the distribution of $\hat{\theta}$ around its mean or expected value, whereas MSE$(\hat{\theta})$ measures dispersion around the true value of the parameter. The relationship between the two is as follows:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\
&= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + E[E(\hat{\theta}) - \theta]^2 \qquad \text{since the last term is zero}^6 \\
&= \operatorname{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2 \\
&= \text{variance of } \hat{\theta} \textit{ plus } \text{square bias}
\end{aligned}$$

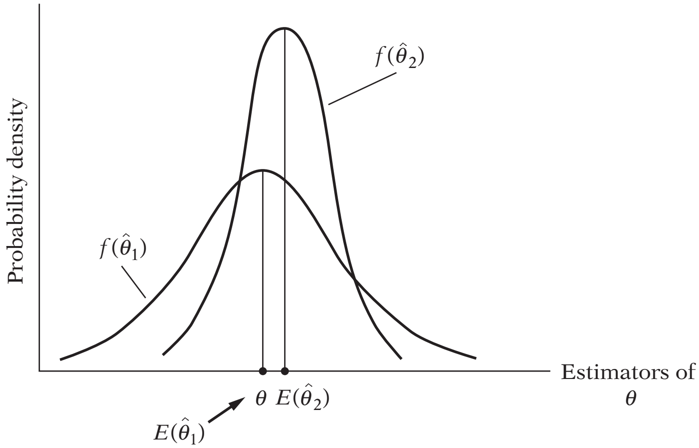
Of course, if the bias is zero, MSE$(\hat{\theta}) = \operatorname{var}(\hat{\theta})$.

The minimum MSE criterion consists in choosing an estimator whose MSE is the least in a competing set of estimators. But notice that even if such an estimator is found, there is a tradeoff involved—to obtain minimum variance you may have to accept some bias. Geometrically, the situation is as shown in Figure A.10. In this figure, $\hat{\theta}_2$ is slightly biased, but its variance is smaller than that of the unbiased estimator $\hat{\theta}_1$. In practice, however, the minimum MSE criterion is used when the best unbiased criterion is incapable of producing estimators with smaller variances.

Large-Sample Properties

It often happens that an estimator does not satisfy one or more of the desirable statistical properties in small samples. But as the sample size increases indefinitely, the estimator possesses several desirable statistical properties. These properties are known as the **large-sample,** or **asymptotic, properties.**

FIGURE A.10
Tradeoff between bias and variance.



[6] The last term can be written as $2\{[E(\hat{\theta})]^2 - [E(\hat{\theta})]^2 - \theta E(\hat{\theta}) + \theta E(\hat{\theta})\} = 0$. Also note that $E[E(\hat{\theta}) - \theta]^2 = [E(\hat{\theta}) - \theta]^2$, since the expected value of a constant is simply the constant itself.

Asymptotic Unbiasedness

An estimator $\hat{\theta}$ is said to be an asymptotically unbiased estimator of θ if

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

where $\hat{\theta}_n$ means that the estimator is based on a sample size of n and where lim means limit and $n \to \infty$ means that n increases indefinitely. In words, $\hat{\theta}$ is an asymptotically unbiased estimator of θ if its expected, or mean, value approaches the true value as the sample size gets larger and larger. As an example, consider the following measure of the sample variance of a random variable X:

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

It can be shown that

$$E(S^2) = \sigma^2\left(1 - \frac{1}{n}\right)$$

where σ^2 is the true variance. It is obvious that in a small sample S^2 is biased, but as n increases indefinitely, $E(S^2)$ approaches true σ^2; hence it is asymptotically unbiased.

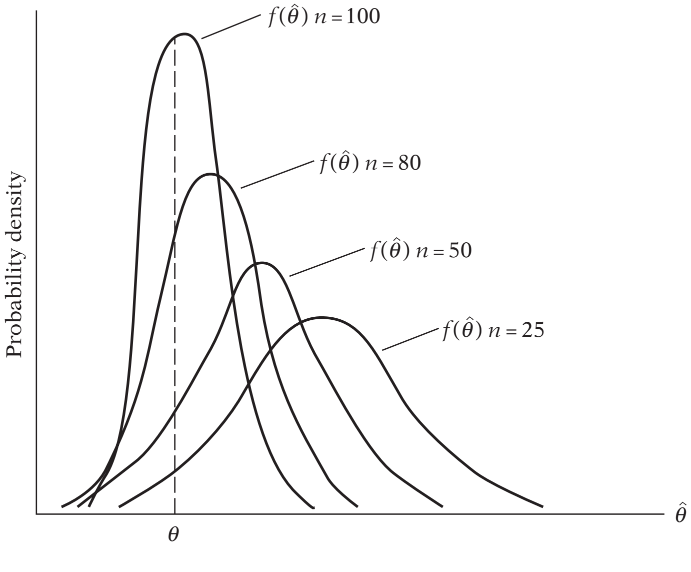
Consistency

$\hat{\theta}$ is said to be a consistent estimator if it approaches the true value θ as the sample size gets larger and larger. Figure A.11 illustrates this property.

In this figure we have the distribution of $\hat{\theta}$ based on sample sizes of 25, 50, 80, and 100. As the figure shows, $\hat{\theta}$ based on $n = 25$ is biased since its sampling distribution is not centered on the true θ. But as n increases, the distribution of $\hat{\theta}$ not only tends to be more closely centered on θ (i.e., $\hat{\theta}$ becomes less biased) but its variance also becomes smaller. If in the limit (i.e., when n increases indefinitely) the distribution of $\hat{\theta}$ collapses to the single point θ, that is, if the distribution of $\hat{\theta}$ has zero spread, or variance, we say that $\hat{\theta}$ is a **consistent estimator** of θ.

FIGURE A.11

The distribution of $\hat{\theta}$ as sample size increases.



More formally, an estimator $\hat{\theta}$ is said to be a consistent estimator of θ if the probability that the absolute value of the difference between $\hat{\theta}$ and θ is less than δ (an arbitrarily small positive quantity) approaches unity. Symbolically,

$$\lim_{n\to\infty} P\{|\hat{\theta} - \theta| < \delta\} = 1 \qquad \delta > 0$$

where P stands for probability. This is often expressed as

$$\underset{n\to\infty}{\text{plim}}\, \hat{\theta} = \theta$$

where plim means probability limit.

Note that the properties of unbiasedness and consistency are conceptually very different. The property of unbiasedness can hold for any sample size, whereas consistency is strictly a large-sample property.

A *sufficient condition* for consistency is that the bias and variance both tend to zero as the sample size increases indefinitely.[7] Alternatively, a sufficient condition for consistency is that the MSE($\hat{\theta}$) tends to zero as n increases indefinitely. (For MSE[$\hat{\theta}$], see the discussion presented previously.)

EXAMPLE 26

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Show that the sample mean $\bar{X}$ is a consistent estimator of μ.

From elementary statistics it is known that $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \sigma^2/n$. Since $E(\bar{X}) = \mu$ regardless of the sample size, it is unbiased. Moreover, as n increases indefinitely, $\text{var}(\bar{X})$ tends toward zero. Hence, $\bar{X}$ is a consistent estimator of μ.

The following rules about probability limits are noteworthy.

1. *Invariance (Slutsky property).* If $\hat{\theta}$ is a consistent estimator of θ and if $h(\hat{\theta})$ is any continuous function of $\hat{\theta}$, then

$$\underset{n\to\infty}{\text{plim}}\, h(\hat{\theta}) = h(\theta)$$

What this means is that if $\hat{\theta}$ is a consistent estimator of θ, then $1/\hat{\theta}$ is also a consistent estimator of $1/\theta$ and that $\log(\hat{\theta})$ is also a consistent estimator of $\log(\theta)$. Note that this property does not hold true of the expectation operator E; that is, if $\hat{\theta}$ is an unbiased estimator of θ (that is, $E[\hat{\theta}] = \theta$), it is *not true* that $1/\hat{\theta}$ is an unbiased estimator of $1/\theta$; that is, $E(1/\hat{\theta}) \neq 1/E(\hat{\theta}) \neq 1/\theta$.

2. If b is a constant, then

$$\underset{n\to\infty}{\text{plim}}\, b = b$$

That is, the probability limit of a constant is the same constant.

3. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are consistent estimators, then

$$\text{plim}\,(\hat{\theta}_1 + \hat{\theta}_2) = \text{plim}\,\hat{\theta}_1 + \text{plim}\,\hat{\theta}_2$$

$$\text{plim}\,(\hat{\theta}_1\hat{\theta}_2) = \text{plim}\,\hat{\theta}_1\,\text{plim}\,\hat{\theta}_2$$

$$\text{plim}\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) = \frac{\text{plim}\,\hat{\theta}_1}{\text{plim}\,\hat{\theta}_2}$$

[7] More technically, $\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$ and $\lim_{n\to\infty} \text{var}(\hat{\theta}_n) = 0$.

The last two properties, in general, do not hold true of the expectation operator E. Thus, $E(\hat{\theta}_1/\hat{\theta}_2) \neq E(\hat{\theta}_1)/E(\hat{\theta}_2)$. Similarly, $E(\hat{\theta}_1\hat{\theta}_2) \neq E(\hat{\theta}_1)E(\hat{\theta}_2)$. If, however, $\hat{\theta}_1$ and $\hat{\theta}_2$ are independently distributed, $E(\hat{\theta}_1\hat{\theta}_2) = E(\hat{\theta}_1)E(\hat{\theta}_2)$, as noted previously.

Asymptotic Efficiency

Let $\hat{\theta}$ be an estimator of θ. The variance of the asymptotic distribution of $\hat{\theta}$ is called the **asymptotic variance** of $\hat{\theta}$. If $\hat{\theta}$ is consistent and its asymptotic variance is smaller than the asymptotic variance of all other consistent estimators of θ, $\hat{\theta}$ is called **asymptotically efficient.**

Asymptotic Normality

An estimator $\hat{\theta}$ is said to be asymptotically normally distributed if its sampling distribution tends to approach the normal distribution as the sample size n increases indefinitely. For example, statistical theory shows that if $X_1, X_2, \ldots, X_n$ are independent normally distributed variables with the same mean μ and the same variance σ^2, the sample mean $\bar{X}$ is also normally distributed with mean μ and variance σ^2/n in small as well as large samples. But if the X_i are independent with mean μ and variance σ^2 but are not necessarily from the normal distribution, then the sample mean $\bar{X}$ is asymptotically normally distributed with mean μ and variance σ^2/n; that is, as the sample size n increases indefinitely, the sample mean tends to be normally distributed with mean μ and variance σ^2/n. That is in fact the central limit theorem discussed previously.

A.8 Statistical Inference: Hypothesis Testing

Estimation and hypothesis testing constitute the twin branches of classical statistical inference. Having examined the problem of estimation, we briefly look at the problem of testing statistical hypotheses.

The problem of hypothesis testing may be stated as follows. Assume that we have an rv X with a known PDF $f(x;\theta)$, where θ is the parameter of the distribution. Having obtained a random sample of size *n,* we obtain the point estimator $\hat{\theta}$. Since the true θ is rarely known, we raise the question: Is the estimator $\hat{\theta}$ "compatible" with some hypothesized value of θ, say, $\theta = \theta^*$, where θ^* is a specific numerical value of θ? In other words, could our sample have come from the PDF $f(x;\theta) = \theta^*$? In the language of hypothesis testing $\theta = \theta^*$ is called the **null** (or maintained) **hypothesis** and is generally denoted by H_0. The null hypothesis is tested against an **alternative hypothesis,** denoted by H_1, which, for example, may state that $\theta \neq \theta^*$. (*Note:* In some textbooks, H_0 and H_1 are designated by H_1 and H_2, respectively.)

The null hypothesis and the alternative hypothesis can be **simple** or **composite.** A hypothesis is called *simple* if it specifies the value(s) of the parameter(s) of the distribution; otherwise it is called a *composite* hypothesis. Thus, if $X \sim N(\mu, \sigma^2)$ and we state that

$$H_0\colon \mu = 15 \qquad \text{and} \qquad \sigma = 2$$

it is a simple hypothesis, whereas

$$H_0\colon \mu = 15 \qquad \text{and} \qquad \sigma > 2$$

is a composite hypothesis because here the value of σ is not specified.

To test the null hypothesis (i.e., to test its validity), we use the sample information to obtain what is known as the **test statistic.** Very often this test statistic turns out to be the point estimator of the unknown parameter. Then we try to find out the *sampling,* or *probability, distribution* of the test statistic and use the **confidence interval** or **test of significance** approach to test the null hypothesis. The mechanics are illustrated below.

To fix the ideas, let us revert to Example 24, which was concerned with the height (X) of men in a population. We are told that

$$X_i \sim N(\mu, \sigma^2) = N(\mu, 2.5^2)$$

$$\bar{X} = 67 \qquad n = 100$$

Let us assume that

$$H_0\colon \mu = \mu^* = 69$$

$$H_1\colon \mu \neq 69$$

The question is: Could the sample with $\bar{X} = 67$, the test statistic, have come from the population with the mean value of 69? Intuitively, we may not reject the null hypothesis if $\bar{X}$ is "sufficiently close" to μ^*; otherwise we may reject it in favor of the alternative hypothesis. But how do we decide that $\bar{X}$ is "sufficiently close" to μ^*? We can adopt two approaches, (1) confidence interval and (2) test of significance, both leading to identical conclusions in any specific application.

The Confidence Interval Approach

Since $X_i \sim N(\mu, \sigma^2)$, we know that the test statistic $\bar{X}$ is distributed as

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

Since we know the probability distribution of $\bar{X}$, why not establish, say, a $100(1-\alpha)$ confidence interval for μ based on $\bar{X}$ and see whether this confidence interval includes $\mu = \mu^*$? If it does, we may not reject the null hypothesis; if it does not, we may reject the null hypothesis. Thus, if $\alpha = 0.05$, we will have a 95 percent confidence interval and if this confidence interval includes μ^*, we may not reject the null hypothesis—95 out of 100 intervals thus established are likely to include μ^*.

The actual mechanics are as follows: since $\bar{X} \sim N(\mu, \sigma^2/n)$, it follows that

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

that is, a standard normal variable. Then from the normal distribution table we know that

$$\Pr(-1.96 \leq Z_i \leq 1.96) = 0.95$$

That is,

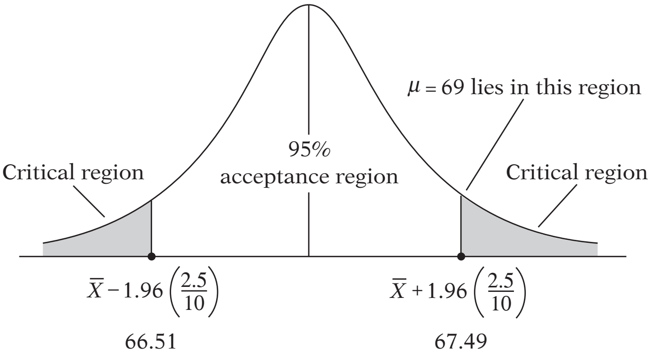
$$\Pr\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

which, on rearrangement, gives

$$\Pr\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

This is a 95 percent confidence interval for μ. Once this interval has been established, the test of the null hypothesis is simple. All that we have to do is to see whether $\mu = \mu^*$ lies in this interval. If it does, we may not reject the null hypothesis; if it does not, we may reject it.

FIGURE A.12
95 percent confidence interval for μ.



Returning to Example 24, we have already established a 95 percent confidence interval for μ, which is

$$66.51 \leq \mu \leq 67.49$$

This interval obviously does not include $\mu = 69$. Therefore, we can reject the null hypothesis that the true μ is 69 with a 95 percent confidence coefficient. Geometrically, the situation is as depicted in Figure A.12.

In the language of hypothesis testing, the confidence interval that we have established is called the **acceptance region** and the area(s) outside the acceptance region is (are) called the **critical region(s),** or **region(s) of rejection** of the null hypothesis. The lower and upper limits of the acceptance region (which demarcate it from the rejection regions) are called the **critical values.** In this language of hypothesis testing, if the hypothesized value falls inside the acceptance region, one may not reject the null hypothesis; otherwise one may reject it.

It is important to note that in deciding to reject or not reject H_0, we are likely to commit two types of errors: (1) we may reject H_0 when it is, in fact, true; this is called a **type I error** (thus, in the preceding example $\bar{X} = 67$ could have come from the population with a mean value of 69), or (2) we may not reject H_0 when it is, in fact, false; this is called a **type II error.** Therefore, a hypothesis test does not establish the value of true μ. It merely provides a means of deciding whether we may act as if $\mu = \mu^*$.

Type I and Type II Errors

Schematically, we have

	State of Nature	
Decision	**H_0 Is True**	**H_0 Is False**
Reject	Type I error	No error
Do not reject	No error	Type II error

Ideally, we would like to minimize both type I and type II errors. But unfortunately, for any given sample size, it is not possible to minimize both the errors simultaneously. The classical approach to this problem, embodied in the work of Neyman and Pearson, is to assume that a type I error is likely to be more serious in practice than a type II error. Therefore, one should try to keep the probability of committing a type I error at a fairly low level, such as 0.01 or 0.05, and then try to minimize the probability of having a type II error as much as possible.

In the literature, the probability of a type I error is designated as α and is called the **level of significance,** and the probability of a type II error is designated as β. The probability of *not* committing a type II error is called the **power of the test.** *Put differently, the power of a test is its ability to reject a false null hypothesis.* The classical approach to hypothesis testing is to fix α at levels such as 0.01 (or 1 percent) or 0.05 (5 percent) and then try to maximize the power of the test; that is to minimize β.

It is important that the reader understand the concept of the power of a test, which is best explained with an example.[8]

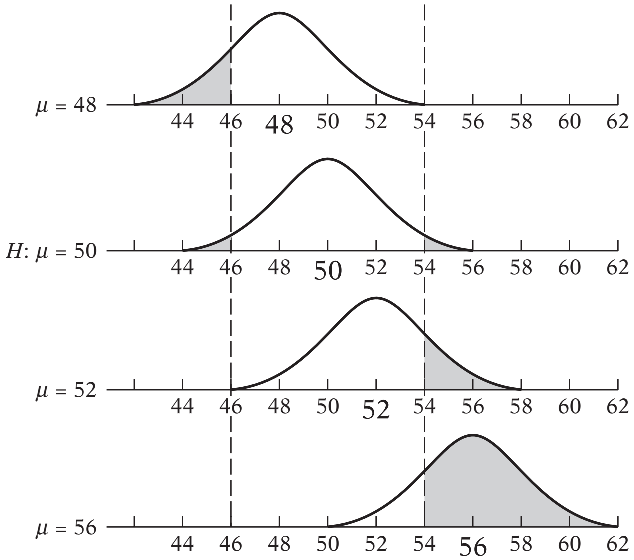
Let $X \sim N(\mu, 100)$; that is, X is normally distributed with mean μ and variance 100. Assume that $\alpha = 0.05$. Suppose we have a sample of 25 observations, which gives a sample mean value of $\bar{X}$. Suppose further we entertain the hypothesis H_0: $\mu = 50$. Since X is normally distributed, we know that the sample mean is also normally distributed as: $\bar{X} \sim N(\mu, 100/25)$. Hence under the stated null hypothesis that $\mu = 50$, the 95 percent confidence interval for $\bar{X}$ is $(\mu \pm 1.96(\sqrt{100/25}) = \mu \pm 3.92$, that is, (46.08 to 53.92). Therefore, the critical region consists of all values of $\bar{X}$ less than 46.08 or greater than 53.92. That is, we will reject the null hypothesis that the true mean is 50 if a sample mean value is found below 46.08 or greater than 53.92.

But what is the probability that $\bar{X}$ will lie in the preceding critical region(s) if the true μ has a value different from 50? Suppose there are three alternative hypotheses: $\mu = 48$, $\mu = 52$, and $\mu = 56$. If any of these alternatives is true, it will be the actual mean of the distribution of $\bar{X}$. The standard error is unchanged for the three alternatives since σ^2 is still assumed to be 100.

The shaded areas in Figure A.13 show the probabilities that $\bar{X}$ will fall in the critical region if each of the alternative hypotheses is true. As you can check, these probabilities

FIGURE A.13 Distribution of X when $N = 25$, $\sigma = 10$, and $\mu = 48, 50, 52$, or 56. Under H: $\mu = 50$, the critical region with $\alpha = 0.05$ is $\bar{X} < 46.1$ and $\bar{X} > 53.9$. The shaded area indicates the probability that $\bar{X}$ will fall into the critical region. This probability is:

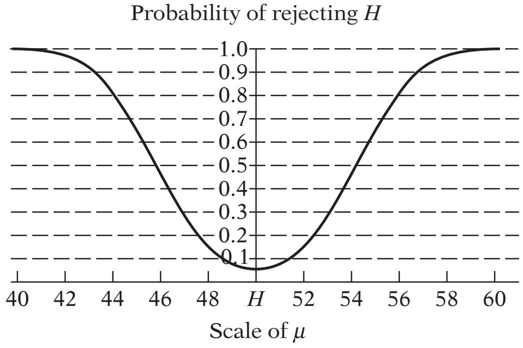
0.17 if $\mu = 48$ 0.17 if $\mu = 52$
0.05 if $\mu = 50$ 0.85 if $\mu = 56$



[8] The following discussion and the figures are based on Helen M. Walker and Joseph Lev, *Statistical Inference,* Holt, Rinehart and Winston, New York, 1953, pp. 161–162.

FIGURE A.14

Power function of test of hypothesis $\mu = 50$ when $N = 25$, $\sigma = 10$, and $\alpha = 0.05$.



are 0.17 (for $\mu = 48$), 0.05 (for $\mu = 50$), 0.17 (for $\mu = 52$) and 0.85 (for $\mu = 56$). As you can see from this figure, whenever the true value of μ differs substantially from the hypothesis under consideration (which here is $\mu = 50$), the probability of rejecting the hypothesis is high but when the true value is not very different from the value given under the null hypothesis, the probability of rejection is small. Intuitively, this should make sense if the null and alternative hypotheses are very closely bunched.

This can be seen further if you consider Figure A.14, which is called the **power function graph;** the curve shown there is called the **power curve.**

The reader will by now realize that the confidence coefficient $(1 - \alpha)$ discussed earlier is simply 1 minus the probability of committing a type I error. Thus a 95 percent confidence coefficient means that we are prepared to accept at the most a 5 percent probability of committing a type I error—we do not want to reject the true hypothesis by more than 5 out of 100 times.

The p Value, or Exact Level of Significance

Instead of preselecting α at arbitrary levels, such as 1, 5, or 10 percent, one can obtain the ***p*** **(probability) value,** or **exact level of significance** of a test statistic. The p value is defined as *the lowest significance level at which a null hypothesis can be rejected*.

Suppose that in an application involving 20 df we obtain a t value of 3.552. Now the p value, or the exact probability, of obtaining a t value of 3.552 or greater can be seen from Table D.2 as 0.001 (one-tailed) or 0.002 (two-tailed). We can say that the observed t value of 3.552 is statistically significant at the 0.001 or 0.002 level, depending on whether we are using a one-tail or two-tail test.

Several statistical packages now routinely print out the p value of the estimated test statistics. Therefore, the reader is advised to give the p value wherever possible.

Sample Size and Hypothesis Tests

In survey-type data involving hundreds of observations, the null hypothesis seems to be rejected more frequently than in small samples. It is worth quoting Angus Deaton here:

> As the sample size increases, and provided we are using a consistent estimation procedure, our estimates will be closer to the truth, and less dispersed around it, so that discrepancies that are undetectable with small sample size will lead to rejection in large samples. Large sample sizes are like greater resolving power on a telescope; features that are not visible from a distance become more and more sharply delineated as the magnification is turned up.[9]

[9] Angus Deaton, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy,* The Johns Hopkins University Press, Baltimore, 2000, p. 130.

Following Leamer and Schwarz, Deaton suggests adjusting the standard critical values of the F and χ^2 tests as follows: *Reject the null hypothesis when the computed F value exceeds the logarithm of the sample size, that is, ln, and when the computed χ^2 statistic for q restriction exceeds* ***qln****, where l is the natural logarithm and where n is the sample size.* These critical values are known as **Leamer–Schwarz** critical values.

Using Deaton's example, if $n = 100$, the null hypothesis would be rejected only if the computed F value were greater than 4.6, but if $n = 10{,}000$, the null hypothesis would be rejected when the computed F value exceeded 9.2.

The Test of Significance Approach

Recall that

$$Z_i = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

In any given application, $\bar{X}$ and n are known (or can be estimated), but the true μ and σ are not known. But if σ is specified and we assume (under H_0) that $\mu = \mu^*$, a specific numerical value, then Z_i can be directly computed and we can easily look at the normal distribution table to find the probability of obtaining the computed Z value. If this probability is small, say, less than 5 percent or 1 percent, we can reject the null hypothesis—if the hypothesis were true, the chances of obtaining the particular Z value should be very high. This is the general idea behind the test of significance approach to hypothesis testing. The key idea here is the test statistic (here the Z statistic) and its probability distribution under the assumed value $\mu = \mu^*$. Appropriately, in the present case, the test is known as the ***Z* test,** since we use the Z (standardized normal) value.

Returning to our example, if $\mu = \mu^* = 69$, the Z statistic becomes

$$\begin{aligned} Z &= \frac{\bar{X} - \mu^*}{\sigma/\sqrt{n}} \\ &= \frac{67 - 69}{2.5/\sqrt{100}} \\ &= -2/0.25 = -8 \end{aligned}$$

If we look at the normal distribution table (Table D.1), we see that the probability of obtaining such a Z value is extremely small. (*Note:* The probability of a Z value exceeding 3 or -3 is about 0.001. Therefore, the probability of Z exceeding 8 is even smaller.) Therefore, we can reject the null hypothesis that $\mu = 69$; given this value, our chance of obtaining $\bar{X}$ of 67 is extremely small. We therefore doubt that our sample came from the population with a mean value of 69. Diagrammatically, the situation is depicted in Figure A.15.

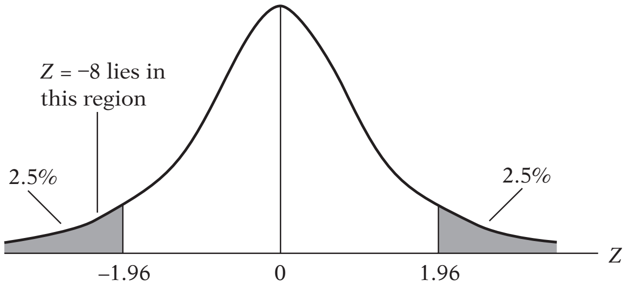


FIGURE A.15
The distribution of the Z statistic.

In the language of test of significance, when we say that a test (statistic) is significant, we generally mean that we can reject the null hypothesis. And the test statistic is regarded as significant if the probability of our obtaining it is equal to or less than α, the probability of committing a type I error. Thus if $\alpha = 0.05$, we know that the probability of obtaining a Z value of -1.96 or 1.96 is 5 percent (or 2.5 percent in each tail of the standardized normal distribution). In our illustrative example Z was -8. Hence the probability of obtaining such a Z value is much smaller than 2.5 percent, well below our prespecified probability of committing a type I error. That is why the computed value of $Z = -8$ is statistically significant; that is, we reject the null hypothesis that the true μ^* is 69. Of course, we reached the same conclusion using the confidence interval approach to hypothesis testing.

We now summarize the steps involved in testing a statistical hypothesis:

Step 1. State the null hypothesis H_0 and the alternative hypothesis H_1 (e.g., H_0: $\mu = 69$ and H_1: $\mu \neq 69$).

Step 2. Select the test statistic (e.g., $\bar{X}$).

Step 3. Determine the probability distribution of the test statistic (e.g.,$\bar{X} \sim N(\mu, \sigma^2/n)$.

Step 4. Choose the level of significance (i.e., the probability of committing a type I error) α.

Step 5. Using the probability distribution of the test statistic, establish a $100(1 - \alpha)\%$ confidence interval. If the value of the parameter under the null hypothesis (e.g., $\mu = \mu^* = 69$) lies in this confidence region, the region of acceptance, do not reject the null hypothesis. But if it falls outside this interval (i.e., it falls into the region of rejection), you may reject the null hypothesis. Keep in mind that in not rejecting or rejecting a null hypothesis you are taking a chance of being wrong α percent of the time.

References

For the details of the material covered in this appendix, the reader may consult the following references:

Hoel, Paul G., *Introduction to Mathematical Statistics,* 4th ed., John Wiley & Sons, New York, 1974. This book provides a fairly simple introduction to various aspects of mathematical statistics.

Freund, John E., and Ronald E. Walpole, *Mathematical Statistics,* 3d ed., Prentice Hall, Englewood Cliffs, NJ, 1980. Another introductory textbook in mathematical statistics.

Mood, Alexander M., Franklin A. Graybill, and Duane C. Boes, *Introduction to the Theory of Statistics,* 3d ed., McGraw-Hill, New York, 1974. This is a comprehensive introduction to the theory of statistics but is somewhat more difficult than the preceding two textbooks.

Newbold, Paul, *Statistics for Business and Economics,* Prentice Hall, Englewood Cliffs, NJ, 1984. A comprehensive nonmathematical introduction to statistics with lots of worked-out problems.