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

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

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

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

A particular case is

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

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

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

and

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

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

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

APPENDIX C

ESTIMATION AND INFERENCE

C.1 INTRODUCTION

The probability distributions discussed in Appendix B serve as models for the underlying data generating processes that produce our observed data. The goal of statistical inference in econometrics is to use the principles of mathematical statistics to combine these theoretical distributions and the observed data into an empirical model of the economy. This analysis takes place in one of two frameworks, classical or Bayesian. The overwhelming majority of empirical study in

FN 1

econometrics has been done in the classical framework. Our focus, therefore, will be on classical methods of inference. Bayesian methods are discussed in Chapter ~~18~~.[1] 16

C.2 SAMPLES AND RANDOM SAMPLING

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

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

C.3 DESCRIPTIVE STATISTICS

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

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

$$\textbf{median: } M = \text{middle ranked observation}, \qquad \text{(C-1)}$$

$$\textbf{sample midrange: } \text{midrange} = \frac{\text{maximum} + \text{minimum}}{2}.$$

The dispersion of the sample observations is usually measured by the

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

Other measures, such as the average absolute deviation from the sample mean, are also used, although less frequently than the standard deviation. The shape of the distribution of values is often of interest as well. Samples of income or expenditure data, for example, tend to be highly

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

skewed while financial data such as asset returns and exchange rate movements are relatively more symmetrically distributed but are also more widely dispersed than other variables that might be observed. Two measures used to quantify these effects are the

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

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

FIG C.1

If the sample contains data on more than one variable, we will also be interested in measures of association among the variables. A **scatter diagram** is useful in a bivariate sample if the sample contains a reasonable number of observations. Figure C.1 shows an example for a small data set. If the sample is a multivariate one, then the degree of linear association among the variables can be measured by the pairwise measures

$$\textbf{covariance:}\, s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}, \tag{C-3}$$

$$\textbf{correlation:}\, r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

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

$$\textbf{covariance matrix:}\, \mathbf{S} = [s_{ij}], \tag{C-4}$$

or

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

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

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

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

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

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

$$\begin{aligned} s_{ax} &= |a| s_x, \\ s_{ax,by} &= (ab) s_{xy}. \end{aligned} \tag{C-8}$$

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

Example C.1 Descriptive Statistics for a Random Sample

Appendix Table FC1 contains a (hypothetical) sample of observations on income and education (The observations all appear in the calculations of the means below.) A scatter diagram appears in Figure C.1. It suggests a weak positive association between income and education in these data. The box and whisker plot for income at the left of the scatter plot shows the distribution of the income data as well.

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

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

Standard deviations:

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

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

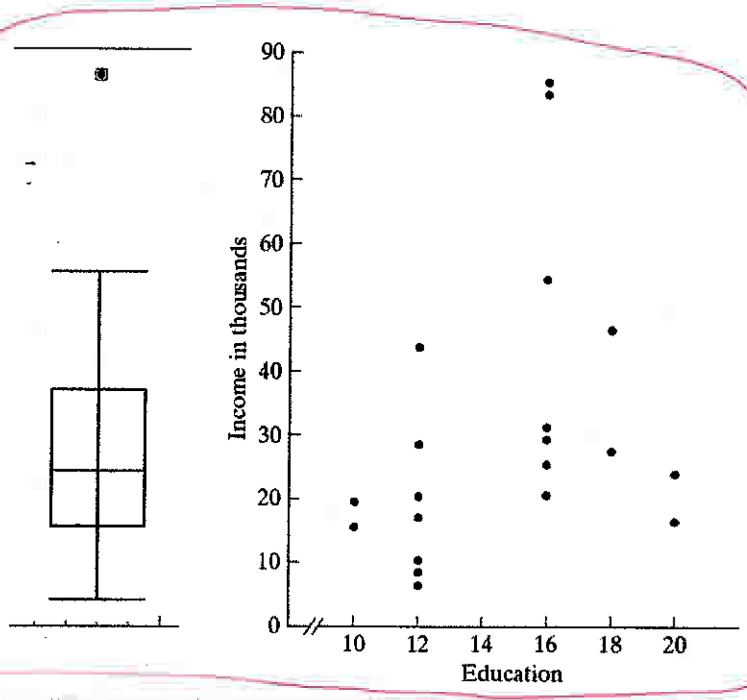


FIGURE C.1 Box and Whisker Plot for Income and Scatter Diagram for Income and Education.

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

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

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

The statistics just described will provide the analyst with a more concise description of the data than a raw tabulation. However, we have not, as yet, suggested that these measures correspond to some underlying characteristic of the process that generated the data. We do assume that there is an underlying mechanism, the data-generating process, that produces the data in hand. Thus, these serve to do more than describe the data; they characterize that process, or population. Because we have assumed that there is an underlying probability distribution, it might be useful to produce a statistic that gives a broader view of the DGP. The **histogram** is a simple graphical device that produces this result—see Examples C.3 and C.4 for applications. For small samples or widely dispersed data, however, histograms tend to be rough and difficult to make informative. A burgeoning literature [see, e.g., Pagan and Ullah (1999) and Li and Racine (2007)] has demonstrated the usefulness of the **kernel density estimator** as a substitute for the histogram as a descriptive tool for the underlying distribution that produced a sample of data. The underlying theory of the kernel density estimator is fairly complicated, but the computations are surprisingly simple. The estimator is computed using

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

where $x_1, \ldots, x_n$ are the n observations in the sample, $\hat{f}(x^*)$ denotes the estimated density function, x^* is the value at which we wish to evaluate the density, and h and $K[\cdot]$ are the "bandwidth" and "kernel function" that we now consider. The density estimator is rather like a histogram, in which the bandwidth is the width of the intervals. The kernel function is a weight function which is generally chosen so that it takes large values when x^* is close to x_i and tapers off to zero in as they diverge in either direction. The weighting function used in the example ~~below~~ is the logistic density discussed in Section B.4.7. The bandwidth is chosen to be a function of $1/n$ so that the intervals can become narrower as the sample becomes larger (and richer). The one used for Figure C.2 is $h = 0.9\text{Min}(s, \text{range}/3)/n^{.2}$. (We will revisit this method of estimation in Chapter 14.) Example C.2 illustrates the computation for the income data used in Example C.1.

following/

AQ: Confirm x-ref to Chap. 14

FIG C.2

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

C.4 STATISTICS AS ESTIMATORS—SAMPLING DISTRIBUTIONS

TB C.1

The measures described in the preceding section summarize the data in a random sample. Each measure has a counterpart in the population, that is, the distribution from which the data were drawn. Sample quantities such as the means and the correlation coefficient correspond to population expectations, whereas the kernel density estimator and the values in Table C.1 parallel the population **pdf** and **cdf.** In the setting of a random sample, we expect these quantities to mimic

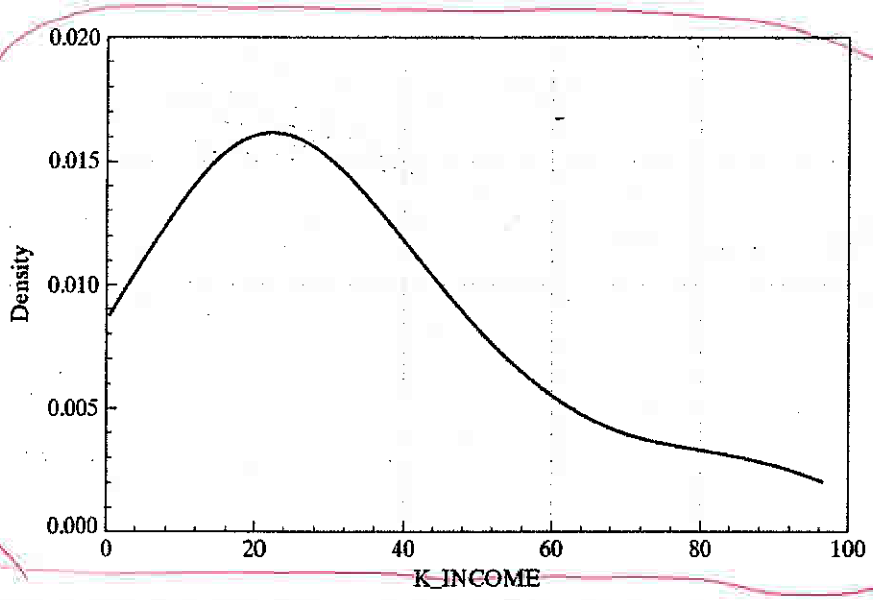


FIGURE C.2 Kernel Density Estimate for Income.

TABLE C.1 Income Distribution

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

the population, although not perfectly. The precise manner in which these quantities reflect the population values defines the sampling distribution of a sample statistic.

DEFINITION C.1 Statistic

A statistic is any function computed from the data in a sample.

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

THEOREM C.1 Sampling Distribution of the Sample Mean

If $x_1, \ldots, x_n$ are a random sample from a population with mean μ and variance σ^2, then $\bar{x}$ is a random variable with mean μ and variance σ^2/n.

Proof: *$\bar{x} = (1/n)\Sigma_i x_i$. $E[\bar{x}] = (1/n)\Sigma_i \mu = \mu$. The observations are independent, so* $\text{Var}[\bar{x}] = (1/n)^2\ \text{Var}[\Sigma_i x_i] = (1/n^2)\Sigma_i \sigma^2 = \sigma^2/n$.

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

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

FIG
C.3

Example C.3 Sampling Distribution of a Sample Mean

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

FIG
C.4

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

Example C.4 Sampling Distribution of the Sample Minimum

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

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

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

Figure C.4 shows the results of a simple sampling experiment you can do to demonstrate this effect. It requires software that will allow you to produce pseudorandom numbers uniformly distributed in the range zero to one and that will let you plot a histogram and control the axes. (We used *NLOGIT*. This can be done with *Stata*, *Excel*, or several other packages.) The experiment consists of drawing 1,000 sets of nine random values, $U_{ij}, i = 1, \ldots 1{,}000, j = 1, \ldots, 9$. To transform these uniform draws to exponential with parameter θ—we used $\theta = 1.5$, use the inverse probability transform—see Section E.2.3. For an exponentially distributed variable, the transformation is $z_{ij} = -(1/\theta)\log(1 - U_{ij})$. We then created $z_{(1)}|3$ from the first three draws and $z_{(1)}|6$ from the other six. The two histograms show clearly the effect on the sampling distribution of increasing sample size from just 3 to 6.

Sampling distributions are used to make inferences about the population. To consider a perhaps obvious example, because the sampling distribution of the mean of a set of normally distributed observations has mean μ, the sample mean is a natural candidate for an estimate of μ. The observation that the sample "mimics" the population is a statement about the sampling

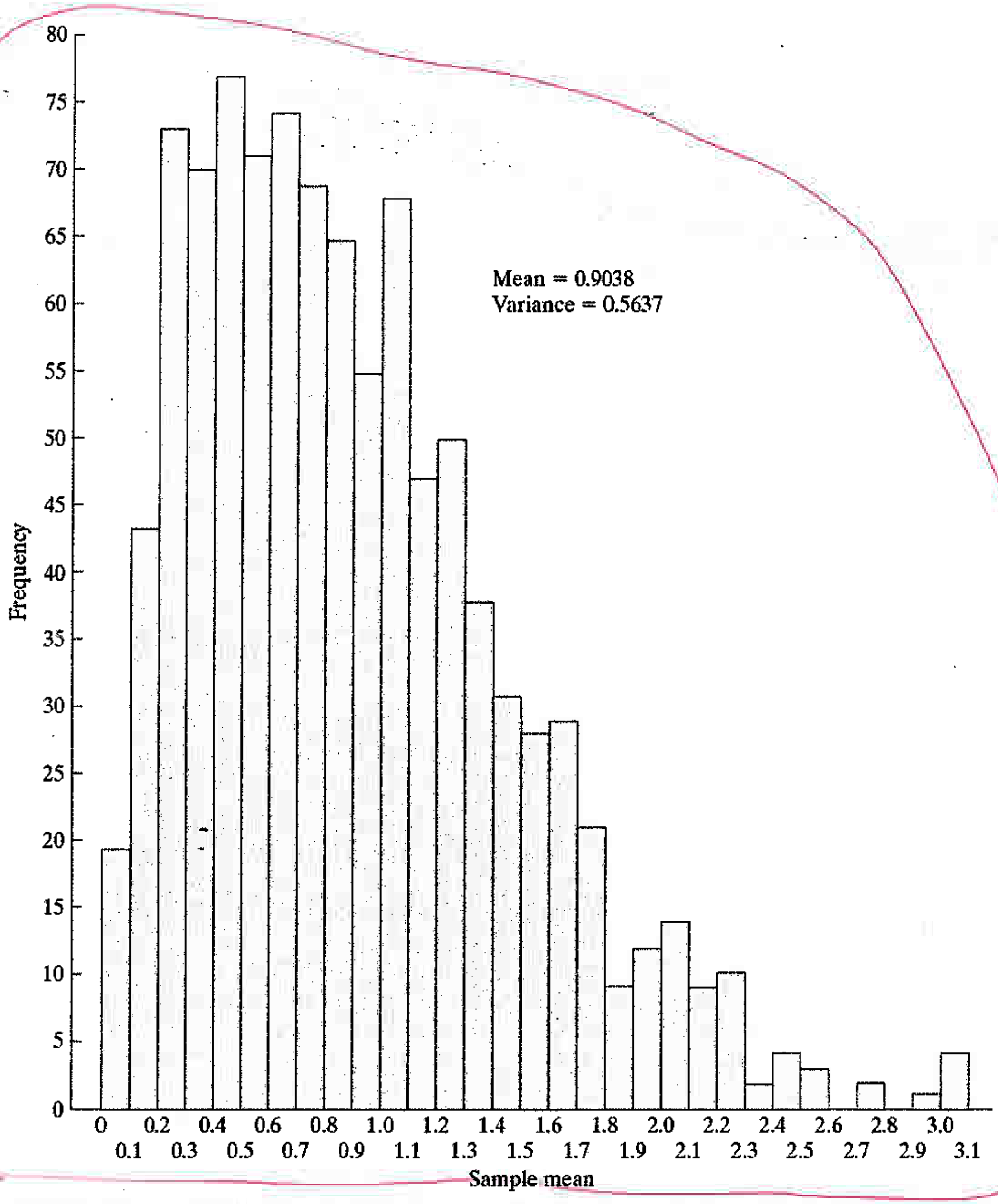


FIGURE C.3 Sampling Distribution of Means of 1,000 Samples of Size 4 from Chi-Squared [1].

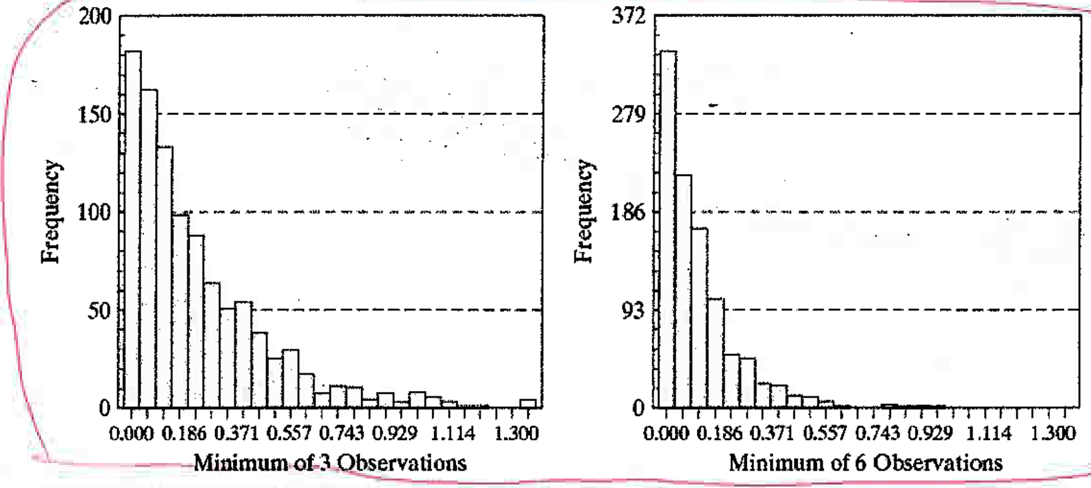


FIGURE C.4 Histograms of the Sample Minimum of 3 and 6 Observations.

distributions of the sample statistics. Consider, for example, the sample data collected in Figure C.3. The sample mean of four observations clearly has a sampling distribution, which appears to have a mean roughly equal to the population mean. Our theory of parameter estimation departs from this point.

C.5 POINT ESTIMATION OF PARAMETERS

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

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

C.5.1 ESTIMATION IN A FINITE SAMPLE

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

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

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

or

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

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

If samples of size n are drawn repeatedly and $\hat{\theta}$ is computed for each one, then the average value of these estimates will tend to equal θ. For example, the average of the 1,000 sample means underlying Figure C.2 is 0.9038, which is reasonably close to the population mean of one. The sample minimum is clearly a biased estimator of the mean; it will almost always underestimate the mean, so it will do so on average as well.

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

DEFINITION C.3 Efficient Unbiased Estimator
An unbiased estimator $\hat{\theta}_1$ is more efficient than another unbiased estimator $\hat{\theta}_2$ if the sampling variance of $\hat{\theta}_1$ is less than that of $\hat{\theta}_2$. That is,

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

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

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

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

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

DEFINITION C.4 Mean Squared Error

The mean squared error of an estimator is

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

FIG C.5

Figure C.5 illustrates the effect. In this example, on average, the biased estimator will be closer to the true parameter than will the unbiased estimator.

Which of these criteria should be used in a given situation depends on the particulars of that setting and our objectives in the study. Unfortunately, the MSE criterion is rarely operational; minimum mean squared error estimators, when they exist at all, usually depend on unknown parameters. Thus, we are usually less demanding. A commonly used criterion is **minimum variance unbiasedness.**

Example C.5 Mean Squared Error of the Sample Variance

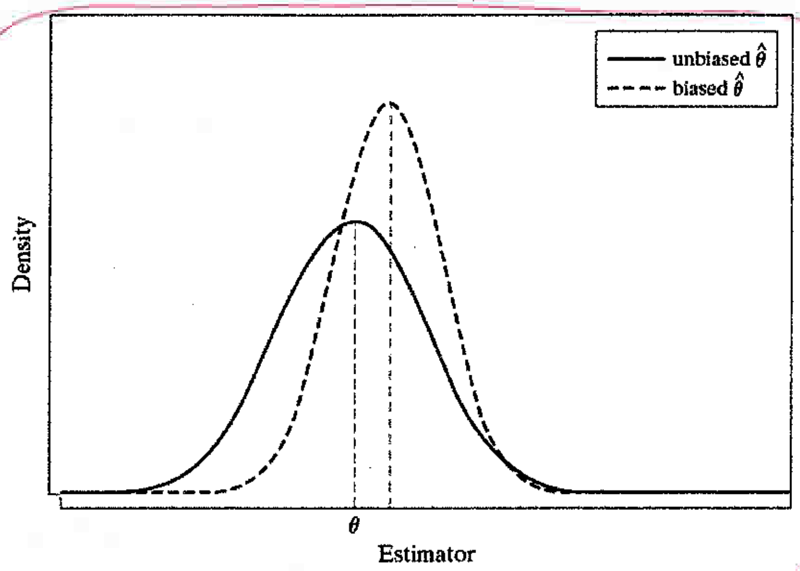
In sampling from a normal distribution, the most frequently used estimator for σ^2 is

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

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

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

FIGURE C.5 Sampling Distributions.



[A proof is based on the distribution of the idempotent quadratic form $(\mathbf{x} - \mathbf{i}\mu)'\mathbf{M}^0(\mathbf{x} - \mathbf{i}\mu)$, which we discussed in Section B11.4.] A less frequently used estimator is

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

This estimator is slightly biased downward:

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

so its bias is

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

But it has a smaller variance than s^2:

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

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

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

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

C.5.2 EFFICIENT UNBIASED ESTIMATION

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

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

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

Example C.6 Likelihood Functions for Exponential and Normal Distributions

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

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

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

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

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

acute

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

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

The quantity $I(\theta)$ is the information number for the sample. We will prove the result that the negative of the expected second derivative equals the expected square of the first derivative in Chapter 16. Proof of the main result of the theorem is quite involved. See, for example, Stuart and Ord (1989).

AQ: Confirm x-ref to Chap. 16

14

The regularity conditions are technical in nature. (See Section 16.4.1.) Loosely, they are conditions imposed on the density of the random variable that appears in the likelihood function; these conditions will ensure that the Lindeberg–Levy central limit theorem will apply to moments of the sample of observations on the random vector $\mathbf{y} = \partial \ln f(x_i \mid \theta)/\partial\theta, i = 1, \ldots, n$. Among the conditions are finite moments of x up to order 3. An additional condition normally included in the set is that the range of the random variable be independent of the parameters.

acute

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

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

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

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

Example C.7 Variance Bound for the Poisson Distribution
For the Poisson distribution,

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

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

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

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

The sum of n identical Poisson variables has a Poisson distribution with parameter equal to n times the parameter of the individual variables. Therefore, the actual distribution of the first derivative will be that of a linear function of a Poisson distributed variable. Because $E[\sum_{i=1}^{n} x_i] = nE[x_i] = n\theta$, the variance bound for the Poisson distribution is $[I(\theta)]^{-1} = \theta/n$. (Note also that the same result implies that $E[\partial \ln L/\partial\theta] = 0$, which is a result we will use in Chapter 16. The same result holds for the exponential distribution.)

AU: Confirm x-ref to Chap. 16

acute

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

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

will be a nonnegative definite matrix.

acute

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

acute

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

DEFINITION C.5 Minimum Variance Linear Unbiased Estimator (MVLUE)

An estimator is the minimum variance linear unbiased estimator or best linear unbiased estimator (BLUE) if it is a linear function of the data and has minimum variance among linear unbiased estimators.

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

C.6 INTERVAL ESTIMATION

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

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

Example C.8 Confidence Intervals for the Normal Mean

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

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

and

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

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

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

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

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

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

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

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

$$\text{lower}(\mathbf{X}) = \hat{\theta} - e_1,$$
$$\text{upper}(\mathbf{X}) = \hat{\theta} + e_2,$$

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

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

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

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

In a sample of 25, $\bar{x} = 1.63$ and $s = 0.51$. Construct a 95 percent confidence interval for μ. Assuming that the sample of 25 is from a normal distribution,

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

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

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

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

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

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

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

C.7 HYPOTHESIS TESTING

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

C.7.1 CLASSICAL TESTING PROCEDURES

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

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

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

DEFINITION C.6 Size of a Test

The probability of a type I error is the size of the test. This is conventionally denoted α and is also called the significance level.

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

DEFINITION C.7 Power of a Test

The power of a test is the probability that it will correctly lead to rejection of a false null hypothesis:

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

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

Example C.10 Testing a Hypothesis About a Mean

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

FIG C.6

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

DEFINITION C.8 Most Powerful Test

A test is most powerful if it has greater power than any other test of the same size.

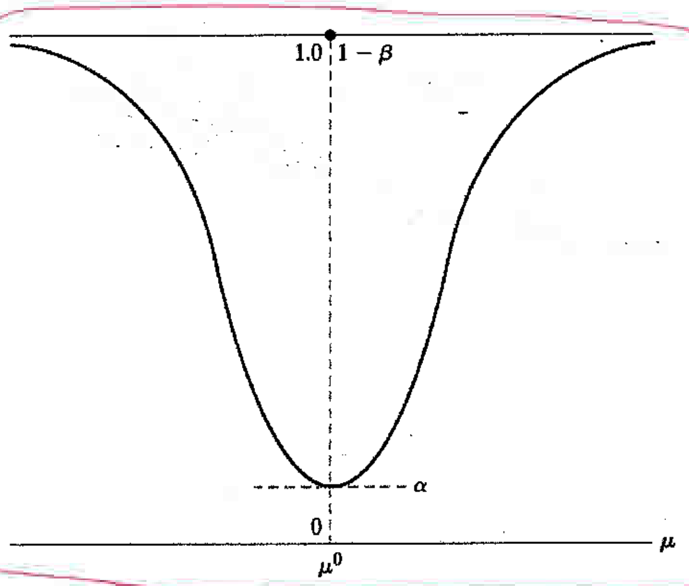


FIGURE C.6 Power Function for a Test.

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

DEFINITION C.9 Unbiased Test

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

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

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

DEFINITION C.10 Consistent Test

A test is consistent if its power goes to one as the sample size grows to infinity.

Example C.11 Consistent Test About a Mean

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

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

C.7.2 TESTS BASED ON CONFIDENCE INTERVALS

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

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

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

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

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

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

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

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

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

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

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

If the test is one-sided, as in

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

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