An Introduction to Asymptotic Theory

Ping Yu

School of Economics and Finance
The University of Hong Kong

Five Weapons in Asymptotic Theory

Five Weapons

- The weak law of large numbers (WLLN, or LLN)
- The central limit theorem (CLT)
- The continuous mapping theorem (CMT)
- Slutsky's theorem
- The Delta method

- **Notations**:
  - In nonlinear (in parameter) models, the capital letters such as X denote random variables or random vectors and the corresponding lower case letters such as x denote the potential values they may take.
  - Generic notation for a parameter in nonlinear environments (e.g., nonlinear models or nonlinear constraints) is θ, while in linear environments is β.

The WLLN

Definition

A random vector Z_n **converges in probability** to Z as $n \to \infty$, denoted as $Z_n \xrightarrow{p} Z$, if for any $\delta > 0$,

$$\lim_{n\to\infty} P(\|Z_n - Z\| > \delta) = 0.$$

- Although the limit Z can be random, it is usually constant. [intuition]
- The probability limit of Z_n is often denoted as $\text{plim}(Z_n)$. If $Z_n \xrightarrow{p} 0$, we denote $Z_n = o_p(1)$.
- When an estimator converges in probability to the true value as the sample size diverges, we say that the estimator is **consistent**.
- Consistency is an important preliminary step in establishing other important asymptotic approximations.

Theorem (WLLN)

Suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random vectors, and $E[\|X\|] < \infty$; then as $n \to \infty$,

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E[X].$$

The CLT

Definition

A random k vector Z_n *converges in distribution* to Z as $n \to \infty$, denoted as $Z_n \xrightarrow{d} Z$, if

$$\lim_{n\to\infty} F_n(z) = F(z),$$

at all z where $F(\cdot)$ is continuous, where F_n is the cdf of Z_n and F is the cdf of Z.

- Usually, Z is normally distributed, so all $z \in \mathbb{R}^k$ are continuity points of F.
- If Z_n converges in distribution to Z, then Z_n is **stochastically bounded** and we denote $Z_n = O_p(1)$.
- Rigorously, $Z_n = O_p(1)$ if $\forall \varepsilon > 0$, $\exists M_\varepsilon < \infty$ such that $P(\|Z_n\| > M_\varepsilon) < \varepsilon$ for any n. If $Z_n = o_p(1)$, then $Z_n = O_p(1)$.
- We can show that $o_p(1) + o_p(1) = o_p(1)$, $o_p(1) + O_p(1) = O_p(1)$, $O_p(1) + O_p(1) = O_p(1)$, $o_p(1)o_p(1) = o_p(1)$, $o_p(1)O_p(1) = o_p(1)$, and $O_p(1)O_p(1) = O_p(1)$.

Theorem (CLT)

suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random k vectors, $E[X] = \mu$, and $Var(X) = \Sigma$; then $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N(\mathbf{0}, \Sigma)$.

Comparison Between the WLLN and CLT

- The CLT tells more than the WLLN.
- $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} N(\mathbf{0}, \Sigma)$ implies $\overline{X}_n \xrightarrow{p} \mu$, so the CLT is stronger than the WLLN.
- $\overline{X}_n \xrightarrow{p} \mu$ means $\overline{X}_n - \mu = o_p(1)$, but does not provide any information about $\sqrt{n}\left(\overline{X}_n - \mu\right)$. The CLT tells that $\sqrt{n}\left(\overline{X}_n - \mu\right) = O_p(1)$ or $\overline{X}_n - \mu = O_p(n^{-1/2})$.
- But the WLLN does not require the second moment finite; that is, a stronger result is not free.

The CMT

Theorem (CMT)

Suppose $X_1, \cdots, X_n, \cdots$ are random k vectors, and g is a continuous function on the support of X (to $\mathbb{R}^l$) a.s. P_X; then

$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X);$$
$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X).$$

- The CMT allows the function g to be discontinuous but the probability of being at a discontinuity point is zero.
- For example, the function $g(u) = u^{-1}$ is discontinuous at $u = 0$, but if $X_n \xrightarrow{d} X \sim N(0,1)$ then $P(X = 0) = 0$ so $X_n^{-1} \xrightarrow{d} X^{-1}$.

Slutsky's Theorem

- In the CMT, X_n converges to X *jointly* in various modes of convergence.
- For the convergence in probability ($\xrightarrow{p}$), marginal convergence implies joint convergence, so there is no problem if we substitute joint convergence by marginal convergence.
- But for the convergence in distribution ($\xrightarrow{d}$), $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ does not imply $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$.
- Nevertheless, there is a special case where this result holds, which is Slutsky's theorem.

Theorem (Slutsky's Theorem)

If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$ $\left(\Longleftrightarrow Y_n \xrightarrow{p} c\right)$, *where* c *is a constant, then* $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$. *This implies* $X_n + Y_n \xrightarrow{d} X + c$, $Y_n X_n \xrightarrow{d} cX$, $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$ *when* $c \neq 0$. *Here* X_n, Y_n, X, c *can be understood as vectors or matrices as long as the operations are compatible.*

Applications of the CMT and Slutsky's Theorem

Example

Suppose $X_n \xrightarrow{d} N(\mathbf{0}, \Sigma)$, and $Y_n \xrightarrow{p} \Sigma$; then $Y_n^{-1/2} X_n \xrightarrow{d} \Sigma^{-1/2} N(0, \Sigma) = N(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. (why?)

Example

Suppose $X_n \xrightarrow{d} N(\mathbf{0}, \Sigma)$, and $Y_n \xrightarrow{p} \Sigma$; then $X_n' Y_n^{-1} X_n \xrightarrow{d} \chi_k^2$, where k is the dimension of X_n. (why?)

- Another important application of Slutsky's theorem is the Delta method.

The Delta Method

Theorem

Suppose $\sqrt{n}\left(Z_n - \mathbf{c}\right) \xrightarrow{d} Z \sim N(\mathbf{0}, \Sigma)$, $\mathbf{c} \in \mathbb{R}^k$, *and* $g(z) : \mathbb{R}^k \to \mathbb{R}$. *If* $\frac{dg(z)}{dz'}$ *is continuous at* c, *then* $\sqrt{n}\left(g(Z_n) - g(\mathbf{c})\right) \xrightarrow{d} \frac{dg(\mathbf{c})}{dz'} Z$.

Proof.

$$\sqrt{n}\left(g(Z_n) - g(\mathbf{c})\right) = \sqrt{n}\frac{dg(\overline{\mathbf{c}})}{dz'}\left(Z_n - \mathbf{c}\right),$$

where $\overline{\mathbf{c}}$ is between Z_n and $\mathbf{c}$. $\sqrt{n}\left(Z_n - \mathbf{c}\right) \xrightarrow{d} Z$ implies that $Z_n \xrightarrow{p} \mathbf{c}$, so by the CMT, $\frac{dg(\overline{\mathbf{c}})}{dz'} \xrightarrow{p} \frac{dg(\mathbf{c})}{dz'}$. By Slutsky's theorem, $\sqrt{n}\left(g(Z_n) - g(\mathbf{c})\right)$ has the asymptotic distribution $\frac{dg(\mathbf{c})}{dz'} Z$. □

- The Delta method implies that asymptotically, the randomness in a transformation of Z_n is completely controlled by that in Z_n.

Asymptotics for the MoM Estimator

The MoM Estimator

- Recall that the MoM estimator is defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i|\theta) = \mathbf{0}.$$

- We can prove the MoM estimator is consistent and asymptotically normal (CAN) under some regularity conditions.
- Specifically, the asymptotic distribution of the MoM estimator is

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{M}^{-1}\Omega\mathbf{M}'^{-1}\right),$$

  where $\mathbf{M} = \frac{dE[m(X|\theta_0)]}{d\theta'}$ and $\Omega = E\left[m(X|\theta_0)m(X|\theta_0)'\right]$.
- The asymptotic variance takes a *sandwich* form and can be estimated by its sample analog.

Derivation of the Asymptotic Distribution of the MoM Estimator

- $$\frac{1}{n}\sum_{i=1}^{n} m(X_i|\widehat{\theta}) = \mathbf{0}$$
  $$\Longrightarrow \frac{1}{n}\sum_{i=1}^{n} m(X_i|\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{dm(X_i|\overline{\theta})}{d\theta'}\left(\widehat{\theta}-\theta_0\right) = \mathbf{0}$$
  $$\Longrightarrow \sqrt{n}\left(\widehat{\theta}-\theta_0\right) = -\left(\frac{1}{n}\sum_{i=1}^{n} \frac{dm(X_i|\overline{\theta})}{d\theta'}\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(X_i|\theta_0)$$
  $$\underset{?}{\overset{d}{\longrightarrow}} -\mathbf{M}^{-1} N\left(\mathbf{0},\Omega\right)$$
- $\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} -\mathbf{M}^{-1} m(X_i|\theta_0)$, so $-\mathbf{M}^{-1} m(X_i|\theta_0)$ is called the **influence function**.
- We use $\frac{dE[m(X|\theta_0)]}{d\theta'}$ instead of $E\left[\frac{dm(X|\theta_0)}{d\theta'}\right]$ because $E\left[m(X|\theta)\right]$ is more smooth than $m(X|\theta)$ and can be applied to such situations as quantile estimation where $m(X|\theta)$ is not differentiable at θ_0. In this course, we will not meet such cases.

Intuition for the Asymptotic Distribution of the MoM Estimator

- Suppose $E[X] = g(\theta_0)$ with $g \in C^{(1)}$ in a neighborhood of θ_0; then $\theta_0 = g^{-1}(E[X]) \equiv h(E[X])$. (what are $m, \mathbf{M}$ and Ω here?)
- The MoM estimator of θ is to set $\overline{X} = g(\theta)$, so $\widehat{\theta} = h(\overline{X})$.
- By the WLLN, $\overline{X} \xrightarrow{p} E[X]$; then by the CMT, $\widehat{\theta} \xrightarrow{p} h(E[X]) = \theta_0$ since $h(\cdot)$ is continuous.
- Now, $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \sqrt{n}\left(h(\overline{X}) - h(E[X])\right) = \sqrt{n}h'\left(\overline{X}^*\right)\left(\overline{X} - E[X]\right) = h'\left(\overline{X}^*\right)\sqrt{n}\left(\overline{X} - E[X]\right)$, where the second equality is from the mean value theorem (MVT).
- Because $\overline{X}^*$ is between $\overline{X}$ and $E[X]$ and $\overline{X} \xrightarrow{p} E[X]$, $\overline{X}^* \xrightarrow{p} E[X]$.
- By the CMT, $h'\left(\overline{X}^*\right) \xrightarrow{p} h'(E[X])$. By the CLT, $\sqrt{n}\left(\overline{X} - E[X]\right) \xrightarrow{d} N(0, Var(X))$. Then by Slutsky's theorem,

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta} - \theta_0\right) &\xrightarrow{d} h'(E[X])\,N(0, Var(X)) \\ &= N\left(0, h'(E[X])^2 Var(X)\right) \overset{?}{=} N\left(0, \frac{Var(X)}{g'(\theta_0)^2}\right).\end{aligned}$$

continue...

- The larger $g'(\theta_0)$ is, the smaller the asymptotic variance of $\widehat{\theta}$ is.
- Consider a more specific example. Suppose the density of X is $\frac{2}{\theta}x\exp\left\{-\frac{x^2}{\theta}\right\}$, $\theta > 0$, $x > 0$, that is, X follows the Weibull $(2, \theta)$ distribution.
- We can show $E[X] = g(\theta) = \frac{\sqrt{\pi}}{2}\theta^{1/2}$, and $Var(X) = \theta\left(1 - \frac{\pi}{4}\right)$.
- So $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \frac{\theta\left(1-\frac{\pi}{4}\right)}{\left(\frac{\sqrt{\pi}}{2}\frac{1}{2}\theta^{-1/2}\right)^2}\right) = N\left(0, 16\theta^2\left(\frac{1}{\pi} - \frac{1}{4}\right)\right)$.
- Figure 1 shows $E[X]$ and the asymptotic variance of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ as a function of θ.
- Intuitively, the larger the derivative of $E[X]$ with respect to θ, the easier to identify θ from $\overline{X}$, so the smaller the asymptotic variance.

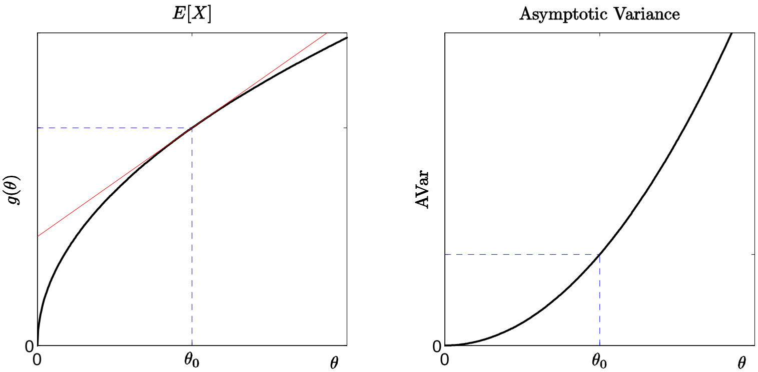

Figure: $E[X]$ and Asymptotic Variance as a Function of θ

An Example

- Suppose the moment conditions are

$$E\begin{bmatrix} X-\mu \\ (X-\mu)^2-\sigma^2 \end{bmatrix}=0.$$

- Then the sample analog is

$$\frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} X_i - n\mu \\ \sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2 \end{pmatrix}=0,$$

- so the solution is

$$\begin{aligned} \widehat{\mu} &= \overline{X} \\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\overline{X^2}-\overline{X}^2. \end{aligned}$$

continue...

- Consistency: $\widehat{\mu} = \overline{X} \xrightarrow{p} \mu$, $\widehat{\sigma}^2 = \overline{X^2} - \overline{X}^2 \xrightarrow{p} \left(\mu^2 + \sigma^2\right) - \mu^2 = \sigma^2$.
- Asymptotic Normality: $\mathbf{M} = E\left[\begin{pmatrix} -1 & 0 \\ -2(X-\mu) & -1 \end{pmatrix}\right] = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$,

$$\begin{aligned} \Omega &= E\left[\begin{pmatrix} (X-\mu)^2 & (X-\mu)^3 - \sigma^2(X-\mu) \\ (X-\mu)^3 - \sigma^2(X-\mu) & (X-\mu)^4 - 2\sigma^2(X-\mu)^2 + \sigma^4 \end{pmatrix}\right] \\ &= \begin{pmatrix} \sigma^2 & E\left[(X-\mu)^3\right] \\ E\left[(X-\mu)^3\right] & E\left[(X-\mu)^4\right] - \sigma^4 \end{pmatrix}, \end{aligned}$$

so

$$\sqrt{n}\begin{pmatrix} \widehat{\mu} - \mu \\ \widehat{\sigma}^2 - \sigma^2 \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Omega).$$

- If $X \sim N\left(\mu, \sigma^2\right)$, then what is Ω?

Another Example: Empirical Distribution Function

- Suppose we want to estimate $\theta = F(x)$ for a *fixed* x, where $F(\cdot)$ is the cdf of a random variable X.
- An intuitive estimator is the ratio of samples below x, $n^{-1}\sum_{i=1}^{n} 1(X_i \le x)$, which is called the **empirical distribution function** (EDF), while it is a MoM estimator.
- Why? note that the moment condition for this problem is

$$E\left[1(X \le x) - F(x)\right] = 0.$$

- Its sample analog is

$$\frac{1}{n}\sum_{i=1}^{n}\left(1(X_i \le x) - F(x)\right) = 0,$$

  so

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x).$$

- By the WLLN, it is consistent. By the CLT,

$$\sqrt{n}\left(\widehat{F}(x) - F(x)\right) \xrightarrow{d} N\left(0, F(x)\left(1 - F(x)\right)\right).\text{(why?)}$$

- An interesting phenomenon is that the asymptotic variance reaches its maximum at the median of the distribution of X.

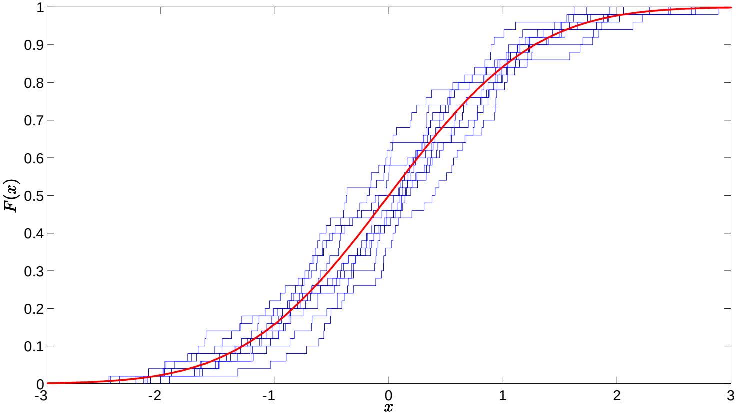

Figure: Empirical Distribution Functions: 10 samples from $N(0,1)$ with sample size $n=50$