STA 355 Lecture 2

· Convergence of random variables.

- a.s. convergence
- convergence in prob
- L^p convergence
- Convergence in distribution.

$(X_n)_{n\geq 1}$ and X living in the same sample space. (e.g. It makes sense to talk about $|X_n - X|$).

Don't require anything about joint distribution of $(X_n)_{n\geq 1}$ and X.

Convergence is only in probability law.

Def. $X_1, X_2, \cdots X_n \cdots$ a sequence of random variables

Let F_n be cdf of X_n.

We say $X_n \xrightarrow{d} X$ if $\lim_{n\to+\infty} F_n(t) = F(t)$ $(\forall t)$

where F is the cdf of X, F is continuous.

(only comparing marginal distributions of X_n and X).

Equivalent definition. for any bounded continuous function h

$$\lim_{n\to+\infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)].$$

(the cdf definition takes $h(x) = \mathbb{1}\{X \leq x\}$).

$$
\begin{tikzcd}
L^p \ (p>0) \arrow[dr, "(i)"] & & \\
 & p \arrow[r, shift left, "(iii)"] & d \arrow[l, shift left, dashed, "(iv)"] \\
a.s. \arrow[ur, "(ii)"] & &
\end{tikzcd}
$$

" $\longleftarrow$ - - - - " only when X is a constant (deterministic).

(i). $$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p)$$

$$\overset{\text{Markov ineq}}{\leq} \frac{\mathbb{E}|X_n - X|^p}{\varepsilon^p} \to 0$$

(for any fixed $\varepsilon > 0$)

(ii) $\forall \varepsilon > 0$, with probability 1, $\{n : |X_n - X| > \varepsilon\}$ is finite

$M = \max\{n : |X_n - X| > \varepsilon\}$, $\mathbb{P}(M < +\infty) = 1$.

$$\lim_{k \to +\infty} \mathbb{P}(M > k) = 0.$$

$$\{|X_n - X| > \varepsilon\} \subseteq \{M \geq n\}$$

So $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$.

(iii). Recall F_n, F are cdf's of X_n and X.

$$F_n(t) = \mathbb{P}(X_n \leq t)$$
$$\leq \mathbb{P}(X \leq t + \varepsilon) + \mathbb{P}(|X_n - X| \geq \varepsilon)$$

$$\{X_n \le t\} \subseteq \{|X_n - X| \ge \varepsilon\} \cup \{X \le t+\varepsilon\}$$

$$\| \qquad\qquad \rotatebox{90}{$\subseteq$} \qquad\qquad \rotatebox{90}{$\subseteq$}$$

$$\{X_n \le t, |X_n - X| \ge \varepsilon\} \cup \{X_n \le t, |X_n - X| < \varepsilon\}$$

So
$$\limsup_{n\to+\infty} F_n(t) \le F(t+\varepsilon)$$

Repeating the arguments on the other side,

$$F_n(t) \ge F(t-\varepsilon) - \mathbb{P}(|X_n - X| \ge \varepsilon).$$

$$\liminf_{n\to+\infty} F_n(t) \ge F(t-\varepsilon).$$

This holds true for any $\varepsilon > 0$. F is continuous.

So $\lim_{n\to+\infty} F_n(t) = F(t)$

(iv). Fact. If $X_n \xrightarrow{d} c$ (c is deterministic)

then $X_n \xrightarrow{p} c$.

Proof. cdf of a constant c is $\mathbb{1}\{x \ge c\}$.

(ignoring discontinuity issue).

$$F_n(t) = \mathbb{P}(X_n \le t) \longrightarrow \begin{cases} 0 & \text{when } t < c \\ 1 & \text{when } t \ge c. \end{cases}$$

$$\mathbb{P}(|X_n - c| \geq \varepsilon) = \mathbb{P}(X_n \geq c+\varepsilon) + \mathbb{P}(X_n \leq c-\varepsilon) \to 0.$$

Operations and convergence.

Thm. $(X_n)_{n\geq 1}$, X, $(Y_n)_{n\geq 1}$, Y be random variables.

- If $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X+Y$

  $X_n Y_n \xrightarrow{P} XY$.

- If $X_n \xrightarrow{L^p} X$, $Y_n \xrightarrow{L^p} Y$, then $X_n + Y_n \xrightarrow{L^p} X+Y$.

  (not necessarily for product, because $\mathbb{E}[|X_n Y_n|^p]$ may not be finite).

- If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} c$. (equiv. $Y_n \xrightarrow{d} c$).

  "Slutzky theorem".

  then we have $\begin{cases} X_n + Y_n \xrightarrow{d} c+X \\ X_n Y_n \xrightarrow{d} cX. \end{cases}$

  When $Y_n \xrightarrow{d} Y$ (or even in probability) with Y random, this does not hold true.

- For any continuous function g

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

Notation :

- For limit of nonrandom seq. $(b_n > 0)$.

Use $a_n = O(b_n)$ to denote $\frac{|a_n|}{b_n}$ is bounded.

Use $a_n = o(b_n)$ to denote $\lim\limits_{n \to +\infty} \frac{|a_n|}{b_n} = 0$.

eg. $a_n = 2n^2 + n$ $b_n = n^2$ $a_n = O(b_n) = O(n^2)$

eg. $a_n = \frac{1}{n^2+n}$ $a_n = O(\frac{1}{n})$

- Probabilistic versions. $(X_n, Y_n)_{n \geq 0}$, $Y_n > 0$

We say $X_n = O_p(Y_n)$ when

$\forall \varepsilon > 0$, $\exists k$ indp of n

s.t. $\mathbb{P}\left(\frac{|X_n|}{Y_n} > k\right) \leq \varepsilon$ $(\forall n)$.

We say $X_n = o_p(Y_n)$ when

$$X_n / Y_n \xrightarrow{p} 0.$$

- $O_p(1) + o_p(1) = O_p(1)$
- $O_p(1) \cdot o_p(1) = o_p(1)$.

- Law of large numbers.

Thm (WLLN) $X_1, X_2, \cdots X_n \overset{iid}{\sim} \mathbb{P}$. $\mathbb{E}|X_i| < +\infty$.

then $\frac{1}{n}\sum_1^n X_i \xrightarrow{p} \mathbb{E}[X_1]$.

Thm (SLLN). under the same conditions.

$$\frac{1}{n}\sum_1^n X_i \xrightarrow{a.s.} \mathbb{E}[X_1].$$

Proof under stronger assumption:

Assuming $\mathbb{E}[|X_1|^2] < +\infty$. then

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\right|^2\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[|X_i - \mathbb{E}[X_i]|^2\right]$$

$$\leq \frac{1}{n}\mathbb{E}[|X_1|^2] \to 0.$$

So $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{L^2} \mathbb{E}[X]$.

We also know that L^2 convergence implies $\xrightarrow{p}$.

(In general: use some truncation).

- Central limit thm.

Thm $X_1, \cdots, X_n \cdots \overset{iid}{\sim} \mathbb{P}$. $\mathbb{E}[|X_1|^2] < +\infty$.

then $\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1]) \xrightarrow{d} N(0, var(X_1))$.

Multivariate version. $X_i \in \mathbb{R}^d$

$X_1, X_2, \cdots X_n, \cdots \overset{iid}{\sim} \mathbb{P}$ random vectors.

$\mathbb{E}[\|X_1\|_2^2] < +\infty$. then

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \Sigma), \quad \mu := \mathbb{E}[X_1]$$

where $\Sigma \in \mathbb{R}^{d \times d}$ defined as

$$\Sigma = \mathbb{E}\left[(X_1-\mu)(X_1-\mu)^T\right].$$

Covariance matrix of X_1.

---

e.g. $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \mathbb{P}$ $\quad \mu = \mathbb{E}[X_1]$

$\sigma^2 = var(X_1)$.

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2)$$

In practice, σ^2 is unknown in statistics.

Replace by the "estimated variance".

$$S_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

(sometimes, we use normalization $\frac{1}{n-1}$ not changing the result).

Thm. $$\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} \xrightarrow{d} N(0,1).$$

(S_n is computed from data — useful in stats).

Proof: CLT $$\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \xrightarrow{d} N(0,1).$$

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}X_i^2}_{} - \underbrace{(\bar{X}_n)^2}_{}.$$

$\mathbb{E}[X_i^2]<+\infty$, by WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}X_i^2 \xrightarrow{p} \mathbb{E}[X_1^2]$$

$\mathbb{E}[|X|]<+\infty$, by WLLN

$$\bar{X}_n \xrightarrow{p} \mathbb{E}[X_1]$$

and therefore

$$\bar{X}_n^2 \xrightarrow{p} (\mathbb{E}[X_1])^2$$

Taking the difference

$$S_n^2 \xrightarrow{p} var(X_1).$$

Applying Slutzky theorem.

$$\frac{\sqrt{n}(\bar{X}_n-\mu)}{S_n} = \underbrace{\frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma}}_{\xrightarrow{d} N(0,1)} \cdot \underbrace{\frac{\sigma}{S_n}}_{\xrightarrow{p} 1}.$$

"Delta method."

Thm: Suppose that $(Y_n)_{n\geq 1}$ satisfies

$$\frac{\sqrt{n}\,(Y_n-\mu)}{\sigma} \xrightarrow{d} N(0,1).$$

(Y_n does not have to be iid average).

Let g be a differentiable function. $g'(\mu)\neq 0$

then we have

$$\frac{\sqrt{n}\,(g(Y_n)-g(\mu))}{g'(\mu)\,\sigma} \xrightarrow{d} N(0,1).$$

Proof. $g(Y_n)-g(\mu)$

$= g'(\mu)\cdot(Y_n-\mu) + o_p(|Y_n-\mu|)$ (in op notation).

More rigorously, we can get

$$R_n = g(Y_n)-g(\mu) - g'(\mu)\cdot(Y_n-\mu).$$

we know $\frac{|R_n|}{|Y_n-\mu|} \xrightarrow{p} 0$ by calculus.

$\sqrt{n}(Y_n-\mu) \xrightarrow{d} N(0,\sigma^2)$. (by assumption).

$$\sqrt{n}\left(g(Y_n)-g(\mu)\right) = \underbrace{\sqrt{n}\cdot g'(\mu)(Y_n-\mu)}_{\xrightarrow{d} N(0,\,\sigma^2 g'(\mu)^2)} + \sqrt{n}\cdot R_n$$

$$\sqrt{n}\,|R_n| = \underbrace{\frac{|R_n|}{|Y_n-\mu|}}_{\xrightarrow{P} 0} \cdot \underbrace{\sqrt{n}\,|Y_n-\mu|}_{\xrightarrow{d} \text{something}} \xrightarrow{P} 0 \quad \text{by Slutzky}$$

Once again by Slutzky

$$\sqrt{n}\left(g(Y_n)-g(\mu)\right) \xrightarrow{d} N(0,\,\sigma^2 g'(\mu)^2).$$

Remark: different limit when one of assumptions break.

e.g. $\sqrt{n}\,Y_n \xrightarrow{d} N(0,\sigma^2)$. $g(y)=|y|$.

then $\sqrt{n}\,|Y_n| \xrightarrow{d} |Y|$, where $Y\sim N(0,\sigma^2)$.

e.g. If g is twice differentiable, but $g'(\mu)=0$.

$g''(\mu)\neq 0$.

$$g(Y_n)-g(\mu) = \cancel{g'(\mu)\cdot(Y_n-\mu)} + \frac{1}{2}g''(\mu)\cdot(Y_n-\mu)^2 + R_n.$$

this leads to the limit

Thm. $$\frac{2n\left(g(Y_n)-g(\mu)\right)}{g''(\mu)\cdot\sigma^2} \xrightarrow{d} \chi_2(1).$$

"second-order Delta method".

Another extension: multivariate Delta method.

$Y_1, Y_2, \cdots Y_n, \cdots \in \mathbb{R}^d$, random vectors.

Suppose that $\sqrt{n}(Y_n-\mu) \xrightarrow{d} N(0,\Sigma)$.

Let $g: \mathbb{R}^d \to \mathbb{R}$. differentiable, $\nabla g(\mu) \neq 0$.

Then we have

$$\sqrt{n}\left(g(Y_n)-g(\mu)\right) \xrightarrow{d} N\left(0, \nabla g(\mu)^T \Sigma \nabla g(\mu)\right).$$

More generally, g can be vector-valued functions.

there's also multivariate, second-order Delta method.

Now, back to statistics.

"Statistical model": a class of distributions

Goal: learn something about the true distribution using data.

parametric: $\mathcal{P}_{\Theta} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ for $\Theta \subseteq \mathbb{R}^d$

"parametrized by θ"

nonparametric: also "parametrized", but θ can live in an infinite-dimensional space.

e.g. $\{N(\mu,\sigma^2) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ is a parametric model

e.g. $\mathcal{F} := \{$the set of all cdfs$\}$.

for $F \in \mathcal{F}$. we observe $X_1, \dots X_n \overset{iid}{\sim} F$.

nonparametric model

e.g. $\mathcal{P} := \{ p$ is a pdf, s.t. $|p'(x)| \leq 1 \quad \forall x\}$

nonparametric.

e.g. Let $\mathcal{P} := \{$class of distributions s.t. $\mathbb{E}|X|^2 \leq \sigma^2\}$

is a nonparametric model.

But we may want to estimate a "statistical functional".

Estimate: e.g. $\mu = T(\mathbb{P}) := \mathbb{E}_{\mathbb{P}}[X]$.

e.g. $\mu = T(\mathbb{P}) =$ median of $\mathbb{P}$.