

# Asymptotic Theory

Zhentaο Shi

August 15, 2020

## 1 Asymptotics

Asymptotic theory is concerned about the behavior of statistics when the sample size is arbitrarily large. It is a useful approximation technique to simplify complicated finite-sample analysis.

### 1.1 Modes of Convergence

Convergence of a deterministic sequence  $z_1, z_2, \dots$  means that for any  $\varepsilon > 0$ , there exists an  $N(\varepsilon)$  such that for all  $n > N(\varepsilon)$ , we have  $|z_n - z| < \varepsilon$ . We say  $z$  is the limit of  $z_n$ , and write as  $z_n \rightarrow z$ .

Instead of a deterministic sequence, here we are interested in the convergence of a sequence of random variables. Since a random variable has a probability measure associated to any event, we must be clear what *convergence* means. Several modes of convergence are often encountered.

- Convergence almost surely\*
- Convergence in probability: for any  $\varepsilon > 0$ , as  $n \rightarrow \infty$  the probability  $P(\omega : |z_n(\omega) - z| < \varepsilon) \rightarrow 1$  (or equivalently  $P(\omega : |z_n(\omega) - z| \geq \varepsilon) \rightarrow 0$ ).
- Squared-mean convergence:  $\lim_{n \rightarrow \infty} E[(z_n - z)^2] = 0$ .

**Example:**  $(z_n)$  is a sequence of binary random variables:  $z_n = \sqrt{n}$  with probability  $1/n$ , and  $z_n = 0$  with probability  $1 - 1/n$ . Then  $z_n \xrightarrow{p} 0$  but  $z_n \not\xrightarrow{m.s.} 0$ .

*Justification:* For any  $\varepsilon > 0$ , we have  $P(\omega : |z_n(\omega) - 0| < \varepsilon) = P(\omega : z_n(\omega) = 0) = 1 - 1/n \rightarrow 1$ . Thus  $z_n \xrightarrow{p} 0$ . On the other hand,  $E[(z_n - 0)^2] = n \cdot 1/n + 0 \cdot (1 - 1/n) = 1 \not\rightarrow 0$ . Thus  $z_n \not\xrightarrow{m.s.} 0$ .

Convergence in probability does not count what happens on a subset in the sample space of small probability. Squared-mean convergence deals with the average over the entire probability space. If a random variable can take a wild value, even with small probability, it may blow away the squared-mean convergence. On the contrary, such irregularity does not undermine convergence in probability.

- Convergence in distribution:  $x_n \xrightarrow{d} x$  if  $F(x_n) \rightarrow F(x)$  for each  $x$  on which  $F(x)$  is continuous.

Convergence in distribution is about *pointwise* convergence of CDF, not the random variables themselves.

**Example:** Let  $x \sim N(0, 1)$ . If  $z_n = x + 1/n$ , then  $z_n \xrightarrow{p} x$  and of course  $z_n \xrightarrow{d} x$ . However, if  $z_n = -x + 1/n$ , or  $z_n = y + 1/n$  where  $y \sim N(0, 1)$  is independent of  $x$ , then  $z_n \xrightarrow{d} x$  but  $z_n \not\xrightarrow{p} x$ .

**Example:**  $(z_n)$  is a sequence of binary random variables:  $z_n = \sqrt{n}$  with probability  $1/n$ , and  $z_n = 0$  with probability  $1 - 1/n$ . Then  $z_n \xrightarrow{d} z = 0$ .

*Justification:*

$$F(z_n) = \begin{cases} 0 & z_n < 0 \\ 1 - 1/n & 0 \leq z_n \leq \sqrt{n} \\ 1 & z_n \geq \sqrt{n} \end{cases}$$

$$F(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}.$$

It is obvious that  $F(z_n)$  converges to  $F(z)$  pointwise on the set where  $F(z)$  is continuous.

Squared-mean convergence implies convergence in probability. Convergence in probability implies convergence in distribution.

*Cramér-Wold device* handles convergence in distribution for random vectors. We say a sequence of  $K$ -dimensional random vectors  $(X_n)$  converge in distribution to  $X$  if  $\lambda' X_n \xrightarrow{d} \lambda' X$  for any  $\lambda \in \mathbb{R}^K$ .

## 1.2 Law of Large Numbers

(Weak) law of large numbers (LLN) is a collection of statements about convergence in probability of the sample average to its population counterpart. The basic form of LLN is:

$$\frac{1}{n} \sum_{i=1}^n (z_i - E[z_i]) \xrightarrow{p} 0$$

as  $n \rightarrow \infty$ . Various versions of LLN work under different assumptions about the distributions and dependence of the random variables.

- Chebyshev LLN: if  $(z_1, \dots, z_n)$  is a sample of i.i.d. observations,  $E[z_1] = \mu$ , and  $\sigma^2 = \text{var}[z_1] < \infty$  exists, then  $\frac{1}{n} \sum_{i=1}^n z_i - \mu \xrightarrow{p} 0$ .

Chebyshev LLN utilizes *Chebyshev inequality*.

- *Chebyshev inequality*: for any random variable  $x$ , we have  $P(|x| > \varepsilon) \leq E[x^2] / \varepsilon^2$  for any  $\varepsilon > 0$ , if  $E[x^2] < \infty$ .

Chebyshev inequality is a special case of *Markov inequality*.

- *Markov inequality*:  $P(|x| > \varepsilon) \leq E[|x|^r] / \varepsilon^r$  for  $r \geq 1$  and any  $\varepsilon > 0$ , if  $E[|x|^r] < \infty$ .

It is easy to verify Markov inequality.

$$\begin{aligned} E[|x|^r] &= \int_{|x|>\varepsilon} |x|^r dF_X + \int_{|x|\leq\varepsilon} |x|^r dF_X \\ &\geq \int_{|x|>\varepsilon} |x|^r dF_X \geq \varepsilon^r \int_{|x|>\varepsilon} dF_X = \varepsilon^r P(|x| > \varepsilon). \end{aligned}$$

Consider a partial sum  $S_n = \sum_{i=1}^n x_i$ , where  $\mu_i = E[x_i]$  and  $\sigma_i^2 = \text{var}[x_i]$ . We apply the Chebyshev inequality to the sample mean  $\bar{x} - \bar{\mu} = n^{-1}(S_n - E[S_n])$ .

$$\begin{aligned} P(|\bar{x} - \bar{\mu}| \geq \varepsilon) &= P(|S_n - E[S_n]| \geq n\varepsilon) \\ &\leq (n\varepsilon)^{-2} E \left[ \sum_{i=1}^n (x_i - \mu_i)^2 \right] \\ &= (n\varepsilon)^{-2} \text{var} \left( \sum_{i=1}^n x_i \right) \\ &= (n\varepsilon)^{-2} \left[ \sum_{i=1}^n \text{var}(x_i) + \sum_{i=1}^n \sum_{j \neq i} \text{cov}(x_i, x_j) \right]. \end{aligned}$$

From the above derivation, convergence in probability holds as long as the right-hand side shrinks to 0 as  $n \rightarrow \infty$ . Actually, the convergence can be maintained under much more general conditions than just under the i.i.d. assumption. The random variables in the sample do not have to be identically distributed, and they do not have to be independent either.

Another useful LLN is *Kolmogorov LLN*. Since its derivation requires advanced knowledge of probability theory, we state the result without proof.

- Kolmogorov LLN: if  $(z_1, \dots, z_n)$  is a sample of i.i.d. observations and  $E[z_1] = \mu$  exists, then  $\frac{1}{n} \sum_{i=1}^n z_i - \mu \xrightarrow{p} 0$ .

Compared to Chebyshev LLN, Kolmogorov LLN only requires the existence of the population mean, but not any higher moments. On the other

hand, i.i.d. is essential for Kolmogorov LLN.

### 1.3 Central Limit Theorem

The central limit theorem (CLT) is a collection of probability results about the convergence in distribution to a stable law, usually the normal distribution. The basic form of the CLT is: for a sample  $(z_1, \dots, z_n)$  of *zero-mean* random variables,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \xrightarrow{d} N(0, \sigma^2). \quad (1)$$

Various versions of CLT work under different assumptions about the random variables.

*Lindeberg-Levy CLT* is the simplest CLT.

- If the sample is i.i.d.,  $E[x_1] = 0$  and  $\text{var}[x_1^2] = \sigma^2 < \infty$ , then CLT holds.

Lindeberg-Levy CLT is easy to verify by the characteristic function. For any random variable  $x$ , the function  $\varphi_x(t) = E[\exp(ixt)]$  is called its *characteristic function*. The characteristic function fully describes a distribution, just like PDF or CDF. For example, the characteristic function of  $N(\mu, \sigma^2)$  is  $\exp(it\mu - \frac{1}{2}\sigma^2 t^2)$ .

Here is a very heuristic argument. If  $E[|x|^k] < \infty$  for a positive integer  $k$ , then

$$\varphi_X(t) = 1 + itE[X] + \frac{(it)^2}{2}E[X^2] + \dots + \frac{(it)^k}{k!}E[X^k] + o(t^k).$$

Under the assumption of Lindeberg-Levy CLT,

$$\varphi_{X_i/\sqrt{n}}(t) = 1 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)$$

for all  $i$ , and by independence we have

$$\begin{aligned}\varphi_{\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i}(t) &= \prod_{i=1}^n \varphi_{x_i/\sqrt{n}}(t) = \left(1 - \frac{t^2}{2n}\sigma^2 + o\left(\frac{t^2}{n}\right)\right)^n \\ &\rightarrow \exp\left(-\frac{\sigma^2}{2}t^2\right),\end{aligned}$$

where the limit is exactly the characteristic function of  $N(0, \sigma^2)$ .

- Lindeberg-Feller CLT: i.n.i.d., and *Lindeberg condition*: for any fixed  $\varepsilon > 0$ ,

$$\frac{1}{s_n^2} \sum_{i=1}^n E \left[ x_i^2 \cdot \mathbf{1} \{ |x_i| \geq \varepsilon s_n \} \right] \rightarrow 0$$

where  $s_n = (\sum_{i=1}^n \sigma_i^2)^{1/2}$ .

- Lyapunov CLT: i.n.i.d.:  $\max_{i \leq n} E \left[ |x_i|^3 \right] < C < \infty$ .

## 1.4 Tools for Transformations

In their original forms, LLN deals with the sample mean, and CLT handles the scaled (by  $\sqrt{n}$ ) and/or standardized (by standard deviation) sample mean. However, most of the econometric estimators of interest are functions of sample means. Therefore, we need tools to handle transformations.

- Small op:  $x_n = o_p(r_n)$  if  $x_n/r_n \xrightarrow{p} 0$ .
- Big Op:  $x_n = O_p(r_n)$  if for any  $\varepsilon > 0$ , there exists a  $c > 0$  such that  $P(|x_n|/r_n > c) < \varepsilon$ .
- Continuous mapping theorem 1: If  $x_n \xrightarrow{p} a$  and  $f(\cdot)$  is continuous at  $a$ , then  $f(x_n) \xrightarrow{p} f(a)$ .
- Continuous mapping theorem 2: If  $x_n \xrightarrow{d} x$  and  $f(\cdot)$  is continuous almost surely on the support of  $x$ , then  $f(x_n) \xrightarrow{d} f(x)$ .
- Slutsky's Theorem: If  $x_n \xrightarrow{d} x$  and  $y_n \xrightarrow{p} a$ , then
  - $x_n + y_n \xrightarrow{d} x + a$
  - $x_n y_n \xrightarrow{d} ax$
  - $x_n / y_n \xrightarrow{d} x/a$  if  $a \neq 0$ .
- Delta method: if  $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$ , and  $f(\cdot)$  is continuously differentiable at  $\theta_0$ , then

$$\sqrt{n} \left( f(\hat{\theta}) - f(\theta_0) \right) \xrightarrow{d} N \left( 0, \frac{\partial f}{\partial \theta'}(\theta_0) \Omega \left( \frac{\partial f}{\partial \theta}(\theta_0) \right)' \right).$$

## 2 Asymptotic Properties of OLS

We apply large sample theory to study the OLS estimator  $\hat{\beta} = (X'X)^{-1} X'Y$ .



## 2.1 Consistency

We say  $\hat{\beta}$  is *consistent* if  $\hat{\beta} \xrightarrow{p} \beta$  as  $n \rightarrow \infty$ . To verify consistency, we write

$$\hat{\beta} - \beta = (X'X)^{-1} X'e = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n x_i e_i. \quad (2)$$

The first term

$$\hat{Q} = \frac{1}{n} \sum_{i=1}^n x_i x_i' \xrightarrow{p} Q = E[x_i x_i'],$$

and the second term

$$\frac{1}{n} \sum_{i=1}^n x_i e_i \xrightarrow{p} 0.$$

No matter whether  $(y_i, x_i)_{i=1}^n$  is an i.i.d., i.n.i.d., or dependent sample, as long as the convergence in probability holds for the above two expressions and  $Q$  is an invertible matrix, we have  $\hat{\beta} - \beta \xrightarrow{p} Q^{-1}0 = 0$  by the continuous mapping theorem. In other words,  $\hat{\beta}$  is a consistent estimator of  $\beta$ .

## 2.2 Asymptotic Normality

In finite sample,  $\hat{\beta}$  is a random variable. We have shown the distribution of  $\hat{\beta}$  under normality in the previous lecture. Without the restrictive normality assumption, how can we characterize the randomness of the OLS estimator?

We know from the previous section that  $\hat{\beta} - \beta \xrightarrow{p} 0$  degenerates to a constant. To study its distribution, we must scale it up by a proper multi-

plier so that in the limit it neither degenerates nor explodes. The suitable scaling factor is  $\sqrt{n}$ :

$$\sqrt{n} (\hat{\beta} - \beta) = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i e_i.$$

Since  $E [x_i e_i] = 0$ , we apply a CLT to obtain

$$n^{-1/2} \sum_{i=1}^n x_i e_i \xrightarrow{d} N(0, \Sigma)$$

where  $\Sigma = E [x_i x_i' e_i^2]$ . By the continuous mapping theorem,

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} Q^{-1} \times N(0, \Sigma) \sim N(0, \Omega)$$

where  $\Omega = Q^{-1} \Sigma Q^{-1}$  is called the *asymptotic variance*. This is the *asymptotic normality* of the OLS estimator.

Up to now we have derived the asymptotic distribution of  $\hat{\beta}$ . However, to make it feasible, we still have to estimator the asymptotic variance  $\Omega$ . If  $\hat{\Sigma}$  is a consistent estimator of  $\Sigma$ , then  $\hat{\Omega} = \hat{Q}^{-1} \hat{\Sigma} \hat{Q}^{-1}$  is a consistent estimator of  $\Omega$ . (Of course, there are other ways to estimate the asymptotic variance.) A feasible version about the distribution of  $\hat{\beta}$  is

$$\hat{\Omega}^{-1/2} \sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} N(0, I_K).$$

In the special case of homoskedasticity,  $\Sigma = E [x_i x_i' e_i^2] = E [x_i x_i' E [e_i^2 | X]] =$

$\sigma^2 E [x_i x_i'] = \sigma^2 Q$  so that

$$n^{-1/2} \sum_{i=1}^n x_i e_i \xrightarrow{d} N(0, \sigma^2 Q).$$

Again by the continuous mapping theorem,

$$\sqrt{n} (\hat{\beta} - \beta) \xrightarrow{d} Q^{-1} \times N(0, \sigma^2 Q) \sim N(0, \sigma^2 Q^{-1}).$$

We can estimate the unknown parameter  $\sigma^2$  by either  $\hat{\sigma}^2 = \hat{e}'\hat{e} / (n - K)$  or  $\hat{\sigma}^2 = \hat{e}'\hat{e} / n$ .

## 2.3 Estimation of the Variance

We show all elements of  $\Sigma = E [x_i x_i' e_i^2]$  is finite. That is,  $\|\Sigma\|_\infty := \max_{i,j \leq K} |\sigma_{ij}| < \infty$ . Let  $z_i = x_i e_i$ , so  $\Sigma = E [z_i z_i']$ . Because of the Cachy-Schwarz inequality,

$$\|\Sigma\|_\infty = \max_{k=1, \dots, K} E [z_{ik}^2].$$

For each  $k$ ,  $E [z_{ik}^2] = E [x_{ik}^2 e_i^2] \leq (E [x_{ik}^4] E [e_i^4])^{1/2}$ .

For the estimation of variance, if the error is homoskedastic,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 &= \frac{1}{n} \sum_{i=1}^n \left( e_i + x_i' (\hat{\beta} - \beta) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 + \left( \frac{2}{n} \sum_{i=1}^n e_i x_i \right)' (\hat{\beta} - \beta) + \frac{1}{n} \sum_{i=1}^n e_i^2 (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta). \end{aligned}$$

The second term

$$\left( \frac{2}{n} \sum_{i=1}^n e_i x_i \right)' (\hat{\beta} - \beta) = o_p(1) o_p(1) = o_p(1).$$

The third term

$$(\hat{\beta} - \beta) \left( \frac{1}{n} \sum_{i=1}^n e_i^2 x_i x_i' \right) (\hat{\beta} - \beta) = o_p(1) O_p(1) o_p(1) = o_p(1).$$

As  $\frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 + o_p(1)$  and  $\frac{1}{n} \sum_{i=1}^n e_i^2 = \sigma_e^2 + o_p(1)$ , we have  $\frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 = \sigma_e^2 + o_p(1)$ . In other words,  $\frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 \xrightarrow{p} \sigma_e^2$ .

For general heteroskedasticity,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 &= \frac{1}{n} \sum_{i=1}^n x_i x_i' \left( e_i + x_i' (\hat{\beta} - \beta) \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i x_i' e_i^2 + \frac{1}{n} \sum_{i=1}^n x_i x_i' e_i x_i' (\hat{\beta} - \beta) + \frac{1}{n} \sum_{i=1}^n x_i x_i' \left( (\hat{\beta} - \beta)' x_i \right)^2. \end{aligned}$$

The third term is bounded by

$$\begin{aligned}
& \text{trace} \left( \frac{1}{n} \sum_{i=1}^n x_i x_i' \left( (\hat{\beta} - \beta)' x_i \right)^2 \right) \\
& \leq K \max_k \frac{1}{n} \sum_{i=1}^n x_{ik}^2 \left[ (\hat{\beta} - \beta)' x_i \right]^2 \\
& \leq K \left\| \hat{\beta} - \beta \right\|_2^2 \max_k \frac{1}{n} \sum_{i=1}^n x_{ik}^2 \|x_i\|_2^2 \\
& \leq K \left\| \hat{\beta} - \beta \right\|_2^2 \frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2 \|x_i\|_2^2 \\
& = K \left\| \hat{\beta} - \beta \right\|_2^2 \frac{1}{n} \sum_{i=1}^n \left( \sum_{k=1}^K x_{ik}^2 \right)^2 \\
& \leq K \left\| \hat{\beta} - \beta \right\|_2^2 K \sum_{k=1}^K \frac{1}{n} \sum_{i=1}^n x_{ik}^4 = o_p(1) O_p(1) = o_p(1).
\end{aligned}$$

where the third inequality follows by  $(a_1 + \dots + a_K)^2 \leq K(a_1^2 + \dots + a_K^2)$ .

The second term is bounded by

$$\begin{aligned}
& \left| \frac{1}{n} \sum_{i=1}^n x_{ik} x_{ik'} e_i x_i' (\hat{\beta} - \beta) \right| \\
& \leq \max_k \left| \hat{\beta}_k - \beta_k \right| K \max_{k,k',k''} \left| \frac{1}{n} \sum_{i=1}^n e_i x_{ik} x_{ik'} x_{ik''} \right| \\
& \leq \left\| \hat{\beta} - \beta \right\|_2 \left( \frac{1}{n} \sum_{i=1}^n e_i^4 \right)^{1/4} K \max_{k,k',k''} \left( \frac{1}{n} \sum_{i=1}^n (x_{ik} x_{ik'} x_{ik''})^{4/3} \right)^{3/4} \\
& \leq \left\| \hat{\beta} - \beta \right\|_2 K \max_k \left( \frac{1}{n} \sum_{i=1}^n x_{ik}^4 \right)^{3/4} = o_p(1) O_p(1)
\end{aligned}$$

where the second and the third inequality hold by the Holder's inequality.