Modern Statistics

Xiangyu Chang

March 31, 2026

Abstract

To be undated.

1 Lecture 8: Convergence of Random Variables

In Lecture 6 we introduced three modes of convergence for sequences of random variables—convergence in probability ($\xrightarrow{P}$), in mean square ($\xrightarrow{L^2}$), and in distribution ($\xrightarrow{d}$)—and saw that the sample mean satisfies $\bar{X}_n \xrightarrow{L^2} \mu$. This lecture deepens that theory. We begin with a classical example to build intuition, then develop two fundamental concentration inequalities (Markov and Chebyshev) that serve as workhorses for convergence proofs. We next establish a hierarchy of implications among the three modes. Finally, we state and prove the central limit theorems of statistics—the Weak Law of Large Numbers, the Central Limit Theorem, Slutsky's Theorem, the Continuous Mapping Theorem, and the Delta Method—culminating in a concrete application.

1.1 Recap: Modes of Convergence

We briefly recall the three definitions from Lecture 6. Let $\{X_n\}$ be a sequence of random variables and X a target random variable.

- **Convergence in probability**: $X_n \xrightarrow{P} X$ if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\mathrm{r}(|X_n - X| > \varepsilon) = 0.$$

- **Convergence in distribution**: $X_n \xrightarrow{d} X$ if

$$\lim_{n\to\infty} F_n(t) = F_X(t)$$

  at every continuity point t of F_X.

- **Convergence in mean square (L^2)**: $X_n \xrightarrow{L^2} X$ if

$$\lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0.$$

A natural question is how these three notions relate to each other. We will establish the full hierarchy in Section 1.4.

1.2 Concentration Inequalities

Before working through a concrete example, we establish two fundamental inequalities that will appear repeatedly in convergence proofs. Both bound the probability of a random variable being far from its mean using only low-order moments.

Theorem 1.1 (Markov's Inequality)**.** *If $X \geqslant 0$ almost surely, then for all $\varepsilon > 0$,*

$$\Pr(X > \varepsilon) \leqslant \frac{\mathbb{E}[X]}{\varepsilon}.$$

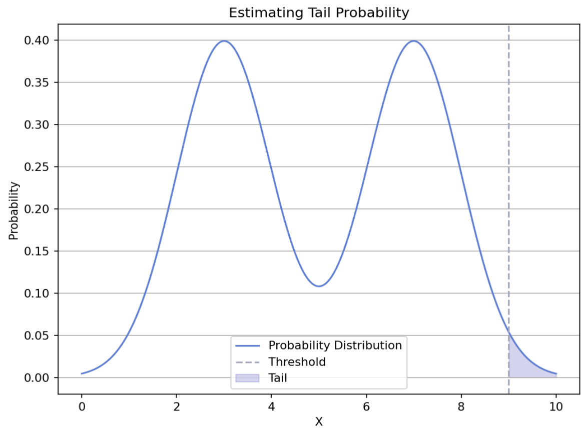

Figure 1: Markov's inequality: the shaded area $\varepsilon \cdot \Pr(X > \varepsilon)$ is at most $\mathbb{E}[X]$.

Proof. Define the indicator $Y = \varepsilon \cdot \mathbf{1}[X \geqslant \varepsilon]$. Since $X \geqslant 0$, we have $Y \leqslant X$ pointwise, so $\mathbb{E}[Y] \leqslant \mathbb{E}[X]$. But $\mathbb{E}[Y] = \varepsilon \cdot \Pr(X \geqslant \varepsilon)$, giving the result. ∎

Theorem 1.2 (Chebyshev's Inequality)**.** *Let X be a random variable with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Then for all $\varepsilon > 0$,*

$$\Pr(|X - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Apply Markov's inequality to the non-negative random variable $(X - \mu)^2$ with threshold ε^2:

$$\Pr(|X - \mu| \geqslant \varepsilon) = \Pr\big((X - \mu)^2 \geqslant \varepsilon^2\big) \leqslant \frac{\mathbb{E}[(X - \mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

∎

Remark 1.3 (Concentration Inequalities)**.** *Chebyshev's inequality is one instance of a broader class of **concentration inequalities**, which bound tail probabilities using moments. For example, applying Markov's inequality to $e^{X-\mu}$ gives the **Chernoff bound**:*

$$\Pr(X - \mu \geqslant \varepsilon) \leqslant \frac{\mathbb{E}[e^{X-\mu}]}{e^{\varepsilon}}.$$

Chernoff bounds are typically much tighter than Chebyshev's inequality when the moment generating function $\mathbb{E}[e^{X-\mu}]$ is finite.

1.3 A Classical Example

The following example is a useful exercise because all three modes of convergence can be verified explicitly, and the proofs use different tools in instructive ways.

Example 1.4 ($X_n \sim N(0, 1/n)$)**.** Let $X_n \sim N(0, 1/n)$. We prove two convergence results.

(i) Convergence in distribution. We claim $X_n \xrightarrow{d} X$ where X is a degenerate random variable satisfying $\mathbb{P}\mathrm{r}(X = 0) = 1$.

Note that $F_X(t) = \mathbf{1}[t \geqslant 0]$, which is discontinuous only at $t = 0$. The CDF of X_n is

$$F_n(t) = \mathbb{P}\mathrm{r}(X_n \leqslant t) = \mathbb{P}\mathrm{r}(\sqrt{n}\, X_n \leqslant \sqrt{n}\, t) = \Phi(\sqrt{n}\, t),$$

where $\sqrt{n}\, X_n \sim N(0, 1)$ and Φ denotes the standard normal CDF. Evaluating the limit at each continuity point of F_X:

$$\lim_{n\to\infty} F_n(t) = \begin{cases} \Phi(+\infty) = 1 = F_X(t) & t > 0, \\ \Phi(-\infty) = 0 = F_X(t) & t < 0. \end{cases}$$

At the discontinuity $t = 0$, $\lim_{n\to\infty} F_n(0) = \Phi(0) = 1/2 \neq F_X(0) = 1$, but we only require convergence at continuity points. Therefore $X_n \xrightarrow{d} X$.

(ii) Convergence in probability. We claim $X_n \xrightarrow{P} 0$.

Since $\mathbb{E}[X_n] = 0$ and $\mathrm{Var}(X_n) = 1/n$, Chebyshev's inequality (Theorem 1.2) gives, for any $\varepsilon > 0$:

$$\mathbb{P}\mathrm{r}(|X_n - 0| \geqslant \varepsilon) = \mathbb{P}\mathrm{r}\big((X_n - \mathbb{E}[X_n])^2 \geqslant \varepsilon^2\big) \leqslant \frac{\mathrm{Var}(X_n)}{\varepsilon^2} = \frac{1}{n\varepsilon^2} \to 0.$$

Therefore $X_n \xrightarrow{P} 0$.

1.4 Hierarchy of Convergence Modes

The three modes of convergence are not equivalent, but they are related by the following implications.

Theorem 1.5 (Hierarchy of Convergence)**.** (i)(i)

1. $X_n \xrightarrow{L^2} X \implies X_n \xrightarrow{P} X$.
2. $X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X$.
3. If $X_n \xrightarrow{d} X$ and $\mathbb{P}\mathrm{r}(X = c) = 1$ for some constant c, then $X_n \xrightarrow{P} c$.

The first implication says that mean-square convergence is the strongest; the third says that convergence in distribution to a *constant* is equivalent to convergence in probability. None of the reverse implications hold in general. The hierarchy is summarized in Figure 2.

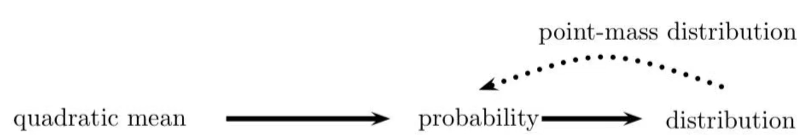

Figure 2: Relationship between modes of convergence. Arrows indicate implications; (iii) is a special case.

Proof. **Proof of (i).** By Markov's inequality applied to $(X_n - X)^2$:

$$\mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon) = \mathbb{P}\mathrm{r}\big((X_n - X)^2 \geqslant \varepsilon^2\big) \leqslant \frac{\mathbb{E}[(X_n - X)^2]}{\varepsilon^2} \xrightarrow{X_n \xrightarrow{L^2} X} 0.$$

Proof of (ii). We show $F_n(x) \to F_X(x)$ at every continuity point x of F_X. Fix $\varepsilon > 0$. For the upper bound:

$$\begin{aligned} F_n(x) &= \mathbb{P}\mathrm{r}(X_n \leqslant x) \\ &= \mathbb{P}\mathrm{r}(X_n \leqslant x,\ X > x + \varepsilon) + \mathbb{P}\mathrm{r}(X_n \leqslant x,\ X \leqslant x + \varepsilon) \\ &\leqslant \mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon) + F_X(x + \varepsilon). \end{aligned}$$

For the lower bound:

$$\begin{aligned} F_X(x - \varepsilon) &= \mathbb{P}\mathrm{r}(X \leqslant x - \varepsilon) \\ &= \mathbb{P}\mathrm{r}(X \leqslant x - \varepsilon,\ X_n \leqslant x) + \mathbb{P}\mathrm{r}(X \leqslant x - \varepsilon,\ X_n > x) \\ &\leqslant F_n(x) + \mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon). \end{aligned}$$

Combining, we get the squeeze:

$$F_X(x - \varepsilon) - \mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon) \leqslant F_n(x) \leqslant F_X(x + \varepsilon) + \mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon).$$

As $n \to \infty$, $\mathbb{P}\mathrm{r}(|X_n - X| \geqslant \varepsilon) \to 0$. Letting $\varepsilon \to 0$ and using continuity of F_X at x, we obtain $F_n(x) \to F_X(x)$.

Proof of (iii). When $\mathbb{P}\mathrm{r}(X = c) = 1$, we have $F_X(t) = \mathbf{1}[t \geqslant c]$. For any $\varepsilon > 0$:

$$\begin{aligned} \mathbb{P}\mathrm{r}(|X_n - c| > \varepsilon) &= \mathbb{P}\mathrm{r}(X_n > c + \varepsilon) + \mathbb{P}\mathrm{r}(X_n < c - \varepsilon) \\ &= \big[1 - F_n(c + \varepsilon)\big] + F_n(c - \varepsilon). \end{aligned}$$

Since $c + \varepsilon$ and $c - \varepsilon$ are continuity points of F_X and $X_n \xrightarrow{d} X$:

$$\lim_{n\to\infty} \big[1 - F_n(c + \varepsilon)\big] + F_n(c - \varepsilon) = \big[1 - F_X(c + \varepsilon)\big] + F_X(c - \varepsilon) = (1 - 1) + 0 = 0.$$ ■

1.5 Main Limit Theorems

With the hierarchy in hand, we now state the principal limit theorems that underpin statistical inference. These theorems justify the use of normal approximations and form the bridge between probability theory and applied statistics.

1.5.1 Weak Law of Large Numbers

Theorem 1.6 (Weak Law of Large Numbers (WLLN)). *Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then*

$$\bar{X}_n \xrightarrow{P} \mu.$$

Proof. From Lecture 5 we know $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. By Chebyshev's inequality, for any $\varepsilon > 0$:

$$\mathbb{P}\mathrm{r}(|\bar{X}_n - \mu| > \varepsilon) \leqslant \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \quad \text{as } n \to \infty.$$

■

1.5.2 Central Limit Theorem

The WLLN tells us that $\bar{X}_n$ is close to μ, but says nothing about how the fluctuations $\bar{X}_n - \mu$ are distributed. The CLT gives a precise answer.

Theorem 1.7 (Central Limit Theorem (CLT))**.** *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance $\sigma^2 \in (0, \infty)$. Then*

$$Z_n \stackrel{\text{def}}{=} \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

The CLT is remarkable: regardless of the shape of the original distribution F, the standardized sample mean always converges to the standard normal. The rate of convergence is $1/\sqrt{n}$, which matches the standard deviation $\sigma/\sqrt{n}$ of $\bar{X}_n$.

1.5.3 Slutsky's Theorem

In applications, the CLT often appears combined with other converging sequences—for instance, when σ is unknown and must be estimated. Slutsky's theorem governs these combinations.

Theorem 1.8 (Slutsky's Theorem)**.** *Let $\{X_n\}$ and $\{Y_n\}$ be random sequences.* (a)(a)

1. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.
2. If $X_n \xrightarrow{L^2} X$ and $Y_n \xrightarrow{L^2} Y$, then $X_n + Y_n \xrightarrow{L^2} X + Y$.
3. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ for a constant c, then $X_n + Y_n \xrightarrow{d} X + c$.
4. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ for a constant c, then $X_n Y_n \xrightarrow{d} cX$.
5. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$.

Parts (c) and (d) are the most useful in practice: they allow a distributional limit to be combined with a term that converges to a constant in probability (which implies convergence in distribution to that constant by Theorem 1.5(iii) reversed).

1.5.4 Continuous Mapping Theorem

Theorem 1.9 (Continuous Mapping Theorem (CMT))**.** *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a continuous function.* (1)(1)

1. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.
2. If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.

Intuitively, continuity of g ensures that if X_n is close to X, then $g(X_n)$ is close to $g(X)$. The CMT is frequently used together with the WLLN to show convergence of transformed sample moments.

1.5.5 Multivariate Central Limit Theorem

The CLT extends naturally to random vectors, a fact that is essential for the analysis of multi-parameter estimators.

Theorem 1.10 (Multivariate CLT)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random vectors in $\mathbb{R}^k$ with mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\Sigma = \mathrm{Var}(X_i)$. Then*

$$\sqrt{n}(\bar{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where $\bar{\mathbf{X}}_n = n^{-1} \sum_{i=1}^n X_i$.

1.5.6 Delta Method

The Delta Method extends the CLT to smooth transformations of the sample mean. It is particularly useful when the parameter of interest is a nonlinear function of the mean.

Theorem 1.11 (Delta Method)**.** *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2, so that $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$ by the CLT. If $g \colon \mathbb{R} \to \mathbb{R}$ is differentiable at μ with $g'(\mu) \neq 0$, then*

$$\frac{\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big)}{\sigma} \xrightarrow{d} N\big(0,\, [g'(\mu)]^2\big).$$

Proof. By the first-order Taylor expansion of g around μ:

$$g(\bar{X}_n) = g(\mu) + g'(\mu)(\bar{X}_n - \mu) + \frac{g''(\xi)}{2}(\bar{X}_n - \mu)^2,$$

where ξ lies between $\bar{X}_n$ and μ. Rearranging:

$$\frac{\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big)}{\sigma} = g'(\mu) \cdot \underbrace{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}_{\xrightarrow{d} N(0,1)} + \underbrace{\frac{g''(\xi)\sqrt{n}}{2\sigma}(\bar{X}_n - \mu)^2}_{R_n}.$$

Since $\bar{X}_n - \mu = O_p(n^{-1/2})$ by the CLT, we have $(\bar{X}_n - \mu)^2 = O_p(n^{-1})$, so $R_n = O_p(n^{-1/2}) \xrightarrow{P} 0$.

By Slutsky's Theorem (parts (c) and (d) applied to the sum):

$$\frac{\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big)}{\sigma} \xrightarrow{d} g'(\mu) \cdot N(0, 1) + 0 = N\big(0,\, [g'(\mu)]^2\big).$$

■

1.6 Application: Studentization

The CLT requires knowledge of the true standard deviation σ, which is typically unknown in practice. The following result shows that replacing σ by the sample standard deviation S_n does not affect the limiting distribution.

Example 1.12 (Student's t-Statistic Convergence)**.** Let $X_1, \ldots, X_n$ be i.i.d. with mean μ, variance $\sigma^2 \in (0, \infty)$, and finite fourth moment. Let

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

We prove that

$$T_n \stackrel{\text{def}}{=} \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1).$$

Step 1: Show $S_n^2 \xrightarrow{P} \sigma^2$. Write

$$S_n^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right].$$

By the WLLN, $\frac{1}{n}\sum X_i^2 \xrightarrow{P} \mathbb{E}[X^2] = \sigma^2 + \mu^2$. By the WLLN and CMT (with $g(x) = x^2$), $\bar{X}_n^2 \xrightarrow{P} \mu^2$. Since $\frac{n}{n-1} \to 1$, Slutsky's Theorem gives

$$S_n^2 \xrightarrow{P} 1 \cdot \left[(\sigma^2 + \mu^2) - \mu^2\right] = \sigma^2.$$

Step 2: Show $S_n \xrightarrow{P} \sigma$. By the CMT with $g(x) = \sqrt{x}$ (continuous on $(0, \infty)$):

$$S_n = \sqrt{S_n^2} \xrightarrow{P} \sqrt{\sigma^2} = \sigma.$$

Step 3: Combine via Slutsky's Theorem. Write

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

By Slutsky's Theorem (part (d)), $T_n \xrightarrow{d} N(0,1) \cdot 1 = N(0,1)$.

This result justifies the use of S_n in place of σ when constructing large-sample confidence intervals in Lecture 8.

References