

**LECTURE 7**  
**LARGE SAMPLE THEORY**

## Limits and convergence concepts: almost sure, in probability, and in mean

Let  $\{a_n : n = 1, 2, \dots\}$  be a sequence of nonrandom real numbers. We say that  $a$  is the limit of  $\{a_n\}$  if for all real  $\delta > 0$  we can find an integer  $N_\delta$  such that for all  $n \geq N_\delta$ ,  $|a_n - a| < \delta$ . When the limit exists, we say that  $\{a_n\}$  converges to  $a$ , and write  $a_n \rightarrow a$  or  $\lim_{n \rightarrow \infty} a_n = a$ . In this case, we can make the elements of  $\{a_n : n \geq N\}$  arbitrarily close to  $a$  by choosing  $N$  sufficiently large. If  $a_n \rightarrow a$ , then  $a_n - a \rightarrow 0$ .

The concept can be extended to vectors or matrices as well. Let  $\{A_n : n = 1, 2, \dots\}$  be a sequence of  $m \times k$  matrices. Then  $A_n \rightarrow A$  if for all  $i = 1, \dots, m$  and  $j = 1, \dots, k$ , the  $(i, j)$ -th element of  $A_n$  converges to the  $(i, j)$ -th element of  $A$ .

The concept of convergence cannot be applied in a straightforward way to sequences of random variables. A random variable is a *function* from the sample space  $\Omega$  to the real line, so point-by-point convergence of the function values is not the natural notion. The solution is to consider convergence of a *nonrandom* sequence derived from the random one. Since there are many ways to derive nonrandom sequences, several stochastic convergence concepts arise. Let  $\{X_n : n = 1, 2, \dots\}$  be a sequence of *random* variables. Let  $X$  be random or nonrandom (i.e., it is possible that  $X(\omega)$  is the same for all  $\omega \in \Omega$ ). We will consider *nonrandom* sequences with the following typical elements: (i)  $X_n(\omega) - X(\omega)$  for a fixed  $\omega \in \Omega$ , (ii)  $E[|X_n - X|^r]$ , and (iii)  $\Pr(|X_n - X| > \varepsilon)$  for some  $\varepsilon > 0$ . These are sequences of nonrandom real numbers, and, consequently, the usual definition of convergence applies to each of them, leading to a corresponding definition of stochastic convergence:

**(i) Almost sure (with probability one) convergence.**  $X_n$  converges to  $X$  almost surely if

$$\Pr(\{\omega : X_n(\omega) \rightarrow X(\omega)\}) = 1.$$

**(ii) Convergence in  $r$ -th mean.**  $X_n$  converges to  $X$  in  $r$ -th mean if  $E[|X_n - X|^r] \rightarrow 0$  as  $n \rightarrow \infty$ .

**(iii) Convergence in probability.**  $X_n$  converges in probability to  $X$  if for all  $\varepsilon > 0$ ,  $\Pr(|X_n - X| \geq \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . It is denoted  $X_n \rightarrow_p X$  or  $\text{plim } X_n = X$ . Alternatively, convergence in probability can be defined as  $\Pr(|X_n - X| < \varepsilon) \rightarrow 1$  for all  $\varepsilon > 0$ . The two definitions are equivalent.

Almost sure convergence is the concept most closely related to convergence of nonrandom sequences. It implies that for almost all outcomes of the random experiment,  $X_n$  converges to  $X$ . Convergence in probability is the most widely used of the three concepts in econometrics.

An equivalent characterization:  $X_n \rightarrow X$  almost surely if and only if for all  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr\left(\sup_{m \geq n} |X_m - X| \geq \varepsilon\right) = 0. \quad (1)$$

Thus, while convergence in probability focuses only on the marginal distribution of  $|X_n - X|$  as  $n \rightarrow \infty$ , almost sure convergence puts restrictions on the joint behavior of all random elements in the sequence  $|X_n - X|, |X_{n+1} - X|, \dots$

Almost sure convergence and convergence in  $r$ -th mean each imply convergence in probability. However, the converse does not hold.

**Example. (a)** Consider a random variable  $X_n$  such that  $\Pr(X_n = 0) = 1 - 1/n$ , and  $\Pr(X_n = n) = 1/n$ . Here  $X_n \rightarrow_p 0$ , since for  $\varepsilon > 0$ ,<sup>1</sup>

$$\begin{aligned} \Pr(|X_n| \geq \varepsilon) &\leq \Pr(X_n = n) \\ &= 1/n \\ &\rightarrow 0. \end{aligned} \quad (2)$$

<sup>1</sup>If one chooses  $\varepsilon$  such that  $\varepsilon > n$ ,  $\Pr(|X_n| \geq \varepsilon)$  is zero. That is why we have “ $\leq$ ” in (2). When establishing convergence results of this kind, only small values of  $\varepsilon$  matter.

However,

$$E[|X_n|] = 1,$$

and, consequently,  $X_n$  does not converge in mean to zero.

(b) In part (a), only the marginal distributions of the elements of  $\{X_n\}$  have been described. Discussing almost sure convergence requires characterizing the joint distribution of the elements of  $\{X_n\}$ . A simple model for the joint distribution: let  $\omega$  be drawn from the uniform distribution on  $[0, 1]$ , so that  $\Pr(a \leq \omega \leq b) = |b - a|$  for  $a, b \in [0, 1]$ . Now,

$$X_n(\omega) = \begin{cases} n, & \omega \in [0, 1/n), \\ 0, & \omega \in [1/n, 1]. \end{cases}$$

so that  $\Pr(X_n = 0) = 1 - 1/n$  and  $\Pr(X_n = n) = 1/n$  as before. Using (1),  $X_n \rightarrow 0$  almost surely. To verify this, note that the condition (1) can be stated equivalently as  $\lim_{n \rightarrow \infty} \Pr(\sup_{m \geq n} |X_m - X| < \varepsilon) = 1$  for all  $\varepsilon > 0$ . Now, with the limit  $X = 0$ , define

$$P_{n,\varepsilon} = \Pr\left(\sup_{m \geq n} |X_m| < \varepsilon\right).$$

For all sufficiently small  $\varepsilon > 0$ ,

$$\begin{aligned} P_{n,\varepsilon} &= \Pr(|X_n| < \varepsilon, |X_{n+1}| < \varepsilon, \dots) \\ &= \Pr(X_n = 0, X_{n+1} = 0, \dots) \\ &= \Pr\left(\omega \geq \frac{1}{n}, \omega \geq \frac{1}{n+1}, \dots\right) \\ &= \Pr\left(\omega \geq \frac{1}{n}\right) \\ &= 1 - \frac{1}{n}. \end{aligned}$$

Thus, for all  $\varepsilon > 0$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr\left(\sup_{m \geq n} |X_m - X| < \varepsilon\right) &= \lim_{n \rightarrow \infty} P_{n,\varepsilon} \\ &= \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \\ &= 1. \end{aligned}$$

(c) Suppose instead that the elements of the sequence  $\{X_n\}$  are independent, but continue to assume that  $\Pr(X_n = 0) = 1 - 1/n$  and  $\Pr(X_n = n) = 1/n$ . For all positive integers  $n$ , we have now that

$$\begin{aligned} P_{n,\varepsilon} &= \Pr(X_n = 0, X_{n+1} = 0, \dots) \\ &= \prod_{m=n}^{\infty} \left(1 - \frac{1}{m}\right) \\ &= \lim_{N \rightarrow \infty} \prod_{m=n}^N \left(1 - \frac{1}{m}\right) \\ &= \lim_{N \rightarrow \infty} \prod_{m=n}^N \frac{m-1}{m} \\ &= \lim_{N \rightarrow \infty} \frac{n-1}{n} \frac{n}{n+1} \cdots \frac{N-2}{N-1} \frac{N-1}{N} \\ &= \lim_{N \rightarrow \infty} \frac{n-1}{N} \\ &= 0, \end{aligned}$$

Therefore, for all sufficiently small  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr \left( \sup_{m \geq n} |X_m| < \varepsilon \right) = \lim_{n \rightarrow \infty} P_{n,\varepsilon} = 0,$$

which violates (1). In this example, almost sure convergence of  $X_n$  to zero fails, as does convergence in mean, while  $X_n \rightarrow_p 0$  still holds.

Next, we show that convergence in  $r$ -th mean implies convergence in probability. The proof requires the following lemma.

**Lemma 1 (Markov's Inequality).** *Let  $X$  be a random variable. For  $\varepsilon > 0$  and  $r > 0$ ,*

$$\Pr(|X| \geq \varepsilon) \leq \mathbb{E}[|X|^r] / \varepsilon^r.$$

**Proof.** Let  $I(\cdot)$  be an indicator function, that is, it is equal to one if the condition inside the parentheses is satisfied, and zero otherwise. For example,

$$I(|x| \geq \varepsilon) = \begin{cases} 1, & |x| \geq \varepsilon, \\ 0, & |x| < \varepsilon. \end{cases}$$

Since  $I(|x| \geq \varepsilon) + I(|x| < \varepsilon) = 1$ ,

$$\begin{aligned} \mathbb{E}[|X|^r] &= \mathbb{E}[|X|^r I(|X| \geq \varepsilon)] + \mathbb{E}[|X|^r I(|X| < \varepsilon)] \\ &\geq \mathbb{E}[|X|^r I(|X| \geq \varepsilon)] \\ &\geq \varepsilon^r \mathbb{E}[I(|X| \geq \varepsilon)] \\ &= \varepsilon^r \Pr(|X| \geq \varepsilon), \end{aligned}$$

where the last equality follows from the fact that for any event  $A$ ,

$$\begin{aligned} \mathbb{E}[I(A)] &= 1 \cdot \Pr(A) + 0 \cdot \Pr(A^c) \\ &= \Pr(A). \end{aligned}$$

□

Now, suppose that  $X_n$  converges to  $X$  in  $r$ -th mean,  $\mathbb{E}[|X_n - X|^r] \rightarrow 0$ . Then,

$$\begin{aligned} \Pr(|X_n - X| \geq \varepsilon) &\leq \mathbb{E}[|X_n - X|^r] / \varepsilon^r \\ &\rightarrow 0. \end{aligned}$$

The following rules govern the manipulation of probability limits. Suppose that  $X_n \rightarrow_p a$  and  $Y_n \rightarrow_p b$ , where  $a$  and  $b$  are some finite constants. Let  $c$  be another constant. Then,

- (i)  $cX_n \rightarrow_p ca$ .
- (ii)  $X_n + Y_n \rightarrow_p a + b$ .
- (iii)  $X_n Y_n \rightarrow_p ab$ .
- (iv)  $X_n / Y_n \rightarrow_p a/b$ , provided that  $b \neq 0$ .

**Proof of (ii).**

$$\begin{aligned} \Pr(|(X_n + Y_n) - (a + b)| \geq \varepsilon) &= \Pr(|(X_n - a) + (Y_n - b)| \geq \varepsilon) \\ &\leq \Pr(|X_n - a| + |Y_n - b| \geq \varepsilon) \\ &\leq \Pr(|X_n - a| \geq \varepsilon/2 \text{ or } |Y_n - b| \geq \varepsilon/2) \\ &\leq \Pr(|X_n - a| \geq \varepsilon/2) + \Pr(|Y_n - b| \geq \varepsilon/2) \\ &\rightarrow 0. \end{aligned}$$

□

The following result shows that if a sequence of random variables converges in probability to a constant, then continuous functions of the sequence converge in probability as well.

**Theorem 2 (Slutsky's Theorem).** Suppose that  $X_n \rightarrow_p c$ , a constant, and let  $h(\cdot)$  be a function continuous at  $c$ . Then,  $h(X_n) \rightarrow_p h(c)$ .

**Proof:** By continuity of  $h(\cdot)$ , given  $\varepsilon > 0$ , there exists  $\delta_\varepsilon > 0$  such that  $|u - c| < \delta_\varepsilon$  implies that  $|h(u) - h(c)| < \varepsilon$ . Consequently, we have the following relation between the two events:

$$\{\omega : |h(X_n(\omega)) - h(c)| < \varepsilon\} \supset \{\omega : |X_n(\omega) - c| < \delta_\varepsilon\},$$

and, therefore,

$$\begin{aligned} \Pr(|h(X_n) - h(c)| < \varepsilon) &\geq \Pr(|X_n - c| < \delta_\varepsilon) \\ &\rightarrow 1. \end{aligned}$$

□

For example, suppose that  $\hat{\beta}_n \rightarrow_p \beta$ . Then  $\hat{\beta}_n^2 \rightarrow_p \beta^2$ , and  $1/\hat{\beta}_n \rightarrow_p 1/\beta$ , provided  $\beta \neq 0$ .

Random vectors/matrices converge in probability if their elements converge in probability. Alternatively, one may consider convergence in probability of norms. Consider the vector case. Let  $\{X_n : n = 1, 2, \dots\}$  be a sequence of random  $k$ -vectors. We will show that  $X_n - X \rightarrow_p 0$  element-by-element, where  $X$  is a possibly random  $k$ -vector, if and only if  $\|X_n - X\| \rightarrow_p 0$ , where  $\|\cdot\|$  denotes the Euclidean norm. First, suppose that  $X_{n,i} - X_i \rightarrow_p 0$  for all  $i = 1, \dots, k$ . Then,

$$\begin{aligned} \|X_n - X\|^2 &= \sum_{j=1}^k (X_{n,j} - X_j)^2 \\ &\rightarrow_p 0, \end{aligned}$$

due to Slutsky's Theorem and property (ii) above. Next, suppose that  $\|X_n - X\|^2 \rightarrow_p 0$ . Since  $\|X_n - X\|^2 = \sum_{j=1}^k (X_{n,j} - X_j)^2$ , and  $(X_{n,j} - X_j)^2 \geq 0$  for all  $n$  and  $j = 1, \dots, k$ , it must be that  $X_{n,j} - X_j \rightarrow_p 0$ .

The rules for manipulation of probability limits in the vector/matrix case are similar to those in the scalar case, (i)–(iv) above, with corresponding definitions of multiplication and division. Slutsky's Theorem is valid in the vector/matrix case as well.

## Weak Law of Large Numbers (WLLN)

The WLLN is one of the most important examples of convergence in probability.

**Theorem 3 (WLLN).** Let  $X_1, \dots, X_n$  be a sample of iid random variables such that  $E[|X_1|] < \infty$ . Then,  $n^{-1} \sum_{i=1}^n X_i \rightarrow_p E[X_1]$  as  $n \rightarrow \infty$ .

By the iid assumption,  $E[X_i] = E[X_1]$  for all  $i = 1, \dots, n$ . We will prove the result assuming instead that  $E[X_1^2] < \infty$ , which implies that  $E[|X_1|] < \infty$ , and  $\text{Var}(X_1) < \infty$ .

**Theorem 4.** Let  $X_1, \dots, X_n$  be a sample of iid random variables such that  $\text{Var}(X_1) < \infty$ . Then,

$$n^{-1} \sum_{i=1}^n X_i \rightarrow_p E[X_1] \quad \text{as } n \rightarrow \infty.$$

**Proof:**

$$\begin{aligned}
\Pr\left(\left|n^{-1}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right) &= \Pr\left(\left|n^{-1}\sum_{i=1}^n (X_i - \mathbb{E}[X_1])\right| \geq \varepsilon\right) \\
&\leq \frac{\mathbb{E}\left[\left|\sum_{i=1}^n (X_i - \mathbb{E}[X_1])\right|^2\right]}{n^2\varepsilon^2} \\
&= \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[(X_i - \mathbb{E}[X_1])(X_j - \mathbb{E}[X_1])]}{n^2\varepsilon^2} \\
&= \frac{\sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_1])^2\right]}{n^2\varepsilon^2} \\
&= \frac{n \operatorname{Var}(X_1)}{n^2\varepsilon^2} \\
&\rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

□

The proof shows that Theorem 4 holds under conditions weaker than iid. For example, instead of the iid part of the assumptions of the theorem, it suffices to assume that  $\mathbb{E}[X_i] = \mathbb{E}[X_1]$  for all  $i$  and  $\operatorname{Cov}(X_i, X_j) = 0$  for all  $i \neq j$  (uncorrelatedness). The proof carries through unchanged, and these conditions can be weakened further.

## Convergence in distribution

Convergence in distribution is another stochastic convergence concept used to approximate the distribution of a random variable  $X_n$  in large samples. Let  $\{X_n : n = 1, 2, \dots\}$  be a sequence of random variables. Let  $F_n(x)$  denote the marginal CDF of  $X_n$ , that is,  $F_n(x) = \Pr(X_n \leq x)$ . Let  $F(x)$  be another CDF. We say that  $X_n$  *converges in distribution* if  $F_n(x) \rightarrow F(x)$  for all  $x$  where  $F(x)$  is continuous. In this case, we write  $X_n \rightarrow_d X$ , where  $X$  is *any* random variable with distribution function  $F(x)$ . Although we say that  $X_n$  converges to  $X$ , convergence in distribution is convergence of distribution functions, not of random variables.

The extension to the vector case is straightforward. Let  $X_n$  and  $X$  be two random  $k$ -vectors. We say that  $X_n \rightarrow_d X$  if the joint CDF of  $X_n$  converges to that of  $X$  at all continuity points, that is,

$$\begin{aligned}
F_n(x_1, \dots, x_k) &= \Pr(X_{n,1} \leq x_1, \dots, X_{n,k} \leq x_k) \\
&\rightarrow \Pr(X_1 \leq x_1, \dots, X_k \leq x_k) \\
&= F(x_1, \dots, x_k)
\end{aligned}$$

for all points  $(x_1, \dots, x_k)$  where  $F$  is continuous. In this case, we say that the elements of  $X_n$ ,  $X_{n,1}, \dots, X_{n,k}$ , *jointly* converge in distribution to  $X_1, \dots, X_k$ , the elements of  $X$ .

The rules for manipulating convergence in distribution are as follows.

**(i) Cramér Convergence Theorem:** Suppose that  $X_n \rightarrow_d X$ , and  $Y_n \rightarrow_p c$ . Then,

- (a)  $X_n + Y_n \rightarrow_d X + c$ .
- (b)  $Y_n X_n \rightarrow_d cX$ ,
- (c)  $X_n/Y_n \rightarrow_d X/c$ , provided that  $c \neq 0$ .

Similar results hold in the vector/matrix case with proper definitions of multiplication and division.

**(ii)** If  $X_n \rightarrow_p X$ , then  $X_n \rightarrow_d X$ . The converse does not hold in general, with one exception:

**(iii)** If  $X_n \rightarrow_d C$ , a constant, then  $X_n \rightarrow_p C$ .

(iv) If  $X_n - Y_n \rightarrow_p 0$ , and  $Y_n \rightarrow_d Y$ , then  $X_n \rightarrow_d Y$ .

The following theorem extends convergence in distribution of random variables/vectors to convergence of their continuous functions.

**Theorem 5 (Continuous Mapping Theorem (CMT)).** *Suppose that  $X_n \rightarrow_d X$ , and let  $h(\cdot)$  be a function continuous on a set  $\mathcal{X}$  such that  $\Pr(X \in \mathcal{X}) = 1$ . Then,  $h(X_n) \rightarrow_d h(X)$ .*

Examples:

- Suppose that  $X_n \rightarrow_d X$ . Then  $X_n^2 \rightarrow_d X^2$ . For example, if  $X_n \rightarrow_d N(0, 1)$ , then  $X_n^2 \rightarrow_d \chi_1^2$ .
- Suppose that  $(X_n, Y_n) \rightarrow_d (X, Y)$  (joint convergence in distribution), and set  $h(x, y) = x$ . Then  $X_n \rightarrow_d X$ . Set  $h(x, y) = x^2 + y^2$ . Then  $X_n^2 + Y_n^2 \rightarrow_d X^2 + Y^2$ . For example, if  $(X_n, Y_n) \rightarrow_d N(0, I_2)$  (bivariate standard normal distribution), then  $X_n^2 + Y_n^2 \rightarrow_d \chi_2^2$ .

Unlike convergence in probability,  $X_n \rightarrow_d X$  and  $Y_n \rightarrow_d Y$  do not imply that  $X_n + Y_n \rightarrow_d X + Y$ , unless a joint convergence result holds. This is due to the fact that the individual convergence in distribution is convergence of the *marginal* CDFs. To characterize the limiting distribution of  $X_n + Y_n$ , one has to consider the limiting behavior of the *joint* CDF of  $X_n$  and  $Y_n$ .

## The Central Limit Theorem (CLT)

Various versions of the CLT are used to establish convergence in distribution of rescaled sums of random variables.

**Theorem 6 (CLT).** *Let  $X_1, \dots, X_n$  be a sample of iid random variables such that  $E[X_1] = 0$  and  $0 < E[X_1^2] < \infty$ . Then, as  $n \rightarrow \infty$ ,  $n^{-1/2} \sum_{i=1}^n X_i \rightarrow_d N(0, E[X_1^2])$ .*

For example, the CLT can be used to approximate the distribution of the average in large samples as follows. Let  $X_1, \dots, X_n$  be a sample of iid random variables with  $E[X_1] = \mu$  and  $\text{Var}(X_1) = \sigma^2 < \infty$ . Define

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i.$$

Consider  $n^{-1/2} \sum_{i=1}^n (X_i - \mu)$ . We have that  $(X_1 - \mu), \dots, (X_n - \mu)$  are iid with the mean  $E[X_1 - \mu] = 0$ , and the variance  $E[(X_1 - \mu)^2] = \sigma^2 < \infty$ . Therefore, by the CLT,

$$\begin{aligned} n^{1/2} (\bar{X}_n - \mu) &= n^{-1/2} \sum_{i=1}^n (X_i - \mu) \\ &\rightarrow_d N(0, \sigma^2). \end{aligned}$$

In practice, we use convergence in distribution as an *approximation*. Let  $\overset{a}{\sim}$  denote ‘‘approximately in large samples’’. Informally, one can say that  $n^{1/2} (\bar{X}_n - \mu) \overset{a}{\sim} N(0, \sigma^2)$  or

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

Under the normality assumption, the result holds *exactly* for any sample size  $n$ .

The CLT can be extended to the vector case by means of the following result.

**Lemma 7 (Cramér-Wold device).** *Let  $X_n$  be a random  $k$ -vector. Then,  $X_n \rightarrow_d X$  if and only if  $\lambda^\top X_n \rightarrow_d \lambda^\top X$  for all nonzero  $\lambda \in \mathbb{R}^k$ .*

**Corollary 8 (Multivariate CLT).** *Let  $X_1, \dots, X_n$  be a sample of iid random  $k$ -vectors such that  $E[X_1] = 0$  and  $E[X_{1,j}^2] < \infty$  for all  $j = 1, \dots, k$ , and  $E[X_1 X_1^\top]$  is positive definite. Then,  $n^{-1/2} \sum_{i=1}^n X_i \rightarrow_d N(0, E[X_1 X_1^\top])$ .*

**Proof:** Let  $\lambda$  be a  $k$ -vector of constants. Consider  $Y_i = \lambda^\top X_i$ . The random variables  $Y_1, \dots, Y_n$  are iid. Further,

$$\begin{aligned} \mathbb{E}[Y_1] &= \lambda^\top \mathbb{E}[X_1] \\ &= 0, \\ \text{Var}(Y_1) &= \mathbb{E}[Y_1^2] \\ &= \lambda^\top \mathbb{E}[X_1 X_1^\top] \lambda. \end{aligned}$$

The variance of  $Y_1$  is finite provided that all the elements of the variance-covariance matrix  $\mathbb{E}[X_1 X_1^\top]$  are finite. To verify this, observe that the  $(r, s)$ -th element of  $\mathbb{E}[X_1 X_1^\top]$  is given by  $\mathbb{E}[X_{1,r} X_{1,s}]$ . By the Cauchy-Schwarz inequality,

$$\mathbb{E}[|X_{1,r} X_{1,s}|] \leq \sqrt{\mathbb{E}[X_{1,r}^2] \mathbb{E}[X_{1,s}^2]},$$

which is finite for all  $r = 1, \dots, k$ ,  $s = 1, \dots, k$  due to the assumption that  $\mathbb{E}[X_{1,j}^2] < \infty$  for all  $j = 1, \dots, k$ . Consequently,  $\mathbb{E}[Y_1^2] < \infty$ , and it follows from the univariate CLT that

$$n^{-1/2} \sum_{i=1}^n Y_i \rightarrow_d N(0, \lambda^\top \mathbb{E}[X_1 X_1^\top] \lambda).$$

Let  $W$  be any  $N(0, \mathbb{E}[X_1 X_1^\top])$  random vector. Since  $\lambda^\top W \sim N(0, \lambda^\top \mathbb{E}[X_1 X_1^\top] \lambda)$ ,

$$\begin{aligned} \lambda^\top \left( n^{-1/2} \sum_{i=1}^n X_i \right) &= n^{-1/2} \sum_{i=1}^n Y_i \\ &\rightarrow_d \lambda^\top W. \end{aligned}$$

Therefore, by the Cramér-Wold device,

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n X_i &\rightarrow_d W \\ &= N(0, \mathbb{E}[X_1 X_1^\top]). \end{aligned}$$

□

## Delta method

The delta method is used to derive the asymptotic distribution of nonlinear functions of estimators. For example, in the case of an iid random sample, by the WLLN the average converges in probability to the population mean:  $\bar{X}_n \rightarrow_p \mathbb{E}[X_1] = \mu$ . Further, it follows from Slutsky's Theorem that  $h(\bar{X}_n) \rightarrow_p h(\mu)$ . However, this does not allow us to approximate the distribution of  $h(\bar{X}_n)$ , since  $h(\mu)$  is a constant (non-random). The CMT cannot be applied to general nonlinear  $h(\bar{X}_n)$ , since we have only a convergence in distribution result for  $n^{1/2}(\bar{X}_n - \mu)$ .

**Theorem 9 (Delta method).** *Let  $\hat{\theta}_n$  be a random  $k$ -vector, and suppose that  $n^{1/2}(\hat{\theta}_n - \theta) \rightarrow_d Y$  as  $n \rightarrow \infty$ , where  $\theta$  is a  $k$ -vector of constants, and  $Y$  is a random  $k$ -vector. Let  $h: \mathbb{R}^k \rightarrow \mathbb{R}^m$  be a function continuously differentiable on some open neighborhood of  $\theta$ . Then,  $n^{1/2}(h(\hat{\theta}_n) - h(\theta)) \rightarrow_d \frac{\partial h(\theta)}{\partial \theta^\top} Y$ .*

**Proof:** First, note that  $n^{1/2}(\hat{\theta}_n - \theta) \rightarrow_d Y$  implies  $\hat{\theta}_n - \theta \rightarrow_p 0$ , that is,  $\hat{\theta}_n \rightarrow_p \theta$ . Indeed, define  $\tau_n = n^{-1/2}$ .

Then  $\tau_n \rightarrow 0$  and, consequently,  $\tau_n \rightarrow_p 0$ . By the Cramér Convergence Theorem (b),

$$\begin{aligned} \left(\widehat{\theta}_n - \theta\right) &= \tau_n n^{1/2} \left(\widehat{\theta}_n - \theta\right) \\ &\rightarrow_d (\text{plim } \tau_n) Y \\ &= 0. \end{aligned}$$

Therefore, by property (iii) of convergence in distribution,  $\widehat{\theta}_n \rightarrow_p \theta$ .

Apply the mean-value theorem to the function  $h\left(\widehat{\theta}_n\right)$  element-by-element (see the Appendix) to obtain

$$h\left(\widehat{\theta}_n\right) = h(\theta) + \frac{\partial h(\theta_n^*)}{\partial \theta^\top} \left(\widehat{\theta}_n - \theta\right), \quad (3)$$

where  $\theta_n^*$  is a random vector that lies between  $\widehat{\theta}_n$  and  $\theta$  (element-by-element), that is,  $\|\theta_n^* - \theta\| \leq \|\widehat{\theta}_n - \theta\|$ . Since  $\widehat{\theta}_n \rightarrow_p \theta$ , it has to be that  $\theta_n^* \rightarrow_p \theta$  as well:

$$\begin{aligned} \Pr(\|\theta_n^* - \theta\| \geq \varepsilon) &\leq \Pr\left(\|\widehat{\theta}_n - \theta\| \geq \varepsilon\right) \\ &\rightarrow 0. \end{aligned}$$

Furthermore, by Slutsky's Theorem,

$$\frac{\partial h(\theta_n^*)}{\partial \theta^\top} \rightarrow_p \frac{\partial h(\theta)}{\partial \theta^\top}. \quad (4)$$

Next, rewrite (3) as follows:

$$n^{1/2} \left(h\left(\widehat{\theta}_n\right) - h(\theta)\right) = \frac{\partial h(\theta_n^*)}{\partial \theta^\top} n^{1/2} \left(\widehat{\theta}_n - \theta\right).$$

Then, it follows from the result in (4), the assumption  $n^{1/2} \left(\widehat{\theta}_n - \theta\right) \rightarrow_d Y$ , and the Cramér Convergence Theorem (b) that

$$n^{1/2} \left(h\left(\widehat{\theta}_n\right) - h(\theta)\right) \rightarrow_d \frac{\partial h(\theta)}{\partial \theta^\top} Y.$$

□

Consider again the example of the average of iid random variables with finite variance. By the CLT,  $n^{1/2} (\overline{X}_n - \mu) \rightarrow_d N(0, \sigma^2)$ . Suppose that  $\mu \neq 0$ . Then, by the delta method,

$$\begin{aligned} n^{1/2} \left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right) &\rightarrow_d -\frac{1}{\mu^2} N(0, \sigma^2) \\ &= N\left(0, \frac{\sigma^2}{\mu^4}\right). \end{aligned}$$

## Appendix A: Mean value theorem

**Theorem 10 (One-Dimensional Mean-Value Theorem).** *Let  $f : [a, b] \rightarrow \mathbb{R}$  be continuous on  $[a, b]$  and differentiable on  $(a, b)$ . Then there is  $c \in (a, b)$  such that*

$$f(b) - f(a) = \frac{df(c)}{dx} (b - a).$$

Now suppose  $h : \Theta \rightarrow \mathbb{R}$ , where  $\Theta \subset \mathbb{R}^k$ . Suppose further that  $h$  is continuously differentiable on some open neighborhood of  $\theta_0$ , say  $N_0$ , and let  $u$  be such that  $\theta_0 + tu \in N_0$  for all  $t \in [0, 1]$ . Define

$f(t) = h(\theta_0 + tu)$ . The function  $f$  is continuous and differentiable on the  $[0, 1]$  interval, and by the one-dimensional mean-value theorem,

$$\begin{aligned} h(\theta_0 + u) - h(\theta_0) &= f(1) - f(0) \\ &= \frac{df(t^*)}{dt} \text{ for some } t^* \in (0, 1) \\ &= \frac{\partial h(\theta_0 + t^*u)}{\partial \theta^\top} u \\ &= \frac{\partial h(\theta^*)}{\partial \theta^\top} u, \end{aligned}$$

where

$$\theta^* = \theta_0 + t^*u,$$

and

$$\begin{aligned} \|\theta^* - \theta_0\| &= t^* \|u\| \\ &< \|u\|. \end{aligned}$$

(The argument follows that of Theorem 10 on page 106 of Magnus and Neudecker (2007): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd Edition.) We have established a mean-value theorem for real-valued functions of several variables:

**Theorem 11.** *Let  $h : \Theta \rightarrow \mathbb{R}$ , where  $\Theta \subset \mathbb{R}^k$ , be continuously differentiable on some open neighborhood  $N_\theta$  of  $\theta$ . If  $\hat{\theta} \in N_\theta$ , then there is  $\theta^* \in N_\theta$  such that*

$$h(\hat{\theta}) - h(\theta) = \frac{\partial h(\theta^*)}{\partial \theta^\top} (\hat{\theta} - \theta),$$

where  $\|\theta^* - \theta\| \leq \|\hat{\theta} - \theta\|$ .

If  $h(\theta) = (h_1(\theta), \dots, h_m(\theta))^\top$  is a vector-valued function with  $h_j : \Theta \rightarrow \mathbb{R}$  for all  $j = 1, \dots, m$ , the above theorem can be applied element-by-element:

$$\begin{aligned} h(\hat{\theta}) - h(\theta) &= \begin{pmatrix} h_1(\hat{\theta}) - h_1(\theta) \\ \vdots \\ h_m(\hat{\theta}) - h_m(\theta) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial h_1(\theta^{*,1})}{\partial \theta^\top} (\hat{\theta} - \theta) \\ \vdots \\ \frac{\partial h_m(\theta^{*,m})}{\partial \theta^\top} (\hat{\theta} - \theta) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial h_1(\theta^*)}{\partial \theta^\top} \\ \vdots \\ \frac{\partial h_m(\theta^*)}{\partial \theta^\top} \end{pmatrix} (\hat{\theta} - \theta), \end{aligned}$$

where

$$\|\theta^{*,j} - \theta\| \leq \|\hat{\theta} - \theta\|, \tag{5}$$

for all  $j = 1, \dots, m$ . To simplify notation, write

$$\begin{pmatrix} \frac{\partial h(\theta^*)}{\partial \theta^\top} \\ \vdots \\ \frac{\partial h(\theta^*)}{\partial \theta^\top} \end{pmatrix} = \frac{\partial h(\theta^*)}{\partial \theta^\top},$$

indicating that  $\theta^*$  may be different across the rows of the matrix  $\partial h(\theta^*)/\partial\theta^\top$ , and that, in each row,  $\theta^*$  satisfies (5).

## Appendix B: Proof of the CLT

The material in this section is adapted from Hogg, McKean, and Craig (2005): *Introduction to Mathematical Statistics*. Let  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ , where the  $X_i$ 's are iid with mean  $\mu$  and variance  $\sigma^2$ . Recall from Lecture 1 that the MGF of an  $N(0, \sigma^2)$  distribution is given by  $\exp(t^2\sigma^2/2)$ . In view of Lemma 2 in Lecture 1, it suffices to show that the MGF of  $n^{1/2}(\bar{X}_n - \mu)$  converges to  $\exp(t^2\sigma^2/2)$ .<sup>2</sup>

Let  $m(t)$  denote the MGF of  $X_1 - \mu$ :

$$m(t) = \text{E}[\exp(t(X_1 - \mu))].$$

Recall from Lecture 1 that

$$\begin{aligned} m(0) &= 1, \\ m^{(1)}(0) &= \text{E}[X_1 - \mu] = 0, \\ m^{(2)}(0) &= \text{E}[(X_1 - \mu)^2] = \sigma^2, \end{aligned}$$

where

$$m^{(s)}(0) = \left. \frac{d^s m(t)}{dt^s} \right|_{t=0}.$$

We have the following expansion of  $m(t)$ :

$$\begin{aligned} m(t) &= m(0) + m^{(1)}(0)t + \frac{m^{(2)}(s)t^2}{2} \\ &= 1 + \frac{m^{(2)}(s)t^2}{2}, \end{aligned} \tag{6}$$

where  $s$  is some value between 0 and  $t$ .

---

<sup>2</sup>If the MGF does not exist, one can replace it with the *characteristic function* of  $X_1 - \mu$ , which is defined as  $\varphi(t) = \text{E}[\exp(it(X_1 - \mu))]$ , where  $i = \sqrt{-1}$ . The characteristic function always exists, and the proof with the characteristic function is essentially the same as the proof that uses the MGF.

Let  $M_n(t)$  denote the MGF of  $n^{1/2}(\bar{X}_n - \mu)$ . We have:

$$\begin{aligned}
M_n(t) &= \mathbb{E} \left[ \exp \left( t \frac{1}{n^{1/2}} \sum_{i=1}^n (X_i - \mu) \right) \right] \\
&= \mathbb{E} \left[ \prod_{i=1}^n \exp \left( \frac{t}{n^{1/2}} (X_i - \mu) \right) \right] \\
&= \prod_{i=1}^n \mathbb{E} \left[ \exp \left( \frac{t}{n^{1/2}} (X_i - \mu) \right) \right] \quad (\text{by independence}) \\
&= \left( \mathbb{E} \left[ \exp \left( \frac{t}{n^{1/2}} (X_1 - \mu) \right) \right] \right)^n \quad (\text{by identical distributions}) \\
&= \left( m \left( \frac{t}{n^{1/2}} \right) \right)^n \quad (\text{by definition of } m(t)) \\
&= \left( 1 + \frac{m^{(2)}(s) \left( \frac{t}{n^{1/2}} \right)^2}{2} \right)^n \quad (\text{by (6)}) \\
&= \left( 1 + \frac{m^{(2)}(s) t^2}{2n} \right)^n \\
&= \left( 1 + \frac{a_n}{n} \right)^n,
\end{aligned}$$

where  $s$  lies between 0 and  $t/n^{1/2}$  and therefore converges to zero as  $n \rightarrow \infty$ , and

$$\begin{aligned}
a_n &= \frac{m^{(2)}(s) t^2}{2} \\
&\rightarrow \frac{m^{(2)}(0) t^2}{2} \quad (\text{as } n \rightarrow \infty) \\
&= \frac{\sigma^2 t^2}{2}.
\end{aligned}$$

We will show next that

$$\log M_n(t) = n \log \left( 1 + \frac{a_n}{n} \right) \rightarrow \lim_{n \rightarrow \infty} a_n = \frac{\sigma^2 t^2}{2}. \quad (7)$$

The result in (7) implies that

$$M_n(t) = \exp(\log M_n(t)) \rightarrow \exp \left( \lim_{n \rightarrow \infty} \log M_n(t) \right) = \exp \left( \frac{\sigma^2 t^2}{2} \right).$$

To show (7), write

$$\begin{aligned}
\lim_{n \rightarrow \infty} n \log \left( 1 + \frac{a_n}{n} \right) &= \lim_{n \rightarrow \infty} \frac{\log(1 + a_n/n)}{1/n} \\
&= \lim_{n \rightarrow \infty} a_n \frac{\log(1 + a_n/n)}{a_n/n} \\
&= \lim_{n \rightarrow \infty} a_n \lim_{\delta \rightarrow 0} \frac{\log(1 + \delta)}{\delta} \quad (\text{by the change of variable } \delta = a_n/n) \\
&= \frac{\sigma^2 t^2}{2} \lim_{\delta \rightarrow 0} \frac{\log(1 + \delta)}{\delta}.
\end{aligned}$$

Lastly, by l'Hôpital's rule,

$$\lim_{\delta \rightarrow 0} \frac{\log(1 + \delta)}{\delta} = \lim_{\delta \rightarrow 0} \frac{1/(1 + \delta)}{1} = 1.$$