Lecture Notes 5

Today we will start off by deriving some of the implications between the different modes of convergence. Then we will prove the CLT.

1 Quadratic mean $\implies$ convergence in probability

Suppose that $X_1, \ldots, X_n$ converges in quadratic mean to X, then fix an $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{P}(|X_n - X|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X_n - X)^2}{\epsilon^2} \to 0,$$

showing convergence in probability.

At a high-level the convergence in qm requirement penalizes X_n for having large deviations from X by both how frequent the deviation is but also by the *magnitude of the deviation.* On the other hand convergence in probability only penalizes you for how frequent the deviation is and hence is a weaker notion of convergence.

Counterexample to reverse: Suppose we take $U \sim U[0, 1]$ and define $X_n = \sqrt{n}\mathbb{I}_{[0,1/n]}(U)$, then X_n converges in probability to 0 but does not converge in quadratic mean to 0.

To see this:

$$\mathbb{P}(|X_n| \geq \epsilon) = \mathbb{P}(\sqrt{n}\mathbb{I}_{[0,1/n]}(U) \geq \epsilon) = \mathbb{P}(U \in [0, 1/n]) = \frac{1}{n} \to 0.$$

On the other hand,

$$\mathbb{E}(X_n - X)^2 = \mathbb{E}X_n^2 = n\mathbb{P}(U \in [0, 1/n]) = 1.$$

Observe that most of the time the RV X_n takes the value 0, but when it does not it takes a huge value.

1.1 Convergence in probability $\implies$ convergence in distribution

This one is a little bit involved but perhaps also useful to know. The idea roughly is to trap the CDF of X_n by the CDF of X with an interval whose length converges to 0.

Suppose that $X_n \rightsquigarrow X$. We fix a point x where the CDF $F_X(x)$ is continuous. Choose an arbitrary $\epsilon > 0$. We have that,

$$\begin{aligned} F_{X_n}(x) = \mathbb{P}(X_n \leq x) &= \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X \geq x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon) \\ &= F_X(x + \epsilon) + \mathbb{P}(|X_n - X| \geq \epsilon). \end{aligned}$$

Now,

$$
\begin{aligned}
F_X(x-\epsilon) = \mathbb{P}(X \le x-\epsilon) &= \mathbb{P}(X \le x-\epsilon, X_n \le x) + \mathbb{P}(X \le x-\epsilon, X_n \ge x) \\
&\le F_{X_n}(x) + \mathbb{P}(|X_n - X| \ge \epsilon).
\end{aligned}
$$

Putting these two together we have,

$$
F_X(x-\epsilon) - \mathbb{P}(|X_n - X| \ge \epsilon) \le F_{X_n}(x) \le F_X(x+\epsilon) + \mathbb{P}(|X_n - X| \ge \epsilon).
$$

Intuitively, now as n gets large the two probabilities converge to 0, and since ϵ was chosen arbitrarily we can let $\epsilon \to 0$ and use the continuity of $F_X(x)$ at x to conclude that $F_{X_n}(x) \to F_X(x)$.

Slightly more rigorously, we cannot assume that the limit of $F_{X_n}(x)$ exists so we instead need to use lim infs and lim sups (do not worry about this if you have not seen it before). Formally, we would take the lim sup of the first half to obtain that,

$$
\limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon),
$$

and similarly that,

$$
\liminf_{n\to\infty} F_{X_n}(x) \ge F_X(x-\epsilon),
$$

and conclude that,

$$
F_X(x-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon).
$$

Now since $\epsilon > 0$ was arbitrary, we can take the limit as $\epsilon \to 0$ and use continuity to conclude the desired convergence in distribution.

Counterexample to reverse: This is easy since two random variables having the same distribution does not in any sense mean that they are close. For example, let $X, X_1, X_2, \sim N(0,1)$. They all have the same cdf so $X_n \rightsquigarrow X$. But $P(|X_n - X| > \epsilon)$ does not go to 0.

An important exception: An important exception is that when X is deterministic then convergence in distribution implies convergence in probability. Suppose that $P(X = c) = 1$. Fix $\epsilon > 0$. Then

$$
\begin{aligned}
\mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n > \epsilon + c) + \mathbb{P}(X_n < c - \epsilon) \\
&= F_{X_n}(c-\epsilon) + 1 - F_{X_n}(c+\epsilon) \\
&\to F_X(c-\epsilon) + 1 - F_X(c+\epsilon) = 0.
\end{aligned}
$$

using convergence in distribution and the fact that at both $c+\epsilon$, and $c-\epsilon$, the distribution function F_X is continuous. So $X_n \rightsquigarrow c$ implies that $X_n \xrightarrow{P} c$.

2 Other things that are very useful to know

1. **Continuous mapping theorem.** If a sequence $X_1, \ldots, X_n$ converges in probability to X then for any continuous function h, $h(X_1), \ldots, h(X_n)$ converges in probability to $h(X)$. The same is true for convergence in distribution.

2. A consequence of the continuous mapping theorem. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$ then $X_n + Y_n \xrightarrow{P} X + Y$. Similarly, $X_n Y_n \xrightarrow{P} XY$.

3. **Slutsky's theorem.** If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ we **cannot** conclude that the sum converges. The one exception is known as Slutsky's theorem. It says that if Y_n converges in distribution to a constant c, and X converges in distribution to X: then $X_n + Y_n$ converges in distribution to $X + c$ and $X_n Y_n$ converges in distribution to cX.

4. **Convergence of moments is not implied by convergence in probability.** Convergence in probability is actually quite weak as a form of convergence. We have seen previously that it does not imply quadratic mean convergence. Now we will see that it does not even imply something much simpler.

   If we have X_n converges in probability to some constant c, then it is not the case that $\mathbb{E}[X_n]$ converges to c. Here is an example of this non-convergence. Let X_n be 0 with probability $1 - 1/n$ and n^2 with probability $1/n$. Then X_n converges to 0 in probability, but $\mathbb{E}[X_n] = n \to \infty$.

   This is a manifestation of the same phenomena as we saw in the counterexample to qm convergence. On the events when $|X_n| \geq \epsilon$ it has a huge value and this affects the moments but does not affect the convergence in probability.

3 The Central Limit Theorem (CLT)

We will now state and prove a form of the central limit theorem, which is one of the most famous and important examples of convergence in distribution. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with mean μ and variance σ^2.

Theorem 1 *Assume that the mgf $\mathbb{E}[\exp(tX_i)]$ is finite for t in a neighborhood around zero. Let $\overline{X}_n = n^{-1} \sum_i X_i$. Let*

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

then Z_n converges in distribution to $Z \sim N(0,1)$, that is $Z_n \rightsquigarrow Z$. Hence, as $n \to \infty$,

$$\mathbb{P}(Z_n \leq t) \to \Phi(t)$$

for all t, where

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-s^2/s} ds.$$

The central limit theorem is incredibly general. It does not matter what the distribution of X_i is, the average S_n converges in distribution to a Gaussian (under fairly mild assumptions). The most general version of the CLT does not require any assumption about the mgf. It just requires that the mean and variance are finite. The interpretation of the CLT is that $Z_n \approx N(0,1)$. In other words,

$$\overline{X}_n \approx N(\mu, \sigma^2/n).$$

It can be shown that

$$\sup_t |\mathbb{P}(Z_n \leq t) - \Phi(t)| \leq \frac{33}{4} \frac{\mu_3}{\sigma^3 \sqrt{n}}$$

where $\mu_3 = \mathbb{E}[|X_i - \mu|^3]$.

We should try to understand why the CLT might be useful. Roughly, the CLT allows to make *approximate* probability statements about averages using corresponding statements about standard normals. Here is an example that we will discuss in detail later: confidence intervals.

Suppose for now that we are averaging iid random variables with known variance σ (and unknown mean μ). Typically one would also estimate the variance but this will not change much. We would like to construct a *confidence interval* for the unknown mean. We specify $\alpha \in (0,1)$ and we find a random set C such that

$$\mathbb{P}(\mu \in C) \geq 1 - \alpha.$$

We might take

$$C = [\widehat{\mu} - t, \widehat{\mu} + t]$$

where $\widehat{\mu} = \overline{X}_n$. Then

$$\mathbb{P}(\mu \in [\widehat{\mu} - t, \widehat{\mu} + t]) = \mathbb{P}(|\widehat{\mu} - \mu| \leq t).$$

So we would like to choose t to make this probability equal to $1 - \alpha$. Now

$$\mathbb{P}(|\widehat{\mu} - \mu| \leq t) = \mathbb{P}\left(\frac{\sqrt{n}|\widehat{\mu} - \mu|}{\sigma} \leq \frac{\sqrt{n}t}{\sigma} \right) \approx \mathbb{P}(|Z| \leq t)$$

where $Z \sim N(0,1)$. In the last step we used the CLT. Let Φ denote the cdf of Z and define

$$z_\alpha = \Phi^{-1}(1 - \alpha).$$

Note that

$$P(Z > z_{\alpha/2}) = P(Z < -z_{\alpha/2}) = \frac{\alpha}{2}$$

so that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. So we want to set

$$\frac{\sqrt{n}t}{\sigma} = z_{\alpha/2}$$

that is,

$$t = \frac{\sigma z_{\alpha/2}}{\sqrt{n}}.$$

To summarize: if we define

$$C = \left[\widehat{\mu} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \widehat{\mu} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right],$$

then

$$\mathbb{P}(\mu \in C) \to 1 - \alpha$$

as $n \to \infty$. The convergence is due to the CLT.

3.1 Preliminaries

First we note that

$$\mathbb{E}[Z_n] = 0$$

and

$$\text{Var}[Z_n] = 1.$$

Also note that if $X_1, \ldots, X_n \sim N(0,1)$ then Z_n is exactly $N(0,1)$.

Calculus with mgfs: We need a few simple facts about mgfs that we will quickly prove.

Fact 1: If X and Y are independent with mgfs M_X and M_Y then $Z = X + Y$ has mgf $M_Z(t) = M_X(t)M_Y(t)$.

Proof: We note that,

$$M_Z(t) = \mathbb{E}[\exp(t(X+Y)] = \mathbb{E}[\exp(tX)]\mathbb{E}[\exp(tY)],$$

using independence.

Fact 2: If X has mgf M_X then $Y = a + bX$ has mgf, $M_Y(t) = \exp(at)M_X(bt)$.

Proof: We just use the definition,

$$M_Y(t) = \mathbb{E}[\exp(at + btX)] = \exp(at)\mathbb{E}[\exp(btX)].$$

Fact 3: We will not prove this one (strictly speaking one needs to invoke the dominated convergence theorem) but it should be familiar to you. The derivative of the mgf at 0 gives us moments, i.e.

$$M_X^{(r)}(0) = \mathbb{E}[X^r].$$

Fact 4: The most important result that we also will not prove is that we can show convergence in distribution by showing convergence of the mgfs. Let $X_1, \ldots, X_n$ be a sequence of random variables with mgfs $M_{X_1}, \ldots, M_{X_n}$. Let X be a random variable with mfg M_X. If for all t in an open interval around 0 we have that, $M_{X_n}(t) \to M_X(t)$, then X_n converges in distribution to X.

Fact 5: If $Z \sim N(0,1)$ then $M_Z(t) = e^{t^2/2}$.

3.2 Proof of the CLT

Note that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_i A_i$$

where

$$A_i = \frac{X_i - \mu}{\sigma}.$$

Let $M(t)$ be the mgf for A_i. Since A_i has mean 0 and variance 1, we have that $M(0) = 1$, $M'(0) = 0$ and $M''(0) = 1$. Now

$$M_{Z_n}(t) = \mathbb{E}[e^{tZ_n}] = \mathbb{E}[e^{\frac{t}{\sqrt{n}} \sum_i A_i}] = \prod_i \mathbb{E}[e^{\frac{t}{\sqrt{n}} A_i}] = M(t/\sqrt{n})^n.$$

Expanding M:

$$M(t/\sqrt{n}) \approx M(0) + \frac{t}{\sqrt{n}} M'(0) + \frac{t^2}{2n} M''(0) = 1 + \frac{t^2}{2n}$$

and so

$$M(t/\sqrt{n})^n \approx \left(1 + \frac{t^2}{2n}\right)^n \to e^{t^2/2}$$

which is the mgf of a N(0,1). Here we used the fact that,

$$\lim_{n \to \infty} (1 + x/n)^n \to \exp(x).$$