Limit Theorems

Consider making repeated measurements of some quantity where each measurement is beset by an unknown error. To estimate the quantity, we can form the average of the measurements. Under a wide variety of conditions concerning the propagation of the errors, we are liable to find that the average converges upon the true value of the quantity.

To illustrate this convergence, let us imagine that each error is propagated independently with a zero expected value and a finite variance. Then there is an upper bound on the probability that the error will exceed a certain size. In the process of averaging the measurements, these bounds are transmuted into upper bounds on the probability of finite deviations of the average from the true value of the unknown quantity; and, as the number of measurements comprised in the average increases indefinitely, this bound tends to zero.

We shall demonstrate this result mathematically. Let $\{x_t; t = 1, \ldots, T, \ldots\}$ be a sequence of measurements, and let μ be the unknown quantity. Then the errors are $x_t - \mu$ and, by our assumptions, $E(x_t - \mu) = 0$ and $E\{(x_t - \mu)^2\} = \sigma_t^2$. Equivalently, $E(x_t) = \mu$ and $V(x_t) = \sigma_t^2$.

We begin by establishing an upper bound for the probability $P(|x_t - \mu| > \epsilon)$. Let $g(x)$ be a non-negative function of $x \sim f(x)$, and let $\mathcal{S} = \{x; g(x) > k\}$ be the set of all values of x for which $g(x)$ exceeds a certain constant. Then

$$
(1)\qquad \begin{aligned} E\{g(x)\} &= \int_x g(x)f(x)dx \\ &\geq \int_{\mathcal{S}} kf(x)dx = kP\{g(x) > k\}; \end{aligned}
$$

and it follows that

(2) If $g(x)$ is a non-negative function of a random variable x, then, for every $k > 0$, we have $P\{g(x) > k\} \leq E\{g(x)\}/k$.

This result is know as Chebyshev's inequality. Now let $g(x_t) = |x_t - \mu|^2$. Then $E\{g(x_t)\} = V(x_t) = \sigma_t^2$ and, setting $k = \epsilon^2$, we have $P(|x_t - \mu|^2 > \epsilon^2) \leq \sigma_t^2/\epsilon^2$. Thus

(3) if $x_t \sim f(x_t)$ has $E(x_t) = \mu$ and $V(x_t) = \sigma_t^2$, then $P(|x_t - \mu| > \epsilon) \leq \sigma_t^2/\epsilon^2$;

and this gives an upper bound on the probability that an error will exceed a certain magnitude.

Now consider the average $\bar{x} = \sum x_t/T$. Since the errors are independently distributed, we have $V(\bar{x}) = \sum V(x_t)/T^2 = \sum \sigma_t^2/T^2$. Also $E(\bar{x}) = \mu$. On replacing x_t, $E(x_t)$ and $V(x_t)$ in the inequality in (3) by $\bar{x}_T$, $E(\bar{x}_T)$ and $V(x_T)$, we get

$$P(|\bar{x}_T - \mu| > \epsilon) \leq \sum \sigma_t^2/(\epsilon T)^2; \tag{4}$$

and, on taking limits as $T \to \infty$, we find that

$$P(|\bar{x}_T - \mu| > \epsilon) = 0. \tag{5}$$

Thus, in the limit, the probability that $\bar{x}$ diverges from μ by any finite quantity is zero. We have proved a version of a fundamental limit theorem known as the law of large numbers.

Although the limiting distribution of $\bar{x}$ is degenerate, we still wish to know how $\bar{x}$ is distributed in large samples. If we are prepared to make specific assumptions about the distributions of the elements x_t, then we may be able to derive the distribution of $\bar{x}$. Unfortunately, the problem is liable to prove intractable unless we can assume that the elements are normally distributed. However, what is remarkable is that, given that the elements are independent, and provided that their sizes are constrained by the condition that

$$\lim(T \to \infty) P\left(\left|(x_t - \mu)\Big/\sum_{t=1}^{T} \sigma_t^2\right| > \epsilon\right) = 0, \tag{6}$$

the distribution of $\bar{x}$ tends to the normal distribution $N(\mu, \sum \sigma_t^2/T^2)$. This result, which we shall prove in a restricted form, is known as the central limit theorem.

The law of large numbers and the central limit theorem provide the basis for determining the asymptotic properties of econometric estimators. In demonstrating these asymptotic properties, we are usually faced with a number of subsidiary complications. To prove the central limit theorem and to dispose properly of the subsidiary complications, we require a number of additional results. Ideally these results should be stated in terms of vectors, since it is mainly to vectors that they will be applied. However, to do so would be tiresome, and so our treatment is largely confined to scalar random variables. A more extensive treatment of the issues raised in the following section can be found in Rao [*]; and an exhaustive treatment is provided by Loeve [*].

Stochastic Convergence

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.

(7) Let $\{x_t\}$ be a sequence of random variables and let c be a constant. Then

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

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

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

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

(c) 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.

(8) A sequence of $\{x_t\}$ random variables is said to converge to a random variable in the sense of (a), (b) or (c) of (7) 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

(9) 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,

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

This follows directly from Chebychev's inequality whereby

$$P(|x_t - c| > \epsilon) \leq \frac{E\{(x_t - c)^2\}}{\epsilon^2}. \tag{11}$$

A result which is often used in establishing the properties of econometric estimators is the following:

(12) If g is a continuous function and if x_t converges in probability to x, then $g(x_t)$ converges in probability to $g(x)$. Thus $x_t \xrightarrow{P} x$ implies $g(x_t) \xrightarrow{P} g(x)$.

Proof. If x is a constant, then the proof is straightforward. Let $\delta > 0$ be an arbitrary value. Then, since g is a continuous function, there exists a value ϵ such that $|x_t - x| \leq \epsilon$ implies $|g(x_t) - g(x)| \leq \delta$. Hence $P(|g(x_t) - g(x)| \leq \delta) \geq P(|x_t - x| \leq \epsilon)$; and so $x_t \xrightarrow{P} x$, which may be expressed as $\lim P(|x_t - x| \leq \epsilon) = 1$, implies $\lim P(|g(x_t) - g(x)| \leq \delta) = 1$ or, equivalently, $g(x_t) \xrightarrow{P} g(x)$.

When x is random, we let δ be an arbitrary value in the interval $(0, 1)$, and we choose an interval $\mathcal{A}$ such that $P(x \in \mathcal{A}) = 1 - \delta/2$. Then, for $x \in \mathcal{A}$, there exists some value ϵ such that $|x_t - x| \leq \epsilon$ implies $|g(x_t) - g(x)| \leq \delta$. Hence

$$\begin{aligned} P(|g(x_t) - g(x)| \leq \delta) &\geq P(\{|x_t - x| \leq \epsilon\} \cap \{x \in \mathcal{A}\}) \\ &\geq P(|x_t - x| \leq \epsilon) + P(x \in \mathcal{A}) - 1. \end{aligned} \tag{13}$$

But there is some value τ such that, for $t > \tau$, we have $P(|x_t - x| \leq \epsilon) > 1 - \delta/2$. Therefore, for $t > \tau$, we have $P(|g(x_t) - g(x)| \leq \delta) > 1 - \delta$, and letting $\delta \to 0$ shows that $g(x_t) \xrightarrow{P} g(x)$.

The proof of such propositions are often considerably more complicated than the intuitive notions to which they are intended to lend rigour. The special case of the proposition above where x_t converges in probability to a constant c is frequently invoked. We may state this case as follows:

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

This is known as Slutsky's theorem.

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.

(15) 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

(16) 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 $x_t \xrightarrow{D} x$.

Proof. Let F and F_t denote the distribution functions of x and x_t respectively, and define $z = x - x_t$. Then $x_t \xrightarrow{P} x$ implies $\lim P(|z_t| > \epsilon) = 0$ for any $\epsilon > 0$. Let y be any continuity point of F. Then

$$
\begin{aligned}
P(x_t < y) &= P(x < y + z_t) \\
&= P(\{x < y + z_t\} \cap \{z_t \leq \epsilon\}) + P(\{x < y + z_t\} \cap \{z_t > \epsilon\}) \\
&\leq P(x < y + \epsilon) + P(z_t > \epsilon),
\end{aligned} \tag{17}
$$

where the inequality follows from the fact that the events in the final expression subsume the events of the preceding expressions. Taking limits at $t \to \infty$ gives $\lim P(x_t < y) \leq P(x, y + \epsilon)$. By a similar argument, we may show that $\lim P(x_t < y) \geq P(x < y - \epsilon)$. By letting $\epsilon \to 0$, we see that $\lim P((x_t < y) = P(x < y)$ or simply that $\lim F_t(y) = F(y)$, which proves the theorem.

A theorem of considerable importance, which lies on our way towards the central limit theorem, is the Helly–Bray theorem as follows:

(18) Let $\{F_t\}$ be a sequence of distribution functions converging to the distribution function F, and let g be any bounded continuous function in the same argument. Then $\int g dF_t \to \int g dF$ as $t \to \infty$.

A proof of this may be found in Rao [*p. 97]. The theorem indicates, in particular, that, if $g(x_t) = \mu_t^r$ is the rth moment of x_t and if $g(x) = \mu^r$ is the rth moment of x, then $x_t \xrightarrow{D} x$ implies $\mu_t^r \to \mu^r$. However, this result must be strongly qualified, for it presumes that the rth moment exists for all elements of the sequence $\{x_t\}$; and this cannot always be guaranteed.

It is one of the bugbears of econometric estimation that whereas, for any reasonable estimator, there is usually a limiting distribution possessing finite moments up to the order r, the small-sample distributions often have no such moments. We must therefore preserve a clear distinction between the moments of the limiting distribution and the limits of the moments of the sampling distributions. Since the small-sample moments often do not exist, the latter concept has little operational validity,

We can establish that a sequence of distributions converges to a limiting distribution by demonstrating the convergence of their characteristic functions.

(19) The characteristic function of a random variable x is defined by $\phi(h) = E(\exp\{ihx\})$ where $i = \sqrt{-1}$.

The essential property of a characteristic function is that it uniquely determined by the distribution function. In particular, if x has a probability density function $f(x)$ so that

$$\phi(h) = \int_{-\infty}^{+\infty} e^{ihx} f(x) dx,$$

then an inversion relationship holds whereby

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ihx} \phi(h) dh,$$

Thus the characteristic function and the probability density function are just Fourier transforms of each other.

Example. The standard normal variate $x \sim N(0,1)$ has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The corresponding characteristic function is

$$\begin{aligned}\phi(h) &= \frac{1}{\sqrt{2\pi}}\int_{\infty}^{+\infty} e^{ihx-x^2/2}dx\\ &= e^{-h^2/2}\frac{1}{\sqrt{2\pi}}\int_{\infty}^{+\infty} e^{-(x-ih)^2/2}dx\\ &= e^{-h^2/2}\frac{1}{\sqrt{2\pi}}\int_{\infty}^{+\infty} e^{-z^2/2}dz\end{aligned}$$

where $z = x - ih$ is a complex variable. The integral of the complex function $\exp\{-z^2/2\}$ can be shown to be equal to the integral of the corresponding function defined on the real line. The latter has a value of $\sqrt{2\pi}$, so

$$\phi(h) = e^{-h^2/2}.$$

Thus the probability density function and the characteristic function of the standard normal variate have the same form. Also, it is trivial to confirm, in this instance, that $f(x)$ and $\phi(h)$ satisfy the inversion relation.

The theorem which is used to establish the convergence of a sequence of distributions states that

(20) If $\phi_t(h)$ is the characteristic function of x_t and $\phi(h)$ is that of x, then x_t converges in distribution to x if and only if $\phi_T(h)$ converges to $\phi(h)$. That is $x_t \xrightarrow{D} x$ if and only if $\phi_t(h) \to \phi(h)$.

Proof. The Helley–Bray theorem establishes that $\phi_t \to \phi$ if $x_t \xrightarrow{D} x$. To establish the converse, let F be the distribution function corresponding to ϕ and let $\{F_t\}$ be a sequence of distribution functions corresponding to the sequence $\{\phi_t\}$. Choose a subsequence $\{F_m\}$ tending to a non-decreasing bounded function G. Now G must be a distribution function; for, by taking limits in the expression $\phi_m(h) = \int e^{ihx}dF_m$, we get $\phi(h) = \int e^{ihx}dG$, and setting $h = 0$ gives $\phi(0) = \int dG = 1$ since, dy definition, $\phi(0) = e^0\int dF = 1$. But the distribution function corresponding to $\phi(h)$ is unique, so $G = F$. All subsequences must necessarily converge to the same distribution function, so $\phi_t \to \phi$ implies $F_t \to F$ or, equivalently $x_t \xrightarrow{D} x$.

We shall invoke this theorem in proving the central limit theorem.

The law of large numbers and the central limit thereon

The theorems of the previous section contribute to the proofs of the two limit theorems which are fundamental to the theory of estimation. The first is the law of large numbers. We have already proved that

(21) If $\{x_t\}$ is a sequence of independent random variables with $E(x_t) = \mu$ and $V(x_t) - \sigma_t^2$, and if $\bar{x} = \sum_{t=1}^{T} x_t/T$, then

$$\lim(T \to \infty)P(|\bar{x} - \mu| > \epsilon) = 0.$$

This theorem states that $\bar{x}$ converges to μ weakly in probability and it is called, for that reason, the weak law of large numbers. In fact, if we assume that the elements of $\{x_t\}$ are independent and identically distributed, we no longer need the assumption that their second moments exist in order to prove the convergence of $\bar{x}$. Thus Khinchine's theorem states that

(22) If $\{x_t\}$ is a sequence of independent and identically distributed random variables with $E(x_t) = \mu$, then $\bar{x}$ tends weakly in probability to μ.

Proof. Let $\phi(h) = E(\exp(ihx_t)\})$ be the characteristic function of x_t. Expanding in a neighbourhood of $h = 0$, we get

$$\phi(h) = E\Big\{1 + ihx_t + \frac{(ihx_t)^2}{2!} + \cdots\Big\}$$

and, since the mean $E(x_t) = \mu$ exists, we can write this as

$$\phi(h) = 1 + i\mu h + o(h),$$

where $o(h)$ is a remainder term of a smaller order than h, so that $\lim(h \to 0)\{o(h)/h\} = 0$. Since $\bar{x} = \sum x_t/T$ is a sum of independent and identically distributed random variables x_t/T, its characteristic function can be written as

$$\begin{aligned}\phi_T^* &= E\Big[\exp\Big\{ih\Big(\frac{x_1}{T} + \cdots + \frac{x_T}{T}\Big)\Big\}\Big]\\ &= \prod_{t=1}^{T} E\Big(\exp\Big\{\frac{ihx_t}{T}\Big\}\Big) = \Big[\phi\Big(\frac{h}{T}\Big)\Big]^T.\end{aligned}$$

On taking limits, we get

$$\begin{aligned}\lim(t \to \infty)\phi_T^* &= \lim\Big\{1 + i\frac{h}{T}\mu + o\Big(\frac{h}{T}\Big)\Big\}^T\\ &= \exp\{ih\mu\}.\end{aligned}$$

which is the characteristic function of a random variable with the probability mass concentrated on μ. This proves the convergence of $\bar{x}$.

It is possible to prove Khinchine's theorem without using a characteristic function as is show for example, by Rao. However, the proof that we have just given has an interesting affinity with the proof of the central limit thereon. The Lindeberg–Levy version of the theorem is as follows:

(23) Let $\{x_t\}$ be a sequence of independent and identically distributed random variables with $E(x_t) = \mu$ and $V(x_t) = \sigma^2$. Then $z_T = (1/\sqrt{T})\sum_{t=1}^{T}(x_t - \mu)/\sigma$ converges in distribution to $z \sim N(0,1)$. Equivalently, the limiting distribution of $\sqrt{T}(\bar{x} - \mu)$ is the normal distribution $N(0, \sigma^2)$.

Proof. First we recall that the characteristic function of the standard normal variate $z \sim N(0,1)$ is $\phi(h) = \exp\{-h^2/2\}$. We must show that the characteristic function ϕ_T of z_T converges to ϕ as $T \to \infty$. Let us write $z_T = T^{-1/2}\sum z_t$ where $z_t = (x_t - \mu)/\sigma$ has $E(z_t) = 0$ and $E(z_t^2) = 1$. The characteristic fuction of z_t can now be written as

$$\begin{aligned}\phi^0(h) &= 1 + ihE(z_t) - \frac{h^2 E(z_t^2)}{2} + o(h^2) \\ &= 1 - \frac{h^2}{2} + o(h^2).\end{aligned}$$

Since $z_T = T^{-1/2}\sum z_t$ is a sum of independent and identically distributed random variables, it follows that its characterstic function can be written, in turn, as

$$\begin{aligned}\phi_T\Big(\frac{h}{\sqrt{T}}\Big) &= \Big[\phi^0\Big(\frac{h}{\sqrt{T}}\Big)\Big]^T \\ &= \Big[1 - \frac{h^2}{2T} + 0\Big(\frac{h^2}{T}\Big)\Big]^T.\end{aligned}$$

Letting $T \to \infty$, we find that $\lim \phi_T = \exp\{-h^2/2\} = \phi$, which proves the theorem.