# ST5215 Advanced Statistical Theory, Lecture 15

HUANG Dongming

National University of Singapore

8 Oct 2020

# Overview

Last time

- Fisher's Information
- Cramér-Rao Lower Bound

Today

- Convergence modes
- Stochastic orders

# Convergence modes

In statistics, we often need to assess the quality of an estimator by its asymptotic convergence rate

- A good estimator should become closer to the true quantity as we collect more and more data
- e.g., $\overline{X}$ gets closer to $\mu$ if $n$ increases
- In math language, $\overline{X}$ converges to $\mu$ "in some sense"
- How to define "convergence" properly?

There are at least four popular definitions of "convergence" in probability

1. almost sure convergence (or convergence with probability 1)
2. convergence in probablity
3. convergence in $L^p$
4. convergence in distribution (also called weak convergence)

# Almost sure convergence

## Definition

We say a sequence of random elements $X_1, X_2, \ldots$ converges almost surely to a random element $X$, denoted by $X_n \stackrel{a.s.}{\to} X$ if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1. \tag{1}$$

- Notation: $P(\lim_{n\to\infty} X_n = X)$ is a shorthand of the following

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) \tag{2}$$

- Note that is a type of pointwise convergence, but allow an exceptional set of probability zero
- Note that we assume a common probability space $(\Omega, \mathcal{F}, P)$ for $X$, $X_1, \ldots$

How to show almost sure convergence in practice?

1. Useful equivalence: (Lemma 1.4 in JS)
   $X_n \xrightarrow{\text{a. s.}} X$ if and only if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} \{|X_m - X| > \epsilon\}\right) = 0$$

2. Borel-Cantelli lemma

## Definition (Infinitely often)

- Let $\{A_n\}_{n=1}^{\infty}$ be an infinite sequence of events
- For an outcome $\omega \in \Omega$, we say the events in the sequence $\{A_n\}_{n=1}^{\infty}$ happen "*infinitely often*" if $A_n$ happens for an infinite number of indices $n$.
- $\{A_n \; i.o.\} = \{\omega \in \Omega : \omega \in A_n \text{ for an infinite number of indices } n\}$ is the collection of outcomes that make the events in the sequence $\{A_n\}_{n=1}^{\infty}$ happen infinitely often.

If $\{A_n \; i.o.\}$ happens, then infinitely many of $\{A_n\}_{n=1}^{\infty}$ happen

$$\{A_n \; i.o.\} = \bigcap_{n \geq 1} \bigcup_{j \geq n} A_j \equiv \limsup_{n \to \infty} A_n \tag{3}$$

This also shows that $\{A_n \; i.o.\}$ is measurable

## Lemma (First Borel-Cantelli)

*For a sequence of events* $\{A_n\}_{n=1}^{\infty}$, *if* $\sum_{n=1}^{\infty} P(A_n) < \infty$, *then* $P(A_n \quad i.o.) = 0$.

- Intuition: because $\sum_{n=1}^{\infty} P(A_n) < \infty$, $P(A_n)$ must be very small for large $n$, and we cannot find a sufficiently number of $\omega$ that make infinitely many $A_n$ happen
- By the continuity of measures, $P(A_n \quad i.o.) = \lim_n P(\bigcup_{j \geq n} A_j)$
- By the subadditivity of measures, $P(\bigcup_{j \geq n} A_j) \leq \sum_{j \geq n} P(A_j)$
- But $\sum_{n=1}^{\infty} P(A_n) < \infty$ implies $\sum_{j \geq n} P(A_j) \to 0$ as $n \to \infty$.

## Lemma (Second Borel-Cantelli)

*For a sequence of pairwise independent events* $\{A_n\}_{n=1}^{\infty}$, *if* $\sum_{n=1}^{\infty} P(A_n) = \infty$, *then* $P(A_n \quad i.o.) = 1$.

- This results is useful to show a sequence happens infinitely often
- A weaker version assumes that the events $A_n$'s are independent, whose proof is simple:

$$\begin{aligned}
P(\bigcup_{n\geq 1}\bigcap_{j\geq n} A_j^c) &= \lim_n P(\bigcap_{j\geq n} A_j^c) \qquad = \lim_n P(\bigcap_{n\leq j\leq m} A_j^c) \\
&= \lim_n \lim_m \prod_{n\leq j\leq m} P(A_j^c) \\
&= \lim_n \lim_m \prod_{n\leq j\leq m} [1 - P(A_j)] \\
(\because 1 - t \leq e^{-t}) \quad &\leq \lim_n \lim_m \prod_{n\leq j\leq m} \exp[-P(A_j)] \\
&= \lim_n \lim_m \exp[-\sum_{n\leq j\leq m} P(A_j)] = 0
\end{aligned}$$

## Theorem

*Let $X$ and $X_1, X_2, \ldots$ are defined on a common probability space. For a constant $\epsilon > 0$, define the sequence of events $\{A_n(\epsilon)\}_{n=1}^{\infty}$ to be $A_n(\epsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$. If $\sum_{n=1}^{\infty} P\{A_n(\epsilon)\} < \infty$ for all $\epsilon > 0$, then $X_n \xrightarrow{a.s.} X$.*

- According to the first Borel-Cantelli lemma, $P(A_n(1/k)\, i.o.) = 0$, for any $k \in \mathcal{N}$
- Therefore

$$0 = P\left(\bigcup_{k\geq 1}\bigcap_{n\geq 1}\bigcup_{j\geq n} A_j(1/k)\right) \tag{4}$$

- For any $\omega$ not in the event in the last display, we have that for all $k \in \mathcal{N}$, there exists some $n \in \mathcal{N}$ such that for all $j \geq n$, $|X_j(\omega) - X(\omega)| \leq \frac{1}{k}$; in other words, $X_n(\omega) \to X(\omega)$

# Convergence in $L^p$

- In statistics, we expect the mean squared error (MSE) of a good estimator to become small as $n$ increases
- More generally, we can consider convergence in $L^p$ for $p > 0$
- $L^p$-norm of $X$: $(E|X|^p)^{1/p}$ (for $p \geq 1$)

**Definition**

A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable $X$ in the $L^p$ sense for some $p > 0$ if $E|X|^p < \infty$ and $E|X_n|^p < \infty$, and

$$\lim_{n\to\infty} E|X_n - X|^p = 0. \tag{5}$$

- Denoted by $X_n \xrightarrow{L^p} X$
- This is not a pointwise convergence
- For $L^2$, it is also called convergence in mean square
- By Lyapunov's inequality, if $0 < q < p$, convergence in $L^p$ sense implies convergence in $L^q$ sense

# Convergence in probability

**Definition**

A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable $X$ in probability if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0, \tag{6}$$

denoted by $X_n \xrightarrow{P} X$.

- Convergence in probability is weaker than almost sure convergence
- But it is not that weak:
  If $X_n \xrightarrow{\text{P}} X$, then there is a subsequence $\left\{X_{n_j}, j = 1, 2, \ldots\right\}$ such that $X_{n_j} \xrightarrow{\text{a.s.}} X$ as $j \to \infty$

# Convergence in distribution

- In statistics, we often need to show that the centralized sample mean of i.i.d. sample $\sqrt{n}(\overline{X} - EX_i)$ is approximately distributed as $N(0, \mathrm{Var}(X_i))$ if $n$ is large

## Definition

A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable $X$ in distribution (or in law or weakly), if

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{7}$$

for every $x \in \mathcal{R}$ at which $F$ is continuous, where $F_n$ and $F$ are CDF of $X_n$ and $X$, respectively. Denoted by $X_n \xrightarrow{D} X$ or $F_n \Rightarrow F$

# Exercise

Suppose $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. sample from $N(\mu, \sigma^2)$. Consider $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

- Show that $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} Z$, where $Z \sim N(0, 1)$
- Prove $\bar{X}_n \xrightarrow{*} \mu$ where $*$ could be $P$, $L_2$, or $a.s.$

# Relations between Convergence Modes

We have the following relations between different modes of convergence

$$
\begin{array}{ccccc}
\xrightarrow{L^p} & \underset{p>q\geq 0}{\Longrightarrow} & \xrightarrow{L^q} & & \\
 & & \Downarrow & & \\
\xrightarrow{a.s.} & \Longrightarrow & \xrightarrow{P} & \Longrightarrow & \xrightarrow{D}
\end{array}
$$

Other relations

- If $X_n \stackrel{D}{\to} c$ for a constant $c$, then $X_n \stackrel{P}{\to} c$. In general, convergence in distribution does not imply convergence in probability
- If $X_n \stackrel{\text{P}}{\to} X$, then there is a subsequence $\left\{X_{n_j}, j = 1, 2, \ldots\right\}$ such that $X_{n_j} \stackrel{\text{a.s.}}{\to} X$ as $j \to \infty$
- Suppose that $X_n \stackrel{\text{D}}{\to} X$. Then, for any $r > 0$

$$\lim_{n\to\infty} E|X_n|^r = E|X|^r < \infty$$

if and only if $\{|X_n|^r\}$ is uniformly integrable in the sense that

$$\lim_{t\to\infty} \sup_n E\left(|X_n|^r I_{\{|X_n|>t\}}\right) = 0$$

# Exercise

Find examples to show why the converse of each of the relationship in the diagram on last slide is false.

- Note that $X_n \overset{\text{D}}{\to} X$ is a weak mode, since it does not even require $\{X_n\}$ and $X$ to be defined on the same probability space
- However, we can construct a duplicate of $(X, X_1, \ldots,)$ such that the *a.s.* convergence holds

**Theorem (Skorohod's theorem)**

*If $X_n \overset{D}{\to} X$, then there are random vectors $Y, Y_1, Y_2, \ldots$ defined on a common probability space such that $P_{Y_n} = P_{X_n}, n = 1, 2, \ldots,\ P_Y = P_X$, and $Y_n \overset{a.s.}{\to} Y$*

- This result is useful because $Y_n \overset{\text{a.s.}}{\to} Y$ is a strong statement
- Proof in Theorem 25.6 in *Probability and Measure* by P. Billingsley
- The high-level idea is simple:
  1. Let $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B} \cap \Omega$, and $P$ is the Lebesgue on $\Omega$
  2. The *inverse of a CDF* $F$ is defined as $F^-(\omega) = \inf\{x \in \mathcal{R} : \omega \leq F(x)\}$
  3. Define $Y(\omega) = F_X^-(\omega)$ and $Y_n(\omega) = F_{X_n}^-(\omega)$
  4. We can show $Y_n \overset{\mathcal{D}}{=} X_n$ and $Y \overset{\mathcal{D}}{=} X$
  5. We can show that $Y_n(\omega) \to Y(\omega)$ for almost every $\omega \in \Omega$

# Stochastic order

In calculus, for two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$

- $a_n = O(b_n)$ iff $|a_n| \leq c|b_n|$ for a constant $c$ and all $n$
- $a_n = o(b_n)$ iff $a_n/b_n \to 0$ as $n \to \infty$

For two sequences of random variables, $\{X_n\}$ and $\{Y_n\}$, we have similar notations

- $X_n = O_{a.s.}(Y_n)$ iff $P\{|X_n| = O(|Y_n|)\} = 1$
    - in other words, there is a subset $A \subset \Omega$ such that $P(A) = 1$, and for each $\omega \in A$, there exists a constant $c$ (depending on $\omega$), and for all $n$, $|X_n(\omega)| \leq c|Y_n(\omega)|$
- $X_n = o_{a.s.}(Y_n)$ iff $X_n/Y_n \overset{a.s.}{\to} 0$
- $X_n = O_P(Y_n)$ iff, for any $\epsilon > 0$, there exist a constant $C_\epsilon > 0$ and $n_0 \in \mathcal{N}$ such that

$$\sup_{n \geq n_0} P(\{\omega \in \Omega : |X_n(\omega)| \geq C_\epsilon |Y_n(\omega)|\}) < \epsilon \tag{8}$$

    - If $X_n = O_P(1)$, we say $\{X_n\}$ is bounded in probability
- $X_n = o_P(Y_n)$ iff $X_n/Y_n \overset{P}{\to} 0$

# Some properties

- if $X_n = O_P(Y_n)$ and $Y_n = O_P(Z_n)$, then $X_n = O_P(Z_n)$
- if $X_n = O_P(Z_n)$, then $X_n Y_n = O_P(Y_n Z_n)$
- if $X_n = O_P(Z_n)$ and $Y_n = O_P(Z_n)$, then $X_n + Y_n = O_P(Z_n)$

The above properties also hold for $O_{a.s.}$

- If $X_n \xrightarrow{D} X$ for a random variable, then $X_n = O_P(1)$
- If $E|X_n| = O(a_n)$, then $X_n = O_P(a_n)$; If $E|X_n| = o(a_n)$, then $X_n = o_P(a_n)$:use Markov's inequality $P(|X| > a) \leq E|X|/a$

# Tutorial

Assume the conditions in Cramér-Rao lower bound hold and $\Theta \subset \mathcal{R}$.

1. Suppose $T$ is an estimator of $g(\theta)$ with bias $b(\theta)$ and $b$ is differentiable. Prove
$$\mathrm{Var}(T) \geq \frac{(g'(\theta) + b'(\theta))^2}{I(\theta)} \tag{9}$$
2. Show that for any fixed $\theta$, there exists a random variable $T$ such that $ET = g'(\theta)$ and $\mathrm{Var}(T)$ attains the Cramér-Rao lower bound if and only if
$$T = [\frac{g'(\theta)}{I(\theta)}]^2 \frac{\partial}{\partial \theta} \log f_\theta(X) + g(\theta) \tag{10}$$
3. Show that there exists an unbiased estimator $T(X)$ of $g(\theta)$ such that $\mathrm{Var}(T)$ attains the Cramér-Rao lower bound if and only if
$$f_\theta(X) = \exp[\eta(\theta)T(x) - \xi(\theta)]h(x), \tag{11}$$
where $\xi(\theta)$ and $\eta(\theta)$ are differentiable functions such that $\xi'(\theta) = g(\theta)\eta'(\theta)$ and $I(\theta) = \eta'(\theta)g'(\theta)$

# Exercise 1

Suppose $T$ is an estimator of $g(\theta)$ with bias $b(\theta)$ and $b$ is differentiable. Prove

$$\mathrm{Var}(T) \geq \frac{\left(g'(\theta) + b'(\theta)\right)^2}{I(\theta)} \tag{13}$$

**Solution:**

- By definition of bias, we have $ET(X) = g(\theta) + b(\theta)$ for any $\theta$
- We basically follow the same proof of the C-R lower bound until the second last step, where we replace $\frac{\partial}{\partial\theta} E\left[T\right] = g'(\theta)$ by

$$\frac{\partial}{\partial\theta} E\left[T\right] = g'(\theta) + b'(\theta) \tag{12}$$

# Exercise 2

Show that for any fixed $\theta$, there exists a random variable $T$ such that $ET = g(\theta)$ and $\mathrm{Var}(T)$ attains the Cramér-Rao lower bound if and only if

$$T = [\frac{g'(\theta)}{I(\theta)}]\frac{\partial}{\partial\theta}\log f_\theta(X) + g(\theta), \quad \text{a.s.} \tag{14}$$

**Solution:**

"$\Leftarrow$":

- Under the regularity condition, $E\frac{\partial}{\partial\theta}\log f_\theta(X) = 0$
- $ET = g(\theta)$and $\mathrm{Var}(T) = [\frac{g'(\theta)}{I(\theta)}]^2\mathrm{Var}(\frac{\partial}{\partial\theta}\log f_\theta(X)) = [\frac{g'(\theta)}{I(\theta)}]^2 I(\theta)$, which equals to the C-R lower bound

"$\Rightarrow$":

- Follows the proof of C-R lower bound but with $T(X)$ replaced by $T$
- The covariance inequality becomes an equation $\Leftrightarrow$ $T$ and $\frac{\partial}{\partial\theta}\log f_\theta(X)$ are linearly dependent
- Since $I(\theta) = \mathrm{Var}(\frac{\partial}{\partial\theta}\log f_\theta(X)) > 0$, we conclude that $T = a\frac{\partial}{\partial\theta}\log f_\theta(X) + b$, a.s., for some constants $a$ and $b$
- Solve $a$ and $b$ using $ET = g(\theta)$ and $\mathrm{Var}(T) = g'(\theta)^2/I(\theta)$

# Exercise 3

Show that there exists an unbiased estimator $T(X)$ of $g(\theta)$ such that $\mathrm{Var}(T)$ attains the Cramér-Rao lower bound if and only if

$$f_\theta(X) = \exp[\eta(\theta)T(x) - \xi(\theta)]h(x), \tag{16}$$

where $\xi(\theta)$ and $\eta(\theta)$ are differentiable functions such that $\xi'(\theta) = g(\theta)\eta'(\theta)$ and $I(\theta) = \eta'(\theta)g'(\theta)$

**Proof:** "$\Rightarrow$":

- We use the result in Exercise 2 to conclude that

$$T(x) = [\frac{g'(\theta)}{I(\theta)}]\frac{\partial}{\partial\theta}\log f_\theta(x) + g(\theta) \tag{15}$$

- For any fixed $x$, view the last display as an ordinary differential equation about $\log f_\theta(x)$ a function of $\theta$
- The solution is $\log f_\theta(x) = c(x) + T(x)\int_{\theta_0}^{\theta}\frac{I(\theta)}{g'(\theta)}\,\mathrm{d}\theta - \int_{\theta_0}^{\theta}\frac{I(\theta)}{g'(\theta)}g(\theta)\,\mathrm{d}\theta$, where $\theta_0$ is a fixed point in a neighborhood of $\theta$
- Let $\xi(\theta) = \int_{\theta_0}^{\theta}\frac{I(\theta)}{g'(\theta)}g(\theta)\,\mathrm{d}\theta$; $\eta(\theta) = \int_{\theta_0}^{\theta}\frac{I(\theta)}{g'(\theta)}\,\mathrm{d}\theta$; and $h(x) = \exp[c(x)]$

# Exercise 3 (Cont.)

"$\Leftarrow$":

- From Exercise 1 in Tutorial 9, we have $ET(X) = \frac{\xi'(\theta)}{\eta'(\theta)}$. So $ET(X) = g(\theta)$
- From that exercise, we also have

$$\mathrm{Var}(T(X)) = \frac{\xi''(\theta)}{[\eta'(\theta)]^2} - \frac{\xi'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3} \tag{17}$$

- The C-R lower bound is $\frac{g'(\theta)^2}{I(\theta)} = \frac{g'(\theta)^2}{\eta'(\theta)g'(\theta)} = \frac{1}{\eta'}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(\frac{\xi'}{\eta'}\right)$, which equals to the RHS of the last display