36-705: Intermediate Statistics **Fall 2019**

Lecture 7: September 11

Lecturer: Siva Balakrishnan

In today's lecture we will continue our discussion of the CLT. Before we do this we will briefly introduce stochastic order notation.

7.1 Stochastic Order Notation

The classical order notation should be familiar to you already.

1. We say that a sequence $a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$. Similarly, $a_n = o(b_n)$ if $a_n/b_n = o(1)$.
2. We say that a sequence $a_n = O(1)$ if the sequence is eventually bounded, i.e. for all n large, $|a_n| \leq C$ for some constant $C \geq 0$. Similarly, $a_n = O(b_n)$ if $a_n/b_n = O(1)$.
3. If $a_n = O(b_n)$ and $b_n = O(a_n)$ then we use either $a_n = \Theta(b_n)$ or $a_n \sim b_n$. Usually in Stats we avoid the Θ notation (which is more common in CS) because we usually use Θ for the parameter space.

When we are dealing with random variables we use stochastic order notation.

1. We say that $X_n = o_p(1)$ if for every $\epsilon > 0$, as $n \to \infty$

$$\mathbb{P}(|X_n| \geq \epsilon) \to 0,$$

   i.e. X_n converges to zero in probability.

2. We say that $X_n = O_p(1)$ if for every $\epsilon > 0$ there is a finite $C(\epsilon) > 0$ such that, for all n large enough:

$$\mathbb{P}(|X_n| \geq C(\epsilon)) \leq \epsilon.$$

The typical use case: suppose we have $X_1, \ldots, X_n$ which are i.i.d. and have finite variance, and we define:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

1. $\widehat{\mu} - \mu = o_p(1)$ (WLLN)
2. $\widehat{\mu} - \mu = O_p(1/\sqrt{n})$ (CLT)

As with the classical order notation, we can do some simple "calculus" with stochastic order notation and observe that for instance: $o_p(1) + O_p(1) = O_p(1)$, $o_p(1)O_p(1) = o_p(1)$ and so on.

7.2 Proof of the CLT

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 $M_{X_1}, \ldots, M_{X_n}$. 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.

7.2.1 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,

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

7.3 Only independence but not identically distributed

The CLT goes through almost exactly as stated, however, we need conditions to ensure that one or a small number of random variables do not dominate the sum. There are many such results but the most classical is called the Lyapunov CLT. I will state something that is slightly weaker than the actual result. Lyapunov is one of the fathers of the theory of dynamical systems and a student of Chebyshev's. As is the case with Chebyshev, there are several foundational concepts that are named for him (the CLT is only one).

Define the variance of the average:

$$s_n^2 = \sum_{i=1}^n \sigma_i^2.$$

Lyapunov CLT: Suppose $X_1, \ldots, X_n$ are independent but not necessarily identically distributed. Let $\mu_i = \mathbb{E}[X_i]$ and let $\sigma_i = \mathrm{Var}(X_i)$. Then if we satisfy the Lyapunov condition:

$$\lim_{n\to\infty} \frac{1}{s_n^3} \sum_{i=1}^n \mathbb{E}|X_i - \mu|^3 = 0,$$

then

$$\frac{1}{s_n} \sum_{i=1}^n [X_i - \mu_i] \xrightarrow{d} N(0,1).$$

First notice that ignoring the Lyapunov condition, in the i.i.d case we just have $\mu_i = \mu$ and $s_n = \sqrt{n}\sigma$ so we get back our usual CLT.

It is worthwhile trying to understand the Lyapunov condition, and when it might be violated. In particular, consider the extreme case, when all the random variables are deterministic, except X_1 which has mean μ_1 and variance $\sigma_1^2 > 0$. Then $s_n^3 = \sigma_1^3$ and the third absolute moment $\mathbb{E}|X_1 - \mu|^3 > 0$ so that the Lyapunov condition fails. Roughly, what can happen in the non-identically distributed case is that only one random variable can dominate the sum in which case you are not really averaging many things so you do not have a CLT.

On the other hand in a more typical case, one might have that the third absolute moments are bounded by some constant $C > 0$ say and the variance of any particular random variable is not too small. In this case,

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2 \geq n\sigma_{\min}^2,$$

and

$$\sum_{i=1}^{n} \mathbb{E}|X_i - \mu|^3 \leq Cn.$$

In this case, we will have that the Lyapunov ratio $\leq \frac{C}{\sqrt{n}\sigma_{\min}^3} \to 0$ so that the condition is indeed satisfied.

7.4 Multivariate CLT

The first important extension is the multivariate CLT.

Multivariate CLT: If $X_1, \ldots, X_n$ are i.i.d with mean $\mu \in \mathbb{R}^d$, and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ (with finite entries) then,

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

Notes:

1. You might wonder what convergence in distribution means for random vectors. A random vector still has a CDF, typically we define this as:

   $$F_X(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots X_d \leq x_d),$$

   so we can still define convergence in distribution via pointwise convergence of the CDF. In order to define points of continuity it turns out that the correct definition is that a point is a point of continuity of the CDF if the boundary of the rectangle whose upper right corner is $(x_1, \ldots, x_d)$ has probability 0.

2. Although d can be larger than 1, it is taken to be fixed as $n \to \infty$. Central limit theorems, when d is allowed to grow, i.e. high-dimensional CLTs are rare and are an active topic of research.

3. The proof of this result follows directly from the proof of the univariate CLT and a powerful result in asymptotic statistics known as the Cramer-Wold device. The Cramer-Wold device roughly asserts that if $a^T X_n \xrightarrow{d} a^T X$ for all vectors $a \in \mathbb{R}^d$ then $X_n \xrightarrow{d} X$.

7.5 CLT with estimated variance

We saw that in our typical use case of the CLT (constructing confidence intervals) we needed to know the variance σ. In practice, we most often do not know this. However, we can estimate this quantity in the usual way,

$$\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2.$$

It turns out that we can replace the standard deviation in the CLT by $\widehat{\sigma}$ and still have the same convergence in distribution, i.e.

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

The proof follows from a sequence of applications of Slutsky's theorem and the continuous mapping theorem.

Proof: First observe that if we can show that $\frac{\sigma}{\widehat{\sigma}_n} \xrightarrow{d} 1$, then an application of Slutsky's theorem and the CLT gives us the desired result.

Since square-root is a continuous map, by the continuous mapping theorem, it suffices to show that $\frac{\sigma^2}{\widehat{\sigma}_n^2} \xrightarrow{d} 1$. We will instead show the stronger statement that,

$$\widehat{\sigma}_n^2 \xrightarrow{p} \sigma^2,$$

which implies the desired statement via the continuous mapping theorem (see Larry's notes for more details). Note that,

$$\begin{aligned}\widehat{\sigma}_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 \\ &\xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2,\end{aligned}$$

using the fact that $\frac{n-1}{n} \to 1$. Now,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{p} \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

using the WLLN. This concludes the proof.

7.6 Rate of convergence in CLT - Berry Esseen

While the central limit theorem is an asymptotic result (i.e. a statement about $n \to \infty$) it turns out that under fairly general conditions we can say how close to a standard normal the average is, in distribution, for finite values n. Such results are known as Berry Esseen bounds. Roughly, they are proved by carefully tracking the remainder terms in our Taylor series proof but we will not do this here.

Berry-Esseen: Suppose that $X_1, \ldots, X_n \sim P$. Let $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \mathbb{E}[(X_1 - \mu)^2]$, and $\mu_3 = \mathbb{E}[|X_1 - \mu|^3]$. Let

$$F_n(x) = \mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu} - \mu)}{\sigma} \leq x\right),$$

denote the CDF of the normalized sample average. If $\mu_3 < \infty$ then,

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{9\mu_3}{\sigma^3\sqrt{n}}.$$

This bound is roughly saying that if μ_3/σ^3 is small then the convergence to normality in distribution happens quite fast.

7.7 The Delta Method

A natural question that arises frequently is the following: suppose we have a sequence of random variables X_n that converges in distribution to a Gaussian distribution then can we characterize the limiting distribution of $g(X_n)$ where g is a smooth function?

We could work this out by using the continuous mapping theorem (indeed, that is at the heart of the proof we are about to give).

Delta Method: Suppose that,

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

and that g is a continuously differentiable function such that $g'(\mu) \neq 0$. Then,

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

Proof: The basic idea is simply to use Taylor's approximation. We know that,

$$g(X_n) \approx g(\mu) + g'(\mu)(X_n - \mu),$$

so that,

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

To be rigorous however we need to take care of the remainder terms. Here is a more formal proof.

By a rigorous application of Taylor's theorem we obtain,

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{\sigma} = g'(\widetilde{\mu})\frac{\sqrt{n}(X_n - \mu)}{\sigma},$$

where $\widetilde{\mu}$ is on the line joining μ to $\widehat{\mu}$. We know by the WLLN that $\widehat{\mu} \xrightarrow{p} \mu$ and so $\widetilde{\mu} \xrightarrow{p} \mu$. Since g is continuously differentiable, we can use the continuous mapping theorem to conclude that,

$$g'(\widetilde{\mu}) \xrightarrow{p} g'(\mu).$$

Now, we apply Slutsky's theorem to obtain that,

$$g'(\widetilde{\mu})\frac{\sqrt{n}(X_n - \mu)}{\sigma} \xrightarrow{d} g'(\mu)N(0,1) \stackrel{d}{=} N(0, [g'(\mu)]^2).$$

An example: Suppose we have $X_1, \ldots, X_n \sim P$ with $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2 < \infty$. Suppose we are interested in the distribution of $Y_n = \exp(\widehat{\mu}_n)$. Using that fact that $g'(\mu) = \exp(\mu)$, applying the Delta method we obtain,

$$\sqrt{n}\left(\frac{\exp(\widehat{\mu}_n) - \exp(\mu)}{\sigma}\right) \xrightarrow{d} N(0, \exp(2\mu)).$$

Multivariate Delta Method: There is a multivariate analogue of the Delta method (which is likely where the name comes from?). Suppose we have random vectors $X_1, \ldots, X_n \in \mathbb{R}^d$, and $g : \mathbb{R}^d \mapsto \mathbb{R}$ is a continuously differentiable function, then

$$\sqrt{n}(g(\widehat{\mu}_n) - g(\mu)) \xrightarrow{d} N(0, \Delta_\mu(g)^T \Sigma \Delta_\mu(g)),$$

where

$$\Delta_g(\mu) = \begin{pmatrix} \frac{\partial g(x)}{\partial x_1} \\ \vdots \\ \frac{\partial g(x)}{\partial x_d} \end{pmatrix}_{x=\mu},$$

is the gradient of g evaluated at μ.