STA 522, Spring 2021
Introduction to Theoretical Statistics II

Lecture 2

Department of Biostatistics
University at Buffalo

08 February, 2021

Review: Convergence of random variables

- A sequence of random variables $X_1, X_2, \ldots$ **converges in probability** to a random variable X (written as $X_n \xrightarrow{P} X$) if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0,$$

  or, equivalently,

$$\lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1.$$

- A sequence of random variables $X_1, X_2, \ldots$ **converges almost surely** to a random variable X (written $X_n \xrightarrow{a.s.} X$) if, for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1, \text{ i.e., } |X_n - X| < \varepsilon \text{ with probability } 1$$

  This is same as saying

$$P(\lim_{n\to\infty} X_n = X) = 1 \text{ i.e., } P(\{s \in \mathcal{S} : \lim_{n\to\infty} X_n(s) = X(s)\}) = 1$$

Example: Let the sample space $\mathcal{S}$ be the closed interval $[0, 1]$ with the uniform probability distribution. Let $X_n(s) = s + s^n$ and $X(s) = s$. Show that $X_n \xrightarrow{a.s.} X$. Does this sequence converge in probability?

a.s. convergence: For every $s \in [0, 1)$,
$n \to \infty \implies s^n \to 0 \implies X_n(s) \to s = X(s)$.

For $s = 1$, $n \to \infty \implies s^n \to 1 \implies X_n(s) \to 2 \neq 1 = X(s)$.

But the convergence occurs on the set $[0, 1)$ and $P([0, 1)) = 1$.

So X_n converges to X almost surely.

convergence in probability: Fix $\varepsilon > 0$. we have

$$\begin{aligned} P(|X_n - X| \geq \varepsilon) &= P(\{s \in [0, 1] : |s^n| \geq \varepsilon\}) \\ &= P(\{s \in [0, 1] : s \geq \varepsilon^{1/n}\}) \\ &= \int_{\varepsilon^{1/n}}^{1} ds = 1 - \varepsilon^{1/n} \to 1 - 1 = 0 \text{ as } n \to \infty \end{aligned}$$

So, yes, X_n does converge to X in probability.

Example: Same $\mathcal{S} = [0,1]$ with the uniform probability distribution as before. Define the sequence $X_1, X_2, \ldots$ as follows:

$$X_1(s) = s + I_{[0,1]}(s) \qquad X_2(s) = s + I_{[0,1/2]}(s)$$

$$X_3(s) = s + I_{[1/2,1]}(s) \qquad X_4(s) = s + I_{[0,1/3]}(s)$$

$$X_5(s) = s + I_{[1/3,2/3]}(s) \qquad X_6(s) = s + I_{[2/3,1]}(s),$$

and so on, and let $X(s) = s$. Show that this sequence converges in probability, but not almost surely. For any $\varepsilon > 0$

$P\left(|X_n - X| \geq \varepsilon\right) = P(\text{interval whose length is going to zero}) \to 0.$

For every s the value $X_n(s)$ alternates between s and $s+1$ infinitely often. For example, if $s = 3/8$, $X_1(s) = 11/8$, $X_2(s) = 11/8$, $X_3(s) = 3/8$, $X_4(s) = 3/8$, $X_5(s) = 11/8$, $X_6(s) = 3/8$ etc. So no point-wise convergence occurs for this sequence. So X_n does not converge almost surely.

Relationship between convergence in probability and convergence almost surely

- convergence almost surely *implies* convergence in probability, but the converse is not true in general
- However, a sequence that converges in probability has a *sub-sequence* that converges almost surely.

Strong Law of Large Numbers (SLLN)

Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon\right) = 1,$$

so that

$$\overline{X}_n \xrightarrow{a.s.} \mu.$$

Remarks

1. "Stronger" analog of WLLN. SLLN $\implies$ WLLN.
2. For both the WLLN and SLLN the assumption of a finite variance is sufficient but not necessary. The only moment condition needed is that $\mathrm{E}\,|X_i| < \infty$.
3. SLLN and WLLN may hold for non-iid random variables under certain regularity conditions. We can also create examples with non-iid random variables where WLLN holds but not SLLN.

Frequentist Definition of Probability

Given an event $A \subseteq \mathcal{S}$, consider an infinite sequence of independent random experiments/trials, and in each trial check whether or not A occurs. Let $f_n(A)$ be the frequency of the event A in the first n trials. Then the frequentist probability of A is defined as $P_n(A) = \lim_{n\to\infty} \frac{f_n(A)}{n}$ ("long-run relative frequency of A").

Justification via SLLN

Let $X_i = I(A \text{ occurs in trial-}i)$, $i = 1, \ldots, n$. Then $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli$(p = P(A))$ with $\mathrm{E}(X_i) = P(A)$ and $\mathrm{Var}(X_i) = P(A)[1 - P(A)] < \infty$. Also, $\sum_{i=1}^{n} X_i =$ frequency of A in first n trials $= f_n(A) \implies \overline{X}_n = \frac{f_n(A)}{n}$.
Hence, by SLLN,

$$P_n(A) = \frac{f_n(A)}{n} = \overline{X}_n \xrightarrow{a.s.} \mathrm{E}(X_1) = P(A)$$

Convergence in Distribution

Definition: A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous. In this case, we write

$$X_n \xrightarrow{d} X.$$

Note

This definition is actually about the cdfs of the random variables converge and not the random variables themselves.

Example: Let $X_1, X_2, \ldots$ be a sequence of continuous random variables with cdf given by

$$F_{X_n}(x) = 1 - \left(1 - \frac{x}{n}\right)^n \text{ for } 0 < x \leq n.$$

Then For $x > 0$, $F_{X_n}(x) = 1 - \left(1 - \frac{x}{n}\right)^n \to 1 - e^{-x} =: F(x)$ as $n \to \infty$. Note that $F(x)$ is the cdf of Exponential(1) distribution, i.e., $X_n \xrightarrow{d}$ Exponential($\lambda = 1$).

Example: Let $X_1, X_2, \ldots$ be a sequence of continuous random variables with cdf given by

$$F_{X_n}(x) = \left(\frac{x}{1+x}\right)^n \text{ for } x > 0.$$

Then $F_{X_n}(x) \to 0$ for all x. But a function equal to 0 everywhere is not a cdf (if F is a cdf then $\lim_{x\to\infty} F(x)$ must be 1), so X_n **does not** converge in distribution.

Example (contd.) Consider $V_n = \frac{X_n}{n}$ in previous example. Does V_n converge in distribution?

$$
\begin{aligned}
F_{V_n}(v) &= F_{X_n}(nv) \\
&= \left(\frac{nv}{1+nv}\right)^n \\
&= \left[\left(1 - \frac{1}{1+nv}\right)^{nv}\right]^{1/v} \to e^{-1/v} \text{ for } v > 0
\end{aligned}
$$

Since $F(v) = e^{-1/v} I(v > 0)$ is a cdf (verify), so V_n converges in distribution to $V \sim F$.

NOTE: F is the cdf of the inverse-Gamma($\alpha = 1$, $\beta = 1$) distribution.

Example: Suppose $X_1, X_2, \ldots$ are iid Uniform$(0,1)$, and let $X_{(n)} = \max_{1 \le i \le n} X_i$. Does $X_{(n)}$ converge in probability? Can we say anything about the convergence of $n(1 - X_{(n)})$?

Heuristically, $X_{(n)}$ will get closer and closer to 1 as $n \to \infty$. To prove this formally, fix $\varepsilon > 0$. We have

$$\begin{aligned} P(|X_{(n)} - 1| \ge \varepsilon) &= P(X_{(n)} \ge 1 + \varepsilon) + P(X_{(n)} \le 1 - \varepsilon) \\ &= 0 + P(X_{(n)} \le 1 - \varepsilon) \\ &= P(X_i \le 1 - \varepsilon, \quad i = 1, \ldots, n) = (1 - \varepsilon)^n \to 0 \end{aligned}$$

as $n \to \infty$. This means $X_{(n)} \xrightarrow{P} 1$.

Let $Y_n = n(1 - X_{(n)})$. Then for $t > 0$

$$\begin{aligned} F_{Y_n}(t) &= P\left(n(1 - X_{(n)}) \le t\right) \\ &= P\left(X_{(n)} \ge 1 - t/n\right) = 1 - (1 - t/n)^n \to 1 - e^{-t} \end{aligned}$$

as $n \to \infty$. This means that $Y_n = n(1 - X_{(n)}) \xrightarrow{d}$ Exponential(1).

Relationship between convergence in probability & convergence in distribution

Theorem 5.5.12

If the sequence of random variables $X_1, X_2, \ldots$ converges in probability to a random variable X, the sequence also converges in distribution to X.

Proof See exercise 5.40 (homework).

Remark

almost sure convergence $\implies$ convergence in probability $\implies$ convergence in distribution. Reverse implications may not hold in general.

Theorem 5.5.13

The sequence of random variables $X_1, X_2, \ldots$ converges in probability to a constant μ if and only if the sequence also converges in distribution to μ. That is, the statement

$$P\left(|X_n - \mu| > \varepsilon\right) \longrightarrow 0 \qquad \text{for every } \varepsilon > 0$$

is equivalent to

$$P(X_n \leq x) \longrightarrow \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x > \mu. \end{cases}$$

Central Limit Theorem

Theorem 5.5.14

Let $X_1, X_2, \dots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $\mathrm{E}\left(X_i\right) = \mu$ and $\mathrm{Var}\left(X_i\right) = \sigma^2 > 0$ (both μ and σ^2 must be finite since the mgf exists).

Define $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. and $Z_n = \dfrac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$.

Let $G_n(x)$ denote the cdf of Z_n. Then for any x,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

In other words,

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim \mathrm{N}(0,1) \quad \text{(or simply } Z_n \xrightarrow{d} \mathrm{N}(0,1)\text{)}.$$

Remarks

- Note that no assumption on the distribution of X_i is being made, only requirement is that they are iid and the mgf exists. Existence of mgf can be relaxed by just assuming $\text{Var}(X_1) = \sigma^2 < \infty$ (next theorem).

- Heuristic idea: normality comes from sums of "small" (finite variance), independent disturbances.

- DOES NOT hold in general if the regularity conditions are not satisfied. Example: $X_1, X_2, \cdots \sim$ iid Cauchy$(0, 1)$. Then $\sum_{i=1}^n X_i \sim$ Cauchy$(0, n)$ (see Example 5.2.10 in CB 2E; discussed last semester) and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \sim$ Cauchy$(0, 1)$.

Proof of Theorem 5.5.14. Using Taylor's series on the mgf. Define $Y_i = (X_i - \mu)/\sigma$, $M_Y(t)$ = common mgf of the Y_is which exists for $|t| < \sigma h$. Then

$$\begin{aligned} M_{\sqrt{n}(\overline{X}_n - \mu)/\sigma}(t) &= M_{\sum_{i=1}^n Y_i/\sqrt{n}}(t) \\ &= \mathrm{E}\left(e^{\frac{(\sum_{i=1}^n Y_i)}{\sqrt{n}} t} \right) \\ &= \left[\mathrm{E}\left(e^{Y_1\left(\frac{t}{\sqrt{n}}\right)} \right) \right] = \left[M_Y\left(\frac{t}{\sqrt{n}} \right) \right]^n \end{aligned}$$

Consider a Taylor series expansion of $M_Y(s)$ around 0.

$$M_Y(s) = M_Y(0) + M_Y^{(1)}(0)\frac{s}{1!} + M_Y^{(2)}(0)\frac{s^2}{2!} + R_n(s)\frac{s^2}{2!}$$

where $M_Y^{(k)}(0) := \frac{d^k}{dt^k} M_Y(t)\Big|_{t=0}$ and $R_n(s) \to 0$ as $n \to \infty$. The expansion is valid for $|s| < \sigma h$.

Note that

$$
\begin{aligned}
M_Y(0) &= 1 \\
M_Y^{(1)}(0) &= 0 \\
M_Y^{(2)}(0) &= 1
\end{aligned}
$$

So that

$$\left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n = \left[1 + \left(1 + R_n\left(\frac{t}{\sqrt{n}}\right)\right)\frac{t^2}{2n}\right]^n \to e^{\frac{t^2}{2}} = M_Z(t)$$

where $M_Z(t)$ denotes the mgf of N$(0,1)$ distribution. This completes the proof since a mgf uniquely identifies a cdf.

Theorem 5.5.15 (Stronger version of Theorem 5.5.14)

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathrm{E}(X_i) = \mu$ and $0 < \mathrm{Var}\,(X_i) = \sigma^2 < \infty$.

Define $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. and $Z_n = \dfrac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$.

Let $G_n(x)$ denote the cdf of Z_n. Then for any x,

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy.$$

In other words,

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1).$$

Proof Exactly similar to Theorem 5.5.15, but uses characteristic function ($\phi(t) := E(e^{it})$) instead of mgf. Omitted.

Slutsky's Theorem and Applications

Theorem 5.5.17 (Slutsky's Theorem)

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$, where a is a constant, then

- $X_n Y_n \xrightarrow{d} aX$; and
- $X_n + Y_n \xrightarrow{d} X + a$.

Proof: Omitted.

Normal Approximation with estimated variance

Suppose that $\dfrac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0, 1)$, with σ unknown. Define $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$. Then $\dfrac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} \mathrm{N}(0, 1)$.

Proof. Relies on two facts: (a) $S_n^2 \xrightarrow{P} \sigma^2$ as long as $\mathrm{Var}(S_n^2) \to 0$ (Lecture 1); and

(b) $\dfrac{\sigma}{S_n} \xrightarrow{P} 1$ (see Exercise 5.32)

Principles of Data Reduction

Data reduction involves the use of statistics to summarize information (data) on a parameter θ. We study methods that retain important information about θ, and/or discard information that is irrelevant to θ.

- sufficiency principle, in which no information about θ is discarded while achieving some summarization of the data (section 6.2);
- likelihood methods, in which we study functions that contain all information about θ available from a sample (section 6.3); and
- the equivariance principle, another method that preserves important features of the data (section 6.4).

Notation: Bold face and/or underline $\underline{X}$ for the entire sample, i.e., $\underline{X} = (X_1, \ldots, X_n)$, $\underline{x} = (x_1, \ldots, x_n)$, $T(\underline{X}) = T(X_1, \ldots, X_n)$, $T(\underline{x}) = T(x_1, \ldots, x_n)$, $T(\underline{X})$, etc.

Homework

- Read pp. 236 – 240 (convergence concepts).
-