THE THEORY OF POINT ESTIMATION

A point estimator uses the information available in a sample to obtain a single number that estimates a population parameter. There can be a variety of estimators of the same parameter that derive from different principles of estimation, and it is necessary to choose amongst them. Therefore, criteria are required that will indicate which are the acceptable estimators and which of these is the best in given circumstances.

Often, the choice of an estimate is governed by practical considerations such as the ease of computation or the ready availability of a computer program. In what follows, we shall ignore such considerations in order to concentrate on the criteria that, ideally, we would wish to adopt. The circumstances that affect the choice are more complicated than one might imagine at first.

Consider an estimator $\hat{\theta}_n$ based on a sample of size n that purports to estimate a population parameter θ, and let $\tilde{\theta}_n$ be any other estimator based on the same sample. If

$$P(\theta - c_1 \leq \hat{\theta}_n \leq \theta + c_2) \geq P(\theta - c_1 \leq \tilde{\theta}_n \leq \theta + c_2)$$

for all values of $c_1, c_2 > 0$, then $\hat{\theta}_n$ would be unequivocally the best estimator. However, it is possible to show that an estimator has such a property only in very restricted circumstances.

Instead, we must evaluate an estimator against a list of partial criteria, of which we shall itemise the principal ones. The first is the criterion of unbiasedness.

(a) $\hat{\theta}$ is an unbiased estimator of the population parameter θ if $E(\hat{\theta}) = \theta$.

On it own, this is an insufficient criterion. Thus, for example, we observe that both a single element x_i from a random sample of size n and the average $\bar{x} = \sum x_i/n$ of all the sample elements constitute unbiased estimators of the population mean $\mu = E(x_i)$. However, the sample average is always the preferred estimator.

This suggests that we must also consider the dispersion of the estimators, which, in the cases of the examples above, are $V(x_i) = \sigma^2$, which is the population variance, and $V(\bar{x}) = \sigma^2/n$, which is some fraction of that variance.

We observe that $V(\bar{x}) \to 0$ and $n \to \infty$. The collapse of its variance, together with the fact that it is an unbiased estimate of μ, ensures that $\bar{x}$ converges on μ as the sample size increases, which is to say that it is a consistent estimator.

Some quite reasonable estimators do not have finite expected values or finite variances, in which case the criteria that make reference to these moments become irrelevant.

For example, the obvious estimate of $1/\mu$, based on a random sample $x_i \sim N(\mu, \sigma^2); i = 1, \ldots, n$, is $1/\bar{x}$. However, for any finite value of μ there is a finite probability that $\bar{x}$ will fall in the arbitrarily small interval $[-\epsilon, \epsilon]$ that contains zero. Division by zero generates infinity; and, for this reason, the integral that would, otherwise, define the expectation of $1/\bar{x}$ does not converge. For a while, such pathologies will be ignored, and we shall persist in itemising the criteria.

(b) $\hat{\theta}$ has minimum variance within a given class of estimators if $E[\{\hat{\theta} - E(\hat{\theta})\}^2] \leq E[\{\tilde{\theta} - E(\tilde{\theta})\}^2]$, where $\tilde{\theta}$ is any other estimator of the same class.

However, it is possible that a quite reasonable estimator has no finite variance. For estimators that do have finite first and second moments, the following criterion

implies a trade-off between (a) and (b); and it suggests that it may be worthwhile to accept some bias in return for a reduction in variance.

(c) $\hat{\theta}$ has minimum mean-square error if $E(\{\hat{\theta}-\theta\}^2) \leq E(\{\tilde{\theta}-\theta\}^2)$, where θ is the true parameter value and $\tilde{\theta}$ is any other estimator of the same class.

To elucidate this criterion consider the following:

$$\begin{aligned} E\left(\{\hat{\theta}-\theta\}^2\right) &= E\left(\left[\{\hat{\theta}-E(\hat{\theta})\}-\{\theta-E(\hat{\theta})\}\right]^2\right) \\ &= E\left[\{\hat{\theta}-E(\hat{\theta})\}^2\right] - 2E\left[\{\hat{\theta}-E(\hat{\theta})\}\{\theta-E(\hat{\theta})\}\right] + E\left[\{\theta-E(\hat{\theta})\}^2\right] \\ &= V(\hat{\theta}) + E\left[\{\theta-E(\hat{\theta})\}^2\right]. \end{aligned}$$

The final equality follows from the fact that the cross product vanishes, and it shows that the mean-square error is the sum of the variance and the square of the bias.

Unfortunately, the criterion on its own is too ambitious, and we must seek the minimum-mean-square error estimator within a restricted class of estimators, such as the class of estimators that are linear functions of the sample observations.

If bias is disallowed, then the result is the following criterion:

(d) $\hat{\theta}$ is minimum-variance unbiased estimator of θ if it has $E(\hat{\theta}) = \theta$ and if $E\big(\{\hat{\theta}-\theta\}^2\big) \leq E\big(\{\tilde{\theta}-\theta\}^2\big)$, where $\tilde{\theta}$ is any other unbiased estimator with $E(\tilde{\theta}) = \theta$.

This criterion is also too ambitious. We can find the minimum variance unbiased estimator or the "best" estimator, as it is sometimes called, only if we happen to know the functional form of the distribution of the population. If we don't know this functional form, then we might be content with finding the "best" linear unbiased estimator, which is also known as the BLUE estimator.

(e) The "best" or, equivalently, the minimum-variance linear unbiased estimator of θ is a linear function $\hat{\theta}$ of the sample data which has $E(\hat{\theta}) = \theta$ and $E(\{\hat{\theta}-\theta\}^2) \leq E(\{\tilde{\theta}-\theta\}^2)$, where $\tilde{\theta}$ is any other linear unbiased estimator with $E(\tilde{\theta}) = \theta$.

The sample mean $\bar{x} = \sum x_i/n$ is the minimum-variance unbiased estimator of μ in the normal population with a p.d.f. $N(\mu, \sigma^2)$. Also, $\bar{x}$ is certainly the best *linear* unbiased estimator, regardless of the functional form of the distribution of the parent population.

Stochastic Convergence

We are also interested in the asymptotic properties of the estimates as $n \to \infty$. In fact, for estimators with complicated sample distributions and for estimators that lack finite moments, we may have to depend entirely upon as assessment of their relative asymptotic properties. Such estimators have distributions that tend to limiting distributions, as $n \to \infty$, that are usually normal. We can compare the estimates by comparing their limiting distributions.

It is a simple matter to define what is meant by the convergence of a sequence $\{a_n\}$ of nonstochastic elements. We say that the sequence is convergent or, equivalently, that it tends to a limiting constant a if, for any small positive number ϵ, there exists a number $N = N(\epsilon)$ such that $|a_n - a| < \epsilon$ for all $n > N$. This is indicated by writing $\lim(n \to \infty)a_n = a$ or, alternatively, by stating that $a_n \to a$ as $n \to \infty$.

The question of the convergence of a sequence of random variables is less straightforward, and there are a variety of modes of convergence. Let $\{x_t\}$ be a sequence of random variables and let c be a constant. Then

(f) x_t converges to c weakly in probability, written $x_t \xrightarrow{P} c$ or $\text{plim}(x_t) = c$, if, for every $\epsilon > 0$,

$$\lim(t \to \infty)P(|x_t - c| > \epsilon) = 0,$$

(g) x_t converges to c strongly in probability or almost certainly, written $x_t \xrightarrow{a.s.} c$, if, for every $\epsilon > 0$,

$$\lim(\tau \to \infty)\Big(\bigcup_{t>\tau} P(|x_t - c| > \epsilon\Big) = 0,$$

(h) x_t converges to c in mean square, written $x_t \xrightarrow{m.s.} c$, if

$$\lim(t \to \infty)E(|x_t - c|^2) = 0.$$

In the same way, we define the convergence of a sequence of random variables to a random variable. Thus

(j) A sequence $\{x_t\}$ of random variables is said to converge to a random variable in the sense of (f), (g) or (h) if the sequence $\{x_t - x\}$ converges to zero in that sense.

Of these three criteria of convergence, weak convergence in probability is the most commonly used in econometrics. The other criteria are too stringent. Consider the criterion of almost sure convergence which can also be written as $\lim(\tau \to \infty)P(\bigcap_{t>\tau} |x_t - c| \leq \epsilon) = 1$. This requires that, in the limit, all the elements of $\{x_t\}$ with $t > \tau$ should lie simultaneously in the interval $[c - \epsilon, c + \epsilon]$ with a probability of one.

The condition of weak convergence in probability requires much less: it requires only that single elements, taken separately, should have a probability of one of lying in this interval. Clearly

If x_t converges almost certainly to c, then it converges to c weakly in probability. Thus $x_t \xrightarrow{a.s.} c$ implies $x_t \xrightarrow{P} c$.

The disadvantage of the criterion of mean-square convergence is that it requires the existence of second-order moments; and, in many econometric applications, it cannot be guaranteed that an estimator will possess such moments. In fact,

If x_t converges in mean square, then it also converges weakly in probability, so that $x_t \xrightarrow{m.s.} c$ implies $x_t \xrightarrow{P} c$.

A theorem that finds considerable use in econometrics is Slutsky's theorem concerning the probability limits of variables that are continuous functions of other random variables that have well-defined probability limits:

If $g(x_t)$ is a continuous function and if $\text{plim}(x_t) = c$ is a constant, then $\text{plim}\{g(x_t)\} = g\{\text{plim}(x_t)\}$.

The concept of convergence in distribution has equal importance in econometrics with the concept of convergence in probability. It is fundamental to the proof of the central limit theorem.

(c) Let $\{x_t\}$ be a sequence of random variables and let $\{F_t\}$ be the corresponding sequence of distribution functions. Then x_t is said to converge in distribution to a random variable x with a distribution function F, written $x_t \xrightarrow{D} x$, if F_t converges to F at all points of continuity of the latter.

This means simply that, if x^* is any point in the domain of F such that $F(x^*)$ is continuous, then $F_t(x^*)$ converges to $F(x^*)$ in the ordinary mathematical sense. We call F the limiting distribution or asymptotic distribution of x_t.

Weak convergence in probability is sufficient to ensure a convergence in distribution. Thus

If x_t converges to a random variable x weakly in probability, it also converges to x in distribution. That is, $x_t \xrightarrow{P} x$ implies $x_t \xrightarrow{D} x$.

The Central Limit Theorem

Under very general conditions, the distribution of an estimator will tend to a normal distribution as the sample size increases. This can be the case even when the distribution of the estimator does not possess finite moments for any finite size of sample. This is a remarkable and counterintuitive fact; and it has caused considerable confusion amongst econometricians in the past.

There a wide variety of central limit theorems which adopt different assumptions. We shall consider only the simplest of such theorems that makes rather strong assumptions. Its proof depends upon showing that the moment generating function of the distribution of the statistic in question converges on that of a normal distribution. Then, it follows that the distribution of the statistic will converge to a normal distribution. Before we can embark on a proof, we need to find the moment generating function of a normally distributed random variable.

Recall that the probability density function of a normally distributed random variable x with a mean of $E(x) = \mu$ and a variance of $V(x) = \sigma^2$ is

$$N(x; \mu, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

We need only consider the case where $\mu = 0$ and $\sigma^2 = 1$. Then, we have a standard normal, denoted by $N(z; 0, 1)$, and the corresponding moment generating function

is defined by

$$M_z(t) = E(e^{zt}) = \int e^{zt} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$
$$= e^{\frac{1}{2}t^2}.$$

To demonstrate this result, the exponential terms may be gathered and rearranged to give

$$\exp\{zt\} \exp\{-\tfrac{1}{2}z^2\} = \exp\{-\tfrac{1}{2}z^2 + zt\}$$
$$= \exp\{-\tfrac{1}{2}(z-t)^2\} \exp\{\tfrac{1}{2}t^2\}.$$

Then

$$M_z(t) = e^{\frac{1}{2}t^2} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} dz$$
$$= e^{\frac{1}{2}t^2},$$

where the final equality follows from the fact that the expression under the integral is the $N(z; \mu = t, \sigma^2 = 1)$ probability density function, which integrates to unity.

We are now is a position to state and to prove the Lindberg–Levy version of the central limit theorem which is as follows:

Let $x_t; t = 1, 2, \ldots, n$ be a sequence of independent and identically distributed random variables with $E(x_t) = \mu$ and $V(x_t) = \sigma^2$. Then

$$z_n = \sqrt{n}\left(\frac{\bar{x} - \mu}{\sigma}\right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left(\frac{x_t - \mu}{\sigma}\right)$$

converges in distribution to $z \sim N(0, 1)$. Equivalently, the limiting distribution of $\sqrt{n}(\bar{x} - \mu)$ is the normal distribution $N(0, \sigma^2)$.

Proof. First we recall that the moment generating function of the standard normal variate $z \sim N(0, 1)$ is $M_z(t) = \exp\{t^2/2\}$. We must show that the moment generating function M_n of z_n converges to M_z as $n \to \infty$. Let us write $z_n = n^{-1/2} \sum z_t$, where $z_t = (x_t - \mu)/\sigma$ has $E(z_t) = 0$ and $E(z_t^2) = 1$. The moment generating function of z_t can now be written as

$$M^0(t) = 1 + tE(z_t) + \frac{t^2 E(z_t^2)}{2} + o(t^2)$$
$$= 1 + \frac{t^2}{2} + o(t^2),$$

where $o(t)^2$ denotes a term that is of an order smaller than t^2. Since $z_n = n^{-1/2} \sum z_t$ is a sum of independent and identically distributed random variables, it follows that its moment-generating function can be written, in turn, as

$$M_n\Big(\frac{t}{\sqrt{n}}\Big) = \Big[M^0\Big(\frac{t}{\sqrt{n}}\Big)\Big]^n$$
$$= \Big[1 + \frac{t^2}{2n} + o\big(\frac{t^2}{n}\big)\Big]^n.$$

Letting $n \to \infty$, and bearing in mind that

$$e^x = \lim(n \to \infty)\left(1 + \frac{x}{n}\right)^n,$$

we find that $\lim(n \to \infty)M_n = \exp\{t^2/2\} = M_z$, which proves the theorem.

Maximum Likelihood Estimation

There are various principles of estimation that can be adopted. Those that are favoured are liable, in general, to produce estimates with desirable properties. The principle of maximum-likelihood estimation is a leading example.

Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with a probability density function $f(x, \theta)$. The joint p.d.f of the sample will be

$$L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x, \theta).$$

The principle of maximum likelihood suggests that we should choose for an estimate the value of θ that maximises the function L given the sample values $x_1, x_2, \ldots, x_n$. That is to say, it proposes that the highest possible probability measure should be attributed to the sample that has actually transpired.

However, since L is regarded as a function of θ rather than a function of the sample values, it is describes as a likelihood function rather than as a probability density function.

Under normal circumstances, the maximum-likelihood estimate $\hat{\theta}$ can be found by solving the equation $dL/d\theta = 0$, which is the first-order condition for a maximum, or by solving the equivalent equation $d\log L/d\theta = 0$. Since $\log L$ is a monotonic transformation of L, the minimising value of θ is the same for both functions.

Example. Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with a normal p.d.f. $N(\mu, \sigma^2)$. Then

$$\begin{aligned} L(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} N(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x_i-\mu)/\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{\sum_i -\frac{1}{2}(x_i - \mu)/\sigma^2\right\}. \end{aligned}$$

Taking natural logarithms gives

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2.$$

To find the ML estimator of μ, we solve

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_i (x_i - \mu) = 0, \qquad \text{from which} \qquad \hat{\mu} = \frac{1}{n}\sum_i x_i = \bar{x}.$$

To find the ML estimator of σ^2, we evaluate

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (x_i - \mu)^2 = 0.$$

Multiplying the equation by $2\sigma^4/n$ and rearranging shows that

$$\sigma^2(\mu) = \frac{1}{n}\sum_i (x_i - \mu)^2,$$

whence, on substituting the ML estimate of $\hat{\mu} = \bar{x}$ for μ, we get

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2.$$