STA 522, Spring 2022
Introduction to Theoretical Statistics II

Lecture 2

Department of Biostatistics
University at Buffalo

AGENDA

- almost sure convergence & SLLN
- convergence in distribution
- central limit theorem
- sufficiency

Review: Convergence in Probability

- 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 $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$ or, equivalently, $\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1$.

Almost Sure Convergence (or Strong Convergence)

Definition (5.5.6): A sequence of random variables $X_1, X_2, \ldots$ **converges almost surely** to a random variable X if, for every $\varepsilon > 0$,

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

To indicate this, we write $X_n \xrightarrow{a.s.} X$.

Notes:

- Contrast this with the definition of convergence in probability.
- Recall that a random variable is a function from a sample space $\mathcal{S}$ into the real numbers: $X_n : \mathcal{S} \longrightarrow \mathbb{R}$. For each $s \in S$, $X_n(s) = r \in \mathbb{R}$.
- Definition 5.5.6 states that $X_n \xrightarrow{a.s.} X$ if the functions $X_n(s) \longrightarrow X(s)$ for all $s \in S$, except perhaps for $s \in N$, where $N \subseteq \mathcal{S}$ and $P(N) = 0$ (point-wise convergence on all but a few "null" points).

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 (or in Law)

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

The definition is actually about the cdfs of the random variables 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 (verify)

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 $\Longrightarrow$ convergence in probability $\Longrightarrow$ 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

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

Central Limit Theorem (CLT)

Theorem 5.5.14

Let $X_1, X_2, \ldots$ 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 $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 > 0$ (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,

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

Remarks

- The CLT is often expressed as $\overline{X}_n \overset{a}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where $\overset{a}{\sim}$ means "approximately distributed as."
- 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 $\mathrm{Var}(X_1) = \sigma^2 < \infty$ (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$(0, 1)$.
- Even if holds, how good the approximation is for a given n must be checked case by case basis.
- The CLT specifies an asymptotic distribution of $\overline{X}_n$; it does NOT say anything about asymptotic distribution of a single random variable X_n.
- Proof using Taylor's expansion and MGF; ommited.

Example: de Moivre–Laplace CLT Let $S_n \sim \text{Binomial}(n, p)$ for some $0 < p < 1$. (In practice this means p is neither too small nor too large.) Then as $n \to \infty$, $\dfrac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0, 1)$, or written alternatively, $S_n \overset{a}{\sim} N(np, np(1-p))$.

To prove this, write $S_n = \sum_{i=1}^n X_i$ where $X_i \sim$ iid Bernoulli(p). $\mathrm{E}(X_i) = p$, $\mathrm{Var}(X_i) = p(1-p)$. Then use CLT on $\overline{X}_n = S_n/n$.

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) $\dfrac{\sigma}{S_n} \xrightarrow{P} 1$ (Exercise 5.32; homework)

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 (will not be covered).

Notation: We will use $\underline{X}$ and $\underline{x}$ to denote the entire sample, i.e., $\underline{X} = (X_1, \ldots, X_n)$, $\underline{x} = (x_1, \ldots, x_n)$, $T(\underline{X}) = T(X_1, \ldots, X_n)$, $T(\underline{x}) = T(x_1, \ldots, x_n)$, etc.

Sufficiency

- A sufficient statistic for a parameter θ is a statistic that in a sense captures all information about θ contained in the sample.
- If $T(\underline{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\underline{X}$ only through the value $T(\underline{X})$. That is, if $\underline{x}$ and $\underline{y}$ are two sample points such that $T(\underline{x}) = T(\underline{y})$, then the inference about θ should be the same whether $\underline{X} = \underline{x}$ or $\underline{X} = \underline{y}$ is observed.

Definition: A statistic $T(\underline{X})$ is a **sufficient statistic** for θ if the conditional distribution of the sample $\underline{X}$ given that $T(\underline{X}) = t$ does not depend on θ.

To use this definition we must show (in the discrete sense) that

$$P_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t) = P(\underline{X} = \underline{x} \mid T(\underline{X}) = t),$$

i.e., no dependence on θ (need discrete assumptions here).

Checking for Sufficiency

Theorem 6.2.2

If $p(\underline{x} \mid \theta)$ is the joint pdf or pmf of $\underline{X}$ and $q(t \mid \theta)$ is the pdf or pmf of $T(\underline{X})$, then $T(\underline{X})$ is a sufficient statistic for θ if, for every $\underline{x}$ in the sample space, the ratio

$$\frac{p(\underline{x} \mid \theta)}{q(T(\underline{x}) \mid \theta)}$$

is constant as a function of θ.

Proof: Let $\underline{X}$ be discrete. We'll show that $P_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t)$ does not depend on θ.

$$\begin{aligned} P_\theta(\underline{X} = \underline{x} \mid T(\underline{X}) = t) &= \frac{P_\theta(\underline{X} = \underline{x} \text{ and } T(\underline{X}) = t)}{P_\theta(T(\underline{X}) = t)} \\ &= \frac{P_\theta(\underline{X} = \underline{x})}{P_\theta(T(\underline{X}) = t)} = \frac{p(\underline{x} \mid \theta)}{q(T(\underline{x}) \mid \theta)}. \end{aligned}$$

The RHS is constant as a function of θ by assumption.

Example Let $X_1, X_2, \ldots, X_n$ be iid Bernoulli random variables with parameter θ, $0 < \theta < 1$. Show that $T(\underline{X}) = \displaystyle\sum_{i=1}^{n} X_i$ is sufficient for θ.

We'll verify that the previous theorem holds. Note on the outset that $T(\underline{X}) = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \theta)$ (verify using mgf). Write $T(\underline{x}) = \sum_{i=1}^{n} x_i = t$. Then

$$\begin{aligned} \frac{p(\underline{x} \mid \theta)}{q(t \mid \theta)} &= \frac{\prod_{i=1}^{n} \theta^{x_i} \, (1-\theta)^{1-x_i}}{\binom{n}{t} \, \theta^t \, (1-\theta)^{n-t}} \\ &= \frac{\theta^t \, (1-\theta)^{n-t}}{\binom{n}{t} \, \theta^t \, (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

and the RHS is free of θ.

Homework

- Convergence: Read p. 235 – 240 Exercises 5.33, 5.34, 5.39b, 5.41.
- Sufficiency: Read p. 271 – 274.