

This handout is based on Bergantino (1998, 3), Newey (2000), Kuersteiner (2002) and a 14.385 recitation handout by Peter Hinrichs. In cases of conflict among these sources, I follow the notation of Bergantino (1998, 3) and the nomenclature of Newey (2000).

1 Notation

Random variables: x, y, z or, to denote n th in a sequence, x_n, y_n, z_n

Constant scalars: x_0, y_0, z_0

Random vectors: X, Y, Z or, to denote n th in a sequence, X_n, Y_n, Z_n

Constant vectors: X_0, Y_0, Z_0

Random matrices: A, B, C or, to denote n th in a sequence, A_n, B_n, C_n

Constant Matrices: A_0, B_0, C_0

Norm (or “length”) of a vector:

$$\|X\| = (X'X)^{1/2} = \left(\sum_{i=1}^K X_i^2 \right)^{1/2}$$

The distance between two vectors X and Y is defined as the “Euclidean norm” of their difference:

$$d(X, Y) = \|X - Y\| = ((X - Y)'(X - Y))^{1/2} = \left(\sum_{i=1}^K (X_i - Y_i)^2 \right)^{1/2}$$

2 Asymptotic (Large Sample) Theory

This section is very much oriented towards application. Anyone interested in seeing proofs or exploring the theory more deeply should consult Kuersteiner (2002).

2.1 Convergence Concepts

In the econometrics sequence, there are two main notions (or “modes”) of convergence.

Convergence in Probability: $X_n \xrightarrow{p} X \iff \lim_{n \rightarrow \infty} P(\|X_n - X\| > \varepsilon) = 0 \forall \varepsilon > 0.$

In words, this means that, for large n , the two random variables will be arbitrarily close together *in every realization*. It is often written

$$p \lim X_n = X$$

Convergence in Distribution: $X_n \xrightarrow{d} X \iff F_{X_n}(X_0) = P(X_n \leq X_0) \xrightarrow{n \rightarrow \infty} P(X \leq X_0) = F_X(X_0)$ for all X_0 where F_X is continuous.

In words, this means that, for large n , the two random variables' *distributions* will be arbitrarily close. However, the observed values of the random variables need not be close for any particular realization. It is often written

$$d \lim X_n = X$$

To see the difference between the two concepts, consider first

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

$$Y \equiv X$$

X and Y have the same distribution and are also exactly the same in each realization. Now consider

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, I_2)$$

Again, X and Y have the same distribution, but in this case need not be close to each other in any particular realization.

From the preceding discussion, the intuition behind the following theorem should be clear.

Theorem 1

$$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{p} X \not\Leftarrow X_n \xrightarrow{d} X$$

Intuitively, the converse fails because even if two random variables have very similar (or even identical) distributions, their values may not be close together for any particular realization. However, if the two random variables are arbitrarily close together in every realization, then their distributions will be very similar. However, there is one important special case in which the converse does hold: when X_n is converging in distribution to a constant.

$$X_n \xrightarrow{p} X_0 \iff X_n \xrightarrow{d} X_0$$

There are two other important modes of convergence, convergence in r th mean and convergence in convergence concepts, but these are not used much in the econometrics sequence, so they are not covered here. See Kuersteiner (2002) for definitions and discussion.

2.2 Joint and Marginal Convergence

2.2.1 Convergence in Probability

Theorem 2

$$(X_n, Y_n) \xrightarrow{p} (X, Y) \iff X_n \xrightarrow{p} X \text{ and } Y_n \xrightarrow{p} Y$$

The “only if” part of the theorem follows in one step from the definitions; the “if” part requires two steps. It may be useful to consider carefully what $(X_n, Y_n) \xrightarrow{p} (X, Y)$ actually means:

$$(X_n, Y_n) \xrightarrow{p} (X, Y) \iff \lim_{n \rightarrow \infty} P(\|(X_n, Y_n) - (X, Y)\| > \varepsilon) = 0 \forall \varepsilon > 0$$

where

$$\|(X_n, Y_n) - (X, Y)\| = \left(\|X_n - X\|^2 + \|Y_n - Y\|^2 \right)^{1/2}$$

by the definition of the Euclidean norm.

In words, the theorem means that the p -convergence of the *joint* random variable implies the p -convergence of the marginal random variables, and the converse.

Theorem 3 *Continuous Mapping Theorem for convergence in probability (pCMT):*

If $p \lim X_n = X$ and $g(\cdot)$ is continuous, then $p \lim g(X_n) = g(X)$

We can combine Theorems 2 and 3 to obtain the following corollaries:

Corollary 4 *If $p \lim x_n = x$ and $p \lim y_n = y$, then*

$$\begin{aligned}x_n + y_n &\xrightarrow{p} x + y \\x_n y_n &\xrightarrow{p} xy \\ \frac{x_n}{y_n} &\xrightarrow{p} \frac{x}{y} \text{ if } P(y = 0) = 0\end{aligned}$$

Note that this corollary invokes both Theorems 2 and 3. For example, for the first corollary, Theorem 2 ensures that $(x_n, y_n) \xrightarrow{p} (x, y)$, while Theorem 3 applies since addition is a continuous function. We can extend this corollary to vectors and matrices.

Corollary 5 *If $p \lim X_n = X$ and $p \lim Y_n = Y$, then*

$$\begin{aligned}X_n + Y_n &\xrightarrow{p} X + Y \\X_n' Y_n &\xrightarrow{p} X' Y\end{aligned}$$

provided conformability.

Corollary 6 *If $p \lim A_n = A$ and $p \lim B_n = B$, then*

$$\begin{aligned}A_n + B_n &\xrightarrow{p} A + B \\A_n B_n &\xrightarrow{p} AB \\A_n^{-1} &\xrightarrow{p} A^{-1} \text{ if } A \text{ is invertible}\end{aligned}$$

provided conformability.

2.2.2 Convergence in Distribution

Theorem 7

$$\begin{aligned}(X_n, Y_n) \xrightarrow{d} (X, Y) &\Rightarrow X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y \\(X_n, Y_n) \xrightarrow{d} (X, Y) &\not\Leftarrow X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y\end{aligned}$$

An example of why the converse fails: suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0, \Sigma), \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \rho \neq 0$$

Note the marginal random variables are

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

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

Now consider

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

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

where X_n and Y_n are independent for all n . Since $X_n \sim X$ and $Y_n \sim Y$ for all n , we trivially have

$$X_n \xrightarrow{d} X$$

$$Y_n \xrightarrow{d} Y$$

However, note that

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \sim N(0, I_2)$$

for all n , so

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$$

There are three special cases in which the converse does hold.

1. If X_n and Y_n are independent for all n and X and Y are independent, then

$$(X_n, Y_n) \xrightarrow{d} (X, Y) \iff X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{d} Y$$

Actually, it is not necessary that X_n and Y_n are independent for all n – if there exists an N such that X_n and Y_n are independent for all $N \geq n$, then the converse holds.

2. If $X_n \xrightarrow{d} X_0$ (a constant) and $Y_n \xrightarrow{d} Y$, then $(X_n, Y_n) \xrightarrow{d} (X_0, Y)$.
3. *Cramér-Wold Device*: If $c'X_n \xrightarrow{d} c'X$ for all $c \in \mathbb{R}^K$ (where K is the dimension of X_n), then $X_n \xrightarrow{d} X$.

The importance of the Cramér-Wold device is the following: recall that, for X_n to converge in distribution to X , it is not sufficient for each component of X_n (a $K \times 1$ vector) to converge in distribution to the corresponding component of X . However, the Cramér-Wold device states that if all linear combinations of the components of X_n converge in distribution to the corresponding linear combination of the components of X , then X_n converges in distribution to X .

To expand a bit, write the k th component of X_n as x_n^k and the k th component of X as x^k , i.e.

$$X_n = \begin{pmatrix} x_n^1 \\ \vdots \\ x_n^K \end{pmatrix}; X = \begin{pmatrix} x^1 \\ \vdots \\ x^K \end{pmatrix}$$

We have seen that

$$x_n^1 \xrightarrow{d} x^1, \dots, x_n^K \xrightarrow{d} x^K \not\Rightarrow X_n \xrightarrow{d} X$$

The Cramér-Wold device states that if

$$c'X_n = \sum_{k=1}^K c_k x_n^k \xrightarrow{d} \sum_{k=1}^K c_k x^k = c'X$$

for all $c \in \mathbb{R}^K$, then in fact

$$X_n \xrightarrow{d} X$$

Note that $c'X_n$ and $c'X$ are scalars.

The main usefulness of the Cramér-Wold device is in extending univariate central limit theorems to multivariate central limit theorems. See Section 4 below.

Theorem 8 *Continuous Mapping Theorem for convergence in distribution (dCMT):*

$$\text{If } d\lim X_n = X \text{ and } g(\cdot) \text{ is continuous, then } d\lim g(X_n) = g(X)$$

We can combine Theorem 8 and special case (2) from above to obtain the following corollaries:

Corollary 9 *If $d\lim x_n = x$ and $d\lim y_n = y_0$, then*

$$\begin{aligned} x_n + y_n &\xrightarrow{d} x + y_0 \\ x_n y_n &\xrightarrow{d} x y_0 \\ \frac{x_n}{y_n} &\xrightarrow{d} \frac{x}{y_0} \text{ if } y_0 \neq 0 \end{aligned}$$

Note that y_0 must be a scalar to obtain the joint convergence in distribution of (x_n, y_n) to (x, y_0) necessary to invoke Theorem 8. If the sequence y_n were converging in distribution to a random variable y rather than a scalar, we would not have the joint convergence in distribution of (x_n, y_n) to (x, y) that we need. We can extend this corollary to vectors and matrices.

Corollary 10 *If $d \lim A_n = A$ and $d \lim B_n = B_0$, then*

$$A_n + B_n \xrightarrow{d} A + B_0$$

$$A_n B_n \xrightarrow{d} A B_0$$

$$A_n^{-1} \xrightarrow{d} A^{-1} \text{ if } A \text{ is invertible}$$

provided conformability.

Corollary 11 *If $d \lim X_n = X$ and $d \lim A_n = A_0$, then*

$$A_n X_n \xrightarrow{d} A_0 X$$

The most useful results from the corollaries above are summarized in a famous result known as *Slutzky's Theorem*.¹

Theorem 12 *Slutzky's Theorem: If $d \lim X_n = X$, $d \lim Y_n = Y_0$, and $d \lim A_n = A_0$ then*

$$X_n + Y_n \xrightarrow{d} X + Y_0$$

$$A_n X_n \xrightarrow{d} A_0 X$$

provided conformability.

One final useful theorem:

Theorem 13 *If $X_n \xrightarrow{d} X$ and $X_n - Y_n \xrightarrow{p} 0$, then*

$$Y_n \xrightarrow{d} X$$

¹There is some disagreement in the various notes and handouts about which of the theorems and corollaries is called the "Slutzky Theorem." These notes will use W. Newey's nomenclature. Also, this Slutzky is the same Slutzky from consumer theory / 14.121 and Frank Fisher's joke book.

Corollary 14 *If $X_n \xrightarrow{d} X_0$ (a constant) and $X_n - Y_n \xrightarrow{p} 0$, then*

$$Y_n \xrightarrow{p} X_0$$

2.3 Applications / Examples

Two important applications of the preceding results follow. As an example of an application of these results, suppose we have an $N \times K$ data matrix X_N of N observations of K variables and an $N \times 1$ disturbance vector ε_N . Suppose further that as N increases,

$$\begin{aligned} \frac{1}{N} X_N' X_N &\xrightarrow{p} Q_{XX} > 0 \\ \frac{1}{N} X_N' \varepsilon_N &\xrightarrow{p} 0 \\ \frac{1}{\sqrt{N}} X_N' \varepsilon_N &\xrightarrow{d} L \sim N(0, V) \end{aligned}$$

where Q_{XX} is a constant, PD $K \times K$ matrix and 0 is $K \times 1$. Then, by the Continuous Mapping Theorem,

$$\left(\frac{1}{N} X_N' X_N \right)^{-1} \xrightarrow{p} Q_{XX}^{-1}$$

since matrix inversion is a continuous function and Q_{XX} PD assures us that the inverse exists. Now we can invoke the CMT:

$$\begin{aligned} \left(\frac{1}{N} X_N' X_N \right)^{-1} &\xrightarrow{p} Q_{XX}^{-1}; \quad \frac{1}{N} X_N' \varepsilon_N \xrightarrow{p} 0 \\ &\Rightarrow \\ \left(\frac{1}{N} X_N' X_N \right)^{-1} \frac{1}{N} X_N' \varepsilon_N &\xrightarrow{p} Q_{XX}^{-1} 0 = 0 \end{aligned}$$

The finiteness of Q_{XX}^{-1} is guaranteed because Q_{XX} is PD. This result will be used to show the *consistency* of the OLS estimator:

$$\hat{\beta} = (X'X)^{-1} X'y \xrightarrow{p} \beta$$

Furthermore, we can use Slutsky's Theorem to prove the following:

$$\begin{aligned} \left(\frac{1}{N} X'_N X_N \right)^{-1} &\xrightarrow{p} Q_{XX}^{-1}; \quad \frac{1}{\sqrt{N}} X'_N \varepsilon_N \xrightarrow{d} L \sim N(0, V) \\ &\Rightarrow \\ \left(\frac{1}{N} X'_N X_N \right)^{-1} \frac{1}{\sqrt{N}} X'_N \varepsilon_N &\xrightarrow{d} Q_{XX}^{-1} L \sim N(0, Q_{XX}^{-1} V Q_{XX}^{-1}) \end{aligned}$$

This result will be used to show the *asymptotic normality* of the OLS estimator:

$$\sqrt{N} (\hat{\beta} - \beta) \xrightarrow{d} N(0, Q_{XX}^{-1} V Q_{XX}^{-1})$$

3 Laws of Large Numbers

Laws of Large Numbers (LLNs) give conditions under which sample averages will converge to population means. For example, suppose

$$\mathbf{X} = (X_1, \dots, X_N)$$

is a random sample of N from a random variable $X \sim? (\mu_X, \sigma_X^2)$ with unknown distribution but finite mean μ_X and variance σ_X^2 . Defining the sample average as usual as

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$$

it is easy to show that

$$E[\bar{X}_N] = \mu_X \text{ and } V[\bar{X}_N] = \frac{\sigma_X^2}{N}$$

This establishes that

$$\bar{X}_N \sim? \left(\mu_X, \frac{\sigma_X^2}{N} \right)$$

Since the variance goes to zero as N increases, we sense that \bar{X}_N is “collapsing” to a constant, μ_X . LLNs formalize this sense and extend to more general situations (for example when the data are not i.i.d.).

Formally, for a sequence of random variables

$$X_1, X_2, \dots$$

a LLN specifies conditions under which the sample mean, defined as above, converges in probability to the population mean, i.e.

$$\bar{X}_N - \mu_X \xrightarrow{p} 0$$

or, more intuitively (and equivalently, by Corollary 14),

$$\bar{X}_N \xrightarrow{p} \mu_X$$

Note that since we are discussing convergence in probability, we do not need to distinguish between random variables and random vectors. We *will* need make this distinction when discussing central limit theorems in the next section, since this concept relates to convergence in distribution.

Some frequently used LLNs follow. We will start with the simplest case, *i.i.d.* (*independent and identically distributed*) data.

Theorem 15 *Weak (Khinchine) Law of Large Numbers (WLLN): If*

$$\mathbf{X} = (X_1, \dots, X_N)$$

is a random sample of N from a random vector X with $E[\|X\|] < \infty$, then

$$\bar{X}_N \xrightarrow{p} E[X]$$

Note that the condition $E[\|X\|] < \infty$ is just $E[X^2] < \infty$ in the univariate case, which is equivalent to $\sigma_X^2 < \infty$. The law is called “weak” because the convergence is in probability, a “strong” law would give almost-sure convergence. A strong LLN can be proved for this case, but we will not pursue it here.

Next, we generalize to *i.n.i.d.* (*independent, not identically distributed*) data.

Theorem 16 *Markov’s LLN: Suppose*

$$\mathbf{X} = (X_1, \dots, X_N)$$

are independent random vectors. Let

$$\bar{X}_N = \sum_{i=1}^N X_i$$

Note that

$$E[\bar{X}_N] = \frac{1}{N} \sum_{i=1}^N \mu_{X_i}$$

If there exists a $\delta > 0$ such that

$$\frac{1}{N} \sum_{i=1}^N E[\|X_i\|^{1+\delta}] < \infty$$

is bounded for all N , then

$$\bar{X}_N \xrightarrow{p} E[\bar{X}_N]$$

To understand the “if there exists a $\delta > 0$ such that...” condition, note that, for the condition to be satisfied, it is necessary (but not sufficient) that

$$\frac{1}{N} \sum_{i=1}^N E[\|X_i\|] < \infty$$

(this is just substituting $\delta = 0$). For the univariate case, writing $V[X_i]$ as σ_i^2 , this means that

$$\frac{1}{N} \sum_{i=1}^N \sigma_i^2 < \infty$$

That is, the variances cannot “run away” to infinity as the sample size increases. Another way to think about this is that the distributions of the random variables cannot get ever more spread out. For technical reasons, it is not sufficient to confirm this for the standard norm ($\delta = 0$), but must find that it holds for some $\delta > 0$.

Next, we allow for some *dependence* among observations. First, we need two definitions.

Definition 17 A random process X_1, X_2, \dots is strictly stationary if for all i and for all j_1, \dots, j_m , the joint distribution of

$$(X_i, X_{i+j_1}, X_{i+j_m})$$

depends only on j_1, \dots, j_m and not i .

In words, this means that any dependence between observations depends only on the distance between them in the sequence, not on their particular place in the sequence. So, for example, if a process is strictly stationary, then

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_3) = \dots$$

Definition 18 A random process is weakly stationary or covariance stationary if $E[X_i] = \mu$ for all i and $\text{Cov}(X_i, X_{i+j}) = \Sigma_j$ for all j .

This is weaker than the first definition because it only restricts covariance, but not other forms of dependence.

Theorem 19 Chebyshev’s LLN: If $\{X_i\}_i^N$ are stationary random vectors with $E[\|X\|^2] < \infty$ and

$$\sum_{j=1}^{\infty} \|\text{Cov}(X_i, X_j)\| < \infty$$

then

$$\bar{X}_N \xrightarrow{p} E[X]$$

Intuitively, so long as there is not “too much” dependence among observations, the LLN will hold.

4 Central Limit Theorems

Central limit theorems (CLTs) give conditions under which sample averages will be approximately normally distributed as the sample size grows. Recall the simplest example of a CLT from undergraduate probability: suppose

$$\mathbf{x} = (x_1, \dots, x_N)$$

is a random sample of N from a random variable $x \sim? (\mu_X, \sigma_X^2)$ with unknown distribution but finite mean μ_X and variance σ_X^2 . Defining the sample average as usual as

$$\bar{x}_N = \frac{1}{N} \sum_{i=1}^N x_i$$

We have already noted that

$$E[\bar{x}_N] = \mu_x \text{ and } V[\bar{x}_N] = \frac{\sigma_x^2}{N}$$

and so

$$\bar{x}_N \sim? \left(\mu_x, \frac{\sigma_x^2}{N} \right)$$

The *central limit theorem* states that, for large N , we have not just the first two moments of \bar{x}_N but also its approximate distribution, that is,

$$\bar{x}_N \sim_A N \left(\mu_x, \frac{\sigma_x^2}{N} \right)$$

which we can re-write as

$$\frac{\bar{x}_N - \mu_x}{\sigma_x / \sqrt{N}} \sim_A N(0, 1)$$

or

$$\sqrt{N}(\bar{x}_N - \mu_x) \sim_A N(0, \sigma_x^2)$$

and, recalling that $V[\bar{x}_N] = \sigma_x^2/N$,

$$\sqrt{N}(\bar{x}_N - \mu_x) \sim_A N(0, NV[\bar{x}_N])$$

To write this in terms of convergence in distribution, we have that, as N increases,

$$\frac{\bar{x}_N - \mu_x}{\sigma_x/\sqrt{N}} \xrightarrow{d} N(0, 1)$$

equivalently

$$\sqrt{N}(\bar{x}_N - \mu_x) \xrightarrow{d} N(0, \sigma_x^2)$$

equivalently

$$\sqrt{N}(\bar{x}_N - \mu_x) \xrightarrow{d} N(0, NV[\bar{x}_N])$$

Note that we “re-scale” this by multiplying by the square root of the sample size, since otherwise the variance would just collapse to zero.

To formalize, a *central limit theorem* specifies conditions under which, for a sequence of random vectors X_1, X_2, \dots

$$\sqrt{N}(\bar{X}_N - E[\bar{X}]) \xrightarrow{d} N\left(0, \lim_{N \rightarrow \infty} NV[\bar{X}]\right)$$

where we assume that $\lim_{N \rightarrow \infty} NV[\bar{X}]$ exists and is finite. We call $\lim_{N \rightarrow \infty} NV[\bar{X}]$ the “asymptotic variance of \bar{X}_N ”.² A few examples of commonly-used CLTs follow.

Theorem 20 *Lindberg-Lévy CLT: if x_i are i.i.d. with $E[x_i] = \mu$ and $V[x_i] = \sigma^2 < \infty$, then*

$$\sqrt{N}(\bar{x}_N - \mu) \xrightarrow{d} N(0, \sigma^2)$$

We would like to extend this to random *vectors*. Imagine that we have an i.i.d. sequence of K -dimensional random vectors X_i such that the conditions of the LLCLT apply to each *component*,

²This has always seemed to be a misnomer to me, since it is actually the asymptotic variance of \bar{X}_N rescaled by the \sqrt{N} factor. The variance of \bar{X}_N goes to zero, so in a strict sense the asymptotic variance of \bar{X}_N is zero. But this is the standard terminology.

i.e. $E[X_i^k] = \mu_k$, $V[X_i^k] = \sigma_k^2 < \infty$ and the data are i.i.d. Can we come up with a central limit theorem result right away? We *cannot* (at least not right away), because we don't know anything about the covariance among the components – as we have seen, convergence in distribution of the marginal components does not imply joint convergence. Fortunately, we *can* obtain the multivariate CLT we want by appealing to the *Cramér-Wold Device* (page 6). For any $c \in \mathbb{R}^K$, note that

$$c'X_i = c_1X_i^1 + \cdots + c_KX_i^K$$

is a random *variable* with mean

$$E[c'X_i] = \sum_{k=1}^K c_k\mu_k$$

and variance

$$V[c'X_i] = \sum_{k=1}^K c_k^2\sigma_k^2 + \sum_{k=1}^K \sum_{j \neq k}^K c_k c_j \sigma_{kj}$$

Since the data $\{X_i\}$ are i.i.d., the sequence $\{c'X_i\}$ will be i.i.d. as well, and by examining the expression for $V[c'X_i]$ we see that this is finite. So the sequence $\{c'X_i\}$ satisfies the conditions of the LLCLT. Therefore,

$$\sqrt{N}(\overline{c'X_i} - E[c'X_i]) \xrightarrow{d} N(0, V[c'X_i])$$

Equivalently,

$$c'\sqrt{N}(\overline{X_i} - E[X_i]) \xrightarrow{d} N(0, c'V[X_i]c)$$

or

$$c'\sqrt{N}(\overline{X_i} - E[X_i]) \xrightarrow{d} c'A$$

where

$$A \sim N(0, V[X_i])$$

By the Cramér-Wold device, then, we have proved that

$$\sqrt{N}(\overline{X_i} - E[X_i]) \xrightarrow{d} N(0, V[X_i])$$

(note that, here, 0 is a $K \times 1$ vector and $V[X_i]$ is a $K \times K$ matrix). We will restate the result more concisely:

Theorem 21 *Lindberg-Lèvy Central Limit Theorem for Random Vectors: If X_i are i.i.d. random vectors with $E[X_i] = \mu$ and $V[X_i] = \Sigma$, then*

$$\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} N(0, \Sigma)$$

We can write this equivalently as

$$\Sigma^{-1/2}\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} N(0, I)$$

We can extend these central limit theorems to *i.n.i.d.* data.

Theorem 22 *Lindberg-Feller Multivariate CLT for i.n.i.d. data: Suppose X_i are independent (but not necessarily identically distributed) random vectors such that*

- (1) $E[X_i] = \mu_i$ is finite for all i
- (2) $V[X_i] = \Sigma_i$ is finite for all i

Define

$$\bar{\mu}_N \equiv \sum_{i=1}^N \mu_i; \bar{\Sigma}_N \equiv \sum_{i=1}^N \Sigma_i$$

Further suppose that

- (3) All mixed third moments of X_i are finite for all i
- (4) $\bar{\Sigma}_N \xrightarrow{p} \Sigma$, a finite and PD matrix
- (5) $(N\bar{\Sigma}_N)^{-1} \Sigma_i \xrightarrow{p} 0$ for all i

Then:

$$\sqrt{N}(\bar{X}_N - \bar{\mu}_N) \xrightarrow{d} N(0, \Sigma)$$

There also exist CLTs for dependent data. Basically, there cannot be “too much” dependence between observations. Saying what is “too much” precisely is fairly difficult. See White (1984) for details.

4.1 Delta Method

Theorem 23 Suppose X_N is a sequence of $K \times 1$ vectors such that

$$\sqrt{N}(Z_N - Z_0) \xrightarrow{d} N(0, \Sigma)$$

for some $K \times 1$ constant vector Z_0 . If the function

$$g : \mathbb{R}^K \rightarrow \mathbb{R}^J$$

is continuously differentiable at Z_0 , then

$$\sqrt{N}(g(Z_N) - g(Z_0)) \xrightarrow{d} N(0, G\Sigma G')$$

where

$$G \equiv \partial g(Z_0) / \partial Z'$$

is the $J \times K$ matrix of partial derivatives of g at Z_0 .

This is typically used when you have an asymptotically normal estimator of some parameter θ_0 but are interested in estimating some function of that parameter. That is, you already have an estimator $\hat{\theta}_N$ such that

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V)$$

for some finite matrix V , but you are really interested in estimating $g(\theta_0)$. You know that

$$\hat{\theta}_N \xrightarrow{p} \theta_0$$

so if $g(\cdot)$ is continuous you can use the continuous mapping theorem and conclude that

$$g(\hat{\theta}_N) \xrightarrow{p} g(\theta_0)$$

but you would also like to know the asymptotic distribution of $g(\hat{\theta}_N)$ so that you can create confidence intervals and so on. The delta-method lets you obtain the asymptotic distribution of $g(\hat{\theta})$, provided $g(\cdot)$ is not just continuous but also continuously differentiable at θ_0 :

$$\sqrt{N} \left(g(\hat{\theta}_N) - g(\theta_0) \right) \xrightarrow{d} N(0, GVG')$$

We will cover the proof for the univariate case. The logic of the multivariate proof is similar, see Kuersteiner (2002) for details.

Proof: By the Taylor approximation theorem, since $g(\cdot)$ is continuous, we have

$$g(\hat{\theta}_N) = g(\theta_0) + g'(\theta_0) (\hat{\theta}_N - \theta_0) + R$$

where R is a remainder term we can disregard for $\hat{\theta}_N \approx \theta_0$ (and we know that as N grows, $\hat{\theta}_N \approx \theta_0$, since $\hat{\theta}_N \xrightarrow{p} \theta_0$). Rearranging, we have

$$\begin{aligned} g(\hat{\theta}_N) - g(\theta_0) &\approx g'(\theta_0) (\hat{\theta}_N - \theta_0) \\ \sqrt{N} \left(g(\hat{\theta}_N) - g(\theta_0) \right) &\approx g'(\theta_0) \sqrt{N} (\hat{\theta}_N - \theta_0) \end{aligned}$$

Since we have assumed that

$$\sqrt{N} (\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, V)$$

we have

$$\begin{aligned} \sqrt{N} \left(g(\hat{\theta}_N) - g(\theta_0) \right) &\xrightarrow{d} g'(\theta_0) N(0, V) \\ \sqrt{N} \left(g(\hat{\theta}_N) - g(\theta_0) \right) &\xrightarrow{d} g'(\theta_0) N \left(0, (g'(\theta_0))^2 V \right) \end{aligned}$$

5 Large Sample Estimation Criteria

Suppose we have an estimator $\hat{\theta}_N$ of a parameter θ_0 (the N subscript reminds us that the estimator is a function of the sample, i.e. $\hat{\theta}_N = \hat{\theta}_N(X_1, \dots, X_N)$). In finite samples, we usually judge the estimator based on *unbiasedness*, *mean squared error* and *efficiency*. Recall that an estimator is unbiased if

$$E_{\theta_0} [\hat{\theta}_N] = \theta_0$$

for all true values θ_0 of the parameter. We also consider the mean squared error

$$\begin{aligned} \text{MSE}(\hat{\theta}_N) &= E \left[(\hat{\theta}_N - \theta_0)^2 \right] \\ &= (\text{bias}(\hat{\theta}))^2 + V[\hat{\theta}_N] \end{aligned}$$

If $\hat{\theta}_N$ is unbiased, then the MSE reduces to

$$\text{MSE}(\hat{\theta}_N) = V[\hat{\theta}_N]$$

Usually, we prefer unbiased estimators with low variance, although we might allow a bit of bias if it were to reduce the MSE greatly.

Finally, if two estimators $\hat{\theta}_N$ and $\tilde{\theta}_N$ are both unbiased, we say $\hat{\theta}_N$ is *more efficient than* $\tilde{\theta}_N$ if

$$V[\hat{\theta}_N] \leq V[\tilde{\theta}_N]$$

If

$$V[\hat{\theta}_N] \leq V[\tilde{\theta}_N]$$

for *all* unbiased estimators $\tilde{\theta}_N$, we say $\hat{\theta}_N$ is *efficient* (among unbiased estimators).

We would like to establish similar criteria for judging the *asymptotic* behavior of an estimator. We will have three: consistency, asymptotic normality and asymptotic efficiency.

5.1 Consistency

Definition 24 $\hat{\theta}_N$ is consistent if and only if

$$\hat{\theta}_N \xrightarrow{p} \theta_0$$

for all possible values of the true parameter, i.e. for all $\theta_0 \in \Theta$.

This is a very weak requirement – all it demands is that if we give the estimator an infinite amount of data, the estimator will give us the right answer. In some sense this is the least we can ask of our estimator – I think Dan McFadden had a quote when he taught at M.I.T. many years ago that (roughly), “If you can’t get it right with an infinite amount of data, you’re in the wrong business.”

5.2 Asymptotic Normality

Definition 25 $\hat{\theta}_N$ is asymptotically normal if and only if

$$\sqrt{N} \left(\hat{\theta}_N - \theta_0 \right) \xrightarrow{d} N(0, V)$$

where the matrix V is called the asymptotic variance of $\hat{\theta}_N$.

This is a useful property because it allows us to do approximate inference based on the normal distribution. However, we also need a consistent estimator of the *asymptotic variance* V , that is

$$\hat{V} \xrightarrow{p} V$$

We can use $\hat{\theta}_N$ and \hat{V} to construct an asymptotic $1 - \alpha$ confidence interval for θ_0 as follows:

$$\mathcal{I}_\alpha = \left[\hat{\theta}_N - \sqrt{N} q_{\alpha/2} \hat{V}^{1/2}, \hat{\theta}_N + \sqrt{N} q_{\alpha/2} \hat{V}^{1/2} \right]$$

where

$$q_{\alpha/2} : P(Z \leq q_{\alpha/2}) = \frac{\alpha}{2}$$

Then

$$P(\theta_0 \in \mathcal{I}_\alpha) \xrightarrow{n \rightarrow \infty} 1 - \alpha$$

Proof: see Newey (2000).

5.3 Asymptotic Efficiency

Definition 26 $\hat{\theta}_N$ is asymptotically efficient in a class of asymptotically normal estimators if and only if every other estimator $\tilde{\theta}_N$ in the class has an asymptotic variance \tilde{V} such that

$$\tilde{V} - \hat{V} \geq 0 \text{ (is PSD)}$$

That is, \hat{V} is “smaller” (in the matrix sense) than any other \tilde{V} . Note that we only compare estimators within a class; for example, linear unbiased estimators.

An asymptotically efficient estimator is preferred because an asymptotic confidence interval, constructed as above, will be shorter.

6 Maximum Likelihood Estimation

Let X be a random k -vector with a parametric density $f_X(x|\theta_0)$, where θ_0 is the unknown parameter (possibly multi-dimensional) of interest. We will have N observations of X from which to estimate θ_0 . Ex-ante, we write the sample as a random vector

$$\mathbf{X} = (X_1, \dots, X_N)$$

Ex-post, we will write the observed value of \mathbf{X} as

$$\mathbf{x} = (x_1, \dots, x_N)$$

In general, we cannot obtain the joint density of the sample

$$f_{\mathbf{X}}(\mathbf{x}|\theta_0)$$

but if the data are i.i.d. (a random sample), we can write the joint density as the product of the marginals:

$$f_{\mathbf{X}}(\mathbf{x}|\theta_0) = \prod_{i=1}^N f_X(x_i|\theta_0)$$

Thinking of this as a function of the parameter given the data rather than the data given the parameter, we have the *likelihood function*

$$L(\theta) = L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^N f_X(x_i|\theta)$$

It is usually convenient to take the log and obtain the *log-likelihood function*

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x}) = \sum_{i=1}^N \log f_X(x_i|\theta)$$

The *maximum likelihood estimator (MLE)* of θ_0 is

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta)$$

Sometimes this can be found analytically; more frequently we will have to resort to numerical methods. We also define the *information matrix* of a sample of size N as

$$\mathcal{I}_N(\theta_0) = -E \left[\frac{\partial^2 \mathcal{L}(\theta_0)}{\partial \theta' \partial \theta} \right] = E \left[\left(\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \right) \left(\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \right)' \right] = V \left[\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \right]$$

The second and third equalities were proven in 14.381 and a brief review is provided below. The third equality

$$-E \left[\frac{\partial^2 \mathcal{L}(\theta_0)}{\partial \theta' \partial \theta} \right] = E \left[\left(\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \right) \left(\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \right)' \right]$$

is called the *Information Equality*. Since the data are i.i.d., it follows that the Information Matrix of a single observation, written $\mathcal{I}_1(\theta_0)$ satisfies the relation

$$\mathcal{I}_1(\theta_0) = \frac{1}{N} \mathcal{I}_N(\theta_0)$$

Theorem 27 *Cramér-Rao Lower Bound: Under commonly satisfied regularity conditions, if $\tilde{\theta}$ is any unbiased estimator of θ_0 , then*

$$V[\tilde{\theta}] - (\mathcal{I}_N(\theta_0))^{-1} \text{ is PSD}$$

$\mathcal{I}_N(\theta_0)$ is called the Cramér-Rao Lower Bound for the unbiased estimation of θ_0 , and we write

$$CRLB = (\mathcal{I}_N(\theta_0))^{-1}$$

If the data are i.i.d., we also have

$$CRLB = (N\mathcal{I}_1(\theta_0))^{-1} = \frac{1}{N} (\mathcal{I}_1(\theta_0))^{-1}$$

Theorem 28 *Under commonly satisfied regularity conditions, we have*

- (a) $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$ (consistency of MLE)
- (b) $\sqrt{N}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, (\mathcal{I}_1(\theta_0))^{-1})$ (asymptotic normality of MLE)
- (c) $\hat{\theta}_{ML}$ is asymptotically efficient
- (d) The MLE of $\gamma = g(\theta_0)$ is $\hat{\gamma}_{ML} = g(\hat{\theta}_{ML})$ (invariance property of MLE)
- (e) If $\hat{\theta}_{ML}$ is unbiased, it achieves the CRLB and so is BUE

Furthermore,

- (f)

$$\frac{1}{\sqrt{N}} \frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, (\mathcal{I}_1(\theta_0))^{-1})$$

(g) for all $\tilde{\theta}$ such that $\tilde{\theta} \xrightarrow[p]{p} \theta_0$,

$$-\frac{1}{N} \frac{\partial^2 \mathcal{L}(\tilde{\theta})}{\partial \theta' \partial \theta} = -\frac{1}{N} \sum_{i=1}^N \left[\frac{\partial^2 \log f(x_i | \tilde{\theta})}{\partial \theta' \partial \theta} \right] \xrightarrow[p]{p} (\mathcal{I}_1(\theta_0))^{-1}$$

(h) for all $\tilde{\theta}$ such that $\tilde{\theta} \xrightarrow[p]{p} \theta_0$,

$$\frac{1}{N} \sum_{i=1}^N \left[\frac{\partial \log f(x_i | \tilde{\theta})}{\partial \theta} \frac{\partial \log f(x_i | \tilde{\theta})}{\partial \theta'} \right] \xrightarrow[p]{p} (\mathcal{I}_1(\theta_0))^{-1}$$

For a proof and discussion of regularity conditions, see Newey and McFadden (1994). Two important regularity conditions that can be informally described are that (1) the true value of the parameter lie in the interior of the parameter space ($\theta_0 \in \text{int}(\Theta)$) and (2) that the interchange of differentiation and integration be permitted. The second condition usually requires that the support of the data not depend on the value of θ_0 . The estimation of a $U[0, \theta]$ random variable, for example, would not satisfy this condition.

It is especially important to note that these are all *asymptotic* results: the *finite sample* properties of $\hat{\theta}_{ML}$ are generally unknown. In particular, Theorem (28) does not imply that $\hat{\theta}_{ML}$ is unbiased or normally distributed in any finite sample in general, although it may be in some cases.

6.1 Review of Derivation of CRLB

This section derives the CRLB in more detail for the simplest case of a scalar θ . Those already comfortable with this result should skim or skip. The logic of the derivation for a vector θ is similar.

Suppose

$$X \sim f_X(x|\theta)$$

and $\hat{\theta}$ is an *unbiased* estimator of θ . Define the *score function*

$$S(x, \theta) = \frac{\partial}{\partial \theta} \log f_X(x|\theta)$$

as the vector of derivatives of the log-density function and the *Hessian*

$$H(x, \theta) = \frac{\partial}{\partial \theta \partial \theta'} \log f_X(x|\theta)$$

as the matrix of second derivatives. Since we are only considering scalar θ , $S(x, \theta)$ and $H(x, \theta)$ are both scalars and

$$H(x, \theta) = \frac{\partial^2}{\partial \theta^2} \log f_X(x|\theta)$$

Claim 29 *If the technical condition that*

$$\frac{\partial}{\partial \theta} \int \hat{\theta}(x) f(x|\theta) dx = \int \frac{\partial}{\partial \theta} \hat{\theta}(x) f(x|\theta) dx$$

is satisfied, then

$$V[\hat{\theta}] \geq \left(E[(S(x, \theta))^2] \right)^{-1}$$

Proof:

For brevity, we will write $S = S(x, \theta)$

Useful fact #1: for any random variables Y and Z

$$(Cov(Y, Z))^2 \leq V[Y]V[Z]$$

Useful fact #2: given our technical condition, $E[S] = 0$

$$\begin{aligned}
 S &= \frac{\partial}{\partial \theta} \log f_X(x|\theta) \\
 E[S] &= E \left[\frac{\partial}{\partial \theta} \log f_X(x|\theta) \right] \\
 &= \int \left(\frac{\partial}{\partial \theta} \log f_X(x|\theta) \right) f_X(x|\theta) dx \\
 &= \int \frac{\partial}{\partial \theta} f_X(x|\theta) dx && \frac{\partial}{\partial \theta} \log f_X(x|\theta) = \frac{\frac{\partial}{\partial \theta} f_X(x|\theta)}{f_X(x|\theta)} \\
 &= \frac{\partial}{\partial \theta} \int f_X(x|\theta) dx && \text{by our technical condition} \\
 &= \frac{\partial}{\partial \theta} 1 \\
 &= 0
 \end{aligned}$$

With these two facts, we can proceed with the proof. To prove:

$$V[\hat{\theta}] \geq 1/E[S^2]$$

This is true if and only if

$$V[\hat{\theta}] E[S^2] \geq 1$$

Since $E[S] = 0$, $V[S] = E[S^2] - (E[S])^2 = E[S^2]$, so it is sufficient to show that

$$V[\hat{\theta}] V[S] \geq 1$$

By our first useful fact, we know that

$$\left(\text{Cov}(\hat{\theta}, S) \right)^2 \leq V[\hat{\theta}] V[S]$$

so if we can prove that

$$\text{Cov}(\hat{\theta}, S) \geq 1$$

we will have

$$V[\hat{\theta}] V[S] \geq \left(\text{Cov}(\hat{\theta}, S) \right)^2 \geq 1$$

the result we want. In fact, we will prove that

$$\text{Cov}(\hat{\theta}, S) = 1$$

as follows:

$$\begin{aligned}
 \text{Cov}(\hat{\theta}, S) &= E[\hat{\theta}S] - E[\hat{\theta}]E[S] \\
 &= E[\hat{\theta}S] && E[S] = 0 \\
 &= \int \hat{\theta}(x) \frac{\partial}{\partial \theta} \log f_X(x|\theta) f_X(x|\theta) dx \\
 &= \int \hat{\theta}(x) \frac{\partial}{\partial \theta} f_X(x|\theta) dx \\
 &= \frac{\partial}{\partial \theta} \int \hat{\theta}(x) f_X(x|\theta) dx \\
 &= \frac{\partial}{\partial \theta} E[\hat{\theta}] \\
 &= \frac{\partial}{\partial \theta} \theta && \hat{\theta} \text{ is unbiased} \\
 &= 1
 \end{aligned}$$

This concludes the proof.

Note that we did not specify whether X was a random variable or random vector.

In the case of i.i.d. data, we can extend the result:

Corollary 30 *Suppose*

$$\mathbf{X} = (X_1, \dots, X_N)$$

is a random sample of N from

$$X \sim f_X(x|\theta)$$

and $\hat{\theta}$ is an unbiased estimator of θ . If our technical condition is satisfied, then

$$V[\hat{\theta}] \geq \frac{1}{NE[S^2]}$$

where

$$S = S(\mathbf{X}, \theta)$$

is the score function for a single observation.

Proof: The previous theorem established that

$$V[\hat{\theta}] \geq \frac{1}{E[(S(\mathbf{X}, \theta))^2]}$$

where $S(\mathbf{X}, \theta)$ is the score function for the full random *vector*. All we need to show is that

$$E[(S(\mathbf{X}, \theta))^2] = NE[(S(X, \theta))^2]$$

Consider the definition

$$\begin{aligned} (S(\mathbf{X}, \theta))^2 &= \left(\frac{\partial}{\partial \theta} \log f_{\mathbf{X}}(\mathbf{x}|\theta) \right)^2 \\ &= \left(\frac{\partial}{\partial \theta} \sum_{i=1}^N \log f_X(x_i|\theta) \right)^2 \\ &= \left(\sum_{i=1}^N \frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right)^2 \end{aligned}$$

So we have

$$\begin{aligned} E[(S(\mathbf{X}, \theta))^2] &= E \left[\left(\sum_{i=1}^N \frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right)^2 \right] \\ &= E \left[\sum_{i=1}^N \left(\frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right)^2 + \sum_{i=1}^N \sum_{j \neq i} \left(\frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right) \left(\frac{\partial}{\partial \theta} \log f_X(x_j|\theta) \right) \right] \\ &= \sum_{i=1}^N E \left[\left(\frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right)^2 \right] + \sum_{i=1}^N \sum_{j \neq i} E \left[\left(\frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right) \left(\frac{\partial}{\partial \theta} \log f_X(x_j|\theta) \right) \right] \\ &= NE[S(X, \theta)] + \sum_{i=1}^N \sum_{j \neq i} E \left[\left(\frac{\partial}{\partial \theta} \log f_X(x_i|\theta) \right) \left(\frac{\partial}{\partial \theta} \log f_X(x_j|\theta) \right) \right] \end{aligned}$$

So we need only prove that, for all i and j ,

$$E \left[\left(\frac{\partial}{\partial \theta} \log f_X (x_i | \theta) \right) \left(\frac{\partial}{\partial \theta} \log f_X (x_j | \theta) \right) \right] = 0$$

By the independence of the data,

$$\begin{aligned} E \left[\left(\frac{\partial}{\partial \theta} \log f_X (x_i | \theta) \right) \left(\frac{\partial}{\partial \theta} \log f_X (x_j | \theta) \right) \right] &= E \left[\frac{\partial}{\partial \theta} \log f_X (x_i | \theta) \right] E \left[\frac{\partial}{\partial \theta} \log f_X (x_j | \theta) \right] \\ &= E [S (X, \theta)] E [S (X, \theta)] \end{aligned}$$

And we know that

$$E [S (X, \theta)] = 0$$

So we have proved that

$$E \left[(S (\mathbf{X}, \theta))^2 \right] = N E [S (X, \theta)]$$

Finally, we prove the *Information Equality*, which is useful in computations. We want to show that

$$-E \left[\frac{\partial^2 \mathcal{L} (\theta_0)}{\partial \theta' \partial \theta} \right] = E \left[\left(\frac{\partial \mathcal{L} (\theta_0)}{\partial \theta} \right) \left(\frac{\partial \mathcal{L} (\theta_0)}{\partial \theta} \right)' \right]$$

or, equivalently

$$-E [H (x, \theta)] = E [S (X, \theta) S (X, \theta)']$$

Since we have a scalar θ , this is equivalent to

$$-E \left[\frac{\partial^2}{\partial \theta^2} \log f (x | \theta) \right] = E \left[(S (X, \theta))^2 \right]$$

or

$$-E \left[\frac{\partial^2}{\partial \theta^2} \log f (x | \theta) \right] = E \left[\left(\frac{\partial}{\partial \theta} \log f (x | \theta) \right)^2 \right]$$

We have already established that

$$\begin{aligned} 0 &= E \left[\frac{\partial}{\partial \theta} \log f(x|\theta) \right] \\ &= \int \frac{\partial}{\partial \theta} \log f(x|\theta) f(x|\theta) dx \end{aligned}$$

We differentiate again to obtain

$$\begin{aligned} \frac{\partial}{\partial \theta} 0 &= \frac{\partial}{\partial \theta} \int \frac{\partial}{\partial \theta} \log f(x|\theta) f(x|\theta) dx \\ 0 &= \int \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} \log f(x|\theta) f(x|\theta) \right) dx \\ 0 &= \int \left(\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right) f(x|\theta) dx + \int \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) \left(\frac{\partial}{\partial \theta} f(x|\theta) \right) dx \\ 0 &= E \left[\left(\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right) \right] + \int \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) \left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) f(x|\theta) dx \\ 0 &= E \left[\left(\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right) \right] + E \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right] \end{aligned}$$

Subtracting, we obtain

$$-E \left[\left(\frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right) \right] = E \left[\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right)^2 \right]$$

To get from the third line to the fourth, recall that

$$\left(\frac{\partial}{\partial \theta} \log f(x|\theta) \right) f(x|\theta) = \frac{\partial}{\partial \theta} f(x|\theta)$$

because

$$\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)}$$

7 Hypothesis Testing

To test a *null hypothesis* H_0 against an *alternative hypothesis* H_A , you must use a *test statistic* based on the data. The test procedure consists of deciding whether or not to reject H_0 by comparing the

realized value of the test statistic against a number, determined in advance of seeing the data, called the *critical value* of the test statistic. More generally, you examine whether or not the realization of the test statistic falls inside some *critical region*. If so, you reject H_0 in favor of H_A . Usually, H_0 is *nested* inside H_A in the sense that it represents some restrictions placed on H_A . The test, therefore, is testing the restrictions imposed by H_0 . The general logic of a test is to see if the data are reasonable when the hypothesis is true. If the observed data are very unlikely under the null hypothesis, we reject the null.

The two errors a testing procedure can make are:

Type I Error : Reject H_0 | H_0 is true

Type II Error : Accept H_0 | H_0 is false

Definitions:

- $\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$
- $\beta = P(\text{type II error}) = P(\text{accept } H_0 \mid H_0 \text{ is false})$
- $1 - \beta = P(\text{reject } H_0 \mid H_0 \text{ is false}) = \text{“power” of the test}$

Usually, the hypothesis has to do with some parameter of the distribution $f_X(x|\theta)$ that generates the data. The null hypothesis will frequently be of the form

$$H_0 : \theta \in \Theta_0 \subseteq \Theta$$

$$H_A : \theta \notin \Theta_0$$

or of the form

$$H_0 : h(\theta_0) = 0$$

$$H_A : h(\theta_A) = 0$$

In either case, α and β will be functions of the true θ :

- $\alpha(\theta) = P(\text{type I error} \mid \theta \in \Theta_0) = P(\text{reject } H_0 \mid \theta \in \Theta_0)$
- $\beta(\theta) = P(\text{type II error} \mid \theta \notin \Theta_0) = P(\text{accept } H_0 \mid \theta \notin \Theta_0)$
- $1 - \beta(\theta) = P(\text{reject } H_0 \mid \theta \notin \Theta_0) = \text{“power” of the test}$

Again, these will depend on which particular value of the parameter is the true one, hence the term “power curve.” This leads to the following definition:

- “size of the test” = $\alpha = \max_{\theta \in \Theta_0} \alpha(\theta)$. The size is the “worst-case” probability of making a type I error, that is, the probability of type I error when the true θ is making your life most difficult. (Typically, this occurs on the “border” of Θ_0 .)
- $1 - \alpha$ is the “confidence level” of the test
- A test with greater power (smaller $\beta(\theta)$) than all other tests with the same size α is called *most powerful* (for that particular θ).
- A test that is most powerful for all possible values of the underlying parameter is called *uniformly most powerful (UMP)*
- UMP tests need not exist
- A test is *unbiased* if its power is always greater than its size, i.e.

$$1 - \beta(\theta) \geq \alpha \text{ for all } \theta \notin \Theta_0$$

- A test is *consistent* if its power goes to 1 as the sample size goes to infinity.

8 The Trinity: Wald, LR and LM Tests and Their Asymptotic Equivalence

Consider maximum likelihood estimation of an unknown parameter vector $\theta_0 \in \mathbb{R}^K$ and a test of the hypothesis

$$H_0 : h(\theta_0) = 0$$

$$H_A : h(\theta_A) = 0$$

where $h(\theta_0)$ is a $J \times 1$ vector (i.e. imposes J restrictions on θ_0) with $J \leq K$. (Typically, $J < K$.)

Let

$$H(\theta) = \frac{\partial h(\theta)}{\partial \theta'}$$

(note: this is not related to the Hessian discussed in the CRLB section) with

$$\text{rank}(H(\theta)) = J$$

(This means that there are no redundant restrictions.) Define the following MLEs:

$$\hat{\theta}_U = \arg \max_{\theta} \mathcal{L}(\theta)$$

$$\hat{\theta}_R = \arg \max_{\theta: h(\theta)=0} \mathcal{L}(\theta)$$

$\hat{\theta}_U$ is the “unrestricted” MLE, since it can assume any value, while $\hat{\theta}_R$ is the “restricted” MLE, since it must conform to $h(\theta) = 0$.

8.1 Wald Test

The Wald test is based on the observation that by the consistency and invariance properties of MLE, in large samples,

$$h(\hat{\theta}_U) \approx h(\theta_0)$$

and, if H_0 is true,

$$h(\theta_0) = 0$$

So, if the null is true, then, in large samples

$$h(\hat{\theta}_U) \approx 0$$

Of course, it is not likely that $h(\hat{\theta}_U)$ will be exactly zero, even when the null is true, so we need to figure out how far from zero we can allow $h(\hat{\theta}_U)$ to be without rejecting the null. We use the *Wald test statistic*

$$W = h(\hat{\theta}_U)' \left[H(\hat{\theta}_U) \left[\hat{\mathcal{I}}_N(\hat{\theta}_U) \right]^{-1} H(\hat{\theta}_U)' \right]^{-1} h(\hat{\theta}_U)$$

where

$$\frac{1}{N} \hat{\mathcal{I}}_N(\hat{\theta}_U) \xrightarrow{p} \mathcal{I}_1(\theta_0)$$

is a consistent estimator of the Fisher Information matrix.

Theorem 31 Under H_0 ,

$$W \xrightarrow{d} \chi_J^2$$

and so the following test procedure has a confidence level approaching $1 - \alpha$:

$$\text{reject } H_0 \iff W \geq c_\alpha$$

where

$$c_\alpha : (\chi_J^2 \geq c_\alpha) = 1 - \alpha$$

To build some intuition, first note that W is increasing in $h(\hat{\theta}_U)$. Higher values of $h(\hat{\theta}_U)$ make the null hypothesis of $h(\theta_0) = 0$ less plausible. Second, note that W is also increasing in $\hat{\mathcal{I}}_N(\hat{\theta}_U)$. If $\hat{\mathcal{I}}_N(\hat{\theta}_U)$ is high, we believe we have a lot of *information* in our data (that's why it's

called the Fisher *Information* matrix), so we will reject H_0 for smaller values of $h(\hat{\theta}_U)$. That is, we are less likely to believe that $h(\hat{\theta}_U)$ is far from zero just because of sampling error when we believe that our data are very informative. Finally, note that W is *decreasing* in $H(\hat{\theta}_U)$. An intuitive interpretation of this is that if $h(\hat{\theta})$ is very sensitive to small changes in $\hat{\theta}$, then we should be more forgiving of deviations from $h(\hat{\theta}_U) = 0$.

8.2 Lagrange Multiplier Test

The LM (or “score” or “efficient score”) test is based on the observation that if the restrictions imposed on θ by $h(\cdot)$ are true, then the slope of the log-likelihood function evaluated at the *restricted* MLE, $\hat{\theta}_R$, should be approximately zero. Equivalently, the $J \times 1$ vector of Lagrange multipliers, λ , from the restricted maximization program are approximately zero. The constrained optimization problem

$$\begin{aligned} \max_{\theta} \mathcal{L}(\theta) \quad \text{s.t.} \quad h(\theta) &= 0 \\ \max_{\theta, \lambda} \mathcal{L}(\theta) + \lambda' h(\theta) \end{aligned}$$

leads to the FOCs

$$\begin{aligned} FOC_{\theta} &: \frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} + H(\hat{\theta}_R)' \lambda = 0 \\ FOC_{\lambda} &: h(\hat{\theta}_R) = 0 \end{aligned}$$

From the first FOC, we have

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} = -H(\hat{\theta}_R)' \lambda$$

Because H is full rank, $\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta}$ can be small only if λ is small, which is why this is called the “Lagrange multiplier test.” If H_0 is true, we expect

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} \approx \frac{\partial \mathcal{L}(\hat{\theta}_U)}{\partial \theta}$$

and since $\hat{\theta}_U$ is unrestricted, we know that

$$\frac{\partial \mathcal{L}(\hat{\theta}_U)}{\partial \theta} = 0$$

so when the null is true we should have

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} \approx 0$$

The Lagrange multiplier statistic, then, is

$$LM = \left[\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} \right]' \left[\hat{\mathcal{I}}_N(\hat{\theta}_R) \right]^{-1} \left[\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta} \right]$$

where

$$\frac{1}{N} \hat{\mathcal{I}}_N(\hat{\theta}_R) \xrightarrow{p} \mathcal{I}_1(\theta_0)$$

when H_0 is true.

Theorem 32 Under H_0 ,

$$LM \xrightarrow{d} \chi_J^2$$

and so the following test procedure has a confidence level approaching $1 - \alpha$:

$$\text{reject } H_0 \iff LM \geq c_\alpha$$

where

$$c_\alpha : (\chi_J^2 \geq c_\alpha) = 1 - \alpha$$

Note that LM is increasing in

$$\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta}$$

which means that we are more likely to reject when the restriction is “costly” (to use the Micro interpretation of the Lagrange multiplier). It is *decreasing* in $\widehat{\mathcal{I}}_N(\hat{\theta}_R)$, which may seem counter-intuitive at first, especially after seeing the Wald statistic. However, when we recall that

$$\mathcal{I}_N(\theta_0) = -E \left[\frac{\partial^2 \mathcal{L}(\theta_0)}{\partial \theta' \partial \theta} \right]$$

we can see that if $\widehat{\mathcal{I}}_N(\hat{\theta}_R)$ is large it is likely that $\frac{\partial \mathcal{L}(\hat{\theta}_R)}{\partial \theta}$ is very sensitive to small changes in θ , so we should be more forgiving.

8.3 Likelihood Ratio Test

The LR test is based in the observation that if the restrictions are true, then we expect the likelihood functions to be roughly equal, i.e.

$$\frac{L(\hat{\theta}_R)}{L(\hat{\theta}_U)} \approx 1$$

and when the restrictions are false we expect

$$\frac{L(\hat{\theta}_R)}{L(\hat{\theta}_U)} \ll 1$$

The LR test is based on the negative of the log of this ratio

$$-\log \left(\frac{L(\hat{\theta}_R)}{L(\hat{\theta}_U)} \right) = \mathcal{L}(\hat{\theta}_U) - \mathcal{L}(\hat{\theta}_R)$$

which will be close to zero when the null is true and far from zero when the null is false:

$$LR = 2 \left[\mathcal{L}(\hat{\theta}_U) - \mathcal{L}(\hat{\theta}_R) \right]$$

Theorem 33 Under H_0 ,

$$LR \xrightarrow{d} \chi_J^2$$

and so the following test procedure has a confidence level approaching $1 - \alpha$:

$$\text{reject } H_0 \iff LM \geq c_\alpha$$

where

$$c_\alpha : (\chi_J^2 \geq c_\alpha) = 1 - \alpha$$

8.4 The Asymptotic Equivalence of W, LM and LR

Theorem 34 Under H_0 , all three tests have an asymptotic chi-squared distribution and are therefore asymptotically equivalent, i.e.

$$W \xrightarrow{d} \chi_J^2$$

$$LM \xrightarrow{d} \chi_J^2$$

$$LR \xrightarrow{d} \chi_J^2$$

Given c_α such that $P(\chi_J^2 \geq c_\alpha) = 1 - \alpha$, we reject with confidence level approximately $1 - \alpha$ with any of the following tests:

$$\text{reject } H_0 \iff W \geq c_\alpha$$

$$\text{reject } H_0 \iff LM \geq c_\alpha$$

$$\text{reject } H_0 \iff LR \geq c_\alpha$$

Remarks

- The Wald test requires only $\hat{\theta}_U$
- The LM test requires only $\hat{\theta}_R$
- The LR test requires both $\hat{\theta}_U$ and $\hat{\theta}_R$ but does not require estimation of the information matrix
- The equivalence of these tests is only asymptotic; they could give conflicting conclusions in finite samples
- It is usually the case that in finite samples

$$W > LR > LM$$

so the Wald is most likely to reject and the LM least likely to reject.

- The proof of these equivalences is arduous, see Bergantino (1998, 3) for details.
- See Greene (2003, Sec. 17.5) for more details on these tests and a nice picture giving intuition for their asymptotic equivalence.

References

Steven M. Bergantino. Handouts (various). Unpublished course notes for 14.382. Handout 1: Matrix Algebra; Handout 2: Distribution Theory; Handout 3: Statistical Inference. Year is approximate., 1998.

William H. Greene. *Econometric Analysis*. Prentice-Hall, 5th edition, 2003. URL <http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm>.

Guido Kuersteiner. Lecture notes on asymptotic theory. Unpublished lecture notes from 14.381, 2002.

Whitney K. Newey. Asymptotic theory of least squares. MIT lecture note, Spring 2000.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Daniel McFadden and Robert Engle, editors, *Handbook of Econometrics*, volume 4, chapter 36. Elsevier, North Holland, 1994.

Halbert White. *Asymptotic Theory for Econometricians*. Academic Press, 1st edition, 1984.