36-705: Intermediate Statistics **Fall 2017**

Lecture 5: September 8

Lecturer: Siva Balakrishnan

5.1 Convergence of random variables continued

In the last lecture we discussed modes of convergence, proved the WLLN, and looked at some examples. Today we will prove some of the relations between different modes of convergence.

5.1.1 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.

5.1.2 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.

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,

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

Putting these two together we have,

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

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,

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

and similarly that,

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

and conclude that,

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

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 of course is almost trivial since two random variables having the same distribution does not in any sense mean that they are close (see Lecture 4 notes for an example).

An important caveat: An important exception is that when X is deterministic then convergence in distribution implies convergence in probability. Concretely, fix $\epsilon > 0$, consider the case when $X = c$, 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.

5.2 Other things that are very useful to know

5.2.1 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.

This is useful because often we will have a consistent estimator for some parameter, and this theorem allows to construct estimators for some function of the parameter in a straightforward way.

5.2.2 Slutsky's theorem

There are some important consequences of the fact that convergence in distribution is weaker than convergence in probability.

Concretely, for convergence in probability (and stronger forms of convergence) it is the case that, if X_n converges in probability to X and Y_n converges in probability to Y then $X_n + Y_n$ converges in probability to $X + Y$, and the same is true of products, i.e. $X_n Y_n$ converges in probability to XY.

These statements are not true for convergence in distribution, i.e. if X_n converges in distribution to X and Y_n converges in distribution to Y then $X_n + Y_n$ does not necessarily converge in distribution to $X + Y$.

The one exception to this 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.

5.2.3 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 (which only cares about how frequent this violation is).

5.3 The central limit theorem

We will now state and prove a form of the central limit theorem, which is one of the most famous 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. Assume that the mgf $\mathbb{E}[\exp(tX_i)]$ is finite for t in a neighborhood around zero. Let

$$S_n = \frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma},$$

then S_n converges in distribution to $Z \sim N(0, 1)$.

Comments:

1. 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).

2. 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. We will prove this weaker version in lecture.

5.3.1 Use Case

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. At a high-level instead of using a different tail bound for different types of averages (sub-Gaussian, sub-exponential, bounded etc.) we can now just use the Gaussian CDF although our results will only be approximate.

I will introduce a simple use case: we will discuss this idea again later on in more detail when we discuss confidence intervals.

Suppose for now that we are averaging i.i.d. RVs with known variance (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. For some parameter α this is an interval C_α such that,

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

One might guess that we would center such an interval around the sample average $\widehat{\mu}$ but the main difficulty is that we do not know the distribution of $\widehat{\mu}$. We can see that,

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

So we would like to choose t to make this probability at least $1 - \alpha$. One can construct such intervals using tail bounds (see HW3) but we will instead construct an approximate interval using the CLT. Using the CLT we know that distribution of $\widehat{\mu} - \mu$ converges to a normal with mean 0, and variance σ^2/n, i.e.

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

Now, if we let $\Phi(x) = \mathbb{P}(Z \leq x)$ denote the standard normal CDF, then we can see that we need to choose $t = \sigma\Phi^{-1}(\alpha/2)/\sqrt{n} := \sigma z_{\alpha/2}/\sqrt{n}$.

To summarize, the interval:

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

has the property that,

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

where we appealed to the CLT to justify this construction.

5.3.2 Preliminaries

Sanity Checks: Before we prove the theorem there are two very simple sanity checks that one might consider. The random variable S_n has mean 0, and variance:

$$\mathbb{E}[S_n^2] = \frac{n}{\sigma^2}\mathbb{E}(\widehat{\mu} - \mu)^2 = 1.$$

So in some sense the normalizations (of subtracting μ and dividing by $\sigma/\sqrt{n}$) make sense. You should convince yourself that if you did not multiply by $\sqrt{n}$ this would have a degenerate limit (i.e. would converge in distribution to a point mass at 0). Multiplying by $\sqrt{n}$ is enlarging the fluctuations of the average around the expectations at just the right rate.

The other sanity check is to just notice that if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, then S_n would have distribution exactly equal to that of Z. Roughly, if there was going to be a "universal limit" i.e. if the average was going to converge to a single distribution (irrespective of the distribution of X) then it has a to be a Gaussian distribution (just because we know that the average of Gaussians is Gaussian).

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.

Formally, let $X_1, \ldots, X_n$ be a sequence of RVs with mgfs $F_{X_1}, \ldots, F_{X_n}$. If for all t in an open interval around 0 we have that, $F_{X_n}(t) \to F_X(t)$, then X_n converges in distribution to X.

5.3.3 Proof

We will follow the proof from John Rice's (Math Stat and Data Analysis) textbook. Larry's notes have a nearly identical proof. First we recall that the mgf of a standard normal is simply $M_Z(t) = \exp(t^2/2)$.

Note that,

$$M_{S_n}(t) = \left[M_{(X-\mu)}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n,$$

using Facts 1 and 2. Now, one should imagine t as small and fixed so $t/(\sigma\sqrt{n})$ is quite close to 0. Taylor expanding the mgf around 0, and using Fact 3 we obtain

$$\begin{aligned} M_{S_n}(t) &= \left[1 + \frac{t}{\sigma\sqrt{n}}\mathbb{E}(X-\mu) + \frac{t^2}{2n\sigma^2}\mathbb{E}(X-\mu)^2 + \frac{t^3}{6n^{3/2}\sigma^3}\mathbb{E}(X-\mu)^3 + \ldots\right]^n \\ &\approx \left[1 + \frac{t^2}{2n}\right]^n \to \exp(t^2/2), \end{aligned}$$

using the fact that,

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