STAT 203 Notes

Ryan Anderson

2024-01-17

Lecture 1/17: Convergence of Sequences of Random Variables

We distinguish four modes of convergence for sequences of random variables $X_1, X_2, \ldots$:

1. Convergence in law: $X_n \to^L X$ iff $F_{X_n(x)} \to F_X(x)$ at all points x where F_X is continuous.
2. Convergence in probability: $X_n \to^P X$ iff $P(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$.
3. Convergence in the r-th mean: $X_n \to^r X$ iff $E(|X_n - X|^r) \to 0$.
4. Convergence almost surely: $X_n \to^{a.s.} X$ iff $P(\lim_{n\to\infty} X_n = X) = 1$.

Put another way, say we had a set of realizations of random variables $X_1, X_2, \ldots, X_n, X$. That is, for N realizations, each described as $(X_1^i, X_2^i, \ldots, X_n^i, X^i)$, convergence almost surely requires that each of the realizations $X_1^i, X_2^i, \ldots, X_n^i$ converge to X^i.

Convergence in the r-th mean can be thought of as taking the average across all the realizations. We compute $\frac{1}{N}(|X_n^1 - X^1|^r + |X_n^2 - X^2|^r + \cdots + |X_n^N - X^N|^r) \to 0$. Thus, we don't need each sequence to go to 0 - we only need the average to not deviate too far from 0. Hence convergence in the r-th mean is weaker than convergence almost surely.

Convergence in probability is weaker than convergence in the r-th mean. We can see this by modifying the above computation to instead take an average of indicator functions denoting whether each realization ends up more than ϵ away from X^i. That is, we compute $\frac{1}{N}(1(|X_n^1 - X^1| > \epsilon) + 1(|X_n^2 - X^2| > \epsilon) + \cdots + 1(|X_n^N - X^N| > \epsilon)) \to 0$. Hence convergence in probability is weaker than convergence in the r-th mean.

Convergence in law is weakest of all, as it only requires that the distribution of the random variables converge. We can restate this as saying that we can permute entries of each sequence across realizations. That is, we can take $X_n^1, X_n^2, \ldots, X_n^N$ and permute them to $X_n^{\pi(1)}, X_n^{\pi(2)}, \ldots, X_n^{\pi(N)}$ for any permutation π. This is because convergence in law only requires that the distribution of the random variables converge, and the distribution is invariant to permutations.

Because of this system of inclusions, we can say the following:

1. Convergence almost surely implies convergence in probability.
2. Convergence in the r-th mean implies convergence in probability.
3. Convergence in probability implies convergence in law.

We can see why the other statements fail to hold by considering the following counterexamples. Consider $Z \sim U(0,1)$. We will proceed by continuously dividing the interval $(0,1)$ and taking our sequence entries to be indicators of the standard uniform on each component of the partition.

Thus $X_1 = 1_{(0,1)}(Z) = 1$. Then divide $(0,1)$ into two equal parts, and let $X_2 = 1_{(0,\frac{1}{2})}(Z)$ and $X_3 = 1_{(\frac{1}{2},1)}(Z)$. Because we have divided the interval as such, we now have that $X_2 + X_3 = 1$. Continue and create at the next step X_4, X_5, X_6, X_7 for which we know $X_4 + X_5 + X_6 + X_7 = 1$.

This sequence will never converge, as only one of the entries as each stage will be 1 and the rest 0. Because we'll have more entries in the sequence and thus more 0s, the average converges to 0, but $X_n! \to^{a.s.} X$. However, $X_n \to^r X$ and $X_n \to^p X$.

Lecture 1/22: Modes of Convergence

Recall that we distinguish modes of convergence: convergence in law $X_n \to^L X$ (weak convergence), convergence in probability $X_n \to^p X$, convergence in the r-th mean $X_n \to^r X$, and convergence almost surely $X_n \to^{a.s.} X$.

We can see that convergence almost surely implies convergence in probability, which implies convergence in law. Convergence in the r-th mean also implies convergence in probability.

Want to show that convergence in probability implies convergence in law.

Thm: Convergence in probability implies convergence in law We can rewrite this to say that we want to prove $P(X_n \leq X) \to^{n\to\infty} P(X \leq x)$ for all x where F_X is continuous.

Start by defining $C(F_X)$ as the set of all points x where F_X is continuous. We can then rewrite the above as $P(X_n \leq X) \to^{n\to\infty} P(X \leq x)$ for all $x \in C(F_X)$.

Since we have convergence in probability, we know that with high probability, X_n is not far from x. We break up the event in consideration into two components: $\{X_n \leq x\} \subset \{X \leq x+\epsilon\} \cup \{|X_n - x| \geq \epsilon\}$. That is, if $X_n \leq x$, then either $X \leq x+\epsilon$ or X_n is far from x.

We then use a union bound: $P(X_n \leq x) \leq P(\{X \leq x+\epsilon\} \cup \{|X_n - x| \geq \epsilon\}) \leq P(\{X \leq x+\epsilon\}) + P(\{|X_n - x| \geq \epsilon\})$.‘

Rewriting in terms of the CDFs, we get $F_{X_n}(x) \leq F_x(x+\epsilon) + P(|X_n - X| > \epsilon)$.

If X_n approaches X from the right, then we need to rewrite the directions - now we break up the event into $\{X \leq x-\epsilon\} \subset \{X_n \leq x\} \cup \{|X_n - x| > \epsilon\}$.

This gives us $F_X(x-\epsilon) \leq F_{X_n}(x) + P(|X_n - X| > \epsilon)$.

Combining both statements, we have

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

Letting $n \to \infty$ and $\epsilon \to 0$, then two things happen: (1) since x is a continuity point of F_X, we have $F_X(x-\epsilon) \to F_X(x)$ and $F_X(x+\epsilon) \to F_X(x)$, and (2) since $X_n \to^p X$, we have $P(|X_n - X| > \epsilon) \to 0$.

This implies that $F_{X_n}(x) \to F_X(x)$ for all $x \in C(F_X)$, which is convergence in law.

Partial Converses of the Relationships

One interesting partial converse obtains when $X = c$ a constant point mass. Then we have $X_n \to^L c \Rightarrow X_n \to^p c$. In particular, this observation leads to the law of large numbers:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to^{L,p} \mu$$

More to the point, we observe $\bar{X}_n - \mu \to^{L,p} 0$. The extension of this result is the central limit theorem: $\sqrt{n}(\bar{X}_n - \mu) \to^L N(0,1)$.

Thm: Convergence in law to a constant implies convergence in probability to a constant We have

$$\begin{aligned} P(|X_n - c| \leq \epsilon) &= P(c-\epsilon \leq X \leq c+\epsilon) \\ &= P(X_n \leq c+\epsilon) - P(X_n \leq c-\epsilon) \end{aligned}$$

Note that c is not a continuity point for F_c, but $c-\epsilon$ and $c+\epsilon$ are. Thus we have $P(X_n \leq c+\epsilon) \to F_c(c+\epsilon) = 1$ and $P(X_n \leq c-\epsilon) \to F_c(c-\epsilon) = 0$.

Lecture 1/24: Convergence in Expectation

Consider a sequence of random variables which converge in law to an RV X: $X_n \to^L X$.

Theorem: Helly-Bray If $X_n \to^L X$, then for some function $g, E[g(X_n)] \to E[g(x)]$. There are equivalent formulations of the restrictions on g:

1) g is continuous and vanishes outside a compact set, e.g., $g(x) = U(a, b)$
2) g is continuous and bounded, e.g., $g(x) = \sin(x)$
3) g is measurable and bounded such that $P(X \in C(g)) = 1$, the limit RV X must certainly be in the continuity set of g.

Helly-Bray provides the machinery for proving the very powerful continuity theorem, which says that if the characteristic functions of X_n converge to the characteristic function of X, then $X_n \to^L X$.

Proof It is obvious that 3) implies 2) and 1) as it is simply the most general expression of the three. We need to prove 1) implies convergence in law and that convergence in law implies 3).

To prove convergence in law implies 3), start by rewriting convergence in law in terms of the CDFs: $F_{X_n}(x) \to F_X(x)$ for all $x \in C(F_X)$.

We can rewrite statements with CDFs as statements on expectations of indicator functions, so write this as

$$F_{X_n}(x) \to^L F_X(x) \forall x \in C(F_X)$$
$$\Rightarrow E[I(X_n \leq x)] \to^L E[I(X \leq x)] \forall x \in C(F_X)$$

Now let $g(\cdot) = I(\cdot \leq x)$. Now we have a statement $E[g(X_n)] \to^L E[g(X)] \forall x \in C(F_X)$. Note that $C(g) = \mathbb{R}/\{x\}$ as the indicator function is discontinuous at x.

Now we need to show that $P(X \in C(g)) = 1$. From above, $P(X \in C(g)) = P(X \neq x) = 1$, as $x \in C(F_X)$ and $P(X = x) = 0 \forall x \in C(F_X)$. This is because the probability of taking any one value in the continuity set is 0!

Hence convergence in law implies 3).

Now we want to show convergence in law implies 1). We have $E[g(X_n)] \to^L E[g(X)] \forall x \in C(F_X)$, and we want to show that g is continuous and vanishes outside a compact set.

Lecture 1/29: Levy's Continuity Theorem

Levy's continuity theorem converts statements about the convergence of a sequence of RVs to statements about the characteristic functions of such RVs. That is, we have $X_n \to^L X$ iff $\phi_{X_n}(t) \to \phi_X(t)$.

Proof For the forward direction, let $X_n \to^L X$. By the Helly-Bray Theorem, we know that there exists g such that $E[g(X_n)] \to E[g(x)]$.

The characteristic function is given by $\Phi_{X_n}(t) = E[e^{it^T X_n}] = E[g(X_n)]$.

Since we know by Helly-Bray that $g(X_n)$ is continuous and bounded, then we know that $\phi_{X_n}(t) \to \phi_X(t)$.

For the reverse direction, we start again with Helly-Bray and a function g that is continuous and vanishes outside a compact set. We know we get for free that g is bounded, $\forall X |g(X)| \leq B$, and that g is uniformly continuous ($\forall \epsilon > 0, \exists \delta > 0 |x - y| < \delta \Rightarrow |g(x) - g(y)| < \epsilon$).

We want to show that with these assumptions $\phi_{X_n}(t) \to \phi_X(t) \Rightarrow E[g(X_n)] \to E[g(x)]$.

The trick ends up being to introduce a perturbation RV independent of the sequence and limit. Let $Y_\sigma \sim N(0, \sigma^2 I) \perp X_n, X$. Then we have

$$|E[g(X_n)] - E[g(X)]| \leq |E[g(X_n)] - E[g(X_n + Y_\sigma)]| + |E[g(X_n) + Y_\sigma] - E[g(X + Y_\sigma)]| + |E[g(X + Y_\sigma)] - E[g(X)]|$$

Because Y_σ is concentrated at 0, the first and third terms should be small. The second term we can govern by the behavior of the characteristic functions.

Taking each term on its own, we start by writing

$$\begin{aligned}|E[g(X_n)] - E[g(X_n + Y_\sigma)]| &\leq E[|g(X_n) - g(X_n + Y_\sigma)|] \\ &= E[|g(X_n) - g(X_n + Y_\sigma)|I(|Y_\sigma| < \delta)] + E[|g(X_n) - g(X_n + Y_\sigma)|I(|Y_\sigma| \geq \delta)]\end{aligned}$$

We obtain the equality by breaking into two cases: one where $|Y_\sigma \leq \delta|$ and one where it is greater. By uniform continuity, we can bound the first case by ϵ. We then need to handle the second case.

$$\begin{aligned}|E[g(X_n)] - E[g(X_n + Y_\sigma)]| &\leq E[|g(X_n) - g(X_n + Y_\sigma)|I(|Y_\sigma| < \delta)] + E[|g(X_n) - g(X_n + Y_\sigma)|I(|Y_\sigma| \geq \delta)] \\ &\leq \epsilon + 2BP(|Y_\sigma| \geq \delta)\end{aligned}$$

We have obtained the next inequality by bounding g. Since Y_σ is a Gaussian, we can bound the probability of it being greater than δ by $\frac{\epsilon}{2B}$. Hence we have

$$|E[g(X_n)] - E[g(X_n + Y_\sigma)]| \leq 2\epsilon$$

The same thing can be done for the third term. That means we are only left with bounding the middle term.

We proceed by actually calculating the expectation of $g(X_n + Y_\sigma)$.

$$\begin{aligned}E[g(X_n + Y_\sigma)] &= \int\int g(x+y)dF_{X_n}(x)(\frac{1}{\sqrt{2\pi}\sigma})^d \exp(-\frac{y^T y}{2\sigma^2})dy \\ &= \int\int g(u)\exp(-\frac{(u-x)^T(u-x))}{2\sigma^2})dudF_{X_n}(x)\end{aligned}$$

where we obtain the equality by performing a transformation $u = x + y$.

Reading Notes on Continuity Theorem

Continuing from last time, we want to have our expectation look like the characteristic function of X_n, which is given by

$$\phi_{X_n}(t) = E[e^{it^T X_n}] = \int e^{it^T x}dF_{X_n}(x)$$

This means we need to get x in the power of the exponential. We make a change of variables $u = x + y$ and get

$$\begin{aligned}E[g(X_n + Y_\sigma)] &= \int\int g(x+y)dF_{X_n}(x)(\frac{1}{\sqrt{2\pi}\sigma})^d \exp(-\frac{y^T y}{2\sigma^2})dy \\ &= \int g(u)\int \exp(-\frac{(u-x)^T(u-x))}{2\sigma^2})dF_{X_n}(x)du\end{aligned}$$

The trick from here is to rewrite the inner integral to resemble the characteristic function of a standard Gaussian. One fun fact is that transformations of Gaussians are still Gaussians, but your variance gets flipped. This is because of the Heisenberg uncertainty principle!

$$\exp(-\frac{(u-x)^T(u-x))}{2\sigma^2}) = (\frac{\sigma}{\sqrt{2\pi}})^d \int \exp(i(u-x)^T z)\exp(-\frac{\sigma^2}{2}z^T z)dz$$

We can then substitute this into the inner integral to get

$$\begin{aligned}
E[g(X_n + Y_\sigma)] &= \int g(u) \int (\frac{\sigma}{\sqrt{2\pi}})^d \int \exp(i(u-x)^T z) \exp(-\frac{\sigma^2}{2} z^T z) dz dF_{X_n}(x) du \\
&= \int g(u) \int (\frac{\sigma}{\sqrt{2\pi}})^d \int \exp(iu^T z) \exp(-\frac{\sigma^2}{2} z^T z) \exp(-ix^T z) dz dF_{X_n}(x) du \\
&= \int g(u) \int \exp(iu^T z) (\frac{\sigma}{\sqrt{2\pi}})^d \int \exp(-\frac{\sigma^2}{2} z^T z) dF_{X_n}(x) dz du \\
&= \int g(u) \int \exp(iu^T z) \phi_{X_n}(\sigma z) dz du
\end{aligned}$$

From there by dominated convergence we know that pointwise convergence of the characteristic function implies convergence in distribution. Hence we get the full continuity theorem.

Levy's continuity theorem highlights the importance of the characteristic function in mathematical statistics. It is a very powerful tool that allows us to prove convergence in distribution. It is also a very useful tool in proving the central limit theorem.

Law of Large Numbers

Let $f : \mathbb{R}^d \to \mathbb{R}$. Denote the gradient of f, $\dot{f} = (\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_d} f(x))$ as a row vector. We define the Hessian of f, $\ddot{f}$ as

$$\ddot{f} = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(x) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x) & \ldots & \frac{\partial^2}{\partial x_1 \partial x_d} f(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(x) & \frac{\partial^2}{\partial x_2^2} f(x) & \ldots & \frac{\partial^2}{\partial x_2 \partial x_d} f(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} f(x) & \frac{\partial^2}{\partial x_d \partial x_2} f(x) & \ldots & \frac{\partial^2}{\partial x_d^2} f(x) \end{bmatrix}$$

Note if we have $g : \mathbb{R}^d \to \mathbb{R}^k, g(x) = (g_1(x), \ldots, g_k(x))^T$, then we get the gradient of g as

$$\dot{g} = \begin{bmatrix} \dot{g}_1 \\ \dot{g}_2 \\ \vdots \\ \dot{g}_k \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1 & \frac{\partial}{\partial x_2} g_1 & \ldots & \frac{\partial}{\partial x_d} g_1 \\ \frac{\partial}{\partial x_1} g_2 & \frac{\partial}{\partial x_2} g_2 & \ldots & \frac{\partial}{\partial x_d} g_2 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_k & \frac{\partial}{\partial x_2} g_k & \ldots & \frac{\partial}{\partial x_d} g_k \end{bmatrix}$$

Now we turn our attention to statistical questions. We distinguish two forms of consistency: an estimator $\hat{\theta}_n$ for a parameter θ is weakly consistent if $\hat{\theta_n} \xrightarrow{P} \theta$ as $n \to \infty$. It is strongly consistent if $\hat{\theta}_n \xrightarrow{a.s.} \theta$ as $n \to \infty$.

Consider a sample $X_1, dots, X_n$ iid from a population with mean μ. Our estimator is the sample mean $\bar{X_n}$. The law of large numbers gives that

$$\bar{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P,a.s.} \mu \iff E[|X|] \leq \infty$$

A related question we can ask concerns the rate of convergence. Convergence in probability for an estimator implies $P(|\hat{\theta}_n - \theta| > \epsilon) = O(f(n))$. Large deviation theory asks what happens in each case where $f(n) \propto n^{-\alpha}, f(n) \propto \exp(-\alpha n)$, etc.

Instead of considering the sample mean, we can consider other statistics. Consider the sample in terms of the CDFs: $X_1, \ldots, X_n \sim F$ iid. The CDF is given as $F_X(x) = P(X \leq x) = E[I(X \leq x)]$.

In this formulation, we can consider the empirical CDF, $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$. The law of large numbers tells us that for a given x,

$$\hat{F}_n(x) \xrightarrow{P,a.s.} F(x)$$

This is only pointwise convergence. A stronger statement is made by the Glivenko-Cantelli Theorem.

$$\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{P,a.s.} 0$$

Example: let $f_n(x) = \frac{x^2}{n} + x, f(x) = x$. Then $|f_n(x) - f(x)| = |\frac{x^2}{n}|$. For any fixed x, of course this goes to 0 as $n \to \infty$. However,

$$\sup_x |\frac{x^2}{n}| \geq |\frac{n^2}{n}| \to^{n\to\infty} \infty$$

The Glivenko-Cantelli Theorem only works because of useful features of the empirical CDF. The empirical CDF is a bounded step function, and the jumps are at the X_i.

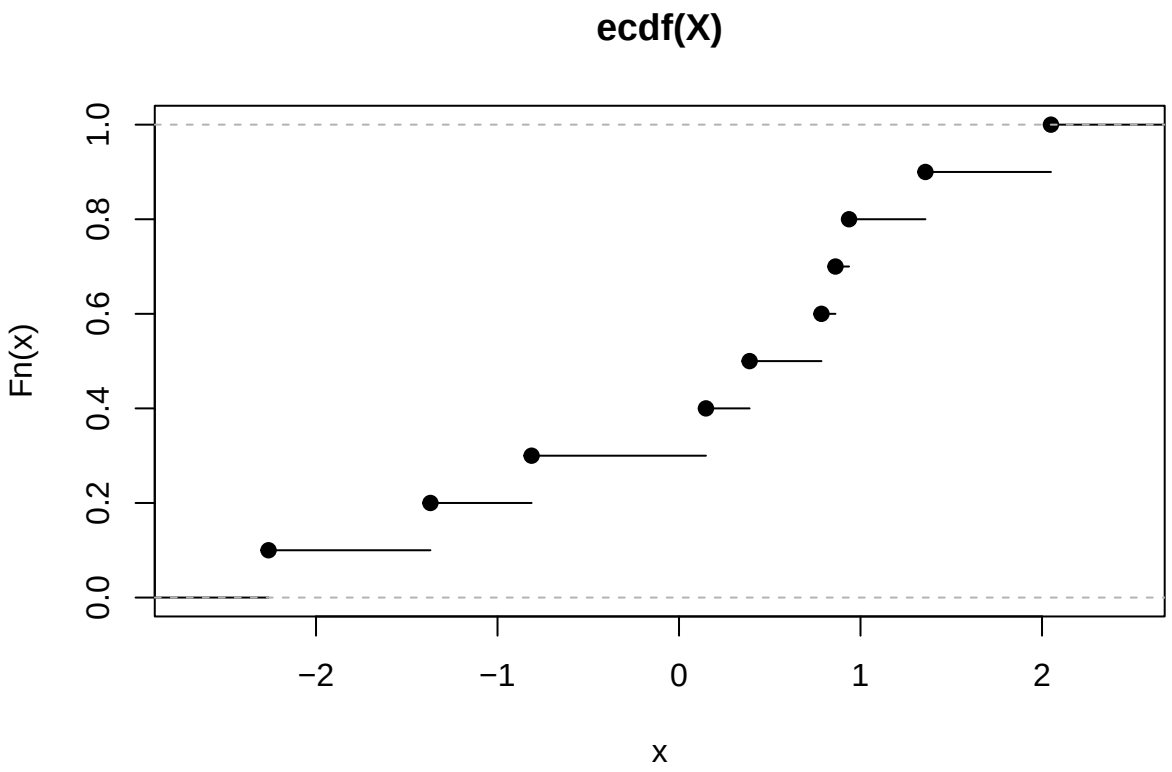
Reading Notes on Glivenko-Cantelli

Recall the definition of the empirical CDF:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

We can visualize this as a step function on $\mathbb{R}$ with n discontinuity points. The jumps are of size $\frac{1}{n}$, and the jumps occur at the X_i. If we sort our sample, then we can visualize the order statistics as

```
n <- 10
X <- rnorm(n)
plot(ecdf(X))
```



The Glivenko-Cantelli Theorem states that the empirical CDF converges uniformly to the true CDF.

By the strong law of large numbers, we know that $\hat{F}_n(x) \xrightarrow{a.s.} F(x)$ for a given x. The Glivenko-Cantelli Theorem states that $\sup_x |\hat{F}_n(x) - F(x)| \to^{a.s.} 0$.

Proof of Glivenko-Cantelli

We need to use the boundedness of both the empirical and true CDF to show that their difference converges to 0. We will want to partition the range of $F(x)$ into intervals of size $\frac{1}{n}$, and then use the strong law of large numbers to show that the empirical CDF converges to the true CDF on each interval.

Let $\epsilon > 0$. Choose $k > \frac{1}{\epsilon}$ and partition the range of $F(x)$ into intervals of size $\frac{1}{k}$, and let $x_0, x_1, x_2, \ldots, x_k$ be the partition points, the minimum of the preimages of each subinterval. Note $F(x_0) = 0$ so x_0 may be $-\infty$, and $F(x_k) = 1$ so x_k may be ∞.

Let $F(x^-) = P(X < x)$. This construction gives us that

$$F(x_j^-) \leq \frac{j}{k} \leq F(x_j)$$
$$P(X < x_j) \leq \frac{j}{k} \leq P(X \leq x_j)$$

and moreover, that $F(x_j^-) - F(x_{j-1}) \leq \frac{1}{k} < \epsilon$.

We can now use the strong law of large numbers to show that $\hat{F}_n(x_j) \xrightarrow{a.s.} F(x_j)$ and $\hat{F}_n(x_j^-) \xrightarrow{a.s.} F(x_j^-)$. Now since our partition is finite, we can take the max over all x_j to show that $\sup_x |\hat{F}_n(x) - F(x)| \xrightarrow{a.s.} 0$.

Let $\Delta n = \max(|\hat{F}_n(x_j) - F(x_j)|, |\hat{F}_n(x_j^-) - F(x_j^-)|)$. Then $\Delta n \xrightarrow{a.s.} 0$.

Now note that for any x we can find j such that $x_{j-1} \leq x \leq x_j$, and so

$$\hat{F}_n(x_{j-1}) \leq \hat{F}_n(x) \leq \hat{F}_n(x_j^-)$$
$$F(x_{j-1}) \leq F(x) \leq F(x_j^-)$$

accounting for potential discontinuity points.

To combine these, we can use the upper bound of $\hat{F}_n(x)$ and the upper bound of $F(x)$ to show that

$$\begin{aligned}\hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_j^-) - F(x_{j-1}) \\ &< \hat{F}_n(x_j^-) - F(x_j^-) + \epsilon\end{aligned}$$

and similarly for the lower bound:

$$\begin{aligned}\hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_{j-1}) - F(x_j^-) \\ &> \hat{F}_n(x_{j-1}) - F(x_{j-1}) - \epsilon\end{aligned}$$

So we have that

$$\begin{aligned}\hat{F}_n(x_{j-1}) - F(x_{j-1}) - \epsilon < \hat{F}_n(x) - F(x) < \hat{F}_n(x_j^-) - F(x_j^-) + \epsilon \\ |\hat{F}_n(x) - F(x)| \leq \max(|\hat{F}_n(x_{j-1}) - F(x_{j-1})|, |\hat{F}_n(x_j^-) - F(x_j^-)|) + \epsilon \\ \sup_x |\hat{F}_n(x) - F(x)| \leq \Delta n + \epsilon\end{aligned}$$

Since $\Delta n \to^{a.s.} 0$, we have that $\sup_x |\hat{F}_n(x) - F(x)| \to^{a.s.} 0$. QED.

Central Limit Theorem via Convergence Laws

To begin, recall the mean value theorem for functions of a vector variable. If $f : \mathbb{R}^n \to \mathbb{R}$ has continuous first derivative in $\{x : \|x - x_0\| < r\}$ then $\forall t \in \mathbb{R}^n, |t| < r$

$$f(x_0 + t) = f(x_0) + (\int_0^1 \dot{f}(x_0 + ut)du)t$$

Taylor's theorem is merely the second derivative version of the mean value theorem. If $\ddot{f}$ is continuous in $\{x : \|x - x_0\| < r\}$ then $\forall t \in \mathbb{R}^n, |t| < r$

$$f(x_0 + t) = f(x_0) + \dot{f}(x_0)t + t^T(\int_0^1 \int_0^1 v\ddot{f}(x_0 + uvt)dudv)t$$

Now consider the characteristic function of an RV X, $\phi_X(t) = E[e^{it^T X}]$. We calculate $\dot{\phi}_X(t) = E[e^{it^T X} iX^T]$, which gives $\dot{\phi}_X(0) = iE[X^T]$.

The second derivative gives $\ddot{\phi}_X(t) = E[e^{it^T X} i^2 XX^T]$, which gives $\ddot{\phi}_X(0) = -E[XX^T]$.

With this in mind we can describe the vector form of the central limit theorem. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. RVs with mean vector $\mu \in \mathbb{R}^d, \mu_i < \infty$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}, \Sigma = E[(Xi - \mu)(X_i - \mu)^T], \Sigma_{ij} < \infty$.

Then we have

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

Proof of Central Limit Theorem We proceed with Levy's continuity theorem.

We have that $\phi_{\sqrt{n}(\bar{X}_n - \mu)}(t) = E[\exp(it^T \sqrt{n}(\bar{X}_n - \mu))]$. Further, we have

$$\sqrt{n}(\bar{X}_n - \mu) = \sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i - \mu) = \frac{1}{\sqrt{n}}\sum(X_i - \mu)$$

We can then calculate the characteristic function of the sum as

$$\begin{aligned}
\phi_{\sqrt{n}(\bar{X}_n - \mu)}(t) &= E[\exp(it^T \frac{1}{\sqrt{n}}\sum(X_i - \mu))] \\
&= \prod_{i=1}^n E[\exp(it^T \frac{1}{\sqrt{n}}(X_i - \mu))] \\
&= E[\exp(it^T \frac{1}{\sqrt{n}}(X_i - \mu))]^n = E[\exp(i(\frac{t}{\sqrt{n}})^T(X_i - \mu))]^n \\
&= \phi_{X_i - \mu}(\frac{t}{\sqrt{n}})^n
\end{aligned}$$

Next we want to perform a Taylor expansion of the characteristic function around $t = 0$. We have

$$\begin{aligned}
(\phi_{X_i - \mu}(\frac{t}{\sqrt{n}}))^n &= (\phi_{X_i - \mu}(0) + \dot{\phi}_{X_i - \mu}(0)\frac{t}{\sqrt{n}} + (\frac{t}{\sqrt{n}})^T[\int_0^1 \int_0^1 v\ddot{\phi}_{X_i - \mu}(uv\frac{t}{\sqrt{n}})dudv]\frac{t}{\sqrt{n}})^n \\
&= (1 + 0 + \frac{1}{n}t^T[\int_0^1 \int_0^1 v\ddot{\phi}_{X_i - \mu}(uv\frac{t}{\sqrt{n}})dudv]t)^n
\end{aligned}$$

Now note that we have

$$n \to \infty, \ddot{\phi}_{X_i - \mu}(uv\frac{t}{\sqrt{n}}) \to \ddot{\phi}_{X_i - \mu}(0) = -E[(X_i - \mu)(X_i - \mu)] = -\Sigma$$

We also have a useful calculus result: $(1 + \frac{a}{n})^n \to^{n \to \infty} \exp(a)$. Combining that, we have

$$\begin{aligned}
(\phi_{X_i - \mu}(\frac{t}{\sqrt{n}}))^n &= (1 + \frac{1}{n}t^T[\int_0^1 \int_0^1 v\ddot{\phi}_{X_i - \mu}(uv\frac{t}{\sqrt{n}})dudv]t)^n \\
&\to^{n \to \infty} \exp(-\frac{1}{2}t^T\Sigma t) = \phi_{N(0,\Sigma)}(t)
\end{aligned}$$

which gives us the result. QED.

Generalizations of the CLT

We can begin to generalize the CLT by first allowing the RVs to come from different distributions. That is, we have a sample of RVs $X_1, X_2, \ldots$ with mean μ_i and covariance Σ_i.

As an example, consider a statistic $Z_n = \sum z_{ni}(X_i - \mu)$ where our X_i share a common mean. The question is for which constants $z_{n1}, \ldots, z_{nn}$ do we still have convergence in law to a Gaussian?

Lecture 2/5: Lindeberg-Feller Theorem

Consider a triangular sequence of RVs with row-wise sums, given by

$$\begin{aligned} X_{11} &= Z_1 \\ X_{21}, X_{22} &= Z_2 \\ X_{31}, X_{32}, X_{33} &= Z_3 \\ &\vdots \\ X_{n1}, X_{n2}, \ldots, X_{nn} &= Z_n \end{aligned}$$

where $E[X_{ni}] = 0, var(X_{ni}) = \sigma_{ni}^2$, and the X_{ni} are independent. Let the Z_n be zero-mean and calculate $var(Z_n) = \sum var(X_{ni}) = \sum \sigma_{ni}^2 = B_n^2$.

Theorem: Lindeberg-Feller Under the Lindeberg condition, the normalized row-wise sums converge in distribution to a Gaussian.

$$\frac{Z_n}{\sqrt{var(Z_n)}} \xrightarrow{L} N(0,1)$$

The Lindeberg condition is given as follows: let $\epsilon > 0$. Then

$$\frac{1}{B_n^2} \sum E[X_{ni}^2 I(|X_{ni}| > \epsilon B_n)] \xrightarrow{n \to \infty} 0$$

We interpret the Lindeberg condition as a statement about no one RV in our sequence dominating the group variance B_n.

In cases where the Lindeberg condition is too onerous to check, we can check a stronger statement known as the Lyapunov condition. This condition is given by

$$\exists \delta > 0 \text{ such that } \frac{1}{B_n^{2+\delta}} \sum E[|X_{ni}|^{2+\delta}] \xrightarrow{n \to \infty} 0$$

Note that although Lindeberg implies that the normalized row-wise sums converge to a Gaussian, the converse is not true. If we have another condition on the maximum of the variances, then we can show that normalized row sum convergence to Gaussian implies Lindeberg condition.

The Lindeberg-Feller theorem allows us to describe asymptotic normality in more situations than the ones given by the Central Limit Theorem.

Ex. 1: Let $X_1, X_2, \ldots$ be iid with $E[X_i] = \mu, var(X_i) < \infty$. This is the same setup as for CLT.

Let $X_{ni} = z_{ni}(X_i - \mu)$, z_{ni} some constant. Then $E[X_{ni}] = 0, var(X_{ni}) = z_{ni}^2 var(X_i)$. Then $B_n^2 = \sigma^2 \sum z_{ni}^2$.

Now we just want to show that if the Lindeberg condition holds then we have asymptotic normality. Let

$\epsilon > 0$ and write

$$\begin{aligned}\frac{1}{B_n^2}\sum E[X_{ni}^2 I(|X_{ni}| > \epsilon B_n)] &= \frac{1}{B_n^2}\sum E[z_{ni}^2(X_i-\mu)^2 I(|z_{ni}||X_i-\mu| \geq \epsilon B_n)] \\ &= \frac{1}{B_n^2}\sum z_{ni}^2 E[(X_i-\mu)^2 I(|X_i-\mu| \geq \frac{\epsilon B_n}{|z_{ni}|})] \\ &\leq \frac{1}{B_n^2}\sum z_{ni}^2 E[(X_i-\mu)^2 I(|X_i-\mu| \geq \frac{\epsilon B_n}{\max|z_{ni}|})]\end{aligned}$$

where the last step follows from the fact that $|z_{ni}| \leq \max|z_{ni}|$.

Next note that we can actually move the expectation out of the summation, as each value is identical for each of the X_i. Then we can use the fact that $B_n^2 = \sigma^2 \sum z_{ni}^2$ to get

$$\frac{1}{B_n^2}\sum E[X_{ni}^2 I(|X_{ni}| > \epsilon B_n)] \leq \frac{1}{\sigma^2} E[(X_i-\mu)^2 I(|X_i-\mu| \geq \frac{\epsilon B_n}{\max|z_{ni}|})]$$

Note that we require one more condition here: namely that $\frac{\max_i \sigma_{ni}^2}{B_n^2} \to^{n\to\infty} 0$. With that in mind, we know that

$$\frac{\epsilon B_n}{\max|z_{ni}|} = \frac{\epsilon}{\max|z_{ni}|}\sqrt{\sum \sigma_{ni}^2} \to^{n\to\infty} \infty$$

Since that value we are comparing in the indicator is going to infinity and since $|X_i - \mu|$ has finite variance, that means the indicator will be 0 for any value of ϵ, and hence the value in the expectation is 0 almost surely. This is not quite enough to guarantee that the expectation goes to 0, but we can use the dominated convergence theorem to show that it does. Hence the additional condition provides the Lindeberg condition.

Stronger Conditions: Lyapunov The Lyapunov condition is stronger than the Lindeberg condition. It is given by

$$\exists \delta > 0 \text{ such that } \frac{1}{B_n^{2+\delta}}\sum E[|X_{ni}|^{2+\delta}] \xrightarrow{n\to\infty} 0$$

Example: let $X_i \sim^{iid} Ber(p_i)$. Under what conditions does the sum of the X_i converge to a Gaussian?

If $p_i = p \quad \forall i$, then we use CLT. Otherwise, we need to contrive a way to use Lindeberg-Feller.

Let $Y_i = X_i - p_i$. Then $E[Y_i] = 0, var(Y_i) = p_i(1-p_i)$. Then we can define $Z_n = \sum_{i=1}^n Y_i$. Then $E[Z_n] = 0, var(Z_n) = \sum p_i(1-p_i) = B_n^2$.

Now our goal is to find a condition such that $Z_n/B_n \to^L N(0,1)$. We can use the Lyapunov condition for this.

Start by $\delta = 1$. Then $E[|Y_i^3|] = p_i(1-p_i)^3 + (1-p_i)p_i^3 = p_i(1-p_i) + [(1-p_i)^2 + p_i^2]$. Then

$$\frac{1}{B_n^3}\sum E[|Y_i^3|] = \frac{1}{B_n^3}\sum p_i(1-p_i) + [(1-p_i)^2 + p_i^2] \leq \frac{1}{B_n^3}\sum p_i(1-p_i) + 1 = \frac{B_n^2}{B_n^3}$$

Now all we need to show is that $B_n^2/B_n^3 \to 0$. This is true if $B_n = \sum p_i(1-p_i) \to \infty$.

Least Squares Estimation via Convergence Laws

Let $Y_i = \alpha + \beta X_i + \epsilon_i$, $x_i \in \mathbb{R}$ fixed, $\epsilon_i \sim^{iid} N(0,\sigma^2)$. Then we have

$$\hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)Y_i}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

The question is under what conditions is $\hat{\beta}$ asymptotically normal? By Lindeberg-Feller, we need to show that

$$\frac{\max_i (x_i - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \to 0$$

Start by defining an empirical estimate $\hat{\beta}_n$ as

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)\epsilon_i}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

And define $X_{ni} = (x_i - \bar{x}_n)\epsilon_i$. Then $E[X_{ni}] = 0, var(X_{ni}) = (x_i - \bar{x}_n)^2\sigma^2$. Then we can use Lindeberg-Feller to show that $\hat{\beta}_n$ is asymptotically normal.

Ultimately we get

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}(\hat{\beta}_n - \beta) \to^L N(0, \sigma^2)$$

Stationary Sequences

Consider a sequence of RVs $X_1, X_2, \ldots$. The sequence is stationary if all joint distributions of any subsequence of fixed length k are the same. That is, for any k, the joint distribution of $(X_1, X_2, \ldots, X_k)$ is the same as the joint distribution of $(X_{t+1}, X_{t+2}, \ldots, X_{t+k})$ for any t.

One immediate implication of this is that $cov(X_i, X_j)$ only depends on their distance $|i - j|$.

Start by considering the variance of the sample mean of a stationary sequence.

$$\begin{aligned} var(\bar{X}_n) &= var(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)) \\ &= \frac{1}{n^2}(\sum_{i=1}^n var(X_i) + \sum_{i \neq j} cov(X_i, X_j)) \\ &= \frac{1}{n^2}(n\sigma^2 + 2\sum_{i<j} cov(X_i, X_j)) \end{aligned}$$

where the last equality follows from the fact that $cov(X_i, X_j) = cov(X_j, X_i)$. Now let $\gamma = cov(X_i, X_k)$ for some fixed k. Then

$$\begin{aligned} \sum_{i<j} cov(X_i, X_j) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} cov(X_i, X_j) \\ &= \sum_{k=1}^{n-1} (n-k)\gamma \end{aligned}$$

Lecture 2/7: Central Limit Theorem for Stationary Sequences

Under stationarity, we can define the autocovariance function $\gamma_k = cov(X_i, X_{i+k})$. Then recall we can calculate the variance of the sample mean as

$$var(\bar{X}_n) = \frac{1}{n^2}(n\sigma^2 + 2\sum_{k=1}^{n-1}(n-k)\gamma_k)$$

Note this means we can describe the variance of the sum of the RVs as being parametrized by $\sigma^2, \gamma_1, \ldots, \gamma_{n-1}$.

M-Dependent Sequences

A sequence is m-dependent if $cov(X_i, X_j) = 0$ for $|i - j| > m$. That is, the sequence is m-dependent if any RVs separated by a distance greater than m are independent.

Example: let $X_1, X_2, \ldots$ be iid. Now consider $S_n = \sum_{i=1} X_i X_{i+m}$ for some fixed m. We can define a new sequence $Y_i = X_i X_{i+m}$. Note that Y_i is m-dependent, as we have

$$Y_1, Y_2, \cdots = X_1 X_{1+m}, X_2 X_{2+m}, \ldots$$

and we will eventually have a pair $Y_1, Y_{1+m} = X_1 X_{1+m}, X_{1+m} X_{1+2m}$, which are not independent. However, thereafter, $Y_1, Y_{2+m} = X_1 X_{1+m}, X_{2+m} X_{2+2m}$ are independent. Hence the sequence is m-dependent.

If we can show that a stationary sequence was m-dependent, then we can parametrize the variance of the sample mean by $\sigma^2, \gamma_1, \ldots, \gamma_m$, as pairs of RVs in the sequence more than m apart have 0 covariance.

Central Limit Theorem for Stationary m-Dependent Sequences

Theorem: Let $X_1, X_2, \ldots$ be a stationary m-dependent sequence with $E[X_i] = \mu, var(X_i) = \sigma^2 < \infty$, and $\gamma_k = cov(X_i, X_{i+k})$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \sigma^2 + 2 \sum_{k=1}^{\infty} \gamma_k)$$

Proof: WLOG assume $\mu = 0$. We will use the sum and not the average to keep things cleaner, so let $S_n = \sum X_i$. The goal is ultimately to create some partial sums out of subsequences of length k that are separated by m-sized gaps allowing us to use the m-dependence to show that the sum is asymptotically normal. The main problem will be dealing with the final terms in the sum of some remainder length.

That is, start by decomposing $n = (k + m)s + r, r < k + m$, where we have s subsequences of length k separated by m-sized gaps, and r remaining terms. Let $V_{k1}, \ldots, V_{k(s-1)}$ be the sums of the $s - 1$ subsequences of length k separated by m-sized gaps. By m-dependence, these are independent. Then we can use the CLT as m, k are fixed so $n \to \infty$ implies $s \to \infty$. So first we have

$$S'_{nk} = \sum_{j=0}^{s-1} V_{kj}$$

We also need to deal with the subsequences of length m separated by gaps of size k. Let $W_{k0}, \ldots, W_{k(s-1)}$ be the sums of the s subsequences of length m. We cannot use the CLT to handle this term asymptotically, as we don't have independence. So write

$$S''_{nk} = \sum_{j=0}^{s-1} W_{kj}$$

Finally we will have the remainder of the sequence R_{nk}. This allows us to rewrite the entire sum as

$$S_n = S'_{nk} + S''_{nk} + R_{nk}$$

The approach will be to allow the asymptotic normality of S'_{nk} to dominate.

First we need the below lemma.

Lemma: Let $T_n = Z_{nk} + X_{nk}$. If

1. $X_{nk} \to^P 0$ uniformly in n, that is, $\forall \epsilon > 0, \exists K$ s.t. $P(|X_{nk}| > \delta) < \epsilon, \forall k > K, n$
2. $Z_{nk} \to^L Z_k$ as $n \to \infty$ for each k
3. $Z_k \to^L Z$ as $k \to \infty$

Then $T_n \to^L Z$ as $n \to \infty$.

Proof idea is to find $z \in C(F_Z)$, the set of continuity points of F_Z and then show that $P(T_n \leq z) \to P(Z \leq z)$. Since $P(T_n \leq z) = P(Z_{nk} + X_{nk} \leq z) = P(Z_{nk} \leq z - X_{nk})$, we can use the uniform convergence of X_{nk} to show that $P(Z_{nk} \leq z - X_{nk}) \to P(Z \leq z)$.

Start by separating into events $A_n = \{X_{nk} \leq \delta\}$ and $A_n^c = \{X_{nk} > \delta\}$. Then we have

$$\begin{aligned} P(Z_{nk} + X_{nk} \leq z) &= P(Z_{nk} + X_{nk} \leq z, A_n) + P(Z_{nk} + X_{nk} \leq z, A_n^c) \\ &\leq P(Z_{nk} \leq z + \delta) + P(|X_{nk}| \geq \delta) \end{aligned}$$

If we let $n \to \infty$, then the first term will converge to $P(Z_k \leq z + \delta)$ as $z + \delta$ is still in the continuity set of F_Z, and the second term will converge to 0 by the uniform convergence of X_{nk}.

Then just let $k \to \infty$ and we have that $P(Z_k \leq z + \delta) \to P(Z \leq z + \delta)$, and then we have the result since δ is arbitrary.

Now let

$$T_n = \sqrt{n}\bar{X}_n = \frac{S_n}{\sqrt{n}} = \frac{S'_{nk} + S''_{nk} + R_{nk}}{\sqrt{n}}, Z_{nk} = \frac{S'_{nk} + R_{nk}}{\sqrt{n}}, X_{nk} = \frac{S''_{nk}}{\sqrt{n}}$$

Now we check the conditions of the lemma.

Lecture 2/12: More on Stationary m-Dependent Sequences

Fix $m \geq 0$ and let $X_1, X_2, \ldots$ be stationary and m-dependent. Let $E[X_i] = \mu, var(X_i) = \sigma^2 < \infty$. Then

$$\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \sigma^2 + 2\sum_{k=1}^{m} \gamma_k)$$

Ex. Let $X_1, X_2, \ldots$ be a stationary sequence with $E[X_i] = \mu, var(X_i) = \tau^2$. Then consider the product moments $S_n = \sum_{i=1}^n X_i X_{i+m}$. Then we have

$$\begin{aligned} E[X_i X_{i+m}] &= E[X_i]E[X_{i+m}] = \mu^2 \\ var(X_i X_{i+m}) &= E[X_i^2 X_{i+m}^2] - \mu^4 \\ &= (E[X_i^2])^2 - \mu^4 + (\mu^2 + \tau^2)^2 - \mu^4 \\ &= \tau^4 + 2\tau^2\mu^2 \end{aligned}$$

Are these product moments themselves m-dependent and stationary? Yes, which allows us to calculate the asymptotic distribution of the sample product moments. We just need to calculate the autocovariance function of the product moments.

$$\begin{aligned} \gamma_k &= cov(Y_i, Y_{i+k}) \to \gamma_1, \ldots \gamma_{m-1} = 0 \\ \gamma_m &= cov(Y_i, Y_{i+m}) = E[Y_i Y_{i+m}] - E[Y_i]E[Y_{i+m}] \\ &= E[X_i X_{i+m}^2 X_{i+2m}] - \mu^4 = E[X_i]E[X_{i+m}^2]E[X_{i+2m}] - \mu^4 \\ &= \mu^2(\mu^2 + \tau^2) - \mu^4 = \mu^2\tau^2 \end{aligned}$$

By our above theorem, then we have

$$\sqrt{n}(S_n - \mu^2) \to^L N(0, \sigma^2 + 2\sum \gamma_k) = N(0, \tau^4 + 2\mu^2\tau^2 + 2\mu^2\tau^2)$$

Further Consideration of the Univariate Central Limit Theorem We have considered the CLT and its extensions to handle non-identical distributions and dependent sequences. We have also considered the multivariate CLT. Now we want to consider rates of convergence and error estimates.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $E[X_i] = \mu, var(X_i) = \sigma^2 < \infty$. Standardize these and take $Y_i = \frac{X_i - \mu}{\sigma}, E[Y_i] = 0, var(Y_i) = 1$. Then we have

$$\sqrt{n}\bar{Y}_n \to^L N(0,1)$$

Theorem: Berry-Esseen. Let G_n be the CDF of $\sqrt{n}\bar{Y}_n$, i.e. $G_n(x) = P(\sqrt{n}\bar{Y}_n \leq x)$, and let Φ be the CDF of the standard normal. Then

$$\sup_x |G_n(x) - \Phi(x)| \to^{n\to\infty} 0$$

We observe the superficial similarity to the Glivenko-Cantelli theorem, which also described the size of the largest error between the empirical CDF and the true CDF. However, that statement involved randomness, which came from the sample and the empirical CDF. Here, there is no randomness!

A further statement of the Berry-Esseen theorem is that there exists a constant C such that

$$\sup_x |G_n(x) - \Phi(x)| \leq \frac{cE[Y_i]^3}{\sqrt{n}}$$

That is, the rate of convergence depends on the third moment but is on the order of $O(n^{-1/2})$.

Edgeworth Expansions

Consider the higher moments of Y_i. We have $\gamma = EY_i^3$ is the *skewness* and $\tau = EY_i^4$ is the *kurtosis*.

The Edgeworth expansion of G_n writes it in terms of Φ, ϕ and the higher moments. That is we have

$$G_n(x) = \Phi(x) + \frac{\gamma(x^2-1)}{6\sqrt{n}}\phi(x) + \frac{f(\tau,\gamma,x)}{n}\phi(x) + o(\frac{1}{n})$$

where $f(\tau, \gamma, x)$ is a function of the higher moments and $\phi(x)$ is the PDF of the standard normal.

Given that skewness is 0 for symmetric distributions, we can observe that the Edgeworth expansion tells us that the CLT converges quicker for symmetric distributions.

Slutsky Theorems

The Slutsky theorems are a set of useful theorems about the convergence of sequences and combinations of sequences of RVs. Let $X_1, X_2, \ldots$ be a sequence of random variables and $Y_1, Y_2, \ldots$ be a sequence of random variables.

1. Continuous Mapping Theorem Let $X_n \in \mathbb{R}^d, X_n \to^L X$ and let $f : \mathbb{R}^d \to \mathbb{R}^k$ such that $P(X \in C(f)) = 1$. Then $f(X_n) \to^L f(X)$.

Proof: We use Helly-Bray. Let g be bounded and continuous. It is sufficient to show that $E[g(f(X_n))] \to E[g(f(X))]$. Let $h = g \circ f$. Then h is bounded and at the very least we have $C(h) \subset C(f) \Rightarrow P(X \in C(h)) = 1$.

Then by Helly-Bray we have $X_n \to^L X \Rightarrow E[h(X_n)] \to^L E[h(X)]$. Then we have

$$E[g(f(X_n))] = E[h(X_n)] \to E[h(X)] = E[g(f(X))]$$

2. Compatibility of Convergence in Law and Convergence in Probability Let $X_n \to^L X$ and $(X_n - Y_n) \to^P 0$. Then $Y_n \to^L X$.

Proof: Also by Helly-Bray. Let g be continuous with compact support. Then we just need to show $E[g(Y_n)] \to E[g(X)]$. Note that continuity of g gives us $\forall \epsilon > 0, \exists \delta > 0$ s.t. $|x - y| < \delta \Rightarrow |g(x) - g(y)| < \epsilon$.

Now consider $|Eg(Y_n) - Eg(X)| \leq |Eg(Y_n) - Eg(X_n)| + |Eg(X_n) - Eg(X)|$ by triangle inequality. The second term goes to 0 by the convergence in law of X_n. Now we just deal with the first term. We have

$$\begin{aligned}|Eg(Y_n) - Eg(X_n)| &\leq E[|g(Y_n) - g(X_n)|I(|X_n - Y_n| < \delta)] + E[|g(Y_n) - g(X_n)|I(|X_n - Y_n| \geq \delta)] \\ &\leq \epsilon + 2||g||P(|X_n - Y_n| \geq \delta) \to 0 \\ &\Rightarrow Eg(Y_n) \to Eg(X)\end{aligned}$$

Then by Helly-Bray, we have $Y_n \to^L X$.

3. Convergence of Vector of Random Variables Let $X_n \in \mathbb{R}^d, Y_n \in \mathbb{R}^k$. If $X_n \to^L X, Y_n \to^L c$ then

$$(X_n, Y_n) \to^L (X, c)$$

Lecture 2/14: Sample Moments Distribution

Let $X_1, X_2, \ldots$ be a sample of size n from a distribution whose moments we are interested in estimating. Let the population moments be $\mu = E[X_1], \mu^2 + \sigma^2 = E[X_1^2]$.

We can define the sample moments as

$$M_x = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$M_{xx} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Define the sample variance $S_x^2 = \frac{1}{n}\sum(X_i - \bar{X}_n)^2$. Then we have $S_x^2 = M_{xx} - M_x^2$. We can show that the sample variance is asymptotically normal via the CLT and the delta method, with

$$\sqrt{n}(S_x^2 - \sigma^2) \to^L N(0, 2\sigma^4)$$

Asymptotic Normality of Functions of Moments

Recall that if $X \sim N(0,1)$ then $X^2 \sim \chi_1^2$, and also that $\sqrt{n}(X_n - \mu) \to^L N(0, \Sigma)$. How do we determine the asymptotic behavior of functions of the sample and population moments $\sqrt{n}(g(X_n) - g(\mu))$?

Start by noting that $g(x) = g(\mu + \int_0^1 g(\mu + \nu(x-u))))d\nu)(x-\mu)$.

With the mean value theorem, we can write

$$g(x) - g(\mu) = g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2 + R(x)$$

Rewriting to factor out the second derivative, we get

$$g(x) - g(\mu) \simeq \frac{g''(\mu)}{2}[(x - \mu + \frac{g'(\mu)}{g''(\mu)})^2 - \frac{g'(\mu)^2}{g''(\mu)}]$$

Let $\gamma_n = \frac{\sqrt{n}}{\sigma}\frac{g'(\mu)}{g''(\mu)}$. Then we have

$$n(g(X_n) - g(\mu)) \simeq \frac{\sigma^2 g''(\mu)}{2}[(\frac{\sqrt{n}(X_n - \mu)}{\sigma} + \gamma_n)^2 - \gamma_n^2]$$

The expression in the brackets is a Chi-squared distribution shifted by γ_n. Such distributions are extremely close to a normal distribution:$\chi_1^2(\gamma_n) - \gamma_n^2 \simeq^d N(0, 4\gamma_n^2)$.

Therefore for $g(x) = x^2, \mu = 0$, we have

$$\sqrt{n}(\bar{X}_n^2 - 0) \to^L \sigma^2\chi_1^2 \Rightarrow n(\frac{\bar{X}_n}{\sigma})^2 \to^L \chi_1^2$$

Lecture: Slutsky Theorems and Delta Method

Recall the three statements referred to as the Slutsky theorems:

1. Continuous mapping theorem: If $X_n \to^L X$ and $\exists g | P(X \in C(g)) = 1$, then $g(X_n) \to^L g(X)$.
2. Surrogate convergence in law: If $X_n \to^L X, X_n - Y_n \to^P 0$, then $Y_n \to^L X$.
3. Joint convergence to a constant: If $X_n \to^L X$ and $Y_n \to^P c$, then $(X_n, Y_n) \to^L (X, c)$.

One corollary of the Slutsky theorems is that if we have vector valued RVs $X_n \in \mathbb{R}^d, X_n \to^L X, Y_n \in \mathbb{R}^k, Y_n \to^L c, f :\in \mathbb{R}^{d+k} \to \mathbb{R}^r, P((X_n, Y_n) \in C(f)) = 1$, then $f((X_n, Y_n)) \to^L f(X, c)$.

T Statistics

Let $X_1, \ldots, X_n \sim^{i.i.d} E[X_i] = \mu, var(X_i) = \sigma^2$. Let the sample variance be $S_n^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$. Consider the *T statistic*

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

We can show that $T \to^L N(0,1)$ by the CLT and Slutsky's theorem. We can also show that $T^2 \to^L \chi_1^2$.

The CLT gives us that $\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \sigma^2)$.

Note that we can use the LLN to show that $(\frac{\sum X_i^2}{n}, \bar{X}_n) \to^{L,P,a.s.} (\sigma^2 + \mu^2, \mu)$. Then we can use the continuous mapping theorem to show that $S_n^2 \to^L \sigma^2$, and since square root is continuous that $S_n \to^L \sigma$.

The T statistic follows the T distribution. $T_n \sim t_{n-1}$, but the above shows that $T \to^L N(0,1)$.

Delta Method (Cramer's Theorem)

Let $X_n \in \mathbb{R}^d$, and let $g : \mathbb{R}^d \to \mathbb{R}^k$ be a function such that

$$g'(\mu) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_k}{\partial x_1} & \cdots & \frac{\partial g_k}{\partial x_d} \end{bmatrix}$$

exists and is continuous at μ. Then if $\sqrt{n}(X_n - \mu) \to^L X$, then $\sqrt{n}(g(X_n) - g(\mu)) \to^L g'(\mu)X$.

That is, applying g to X_n asymptotically just looks like a linear transformation of X.

Ex. 1: Consider the CLT $\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \Sigma)$.

Let us have such a g. Then $\sqrt{n}(g(X_n) - g(\mu)) \to^L g'(\mu)N(0, \Sigma) = N(0, g'(\mu)\Sigma g'(\mu)^T)$.

Proof of Cramer's Theorem: Consider $X \in \mathbb{R}^d, |X - \mu| < \delta$ for some $\delta > 0$. By the mean value theorem, we can expand $g(X)$ around μ:

$$g(X) = g(\mu) + (\int_0^1 g'(\mu + v(x - \mu))dv)(x - \mu)$$

Now consider $g(X_n)$:

$$g(X_n) - g(\mu) = (\int_0^1 g'(\mu + v(X_n - \mu))dv)(X_n - \mu)$$
$$\sqrt{n}(g(X_n) - g(\mu)) = (\int_0^1 g'(\mu + v(X_n - \mu))dv)\sqrt{n}(X_n - \mu)$$

Note that the condition $\sqrt{n}(X_n - \mu) \to^L X$ gives us for free that $X_n \to^{L,P} \mu$. This means in the limit

$$\int_0^1 g'(\mu + v(X_n - \mu))dv \to^P g'(\mu)$$

Hence we have that $\sqrt{n}(g(X_n) - g(\mu)) \to^L g'(\mu)X$.

Ex. 2: Let $X_1, \dots X_n \sim^{iid}, E[X_i] = \mu, var(X_i) = \sigma^2$.

By the CLT, $\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \sigma^2)$. What is the asymptotic distribution of $\bar{X}_n^2$?

Let $g(x) = x^2, g'(x) = 2x$. Then $\sqrt{n}(\bar{X}_n^2 - \mu^2) \to^L 2\mu N(0, \sigma^2) = N(0, 4\mu^2\sigma^2)$.

What about if $\mu = 0$? This implies that $\bar{X}_n^2$ converges to a point mass at a rate faster than $\sqrt{n}$!

The alternative in a problem with Cramer's theorem like this is to refer to the Slutsky theorems. If μ is 0, then the CLT gives $\sqrt{n}\bar{X}_n \to^L N(0, \sigma^2)$. Dividing by σ gives $\frac{\sqrt{n}}{\sigma}\bar{X}_n \to^L N(0, 1)$. Then by Slutsky's theorem, $\frac{\sqrt{n}}{\sigma}\bar{X}_n^2 \to^L \chi_1^2$, or alternately $n(\bar{X}_n)^2 \to^L \sigma^2\chi_1^2$.

Ex. 3: Consider the asymptotic distribution of $1/\bar{X}_n$.

Let $g(x) = 1/x, g'(x) = -1/x^2$. Then $\sqrt{n}(1/\bar{X}_n - 1/\mu) \to^L -\frac{1}{\mu^2}N(0, \sigma^2) = N(0, \frac{\sigma^2}{\mu^4})$. Obviously this only holds for $\mu \neq 0$.

One more remark is that the asymptotic mean of $1/\bar{X}_n$ is $\frac{1}{\mu}$. However, if $X_i \sim^{iid} N(\mu, \sigma^2)$, then $E[1/\bar{X}_n]$ does not exist!

Lecture 2/21: Asymptotics of Order Statistics

Recall that if $c_n \to c, b_n \to \infty$, then $(1 + \frac{c_n}{b_n})_n^b \to e^c$.

Finding Stable Distributions for Extreme Order Statistics

Ex. 1: Maximum of Uniform RVs Let $X_1, \dots, X_n \sim^{i.i.d} U(0, 1)$. What is the asymptotic distribution of $X_{(n)}$?

We showed earlier that $X_{(n)} \to^P 1$. Then consider the quantity $T_n = a_n(X_{(n)} - b_n)$. We know that $T_n \to^L \xi$ for some random variable ξ which does not depend on n.

If we take the CDF of T_n we get

$$\begin{aligned} P(T_n \leq t) &= P(a_n(X_{(n)} - b_n) \leq t) \\ &= P(X_{(n)} \leq \frac{t}{a_n} + b_n) \\ &= (\frac{t}{a_n} + b_n)^n, 0 \leq \frac{t}{a_n} + b_n \leq 1 \end{aligned}$$

with the last equality following from the fact that each $X_i \sim U(0, 1)$.

Setting $a_n = n, b_n = 1$, we have $P(n(X_{(n)} - 1) \leq t) = (1 + \frac{t}{n})^n \to e^t$. This means our ξ has CDF $P(\xi \leq t) = e^t$. Moreover,

$$\begin{aligned} P(-\xi \leq t) &= P(\xi \geq -t) \\ &= 1 - P(\xi \leq -t) \\ &= 1 - e^{-t} \end{aligned}$$

and hence $-\xi \sim Exp(1)$.

One last thing to note is that $n(X_{(n)} - 1) \to^L \xi \Rightarrow n(1 - X_{(n)}) \to^L -\xi = Exp(1)$. The quantity $1 - X_{(n)}$ is the *waiting time*.

Ex. 2: Maximum of Exponential RVs Let $X_1, \ldots, X_n \sim^{i.i.d} Exp(1)$. What is the asymptotic distribution of $X_{(n)}$?

We want to find a_n, b_n such that $T_n = a_n(X_{(n)} - b_n)$ has a stable distribution. Taking the CDF of T_n we get

$$\begin{aligned} P(T_n \leq t) &= P(a_n(X_{(n)} - b_n) \leq t) = P(X_{(n)} \leq \frac{t}{a_n} + b_n) \\ &= (P(X_1 \leq \frac{t}{a_n} + b_n))^n \\ &= (1 - e^{-(\frac{t}{a_n} + b_n)})^n, \frac{t}{a_n} + b_n > 0 \end{aligned}$$

Now we just want to use the result which allows us to eliminate the dependence on n. Rewrite as

$$\begin{aligned} P(T_n \leq t) &= (1 - e^{-(\frac{t}{a_n} + b_n)})^n \\ &= (1 - \frac{e^{-t/a_n}}{e^{b_n}})^n \to e^{-e^{-t}} \end{aligned}$$

where the convergence follows when we have $a_n = 1, b_n = \log(n), t + \log(n) > 0$. Hence we have $X_{(n)} - logn \to^L T, P(T \leq t) = e^{-e^{-t}}$. Such a T is said to have the *Gumbel distribution.*

Ex. 3: Range of Uniform RVs Let $X_1, \ldots, X_n \sim^{i.i.d} U(0,1)$. What is the asymptotic distribution of $R_n = X_{(n)} - X_{(1)}$?

We saw above that $n(1 - X_{(n)}) \to^L Exp(1)$. By exchangability, we would surmise that $n(X_{(1)} - 0) \to^L Exp(1)$. Consider the joint distribution $(nX_{(1)}, n(1 - X_{(n)}))$. We have

$$\begin{aligned} P(nX_{(1)} > x, n(1 - X_{(n)}) > y) &= P(X_{(1)} > \frac{x}{n}, X_{(n)} < 1 - \frac{y}{n}) \\ &= P(\frac{x}{n} < X_{(1)} < X_{(n)} < 1 - \frac{y}{n}), \frac{x}{n} < 1 - \frac{y}{n} \end{aligned}$$

Lecture: Delta Method and Sample Moments

The delta method gives that if we have $X_n \to^L X$ and a function g that at least has continuous derivative in the neighborhood of μ the mean of X_i, then

$$\sqrt{n}(g(X_n) - g(\mu)) \to^L \dot{g(\mu)}X \sim N(0, \dot{g(\mu)}\Sigma(\dot{g(\mu)})^T)$$

Consider the sample variance $S_X^2 = \frac{1}{n}\sum X_i^2 - (\bar{X}_n)^2$. The sample variance is the difference in the average of the squared variables, which we refer to as M_{XX} and the square of the avaerage variables M_X. By the CLT,

$$\sqrt{n}((M_X, M_{XX}) - (\mu, \mu^2 + \sigma^2)) \to^L N(0, \Sigma)$$

Using that result, we can then use the delta method to obtain the distribution of S_X^2.

Asymptotic Distribution of the Pearson Correlation Coefficient

Now consider two sequences of random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. We form

$$\begin{aligned} S_X^2 &= \frac{1}{n}\sum (X_i - \bar{X}_n)^2 \\ S_Y^2 &= \frac{1}{n}\sum (Y_i - \bar{Y}_n)^2 \\ S_{XY} &= \frac{1}{n}\sum (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) \end{aligned}$$

The sample correlation coefficient is given by $R = \frac{S_{XY}}{S_X S_Y}$, whereas the population correlation coefficient is given by $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$.

Note that since $R = g(S_{XY}, S_X, S_Y)$ and $E[R] = \rho$, we can use the delta method to obtain the distribution of R. In particular we need to show that

$$\sqrt{n}((S_X^2, S_Y^2, S_{XY}) - (\sigma_X^2, \sigma_Y^2, \sigma_X\sigma_Y)) \to^L N(0, \Sigma)$$

where Σ is the covariance matrix of (S_X^2, S_Y^2, S_{XY}).

We rewrite the vector of sample covariances as

$$(S_X^2, S_Y^2, S_{XY}) = (M_{XX} - M_X^2, M_{YY} - M_Y^2, M_{XY} - M_X M_Y)$$

If we can then obtain the asymptotic ditribution of the sample moments which determine the sample covariances, we can then use the delta method to obtain the distribution of the sample correlation coefficient.

Let $(X_i, Y_i) \sim N((\mu_1, \mu_2), \Sigma)$, where $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$. Then we have

$$\sqrt{n}(R - \rho) \to^L N(0, (1-\rho^2)^2)$$

We interpret the above to note that the sample correlation coefficient is a better estimate of the population correlation coefficient when the population correlation coefficient is close to 1. This is bizarre, because we end up needing the population correlation coefficient to estimate the distribution of the sample correlation coefficient.

We can get around this via *variance stabilizing transformations*. That is, we seek $g : \mathbb{R} \to \mathbb{R}$ such that the variance term below

$$\sqrt{n}(g(R) - g(\rho)) \to^L N(0, (\dot{g}(\rho)))^2(1-\rho^2)^2)$$

does not depend on ρ.

In particular, pick $\dot{g}(\rho) = \frac{1}{1-\rho^2} \Rightarrow g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$. Then we have

$$\sqrt{n}(\frac{1}{2}\log\frac{1+R}{1-R} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}) \to^L N(0,1)$$

Second-Order Delta Method

Let $g : \mathbb{R} \to \mathbb{R}$ be a function with continuous second derivative in a neighborhood of μ. Then we have

$$g(x) - g(\mu) = g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2$$

We can then use the second-order delta method to obtain the distribution of $g(\bar{X}_n)$.

$$g(\bar{X}_n) - g(\mu) = g'(\mu)(\bar{X}_n - \mu) + \frac{1}{2}g''(\mu)(\bar{X}_n - \mu)^2$$

We effect a transformation of the second order term by multiplying and dividing by n/σ^2. Rewriting, we have

$$n(g(\bar{X}_n) - g(\mu)) \simeq \frac{\sigma^2 g''(\mu)}{2}[(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma^2} - \frac{\sqrt{n}(g'(\mu)}{\sigma g''(\mu)})^2 - \frac{n(g'(\mu)^2}{\sigma^2 g''(\mu)^2}]$$

Note that by the CLT, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma^2} \sim N(0,1)$. Let $\gamma_n = \frac{\sqrt{n}(g'(\mu)}{\sigma g''(\mu)}$. Then we have

$$n(g(\bar{X}_n) - g(\mu)) \simeq \frac{\sigma^2 g''(\mu)}{2}[(N(0,1) - \gamma_n)^2 - \gamma_n^2]$$

The expression $(N(0,1) - \gamma_n)^2$ is known to converge in law to a *non-central chi-squared distribution* with 1 degree of freedom and non-centrality parameter γ_n^2: $(N(0,1) - \gamma_n)^2 \to^L \chi_1^2(\gamma_n)$.

Extreme Order Statistics

Given an iid sample $X_1, \ldots, X_n$, we resort and obtain the order statistics as $X_{(1)} \leq \cdots \leq X_{(n)}$. The *extreme order statistics* are the minimum and maximum, or the next to minimum and next to maximum, etc.

Recall a useful calculus fact: $(1 + \frac{c}{n})^n \to e^c$. More generally, if $c_n \to c, (1 + \frac{c_n}{b_n})_n^b \to e^c$.

Ex. 1: Let $X_1, \ldots, X_n \sim U(0, 1)$. We start with the order statistics of the uniform distribution as under monotonic transformations, the order statistics of any transformation of the uniform are the same as the original order statistics.

What is the asymptotic distribution of the maximum $X_{(n)}$? Guess to start that $X_{(n)} \to^p 1$. Then we need $P(X_{(n)} - 1 > \epsilon)$ to go to 0 for any $\epsilon > 0$. We have

$$P(X_{(n)} - 1 > \epsilon) = P(X_{(n)} < 1 - \epsilon) = \prod_{i=1}^{n} P(X_i < 1 - \epsilon) = (1 - \epsilon)^n \to 0$$

By analogy to the asymptotic distribution of the sample mean, although $\bar{X}_n \to^p \mu \Rightarrow P(\bar{X}_n - \mu > \epsilon) \to 0$, we apply a stabilizing transformation by scaling the difference by $\sqrt{n}$. This gives us a non-degenerate asymptotic distribution for the sample mean: $\sqrt{n}(\bar{X}_n - \mu) \to^L N(0, \sigma^2)$.

We therefore seek a_n such that $a_n(1 - X_{(n)}) \to^L F$ where F is some non-degenerate distribution. Consider the CDF $F(t) = P(a_n(1 - X_{(n)}) \leq t)$. We have

$$F(t) = P(1 - X_{(n)} \leq \frac{t}{a_n}) = P(X_{(n)} \geq 1 - \frac{t}{a_n}) = 1 - P(X_{(n)} \leq 1 - \frac{t}{a_n}) = 1 - (1 - \frac{t}{a_n})^n$$

If we let $a_n = n$, then we get $F(t) = 1 - (1 - \frac{t}{n})^n \to 1 - e^{-t}$. This is the CDF of an exponential distribution with rate 1.

Lecture: Consistency and Asymptotics of the MLE

Consider an iid sample $X_1, \ldots, X_n \sim f(x|\theta_0)$ with $\theta_0 \in \Theta \subseteq \mathbb{R}$. We have the *likelihood function* $L_n(\theta) = f(X_1, \ldots, X_n|\theta) = \prod_{i=1}^n f(X_i|\theta)$ under independence. We also have the *log-likelihood function* $l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^n \log f(X_i|\theta)$.

By the strong law of large numbers, $\frac{1}{n} l_n(\theta) \to^p E_{\theta_0}[\log f(X|\theta)] = \int f(X_i|\theta) f(X_i) dX = g(\theta|\theta_0)$.

The *maximum likelihood estimate* $\hat{\theta}_n$ is the value of θ that maximizes the (log-)likelihood function. We have $\hat{\theta}_n = \arg\max_{\theta \in \Theta} L_n(\theta)$. We also have $\hat{\theta}_n = \arg\max_{\theta \in \Theta} l_n(\theta)$.

By contrast, the true parameter $\theta_0 = \arg\max_{\theta \in \Theta} E_{\theta_0}[\log f(X|\theta)] = \arg\max_{\theta \in \Theta} g(\theta|\theta_0)$.

Take $E_{\theta_0}[\log \frac{f(X|\theta)}{f(X|\theta_0)}]^n$. We have

$$E_{\theta_0}[\log \frac{f(X|\theta)}{f(X|\theta_0)}]^n = g(\theta|\theta_0) - g(\theta_0|\theta_0)$$

By Jensen's inequality, $E_{\theta_0}[\log \frac{f(X|\theta)}{f(X|\theta_0)}] \leq \log E_{\theta_0}[\frac{f(X|\theta)}{f(X|\theta_0)}]$. We have

$$E_{\theta_0}[\frac{f(X|\theta)}{f(X|\theta_0)}] = \int \frac{f(X|\theta)}{f(X|\theta_0)} f(X|\theta_0) dX = \int f(X|\theta) dX = 1$$

Therefore, $E_{\theta_0}[\log \frac{f(X|\theta)}{f(X|\theta_0)}] \leq 0$, which implies that $g(\theta|\theta_0) \leq g(\theta_0|\theta_0)$ for all θ, and so θ_0 maximizes g.

Consistency of the MLE

To review, we have the following:

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} l_n(\theta) = \arg\max_{\theta\in\Theta} \sum \log f(X_i|\theta)$$
$$\theta_0 = \arg\max_{\theta\in\Theta} E_{\theta_0}[\log f(X|\theta)] = \arg\max_{\theta\in\Theta} g(\theta|\theta_0)$$
$$\frac{1}{n}\sum logf(X_i|\theta) \to^p E_{\theta_0}[\log f(X|\theta)] = g(\theta|\theta_0)$$

That is, for any fixed θ, the log-likelihood function $l_n(\theta)$ converges in probability to the expected value of the log-likelihood function under the true parameter θ_0.

The remaining question is does the maximizer of the log-likelihood function converge to the maximizer of $g(\theta|\theta_0)$? The answer is the *uniform strong law of large numbers*:

$$\sup_{\theta\in\Theta} |\frac{1}{n}\sum logf(X_i|\theta) - g(\theta|\theta_0)| \to^p 0$$

With the uniform SLLN, we can show the $\hat{\theta}_n \to^p \theta_0$.

Asymptotic Distribution of the MLE

The SLLN shows that the sample mean converges to the population mean. The *central limit theorem* shows that the sample mean is asymptotically normally distributed for large n. We need to combine these somehow to understand the asymptotic distribution of the MLE.

We need a property of the log-likelihood function: let $l_n(\theta)$ be twice differentiable and have continuous second derivative in the neighborhood of θ_0. Note that $l_n'(\hat{\theta}_n) = 0$ as the MLE maximizes the log-likelihood function. Then by the mean-value theorem there exists $\tilde{\theta}_n \in (\hat{\theta}_n, \theta_0)$ such that

$$l_n'(\hat{\theta}_n) - l_n'(\theta_0) = l_n''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

Since $\hat{\theta}_n$ maximizes the log-likelihood, we should be able to write the gap between the MLE and the true parameter as

$$\hat{\theta}_n - \theta_0 = -\frac{l_n'(\theta_0)}{l_n''(\tilde{\theta}_n)}$$

The next step is to use the Slutsky theorems: the LHS converges to a random variable, but the RHS converges to a constant, so we can use Slutsky to get a limiting distribution for the MLE.

Recall that the RHS of the above expression is also called the *score function*:

$$l_n'(\theta_0) = \frac{d}{d\theta}\sum_i^n \log f(X_i|\theta_0) = \sum_i^n \frac{f'(X_i|\theta_0)}{f(X_i|\theta_0)}$$

The score function is the derivative of the log-likelihood function with respect to θ.

We also have that $E_{\theta_0}[l_n'(\theta_0)] = 0$. Moreover, by CLT,

$$\sqrt{n}(\frac{1}{n}l_n'(\theta_0) - 0) \to^L N(0, I(\theta_0)), I(\theta_0) = var(\frac{f_{\theta_0}'(X_i)}{f_{\theta_0}(X_i)})$$

The variance of the score function $I(\theta)$ is called the *Fisher information*. Two more useful properties of the Fisher information are that $I(\theta) = -E_\theta[l_n''(\theta)]$ and that $I_n(\theta) = nI(\theta)$.

Thus we have the asymptotic distribution of the numerator of the RHS in the expression for the gap between MLE and true parameter. For the denominator, we have that $\frac{1}{n}l_n''(\theta_0) \to^p E_{\theta_0}[l_n''(\theta)] = -I(\theta_0)$.

Finally, we apply the continuous mapping theorem to get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}(\frac{1}{n}l_n'(\theta_0) - 0)}{\frac{1}{n}l_n''(\theta_0)} = \frac{1}{-\frac{1}{n}l_n''(\theta_0)}\sqrt{n}(\frac{1}{n}l_n'(\theta_0) - 0)$$
$$\to^{p,L} \frac{1}{I(\theta_0)}N(0, I(\theta_0)) = N(0, \frac{1}{I(\theta_0)})$$

Asymptotic Distribution of the MLE for Bernoulli RVs Let $X_1, \ldots, X_n$ be iid Bernoulli(p) random variables. Our MLE is $\hat{p} = \frac{1}{n}\sum X_i$. By the CLT,

$$\sqrt{n}(\hat{p} - p) \to^L N(0, p(1-p))$$

Our likelihood function is $f_p(X_i) = p^{X_i}(1-p)^{1-X_i}$. The log-likelihood function is $l_n(p) = \sum X_i \log p + (n - \sum X_i)\log(1-p)$. The first two derivatives of the likelihood functions are

$$\frac{d}{dp}f_p(X_i) = \frac{X_i}{p} - \frac{1-X_i}{1-p}$$
$$\frac{d^2}{dp^2}f_p(X_i) = -\frac{X_i}{p^2} - \frac{1-X_i}{(1-p)^2}$$

This gives the score function $I(p) = -E_p[l_n''(p)] = -E_p[-\frac{X_i}{p^2} - \frac{1-X_i}{(1-p)^2}] = \frac{1}{p(1-p)}$. Thus the asymptotic distribution of the MLE is

$$\sqrt{n}(\hat{p} - p) \to^L N(0, \frac{1}{p(1-p)})$$

Lecture: Extreme Order Statistics, Sample Quantiles

Recall our earlier worked examples:

1. $X_1, \ldots, X_n \sim^{iid} U(0,1)$ then $n(1 - X_{(n)}) \to^L Exp(1)$.
2. $X_1, \ldots, X_n \sim^{iid} Exp(1)$ then $(X_{(n)} - \log n) \to^L$ Gumbel.

Ex. 3: Range of Uniform RVs

Now let $X_1, \ldots, X_n \sim^{iid} U(0,1)$ again. What about the distribution of the range, $R_n = X_{(n)} - X_{(1)}$?

We found the asymptotic distribution of the maximum, and by symmetry we know the minimum has the same distribution $nX_{(1)} \to^L Exp(1)$. However, just knowing the marginals isn't enough - we need the joint distribution of the extreme order statistics.

Take the CDF of the joint with each scaled by n to get

$$F(x,y) = P(nX_{(1)} \le x, n(1 - X_{(n)}) \le y) = P(X_{(1)} \le \frac{x}{n}, X_{(n)} \ge 1 - \frac{y}{n})$$

This last expression implies with the iid assumption that each X_i is in the interval $[\frac{x}{n}, 1 - \frac{y}{n}]$. The probability of this is $(1 - \frac{x}{n} - \frac{y}{n})^n$. Thus we have

$$F(x,y) = (1 - \frac{x}{n} - \frac{y}{n})^n = (1 + \frac{-x-y}{n})^n$$

By the limit equality used before, we know that this converges to $e^{-x-y} = e^{-x}e^{-y}$.

Finally by definition of convergence in law, we have

$$\begin{bmatrix} nX_{(1)} \\ n(1 - X_{(n)}) \end{bmatrix} \to^L \begin{bmatrix} Exp(1) \\ Exp(1) \end{bmatrix}$$

Obviously, the minimum and maximum for a given sample are not independent, but the we have this asymptotic joint distribution which has no dependence between the two order statistics.

Ex. 4: Penultimate Order Statistics

Now let $X_1, \ldots, X_n \sim^{iid} U(0,1)$ again. What about the joint distribution of the penultimate order statistics, (X_{n-1}, X_n)?

We want to calculate

$$P(\begin{bmatrix} n(1-X_{(n-1)}) \\ n(1-X_{(n)}) \end{bmatrix} \leq \begin{bmatrix} x \\ y \end{bmatrix}) = P(1-X_{(n-1)} \leq \frac{x}{n}, 1-X_{(n)} \leq \frac{y}{n})$$

Now need to consider two cases. Case 1: $0 \leq x \leq y \rightarrow 1-x/n > 1-y/n$. Then we have

$$P(1-X_{(n-1)} \leq \frac{x}{n}, 1-X_{(n)} \leq \frac{y}{n}) = P(X_{(n)} < 1-\frac{y}{n}) = P(X_i < 1-\frac{y}{n})^n \rightarrow e^{-y}$$

Case 2: $0 \leq y \leq x \rightarrow 1-x/n < 1-y/n$. Then we have

$$\begin{aligned} P(1-X_{(n-1)} \leq \frac{x}{n}, 1-X_{(n)} \leq \frac{y}{n}) &= P(X_{(n)} < 1-\frac{x}{n}) + P(X_{(n-1)} < 1-\frac{x}{n} \leq X_{(n)} < 1-\frac{y}{n}) \\ &= (1-\frac{x}{n})^n + (1-\frac{x}{n})^{n-1}(1-\frac{y}{n}) \rightarrow e^{-x}(1+x-y) \end{aligned}$$

We can then show that the joint distribution of the penultimate order statistics converges to

$$\begin{bmatrix} n(1-X_{(n-1)}) \\ n(1-X_{(n)}) \end{bmatrix} \rightarrow^L \begin{bmatrix} X+Y \\ X \end{bmatrix}, X, Y \sim^{iid} Exp(1)$$

Sample Quantiles

The *p-th population quantile* is defined as the value x_p such that $F(x_p) = p$. If $X \sim F$, then $P(X \leq x_p) = p$.

For the sample quantiles, start with a sorted sample $X_{(1)} \leq \cdots \leq X_{(n)}$. The *p-th sample quantile* is defined as the order statistic $X_{(\lceil np \rceil)}$.

We want to take advantage of the theory of change of variables with different RVs. Let $X \sim F, P(X = \alpha) = f_X(\alpha)$. Let $Y = g(X), X = h(Y), h = g^{-1}$. Then the PDF of Y is given as

$$f_Y(y) = f_X(h(y))|\det(J(h(y))|$$

where $J(h(y))$ is the Jacobian of the transformation.

Lemma: Let $Y_1, \ldots, Y_{n+1} \sim^{iid} Exp(1)$. Define partial sums $S_0 = 0, S_1 = Y_1, \ldots, S_n = Y_1 + \cdots + Y_n$. Then the joint distribution of the fractional partial sums is given as

$$(\frac{S_1}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}} | S_{n+1}) = (U_{(1)}, \ldots, U_{(n)})$$

where $U_1, \ldots, U_n$ are the order statistics of a sample from $U(0,1)$.

The asymptotic distributions of the p-th, qth quantiles of samples from the uniform are given by

$$\sqrt{n}(\begin{bmatrix} U_{(i)} \\ U_{(j)} \end{bmatrix} - \begin{bmatrix} p \\ q \end{bmatrix}) \rightarrow^L N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} p(1-p) & p(1-q) \\ p(1-q) & q(1-q) \end{bmatrix})$$

Let X_p, X_q be the p-th and q-th sample quantiles of a sample taken from an RV $X \sim F$. Then

$$\sqrt{n}(\begin{bmatrix} X_p \\ X_q \end{bmatrix} - \begin{bmatrix} X_p \\ X_q \end{bmatrix}) \rightarrow^L N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{p(1-p)}{f(X_p)^2} & \frac{p(1-q)}{f(X_p)f(X_q)} \\ \frac{p(1-q)}{f(X_p)f(X_q)} & \frac{q(1-q)}{f(X_q)^2} \end{bmatrix})$$

Ex. 1: Sample Median Let $X_1, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$. The sample median is given as $X_{(\lceil n/2 \rceil)}$. For a normal distribution, the population median is simply μ.

We calculate $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$. Then

$$\sqrt{n}(X_{(\lceil n/2 \rceil)} - \mu) \to^L N(0, \frac{\frac{1}{2}(1-\frac{1}{2})}{f(\mu)^2}) = N(0, \frac{\pi\sigma^2}{2})$$

By contrast, note that $\sqrt{n}(\bar{X} - \mu) \to^L N(0, \sigma^2)$. Therefore we say that $\bar{X}_n$ is more *asymptotically efficient* than $X_{(\lceil n/2 \rceil)}$ for estimating μ.

Lecture 3/4: U-Statistics

Consider a sample $X_1, \ldots, X_n$ of iid random variables with common distribution F. A V-statistic is a statistic of the form

$$V_n = E_{\hat{F}_n}[g(Y_1, \ldots, Y_m)]$$

This is a biased statistic for $m > 1$ due to double counting. The U-statistic is a statistic of the form

$$U_n = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} g(X_{i_1}, \ldots, X_{i_m})$$

Ex. Wilcoxon rank-sum test Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent samples from two populations with CDFs F and G. The Wilcoxon rank-sum test uses a statistic given by

$$W = \frac{n(n-1)}{m} + \sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i < Y_j)$$

The goal here is to test the mixing of the two samples. If the two samples are from the same distribution, then the ranks of the two samples should be mixed together.

Consider the extreme case where $Y_{(1)}, \ldots, Y_{(n)}, X_{(1)}, \ldots, X_{(m)}$ are all distinct. Then $W = \frac{n(n-1)}{m}$, as none of the indicators fire.

Lecture 3/6: Statistical Functionals and the Bootstrap

Recall that the bootstrap is a two-stage process to perform inference on a parameter $\theta = T(F)$ of a distribution from a sample $X_1, \ldots, X_n \sim^{iid} F$. We obtain an estimator $\hat{\theta}_n$.

We also obtain the empirical CDF $\hat{F}_n$, which leads to the *plug-in estimator* $\hat{\theta}_n = T(\hat{F}_n)$.

To evaluate this estimator we define the *performance metric* $\lambda_n(F)$, which could be any one of several measures - the bias or variance of the estimator, the mean squared error, etc. Let $\lambda_n(F) = E_F[\hat{\theta}_n] - \theta$. We then estimate $\lambda_n(F)$ by $\lambda_n(\hat{F}_n)$.

Our two stages are as follows:

1. Estimate $\lambda_n(F)$ by $\lambda_n(\hat{F}_n)$.
2. Approximate $\lambda_n(\hat{F}_n)$ by $\lambda^*_{B,n}$, which is obtained via creating new bootstrap samples

$$X^*_{11}, \ldots, X^*_{1n} \sim^{iid} \hat{F}_n$$
$$\vdots$$
$$X^*_{B1}, \ldots, X^*_{Bn} \sim^{iid} \hat{F}_n$$

$$