18.650: Fundamentals of Statistics Fall 2023

Lecture 3 — September 11, 2023

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

1 Convergence of Random Variables

For a sequence of numbers x_n, there is only one meaning of "$x_n \to x$ as $n \to \infty$". But there are multiple ways that a sequence of random variables X_n can converge to another random variable X. Here we go over two types of convergence.

Definition 1.1: Convergence in probability

We say $X_n \xrightarrow{\mathbb{P}} X$ as $n \to \infty$ (in words, "X_n converges to X in probability") if for every $\epsilon > 0$, it holds

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

Example.

Suppose $X_n \sim \text{Ber}(1/2)$ for all $n = 1, 2, \ldots$. Is it true that $X_n \xrightarrow{\mathbb{P}} X \sim \text{Ber}(1/2)$ for some other random variable X that has the same distribution Ber(1/2)? Let's check this, supposing X is independent of X_n. Note that $|X_n - X|$ is either 0 or 1. So if we take any $\epsilon \in (0, 1)$, the event $\{|X_n - X| > \epsilon\}$ is the same as the event $\{|X_n - X| = 1\}$, and this occurs if $X_n = 0$ and $X = 1$ or if $X_n = 1$ and $X = 0$. Therefore,

$$\begin{aligned}
\mathbb{P}(|X_n - X| > \epsilon) &= \mathbb{P}(\{X_n = 1 \cap X = 0\} \cup \{X_n = 0 \cap X = 1\}) \\
&= \mathbb{P}(X_n = 1, X = 0) + \mathbb{P}(X_n = 0, X = 1) \\
&= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}.
\end{aligned}$$

This does *not* go to zero! So X_n does not converge to X in probability.

Definition 1.2: Convergence in distribution

We say $X_n \rightsquigarrow X$ as $n \to \infty$ (in words, "X_n converges to X in probability") if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \quad \text{as} \quad n \to \infty$$

for all x at which the cdf $x \mapsto \mathbb{P}(X \leq x)$ is continuous.

Example.

> Consider the same set-up as the previous example: $X_n \sim \text{Ber}(1/2)$ for all n. Then indeed, $X_n \rightsquigarrow \text{Ber}(1/2)$. Here we use the convention of indicating the limit X by its distribution $\text{Ber}(1/2)$.

What is the relationship between the two types of convergence? The next theorem shows that convergence in probability is stronger.

Theorem 1.3: Relationship between convergence types

If $X_n \xrightarrow{\mathbb{P}} X$ then $X_n \rightsquigarrow X$.

Note the converse does not hold, as the above two Bernoulli examples demonstrate. CLT uses convergence in distribution. LLN uses convergence in probability.

Lemma 1.4: convergence to a constant

If $X_n \rightsquigarrow c$ for a deterministic constant c, then $X_n \xrightarrow{\mathbb{P}} c$.

Proof.

$$\begin{aligned}\mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n \leq c - \epsilon) + \mathbb{P}(X_n \geq c + \epsilon) \\ &\to \mathbb{P}(X \leq c - \epsilon) + \mathbb{P}(X \geq c + \epsilon) = 0 + 0 = 0,\end{aligned}$$

since $X = c$. $\square$

1.1 Operations which preserve convergence

Theorem 1.5: Convergence of sums and products

If $X_n \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$ then $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$ and $X_n Y_n \xrightarrow{\mathbb{P}} XY$.

If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{\mathbb{P}} c$ then $X_n + Y_n \rightsquigarrow X + c$ and $X_n Y_n \rightsquigarrow Xc$.

The second statement is known as Slutsky's Theorem.

Remark.

In general, $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ does *not* imply $X_n + Y_n \rightsquigarrow X + Y$. In fact, a statement like this does not even make sense, as the next example shows.

Example.

Suppose $X_n \sim \mathcal{N}(0,1)$ for all n so $X_n \rightsquigarrow X$ *for any* X *such that* $X \sim \mathcal{N}(0,1)$. Next, let $Y_n = -X_n$ for all n, so by symmetry of the standard normal, $Y_n \sim \mathcal{N}(0,1)$ as well. Therefore, $Y_n \rightsquigarrow Y$ *for any* $Y \sim \mathcal{N}(0,1)$.

So does $0 = X_n + Y_n$ converge in distribution to $X + Y$? This is true only if $Y = -X$! But it would be equally valid to choose $Y = X$, in which case 0 does not converge to $X + Y = 2X$. The problem is that we have no information about the correlation between the limits X and Y, but we need this information to determine the distribution of $X + Y$.

Theorem 1.6: Continuous Mapping Theorem

If $X_n \xrightarrow{\mathbb{P}} X$ then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$ for continuous functions g. Similarly, if $X_n \rightsquigarrow X$ then $g(X_n) \rightsquigarrow g(X)$ for continuous g.

Theorem 1.7: Delta Method

Suppose $\sqrt{n}(Y_n - \mu)/\sigma \rightsquigarrow Y \sim \mathcal{N}(0,1)$ for a sequence of random variables Y_n. Then for any differentiable g such that $g'(\mu) \neq 0$, we have

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

Remark.

The theorem is typically applied for $Y_n = \bar{X}_n$ (a sample average).

Proof. We Taylor expand g around the point μ: $g(Y_n) - g(\mu) = g'(\mu)(Y_n - \mu) + \dots$, where the dots represent negligible terms. We multiply both sides by $\sqrt{n}/\sigma$ to get

$$\frac{\sqrt{n}}{\sigma}(g(Y_n) - g(\mu)) \approx g'(\mu)\left[\frac{\sqrt{n}}{\sigma}(Y_n - \mu)\right] \rightsquigarrow g'(\mu)Y$$

using that the expression in square brackets converges to $Y \sim \mathcal{N}(0,1)$. But $g'(\mu)Y$ has distribution $\mathcal{N}(0, g'(\mu)^2)$, and we are done. $\square$

2 Slutsky's theorem in statistics: an example

A humble Harvard grad claims that on average, Harvard grads make no more than 120K at graduation. Let's test this hypothesis. Suppose we collect the salaries $X_1, \ldots, X_n$ of $n = 100$ recent grads, and we find that the sample mean is $\bar{X}_n = 121\text{K}$ while the sample standard deviation is $\hat{\sigma} = 0.3\text{K}$.

Assume that our model for this data is that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. We want to know: how likely is it to observe $\bar{X}_n = 121\text{K}$ if $\mu = 120\text{K}$?

By the Central Limit Theorem, $\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$, and we've assumed $\mu = 120\text{K}$. However, we don't know the true value of σ. We only have an estimate for it, namely the sample standard deviation $\hat{\sigma}$. It is tempting to just replace σ by $\hat{\sigma}$ in the CLT, and Slutsky's theorem allows us to do just this:

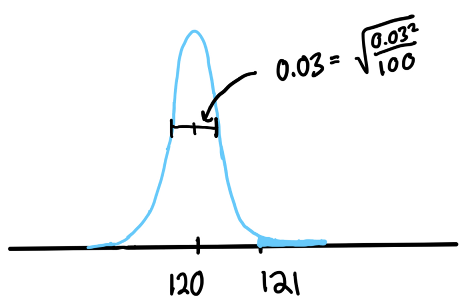
$$\frac{\sqrt{n}}{\hat{\sigma}}(\bar{X}_n - \mu) = \left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] \times \frac{\sigma}{\hat{\sigma}} \rightsquigarrow \mathcal{N}(0,1), \tag{1}$$

because

- $\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0,1)$ by the CLT, and
- $\frac{\sigma}{\hat{\sigma}} \xrightarrow{\mathbb{P}} 1$ by the LLN,

so Slutsky's Theorem (the second part of Theorem 1.5) tells us the product of the two converges in distribution to $\mathcal{N}(0,1)$. From (1) we conclude that

$$\bar{X}_n \approx \mathcal{N}(\mu, \hat{\sigma}^2/n) = \mathcal{N}(120, 0.3^2/100).$$



We see from the figure that $\bar{X}_n = 121\text{K}$ is very unlikely under the distribution $\mathcal{N}(120, 0.03^2)$, so we conclude the Harvard grad's claim that the average income is 120K was an underestimate.

To see why $\sigma/\hat{\sigma}$ converges to 1 in probability, note that

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \approx \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(X_1 - \mu)^2] = \sigma^2$$

by the LLN.