# ST5215 Advanced Statistical Theory, Lecture 18

HUANG Dongming

National University of Singapore

20 Oct 2020

# Overview

Last time

- Weak Law of Large Number (WLLN)
- Weak Convergence of Measures
- Central Limit Theorem

Today

- Consistency
- Asymptotic bias, variance, mean squared error

# Needs for the Asymptotic Approach

- In many applications of statistics, the distribution of given statistic $T_n(X)$ is needed, but the exact distributions of $T_n(X)$ is not available or too complicated to deal with
- The limiting distribution is used as an approximation to the distribution of $T_n(X)$ in the situation with a large but actually finite $n$
  - by using CLT, SLLN, WLLN, $\delta$-method, etc.
  - We treat a sample $X = (X_1, \ldots, X_n)$ as a member of a sequence corresponding to $n = 1, 2, \ldots$
  - Similarly, a statistic $T(X)$, often denoted by $T_n$ to emphasize its dependence on the sample size $n$, is viewed as a member of a sequence $T_1, T_2, \ldots$
- In addition, the asymptotic approach requires less stringent mathematical assumptions than does the exact approach

# Consistency

Intuitively, a good estimator shall be close to the *estimand* (the target parameter) when $n$ is large in some sense

## Definition (Consistency of point estimators)

Let $X = (X_1, \ldots, X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be an estimator of $\theta$ (a parameter related to $P$) for every $n$.

1. $T_n(X)$ is called *consistent* for $\theta$ if and only if $T_n(X) \xrightarrow{P} \theta$ w.r.t. any $P \in \mathcal{P}$.
2. Let $\{a_n\}$ be a sequence of positive constants diverging to $\infty$. $T_n(X)$ is called *$a_n$-consistent* for $\theta$ if and only if $a_n\{T_n(X) - \theta\} = O_P(1)$ w.r.t. any $P \in \mathcal{P}$.
3. $T_n(X)$ is called *strongly consistent* for $\theta$ if and only if $T_n(X) \xrightarrow{a.s.} \theta$ w.r.t. any $P \in \mathcal{P}$
4. $T_n(X)$ is called *$L_r$-consistent* for $\theta$ if and only if $T_n(X) \xrightarrow{L^r} \theta$ w.r.t. any $P \in \mathcal{P}$ for some fixed $r > 0$.

# Remarks

- Consistency is a concept relating to a sequence of estimators, $\{T_n : n = n_0, n_0 + 1, \ldots\}$, but we usually just say "consistency of $T_n$" for simplicity
- Each of the four types of consistency describes the convergence of $T_n$ to $\vartheta$ in some sense, as $n \to \infty$
- Consistency in (1) is the weakest one among the four
  - it is implied by any of the other three types of consistency
  - but it is the most common one in statistics
  - also the most basic requirement for an estimator
- $L_2$-consistency is also called consistency in mse, which is the most useful type of $L_r$-consistency.

# Example

Assume the population mean $\mu$ is finite

- The sample mean $\overline{X}_n$ is strongly consistent for $\mu$ by SLLN
- Consider estimators of the form $T_n = \sum_{i=1}^n c_{ni} X_i$, where $\{c_{ni}\}$ is a double array of constants
  - If $c_{ni} = c_i/n$ satisfying that $n^{-1}\sum_{i=1}^n c_i \to 1$ and $\sup_i |c_i| < \infty$, then $T_n$ is strongly consistent
  - If $P$ has a finite variance, then $T_n$ is consistent in mse iff $\sum_{i=1}^n c_{ni} \to 1$ and $\sum_{i=1}^n c_{ni}^2 \to 0$

# Methods of proving consistency

- Combinations of *the LLN, the CLT, Slutsky's theorem, and the continuous mapping theorem* are typically applied
- In particular, if $T_n$ is (strongly) consistent for $\theta$ and $g$ is continuous at $\theta$, then $g(T_n)$ is (strongly) consistent for $g(\theta)$

Example: $\overline{X}_n^2$ is $\sqrt{n}$-consistent for $\mu^2$ under the assumption that $P$ has a finite variance:

- $\sqrt{n}(\overline{X}_n^2 - \mu^2) = \sqrt{n}(\overline{X}_n - \mu)(\overline{X}_n + \mu)$
- $\overline{X}_n$ is $\sqrt{n}$-consistent for $\mu$ by CLT
- $\overline{X}_n + \mu = O_P(1)$
- If $Y_n = O_P(b_n)$ and $Z_n = O_P(c_n)$, then $Y_n Z_n = O_P(b_n c_n)$

Exercise: Prove the $\sqrt{n}$-consistency of $\overline{X}_n^2$ using CLT and $\delta$-method.

# Example: Estimating the Right End of the Support

- Let $X_1, \ldots, X_n$ be i.i.d. from an unknown $P$ with a continuous c.d.f. $F$.
- Suppose that there is an unknown $\theta \in \mathcal{R}$ s.t. $F(\theta) = 1$ and $F(x) < 1$ for any $x < \theta$.

Consider the largest order statistic $X_{(n)}$ as an estimator of $\theta$

- For any $\epsilon > 0, F(\theta - \epsilon) < 1$
- Note that

$$P\left(\left|X_{(n)} - \theta\right| \geq \epsilon\right) = P\left(X_{(n)} \leq \theta - \epsilon\right) = [F(\theta - \epsilon)]^n \qquad (1)$$

- Therefore, $\sum_{n=1}^{\infty} P\left(\left|X_{(n)} - \theta\right| \geq \epsilon\right) < \infty$, for any $\epsilon > 0$
- By the first Borel-Cantelli lemma, $X_{(n)} \stackrel{a.s.}{\to} \theta$, i.e., $X_{(n)}$ is strongly consistent for $\theta$

# Example (Cont.)

- Note that the CDF of $X_{(n)}$ is $F_n(x) = [F(x)]^n$, which is also continuous
- Suppose $F$ is left-differentiable and $F'(\theta-) \neq 0$, we can show that

$$P(nF'(\theta-)(\theta - X_{(n)}) \geq t) \to \exp(-t), \qquad \forall t > 0,$$

  which means that $X_{(n)}$ is $n$-consistent
  (left for exercise; use the fact that $F_n(X_{(n)}) \sim \text{Unif}(0,1)$ and use Taylor expansion of $F_n(x)$ around $\theta$),
  - More generally, if $F^{(i)}(\theta-) = 0$ for $i < m$ and $F^{(m)}(\theta-) \neq 0$, then $X_{(n)}$ is $n^{\frac{1}{m}}$-consistent.

# Example: Importance of consistent estimators

Suppose that an estimator $T_n$ of $\theta$ satisfies

$$c_n [T_n(X) - \theta] \xrightarrow{\mathcal{D}} \sigma Y,$$

where $Y$ is a random variable with a known distribution, $\sigma > 0$ is an unknown parameter, and $\{c_n\}$ is a sequence of constants

- Ignorance about $\sigma$ makes the asymptotic distribution above useless
- If a consistent estimator $\hat{\sigma}_n$ of $\sigma$ can be found, then, by Slutsky's theorem,

$$c_n [T_n(X) - \theta] / \hat{\sigma}_n \xrightarrow{\mathcal{D}} Y$$

- We may approximate the distribution of $c_n [T_n(X) - \theta] / \hat{\sigma}_n$ by the known distribution of $Y$

# Remarks on Consistency

- There can be many consistent estimators
- Consistency is an essential requirement in the sense that any inconsistent estimators should not be used, but a consistent estimator is not necessarily good
- Consistency should be used together with one or a few more criteria
    1. Asymptotic unbiasedness
    2. Asymptotic efficiency

# Approximate Unbiasedness

- Unbiasedness is a good property, but in many cases it is impossible to have an unbiased estimator
- A slight bias might reduce variability
- Nevertheless, asymptotically, the bias shall be small

## Definition (Approximate unbiasedness)

An estimator $T_n(X)$ for $\theta$ is called *approximately unbiased* if $b_{T_n}(\theta) \equiv E_\theta T_n(X) - \theta \to 0$ as $n \to \infty$.

- If $T_n$ is a consistent estimator, and $\{T_n\}$ is uniformly integrable, then $T_n$ is approximate unbiased (Exercise 1 in Tutorial 17)
- Note that there are many estimators whose expectations are not well defined
  - Consider i.i.d. $X_1, \ldots, X_n$ from a normal distribution $N(\mu, 1)$, $\mu \neq 0$
  - Let $\vartheta = 1/\mu$ be the parameter of interest
  - Then $T_n = 1/\bar{X}$ is consistent but does not have a finite mean

# Asymptotic Unbiasedness

## Definition (Asymptotic Expectation)

Let $\xi, \xi_1, \xi_2, \ldots$ be random variables and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$.
If $a_n \xi_n \xrightarrow{\mathcal{D}} \xi$ and $E|\xi| < \infty$, then $E\xi / a_n$ is called an *asymptotic expectation* of $\xi_n$

## Definition (Asymptotic Bias)

Suppose $\nu$ is a parameter related to $P$. Suppose $T_n$ is a point estimator of $\nu$ for every $n$.

- The asymptotic expectation of $T_n - \nu$, if exists, is called an *asymptotic bias* of $T_n$, denoted by $\tilde{b}_{T_n}(P)$ (or by $\tilde{b}_{T_n}(\theta)$ if $P$ is in a parametric family indexed by $\theta$).
- If $\lim_{n\to\infty} \tilde{b}_{T_n}(P) = 0$ for any $P \in \mathcal{P}$, then $T_n$ is said to be *asymptotically unbiased*.

# Remarks

- Like the consistency, the asymptotic bias is a concept relating to sequences $\{T_n\}$ and $\left\{\tilde{b}_{T_n}(P)\right\}$
- When both the exact bias $b_{T_n}(P)$ and the asymptotic bias $\tilde{b}_{T_n}(P)$ exist, they are not necessarily the same
- If $T_n$ is a consistent estimator of $\vartheta$, then $T_n = \vartheta + o_p(1)$, and thus $T_n$ is asymptotically unbiased
  - $T_n$ may not be approximately unbiased
  - $g(T_n)$ is asymptotically unbiased for $g(\vartheta)$ for any continuous function $g$
  - In the example of estimating $1/\mu$ by $T_n = 1/\bar{X}$, $T_n \overset{a.s.}{\to} 1/\mu$ by the SLLN and the continuous mapping. Hence $T_n$ is asymptotically unbiased, although $ET_n$ is not defined.

# High Order Bias

- Sometimes we are interested in finding a more precise order of the asymptotic bias for asymptotic unbiased estimators
- Suppose $a_n \to \infty$ and $\{\eta_n\}$ is a sequence of random variables such that

$$a_n \eta_n \xrightarrow{\mathcal{D}} Y, \text{ where } EY = 0$$

  and

$$a_n^2 (T_n - \vartheta - \eta_n) \xrightarrow{\mathcal{D}} W,$$

  then we may define $a_n^{-2}$ to be *the order of* $\tilde{b}_{T_n}(P)$ or define $EW/a_n^2$ to be *the* $a_n^{-2}$ *order asymptotic bias of* $T_n$

# Example

Consider i.i.d. $X_1, \ldots, X_n$ with a finite mean $\mu \neq 0$ and finite variance $\sigma^2$. Suppose $\vartheta = 1/\mu$ is the parameter of interest.

- Let $T_n = 1/\bar{X}$. It is asymptotic unbiased.
- Note that $\frac{\mu - \bar{X}}{\mu^2} \xrightarrow{a.s.} 0$ and

$$n\left(T_n - \vartheta - \frac{\mu - \bar{X}}{\mu^2}\right) = n\frac{(\mu - \bar{X})^2}{\bar{X}\mu^2}. \tag{2}$$

  By the CLT and Slutsky's theorem,

$$n\frac{(\mu - \bar{X})^2}{\bar{X}\mu^2} \xrightarrow{\mathcal{D}} \mu^{-3}\sigma^2\chi_1^2,$$

  where the mean of the RHS is $\sigma^2/\mu^3$.
- Therefore, $\sigma^2/(n\mu^3)$ is the $1/n$ order asymptotic bias of $T_n$.

# Asymptotic Mean Squared Error (amse)

Like the bias, the variance and MSE of an estimator is not well defined if its second moment does not exist

## Definition

Let $T_n$ be an estimator of $\vartheta$ for every $n$ and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$ such that

$$a_n (T_n - \vartheta) \xrightarrow{\mathcal{D}} Y$$

with $0 < EY^2 < \infty$.
The *asymptotic mean squared error* of $T_n$, denoted by amse $_{T_n}(P)$ (or $\text{amse}_{T_n}(\theta)$ if $P$ is in a parametric family indexed by $\theta$) is defined to be the asymptotic expectation of $(T_n - \vartheta)^2$. In other words,

$$\text{amse}_{T_n}(P) = EY^2/a_n^2.$$

The asymptotic variance of $T_n$ is defined to be $\sigma^2_{T_n}(P) = \text{Var}(Y)/a_n^2$

# Remarks

- In the definition, the amse and asymptotic variance are the same if and only if $EY = 0$
- In the definition, one can show that

$$EY^2 \leq \lim_{n\to\infty} E\left[a_n^2 (T_n - \vartheta)^2\right]$$

  - Proof is left for exercise: use Skorohod's theorem and Fatou's lemma
  - The equality holds if and only if $\left\{a_n^2 (T_n - \vartheta)^2\right\}$ is uniformly integrable (use Exercise 1 in Tutorial 17).
  - In other words, the amse is no greater than the exact mse and they are equal under a certain condition.

# Asymptotic Relative Efficiency

Let $T_n$ and $T_n'$ be two estimators of $\vartheta$

- The *asymptotic relative efficiency of $T_n'$ w.t.r. $T_n$* is defined to be

$$e_{T_n', T_n}(P) = \operatorname{amse}_{T_n}(P) / \operatorname{amse}_{T_n'}(P)$$

- $T_n$ is said to be *asymptotically more efficient than $T_n'$* if and only if

$$\limsup_n e_{T_n', T_n}(P) \leq 1$$

  for any $P$ and $< 1$ for some $P$

- Historically, the "efficiency" of an estimator $T$ of $\theta$ refers to $1/\left[I(\theta)\mathrm{MSE}_T(\theta)\right]$, where $I(\theta)$ is the Fisher information of $\theta$. So the definition above should be understood as

$$e_{T_n', T_n}(P) = \frac{\text{efficiency of } T_n'}{\text{efficiency of } T_n}$$

# A corollary of $\delta$-method

## Theorem

*Let $U_n$ be a statistic satisfying $a_n (U_n - \theta) \xrightarrow{\mathcal{D}} Y$ for a random variable $Y$ with $0 < EY^2 < \infty$ and a sequence of positive numbers $\{a_n\}$ satisfying $a_n \to \infty$*

*Let $g$ be a function on $\mathcal{R}$ that is differentiable at $\theta \in \mathcal{R}$ and $T_n = g(U_n)$ be an estimator of $\vartheta = g(\theta)$.*

*Then, the amse of $T_n$ is $E\left\{[g'(\theta)Y]^2\right\} / a_n^2$ ;*
*the asymptotic variance of $T_n$ is $\left[g'(\theta)^2 \operatorname{Var}(Y)\right] / a_n^2$*

See Theorem 2.6 in the textbook for the multivariate version.

# Example

Let $X_1, \ldots, X_n$ be i.i.d. from a Poisson distribution Poi$(\theta)$ with $\theta > 0$. Consider the estimation of $\tau = P\left(X_i = 0\right) = e^{-\theta}$.

Let $T_{1n} = \frac{1}{n} \sum_{j=1}^{n} I_{\{X_j = 0\}}$

- $T_{1n}$ is unbiased and has $\mathrm{mse}_{T_{1n}}(\theta) = e^{-\theta}\left(1 - e^{-\theta}\right)/n$
- By CLT, $\sqrt{n}\left(T_{1n} - \tau\right) \xrightarrow{\mathcal{D}} N\left(0, e^{-\theta}\left(1 - e^{-\theta}\right)\right)$
- So $\mathrm{amse}_{T_{1n}}(\theta) = \mathrm{mse}_{T_{1n}}(\theta)$

# Example (Cont.)

Let $X_1, \ldots, X_n$ be i.i.d. from a Poisson distribution Poi$(\theta)$ with $\theta > 0$. Consider the estimation of $\tau = P(X_i = 0) = e^{-\theta}$.

Next, consider $T_{2n} = e^{-\bar{X}}$

- By CLT, $\sqrt{n}\left(\bar{X} - \theta\right) \xrightarrow{\mathcal{D}} N(0, \theta)$
- By $\delta$-method, we have $\sqrt{n}\,(T_{2n} - \tau) \xrightarrow{\mathcal{D}} N\left(0, e^{-2\theta}\theta\right)$
- So $T_{2n}$ is asymptotic unbiased and $\text{amse}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$
- Note that $ET_{2n} = e^{n\theta\left(e^{-1/n}-1\right)}$ and $nb_{T_{2n}}(\theta) \rightarrow \theta e^{-\theta}/2$. The exact bias of $T_{2n}$ is not $o(1/n)$

# Example (Cont.)

$\text{amse}_{T_{1n}}(\theta) = e^{-\theta}\left(1 - e^{-\theta}\right)/n$
$\text{amse}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$

The asymptotic relative efficiency of $T_{1n}$ w.r.t. $T_{2n}$ is

$$e_{T_{1n},T_{2n}}(\theta) = \theta/\left(e^{\theta} - 1\right) < 1, \qquad \forall\theta > 0$$

This shows that $T_{2n}$ is asymptotically more efficient than $T_{1n}$

# Tutorial

1. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with Lebesgue p.d.f.
$$f(x) = \frac{1}{2c} \frac{1}{x^2 \log x} I_{|x|>3},$$
where $c = \int_{x=3}^{\infty} 1/(x^2 \log x)\, \mathrm{d}x$.
Show that $E\,|X_1| = \infty$ but $n^{-1} \sum_{i=1}^{n} X_i \rightarrow_p 0$

2. Suppose that $X_n$ is a random variable having the binomial distribution with size $n$ and probability $\theta \in (0, 1), n = 1, 2, \ldots$.
Define $Y_n = \log\left(X_n/n\right)$ when $X_n \geq 1$ and $Y_n = 1$ when $X_n = 0$.
Show that $\lim_n Y_n = \log\theta$ a.s. and $\sqrt{n}\left(Y_n - \log\theta\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{1-\theta}{\theta}\right)$

3. Let $X_1, X_2, \ldots$ be independent random variables such that $X_j$ has the uniform distribution on $[-j, j], j = 1, 2, \ldots$ Show that
$$\frac{\sum_{j=1}^{n} X_j}{n^{3/2}} \xrightarrow{\mathcal{D}} N(0, 1/3) \tag{3}$$

# Exercise 1

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with Lebesgue p.d.f.

$$f(x) = \frac{1}{2c} \frac{1}{x^2 \log x} I_{|x|>3},$$

where $c = \int_{x=3}^{\infty} 1/(x^2 \log x)\ \mathrm{d}x$.
Show that $E|X_1| = \infty$ but $n^{-1} \sum_{i=1}^{n} X_i \rightarrow_p 0$

**Proof:**

- $E|X_1| = c^{-1} \int_3^\infty \frac{1}{x \log x} dx = \infty$
- For any positive integer $n$, $E\left[X_1 I_{(-n,n)}(X_1)\right] = 0$
- For any $x > 3$, by WLLN and the following

$$\begin{aligned} xP(|X| > x) &= c^{-1} x \int_x^\infty \frac{1}{t^2 \log t}\ \mathrm{d}t \\ &\le \frac{c^{-1} x}{\log(x)} \int_x^\infty \frac{1}{t^2} dt \\ &= \frac{c^{-1} x}{\log(x)} \cdot \frac{1}{x} \rightarrow 0, \text{ as } x \rightarrow \infty \end{aligned}$$

# Exercise 2

Suppose that $X_n$ is a random variable having the binomial distribution with size $n$ and probability $\theta \in (0, 1)$, $n = 1, 2, \ldots$.
Define $Y_n = \log(X_n/n)$ when $X_n \geq 1$ and $Y_n = 1$ when $X_n = 0$.
Show that $\lim_n Y_n = \log\theta$ a.s. and $\sqrt{n}(Y_n - \log\theta) \xrightarrow{\mathcal{D}} N\left(0, \frac{1-\theta}{\theta}\right)$

**Proof:** SLLN does not apply here because the joint distribution of $X_n$'s is unknown

- Let $Z_1, Z_2, \ldots$ be i.i.d. Bern($\theta$). Then the distribution of $X_n$ is the same as that of $\sum_{j=1}^{n} Z_j$
- For any $\epsilon > 0$, by Markov inequality ($f(x) = x^4$),

$$P\left(\left|\frac{X_n}{n} - \theta\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^4} E\left|\frac{X_n}{n} - \theta\right|^4$$
$$= \frac{\theta^4(1-\theta) + (1-\theta)^4\theta}{\epsilon^4 n^3} + \frac{\theta^2(1-\theta)^2(n-1)}{\epsilon^4 n^3}$$

- Hence,

$$\sum_{n=1}^{\infty} P\left(\left|\frac{X_n}{n} - \theta\right| \geq \epsilon\right) < \infty, \quad \forall\epsilon > 0$$

- By the Borel-Cantelli lemma, $\lim_n X_n/n = \theta$ a.s.

- By the continuity of the log function, $\lim_n Y_n = \log\theta$ a.s.
- Since $X_n$ has the same distribution as $\sum_{j=1}^n Z_j$ for each $n$, by the CLT, $\sqrt{n}\,(X_n/n - \theta) \xrightarrow{\mathcal{D}} N(0, \theta(1-\theta))$
- Define $W_n = X_n/n + eI_{\{X_n=0\}}$. Then $Y_n = \log(W_n)$
- $\lim_n X_n/n = \theta$ a.s. implies that $\lim_n \sqrt{n} I_{\{X_n=0\}} = 0$ a.s.
- By Slutsky's theorem,

$$\sqrt{n}\,(W_n - \theta) = \sqrt{n}\left(\frac{X_n}{n} - \theta\right) + e\sqrt{n} I_{\{X_n=0\}}$$

$$\xrightarrow{\mathcal{D}} N(0, \theta(1-\theta))$$

- Then, by the $\delta$-method with $g(t) = \log t$ and $g'(t) = t^{-1}$, $\sqrt{n}\,(\log W_n - \log\theta) \xrightarrow{\mathcal{D}} N\left(0, \frac{1-\theta}{\theta}\right)$, i.e., $\sqrt{n}\,(Y_n - \log\theta) \xrightarrow{\mathcal{D}} N\left(0, \frac{1-\theta}{\theta}\right)$

# Exercise 3

Let $X_1, X_2, \ldots$ be independent random variables such that $X_j$ has the uniform distribution on $[-j, j], j = 1, 2, \ldots$ Show that

$$\frac{\sum_{j=1}^n X_j}{n^{3/2}} \xrightarrow{\mathcal{D}} N(0, 1/3)$$

**Proof:**

- Note that $EX_j = 0$ and $\text{Var}\,(X_j) = \frac{1}{2j}\int_{-j}^{j} x^2 dx = j^2/3$ for all $j$
- Hence

$$\sigma_n^2 = \text{Var}\left(\sum_{j=1}^n X_j\right) = \frac{1}{3}\sum_{j=1}^n j^2 = \frac{n(n+1)(2n+1)}{3 \times 6} \sim \frac{1}{9}n^3$$

- For any $\epsilon > 0$, $n < \epsilon\sigma_n$ for sufficiently large $n$ and $|X_j| \leq j \leq n$, in which case

$$\sum_{j=1}^n E\left(X_j^2 I_{\{|X_j| > \epsilon\sigma_n\}}\right) = 0$$

Thus, Lindeberg's condition holds and by CLT, $\frac{\sum_{j=1}^n X_j}{\sigma_n^2} \xrightarrow{\mathcal{D}} N(0, 1)$