STAT 581: Advanced Theory of Statistical Inference I **Autumn 2026**

Lecture 2: Introduction of large sample theory

Instructor: Yen-Chi Chen

⊙ We thank previous instructors: Jon Wellner, Alex Luedtke, Fang Han, and Andrea Rotnitzky.

⊙ Some of this lecture notes are based on the following book:

> [van der Vaart] Van der Vaart, A. W. (2000). Asymptotic statistics (Vol. 3). Cambridge university press.

In particular, Chapters 2 and 3 are useful references.

In Statistics, we are often facing the problem of estimation or inference. The problem setup is as follows. We observation IID random variables $X_1, \cdots, X_n$ from an unknown (cumulative) distribution F. This is often denoted as $X_1, \cdots, X_n \sim F$. We want to make inference about some characteristic of the underlying distribution function F. For instance, we may want to know the mean of the distribution function F, $\mu(F) = \int x dF(x)$. When we place a parametric model on the distribution, we can write F_θ and we are often interested in estimating the underlying parameter θ.

A *statistic* is a function of the data, which can be expressed as $f(X_1, \cdots, X_n)$. An *estimator* is a statistic that is used to estimate a parameter of interest. For instance, the sample mean $\bar{X}_n$ is an estimator of the population mean (mean of the distribution that generates our data).

How do we know if an estimator is good? Notice that the estimator is a function of n random variables, so it changes with respect to sample size n. Therefore, an estimator can be viewed as a sequence (of random variables) indexed by the sample size. In mathematics, we often use the concept of convergence as a way to argue a sequence is useful (at least in the asymptotic sense). So a natural way to argue that an estimator is useful is to study its convergence. But we immediately face a problem: an estimator is a statistic, which is a function of random variables, so an estimator is a random variable. The conventional concept of convergence of (a sequence of) numbers is not useful here. Thus, we need a new set of convergence.

2.1 Modes of convergence

Let $\{X_n\}$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{B}, P)$. Note that each random variable $X_n \in \mathbb{R}^d$ can be multivariate (formally it should be called a random vector).

Definition 2.1 (Convergence almost surely) *The sequence of random variables $\{X_n\}$ converge almost surely to another random variable Z if*

$$P\left(\lim_{n\to\infty} \|X_n - Z\| = 0\right) = 1.$$

We denote this as $X_n \xrightarrow{a.s.} Z$.

Definition 2.2 (Convergence in probability) *The sequence of random variables $\{X_n\}$ converge in probability to another random variable Z if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} P\left(|X_n - Z| > \epsilon\right) = 0.$$

We denote this as $X_n \xrightarrow{P} Z$.

In most problems of Statistics and Machine Learning, we only need to use the concept of convergence in probability. But some rigorous mathematical result will require almost sure convergence.

Note that an estimator $\hat{\theta}_n$ for the parameter θ is *statistically consistent* or we simply called it a consistent estimator if $\hat{\theta}_n \xrightarrow{P} \theta$.

Example 2.3 (Convergence in probability but not almost surely) *Let* $U \sim \mathsf{Uni}[0,1]$ *be a uniform random variable. Now we define the sequence of random variables as follows.*

$$X_1 = 1, \quad X_2 = I\left(0 < U < \frac{1}{2}\right), \quad X_3 = I\left(\frac{1}{2} \le U < 1\right),$$
$$X_4 = I\left(0 \le U < \frac{1}{3}\right), \quad X_5 = I\left(\frac{1}{3} \le U < \frac{2}{3}\right), \quad X_6 = I\left(\frac{2}{3} \le U < 1\right),$$
$$\dots$$

Let $Z = 0$ *be a point mass at* 0. *Then you can easily show that* $X_n \xrightarrow{P} Z = 0$. *However,* X_n *does not converge almost surely to* 0.

Here is a useful theorem about the relation of the above two convergences.

Theorem 2.4 *The following are true:*

1. $X_n \xrightarrow{a.s.} Z \Rightarrow X_n \xrightarrow{P} Z$.
2. $X_n \xrightarrow{a.s.} Z$ *and* $X_n \xrightarrow{a.s.} Y$, *then* $Y \overset{a.s.}{=} Z$[1].
3. $X_n \xrightarrow{P} Z$ *and* $X_n \xrightarrow{P} Y$, *then* $Y \overset{a.s.}{=} Z$.

2.1.1 Convergence in distribution

Convergence in distribution is a weaker notion than convergence in probability.

We start with two popular definitions of it and then we will show that these two (along with many other definitions) are equivalent.

Definition 2.5 *A sequence of random variable* X_n *converges in distribution to* Z, *denoted as* $X_n \xrightarrow{d} Z$, *if for all bounded continuous function* $f : \mathbb{R}^d \to f$,

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(Z)).$$

Let $F_{X_n}(t) = P(X_{n,1} \le t_1, \cdots, X_{n,d} \le t_d)$ *be the multivariate CDF of* X_n. *We may equivalently say that* $X_n \xrightarrow{d} Z$ *if for any continuous point* t *in* F_Z, *we have*

$$F_{X_n}(t) \to F_Z(t).$$

[1] This means that $P(Y = Z) = 1$.

The convergence in distribution is very useful for constructing confidence interval or perform a hypothesis test. This is because we do not need the limiting random variable Z to be numerically close to X_n ($\|X_n - Z\|$ being small is a notion of convergence in probability). All we need is that the distribution of X_n to similar to the distribution of Z. So we can utilize the distribution of Z to infer how X_n will be like. In statistical applications, we often have $X_n = \sqrt{n}(\hat{\theta}_n - \theta)$ and utilize the central limit theorem to establish convergence in distribution.

The convergence in distribution is also called convergence *in law* or convergence *weakly*.

Now we introduce the famous Portmanteau theorem, which includes 10 different equivalent definitions of convergence in distribution.

Theorem 2.6 (Portmanteau) *The following are equivalent (definition of convergence in distribution):*

1. $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(Z))$ *for all bounded continuous function* f.
2. *For any continuous point* t *in* F_Z, *we have* $F_{X_n}(t) \to F_Z(t)$.
3. $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(Z))$ *for all bounded, Lipschitz-continuous function* f.
4. $\limsup_n \mathbb{E}(f(X_n)) \leq \mathbb{E}(f(Z))$ *for every upper semicontinuous* f *that is bounded from above.*
5. $\liminf_n \mathbb{E}(f(X_n)) \geq \mathbb{E}(f(Z))$ *for every lower semicontinuous* f *that is bounded from below.*
6. $\limsup_n P(X_n \in A) \leq P(Z \in A)$ *for any closed set* A.
7. $\liminf_n P(X_n \in A) \geq P(Z \in A)$ *for any closed set* A.
8. $\lim_n P(X_n \in A) \to P(Z \in A)$ *for any continuous set* A *in the sense that* $P(Z \in \partial A) = 0$.
9. **Levy's continuity theorem.** *For all* $t \in \mathbb{R}^d$, $\mathbb{E}(e^{it^T X_n}) \to \mathbb{E}(e^{it^T Z})$.
10. **Cramer-Wald's device/theorem.** *For all* $t \in \mathbb{R}^d$, $t^T X_n \xrightarrow{d} t^T Z$.

Convergence in distribution is a weaker notion than convergence in probability.

Theorem 2.7 *We have the following results:*

- $X_n \xrightarrow{P} Z \Rightarrow X_n \xrightarrow{d} Z$.
- *If* $X_n \xrightarrow{d} Z$ *and* $X_n \xrightarrow{d} Y$ *then* $Z \overset{d}{=} Y$. $Z \overset{d}{=} Y$ *means that* Z *and* Y *have the same distribution function.*

Example 2.8 (Convergence in distribution does not imply convergence in probability) *This is a trivial case but we still offer an example. Consider* $X_1, \cdots, X_n \sim \mathsf{Ber}(0.5)$ *and let* Z *be a random variable from* $N(0, 1/4)$ *independent of any* X_i. *Then clearly, the quantity* $\sqrt{n}(\bar{X}_n - 0.5) \xrightarrow{d} Z$ *by central limit theorem. However,* $\sqrt{n}(\bar{X}_n - 0.5)$ *does not converge to* Z *in probability. In fact, the difference* $|\sqrt{n}(\bar{X}_n - 0.5) - Z|$ *is asymptotically the difference between two independent Gaussian* $N(0, 1/4)$.

The above example highlights the difference between convergence in distribution versus in probability. Convergence in distribution only requires the distribution functions to match while convergence in probability require the random variables' realization to match. So in most statistical applications, we are often working with $X_n \xrightarrow{P} c$ to a non-random quantity (scalar/vector/matrix).

Note that if $X_n \xrightarrow{d} c$ for a non-random c, then $X_n \xrightarrow{P} c$. However, they do not imply convergence almost surely (Example 2.3).

2.1.2 Convergence in L_p

Consider a sequence of univariate random variables $\{X_n\}$ and a another random variable Z. We may define a convergence in terms of expectation as follows.

Definition 2.9 *X_n converges to Z in L_p-norm, denoted as $X_n \xrightarrow{L_p} Z$, if*

$$\mathbb{E}[|X_n - Z|^p] \to 0.$$

A common application is the convergence in L_2 for an estimator $\hat{\theta}_n$ to its target parameter θ since the quantity $\mathbb{E}(|\hat{\theta}_n - \theta|^2)$ is the mean square error.

Theorem 2.10 *Let $s > p$. Then we have*

- *$X_n \xrightarrow{L_s} Z$ implies $X_n \xrightarrow{L_p} Z$.*
- *$X_n \xrightarrow{L_p} Z$ implies $X_n \xrightarrow{P} Z$ when $p \geq 1$ (via the Markov inequality).*
- *If $X_n \xrightarrow{L_p} Z$ and $X_n \xrightarrow{L_p} Y$, then $Z \overset{a.s.}{=} Y$.*

Example 2.11 (Convergence in probability but not in L_1) *Consider the sequence of random variables X_n such that*

$$X_n = \begin{cases} n^2, & \text{with a probability of } \frac{1}{n} \\ 0, & \text{with a probability of } 1 - \frac{1}{n}. \end{cases}$$

Then you can easily show that $X_n \xrightarrow{P} 0$ but does not converge in L_1 to 0 since the expectation diverges.

2.2 Continuous mapping theorem and Slutsky's lemma

Theorem 2.12 (Continuous mapping theorem) *Consider a function $f : \mathbb{R}^d \to \mathbb{R}^k$ that is continuous on every point in the set C such that $P(Z \in C) = 1$. Then*

1. *If $X_n \xrightarrow{a.s.} Z$, then $f(X_n) \xrightarrow{a.s.} f(Z)$.*
2. *If $X_n \xrightarrow{P} Z$, then $f(X_n) \xrightarrow{P} f(Z)$.*
3. *If $X_n \xrightarrow{d} Z$, then $f(X_n) \xrightarrow{d} f(Z)$.*

Theorem 2.13 (Slutsky's lemma) *Assume that $X_n \xrightarrow{d} Z$ and we have another sequence $Y_n \xrightarrow{P} c \in \mathbb{R}^k$. Then we have*

1. *When $k = d$, we have $X_n + Y_n \xrightarrow{d} Z + c$.*
2. *When $k = 1$, we have $Y_n X_n \xrightarrow{d} c \cdot Z$.*
3. *When $k = 1$ and $c \neq 0$, $X_n / Y_n \xrightarrow{d} Z/c$.*

2.3 Law of large numbers and simple central limit theorem

Consider $X_1, \cdots, X_n \sim F$, where F is some distribution function. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}(X_1)$.

Theorem 2.14 (Law of large numbers (LLN)) *If $\mathbb{E}|X_i| < \infty$, then $\bar{X}_n \xrightarrow{a.s.} \mu$. Therefore, $\bar{X}_n \xrightarrow{P} \mu$.*

Formally, Theorem 2.14 is called the strong law of large numbers, which implies the weak law of large numbers ($\bar{X}_n \xrightarrow{P} \mu$). The weak law of large number requires a slightly weak condition that the characteristic function of X_1 is differentiable at 0.

Theorem 2.15 (Multivariate central limit theorem (CLT)) *Assume that $\mathbb{E}(\|X_1\|^2) < \infty$ and define $\Sigma = \mathsf{Cov}(X_1) = \mathbb{E}((X_1 - \mu)(X_1 - \mu)^T)$ to be the covariance matrix. Then*

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

Proof: Here we show the proof of the multivariate central limit theorem using univariate central limit theorem plus the Cramer-Wald's device in Theorem 2.6.

Due to Cramer-Wald's device, we only need to show that for any $t \in \mathbb{R}^d$, we have

$$\sqrt{n}(t^T\bar{X}_n - t^T\mu) \xrightarrow{d} N(0, t^T\Sigma t).$$

Let $\bar{Y}_n = t^T\bar{X}_n$. Then it is easy to see that $\mathbb{E}(\bar{Y}_n) = t^T\mu$. So we only need to check the second moment of $Y_1 = t^TX_1$.

$$\begin{aligned}
\mathbb{E}(Y_1^2) &= t^T\mathbb{E}\left(X_1X_1^T\right)t \\
&= \sum_{j=1}^d\sum_{k=1}^d t_jt_k\mathbb{E}(X_{1,j}X_{1,k}) \\
&\le \sum_{j=1}^d\sum_{k=1}^d |t_jt_k|\sqrt{\mathbb{E}(X_{1,j}^2)\mathbb{E}(X_{1,k}^2)} \qquad \text{(Cauchy-Schawrz)} \\
&\le \sum_{j=1}^d\sum_{k=1}^d |t_jt_k|\mathbb{E}(\|X_1\|^2) < \infty.
\end{aligned}$$

Thus, the second moment is finite.

Now we analyze the variance of Y_1:

$$\mathsf{Var}(Y_1) = \mathsf{Var}(t^TX_1) = t^T\mathsf{Cov}(X_1)t = t^T\Sigma t,$$

which is the desired quantity.

Thus, by the Cramer-Wald's device, we have completed the proof.

■

Remark 2.16 (Informal notations) *Sometimes, we will informally write*

$$\bar{X}_n \approx N\left(\mu, \frac{1}{n}\Sigma\right)$$

when we want to say the central limit theorem. The use of $\approx$ *notation ease some derivations but it is not a formal mathematical term.*

WARNING: You should NEVER write something like $\bar{X}_n \xrightarrow{d} N\left(\mu, \frac{1}{n}\Sigma\right)$. *This is a wrong expression since the right-hand-side after the limit CANNOT depend on* n.

2.4 Advanced central limit theorems

There are a number of variants of central limit theorems including cases for dependent variables and data from a changing distribution (changing with respect to sample size); see Chapter 3.4 of

Durrett, R. (2019). Probability: theory and examples (Vol. 49). Cambridge university press.

Here we present an advanced central limit theorem under a setup called *triangular array* since it is useful in various statistical applications.

Theorem 2.17 (Lindeberg-Feller and Lyapunov CLT) *For each* n, *let* $X_{n,1}, \cdots, X_{n,n}$ *be independent random variables in* $\mathbb{R}$ *such that each* $\mathbb{E}(X_{n,i}) = \mu_{n,i}$ *and variance* $\mathsf{Var}(X_{n,i}) = \sigma^2_{n,i} < \infty$. *Assume that* $\sigma^2_n = \sum_{i=1}^n \sigma^2_{n,i} > 0$. *Let* $Y_{n,i} = (X_{n,i} - \mu_{n,i})/\sigma_n$.

Then if either one of the following conditions holds:

- **(Lindeberg condition)** *for any* $\epsilon > 0$, *we have* $\sum_{i=1}^n \mathbb{E}\left[Y^2_{n,i} I(|Y_{n,i}| > \epsilon)\right] \to 0$,
- **(Lyapunov condition)** $\sum_{i=1}^n \mathbb{E}\left[Y^{2+\delta}_{n,i}\right] \to 0$ *for some* $\delta > 0$,

we have $\sum_{i=1}^n Y_{n,i} \xrightarrow{d} N(0,1)$.

Note that there is a multivariate version of Theorem 2.17; see Proposition 2.27 of [van der Vaart]. The multivariate version can be obtained via the use of Cramer-Wald's device.

Theorem 2.17 allows every observation to have its own mean and variance so data is not necessarily from an identical distribution. While Theorem 2.17 may seem a bit strange at the first glance since there is no $\sqrt{n}$ nor a division by n in the final result, the dependency on n is implicitly inside $\sigma^2_n = \sum_{i=1}^n \sigma^2_{n,i} > 0$. Under the IID setup, $\sigma^2_{n,i} = \sigma^2_0$, so $\sigma^2_n = n\sigma^2_0$ and then $Y_{n,i} = \frac{X_{n,i} - \mu}{\sqrt{n} \cdot \sigma_0}$, so we recover the conventional CLT setup.

Example 2.18 (Simple linear regression with a fixed design) *Now we consider a simple linear regression where our data consists of independent random vectors*

$$(Y_1, x_1), \cdots, (Y_n, x_n),$$

such that $x_1, \cdots, x_n$ *are non-random (fixed design) and each* Y_i *is generated via*

$$Y_i = \beta_0 + \beta_1^T + e_i, \tag{2.1}$$

where $e_1, \cdots, e_n$ are IID errors with a symmetric distribution (mean 0) and variance $\sigma^2 < \infty$.

Let $\hat{\beta} \in \mathbb{R}^2$ be the least-square estimator of $\beta = (\beta_0, \beta_1)^T$. We want to know what conditions we need for the design points $x_1, \cdots, x_n$ so that we have $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$ for some Σ.

Let $\mathbf{Y}_n = (Y_1, \cdots, Y_n)^T$ be the response vector and $\mathbf{X}_n$ be the design matrix

$$\mathbf{X}_n = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

It is well-known that the least square estimator $\hat{\beta}$ has the following closed-form:

$$\hat{\beta} = [\mathbf{X}_n^T \mathbf{X}_n]^{-1} \mathbf{X}_n^T \mathbf{Y}_n.$$

Using the generative model in equation (2.1), we have $\mathbf{Y}_n = \mathbf{X}_n^T \beta + \mathbf{E}_n$, where $\mathbf{E}_n = (\epsilon_1, \cdots, \epsilon_n)^T$, so we have

$$\hat{\beta} = [\mathbf{X}_n^T \mathbf{X}_n]^{-1} \mathbf{X}_n^T [\mathbf{X}_n^T \beta + \mathbf{E}_n] = \beta + [\mathbf{X}_n^T \mathbf{X}_n]^{-1} \mathbf{X}_n^T \mathbf{E}_n.$$

A simple rearrangement leads to

$$[\mathbf{X}_n^T \mathbf{X}_n]^{1/2} (\hat{\beta} - \beta) = [\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} \mathbf{X}_n^T \mathbf{E}_n \tag{2.2}$$

and our goal is to show that the right-hand-sided converges in distribution to $N(0, \sigma^2 \mathbf{I}_2)$.

We will apply the Cramer-Wald's device. Now pick any $t \in \mathbb{R}^2$ that $t \neq 0$ and denote $a_{n,i}$ to be the i-th column of $[\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} \mathbf{X}_n^T \in \mathbb{R}^{2 \times n}$, i.e.,

$$[\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} \mathbf{X}_n^T = \begin{bmatrix} a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}. \tag{2.3}$$

We immediately have

$$t^T [\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} \mathbf{X}_n^T \mathbf{E}_n = \sum_{i=1}^n [t^T a_{n,i}] e_i.$$

As you can see, the quantity $[t^T a_{n,i}] e_i$ behave like $X_{n,i}$ in Theorem 2.17. So the its variance is

$$\sigma_{n,i}^2 = \mathsf{Var}([t^T a_{n,i}] e_i) = [t^T a_{n,i}]^2 \sigma^2.$$

Thus,

$$\sigma_n^2 = \sum_{i=1}^n \sigma_{n,i}^2 = \sigma^2 \sum_{i=1}^n [t^T a_{n,i}]^2 = \sigma^2 t^T [\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} [\mathbf{X}_n^T \mathbf{X}_n] [\mathbf{X}_n^T \mathbf{X}_n]^{-1/2} t = \sigma^2 \|t\|^2. \tag{2.4}$$

To obtain the variable $Y_{n,i}$ in in Theorem 2.17, we define

$$Z_{n,i} = \frac{[t^T a_{n,i}] e_i}{\sigma_n} = \frac{[t^T a_{n,i}] e_i}{\sigma \|t\|}.$$

Recall that the Lindeberg condition is: for any $\epsilon > 0$,

$$\sum_{i=1}^n \mathbb{E}[Z_{n,i}^2 I(|Z_{n,i}| > \epsilon)] \to 0.$$

Since $t^T a_{n,i}$ is non-random, we have

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}[Z_{n,i}^2 I(|Z_{n,i}|>\epsilon)] &= \sum_{i=1}^{n}\frac{[t^T a_{n,i}]^2}{\sigma^2\|t\|^2}\mathbb{E}\left[e_i^2 I\left(\left|\frac{t^T a_{n,i}}{\|t\|}\right||e_i|>\sigma\epsilon\right)\right]\\ &\le \sum_{i=1}^{n}\frac{[t^T a_{n,i}]^2}{\sigma^2\|t\|^2}\cdot\max_{j=1,\cdots,n}\mathbb{E}\left[e_j^2 I\left(\left|\frac{t^T a_{n,j}}{\|t\|}\right||e_j|>\sigma\epsilon\right)\right]\\ &\overset{(2.4)}{=} \max_{j=1,\cdots,n}\mathbb{E}\left[e_j^2 I\left(\left|\frac{t^T a_{n,j}}{\|t\|}\right||e_j|>\sigma\epsilon\right)\right].\end{aligned}$$

Thus, a sufficient condition is

$$\max_{j=1,\cdots,n}\mathbb{E}\left[e_j^2 I\left(\left|\frac{t^T a_{n,j}}{\|t\|}\right||e_j|>\sigma\epsilon\right)\right]\to 0.$$

First, note that

$$\frac{t^T a_{n,j}}{\|t\|}\le\|a_{n,i}\|,$$

Therefore,

$$\max_{j=1,\cdots,n}\mathbb{E}\left[e_j^2 I\left(\left|\frac{t^T a_{n,j}}{\|t\|}\right||e_j|>\sigma\epsilon\right)\right]\le\max_{j=1,\cdots,n}\mathbb{E}\left[e_j^2 I\left(\|a_{n,i}\||e_j|>\sigma\epsilon\right)\right],$$

which converges to 0 if

$$\max_{i=1,\cdots,n}\|a_{n,i}\|\to 0. \tag{2.5}$$

Thus, we conclude that if the norm of each column of the matrix $[\mathbf{X}_n^T\mathbf{X}_n]^{-1}\mathbf{X}_n^T$ in equation (2.3) converges to 0 (i.e., equation (2.5) holds), then the Lindeberg condition is satisfied so

$$t^T[\mathbf{X}_n^T\mathbf{X}_n]^{1/2}(\hat\beta-\beta)\xrightarrow{d}N(0,\sigma^2)$$

for any t. Since equation (2.5) does not depend on t, it applies for any t. Thus, by Cramer-Wald's device, we conclude that

$$[\mathbf{X}_n^T\mathbf{X}_n]^{1/2}(\hat\beta-\beta)\xrightarrow{d}N(0,\sigma^2\mathbf{I}_2).$$

2.5 Stochastic order: O_p and o_P notations

Definition 2.19 (Big-O and little-o) *Consider a sequence of numbers a_n (indexed by n).*

- *We write $a_n=o(1)$ if $a_n\to 0$ when $n\to\infty$.*
- *For another sequence b_n indexed by n, we write $a_n=o(b_n)$ if $a_n/b_n=o(1)$.*
- *We write $a_n=O(1)$ if for all large n, there exists a constant C such that $|a_n|\le C$.*
- *For another sequence b_n, we write $a_n=O(b_n)$ if $a_n/b_n=O(1)$.*

Example 2.20 *We have the following results:*

- *Let $a_n=\frac{2}{n}$. Then $a_n=o(1)$ and $a_n=O\left(\frac{1}{n}\right)$.*

- *Let* $b_n = n + 5 + \log n$. *Then* $b_n = O(n)$ *and* $b_n = o(n^2)$ *and* $b_n = o(n^3)$.
- *Let* $c_n = 1000n + 10^{-10}n^2$. *Then* $c_n = O(n^2)$ *and* $c_n = o(n^2 \cdot \log n)$.

Essentially, the big O and small o notation give us a way to compare the leading convergence/divergence rate of a sequence of (non-random) numbers.

The O_P and o_P are similar notations to O and o but are designed for random numbers.

Definition 2.21 (Big-O_P and little-o_P) *Consider a sequence of random variables* X_n.

- *We write* $X_n = o_P(1)$ *if for any* $\epsilon > 0$,

$$P(|X_n| > \epsilon) \to 0$$

  when $n \to \infty$. *Namely,* $P(|X_n| > \epsilon) = o(1)$ *for any* $\epsilon > 0$.
- *Let* a_n *be a nonrandom sequence, we write* $X_n = o_P(a_n)$ *if* $X_n/a_n = o_P(1)$.
- *We write* $X_n = O_P(1)$ *if for every* $\epsilon > 0$, *there exists a constant* C *such that*

$$P(|X_n| > C) \leq \epsilon.$$

- *We write* $X_n = O_P(a_n)$ *if* $X_n/a_n = O_P(1)$.

Example 2.22 *We have the following results:*

- *Let* X_n *be an R.V. (random variable) from an Exponential distribution with* $\lambda = n$. *Then* $X_n = O_P(\frac{1}{n})$
- *Let* Y_n *be an R.V from a normal distribution with mean* 0 *and variance* n^2. *Then* $Y_n = O_P(n)$ *and* $Y_n = o_P(n^2)$.
- *Let* A_n *be an R.V. from a normal distribution with mean* 0 *and variance* $10^{100} \cdot n^2$ *and* B_n *be an R.V. from a normal distribution with mean* 0 *and variance* $0.1 \cdot n^4$. *Then* $A_n + B_n = O_P(n^2)$.

The $X_n = o_P(1)$ is essentially the same as converges in probability.

Proposition 2.23 $X_n \xrightarrow{P} 0$ *if and only if* $X_n = o_P(1)$.

Moreover, if a sequence of random variable converges in distribution to another random variable, then it is $O_P(1)$.

Theorem 2.24 (Prokhorov) *We have the following results.*

1. *If* $X_n \xrightarrow{d} Z$, *then* $X_n = O_P(1)$.
2. *If* $X_n = O_P(1)$, *then there exists a subsequence that converges in distribution.*

The property $X_n = O_P(1)$ is also called *uniformly tightness* in the literature.

Here are some useful properties about o_P and O_P.

Proposition 2.25 *For sequences of random variables X_n we have the following properties:*

- $o_P(1) + o_P(1) = o_P(1)$.
- $o_P(1) + O_P(1) = O_P(1)$.
- $O_P(1)O_P(1) = O_P(1)$.
- $O_P(1)o_P(1) = o_P(1)$.
- $[1 + o_P(1)]^{-1} = O_P(1)$.
- $X_n = o_P(1) \Rightarrow X_n = O_P(1)$.

Moreover, when we couple O and O_P, we have

- $o_P(1) + o(1) = o_P(1)$.
- $o_P(1) + O(1) = O_P(1)$.
- $O_P(1) + o(1) = O_P(1)$.
- $O_P(1) + O(1) = O_P(1)$.
- $o_P(1)o(1) = o_P(1)$.
- $o_P(1)O(1) = o_P(1)$.
- $O_P(1)o(1) = o_P(1)$.
- $O_P(1)O(1) = O_P(1)$.

You can see that the use of O_P and o_P notation is very similar to our conventional addition and multiplication rule. One scenario to be cautious is the second result when we combined o_P and O:

$$o_P(1) + O(1) = O_P(1).$$

Even if the randomness is of a smaller order and the non-random part is of a dominating order, we cannot simply drop o_P and use O. We have to respect the randomness, which may be unbounded.

Example 2.26 (Why $o_P(1) + O(1) \neq O(1)$?) *Consider $X_n \sim N(0, 1/n^2)$ and $a_n = 1$. Clearly, $X_n = o_P(1)$ and $a_n = O(1)$. The addition of them is*

$$X_n + a_n \sim N(1, 1/n^2).$$

This quantity is unbounded, so it CANNOT be $O(1)$. However, for any ϵ, we can easily find a constant $C(\epsilon)$ such that

$$P(|X_n + a_n| > C(\epsilon)) < \epsilon.$$

Therefore, $X_n + a_n = O_P(1)$.

Note that in many statistical literature, we will often write something like $\hat{\theta}_n = O(a_n) + O_P(b_n)$. This means that there exists a non-random quantity η_n and a random quantity W_n such that $\hat{\theta}_n = \eta_n + W_n$ with $\eta_n = O(a_n)$ and $W_n = O_P(b_n)$. A common way to obtain such decomposition is via choosing $\eta_n = \mathbb{E}(\hat{\theta}_n)$ but this is not always the case (sometimes we choose η_n to be the asymptotic bias of $\hat{\theta}_n$).

Example 2.27 (Sample mean) *Consider univariate random variables $X_1, \cdots, X_n \sim F$ for some unknown distribution with mean $\mu = \mathbb{E}(X_1)$ and variance $\sigma^2 = \mathsf{Var}(X_1) < \infty$.*

The LLN implies that $\bar{X}_n = o_P(1)$ due to Proposition 2.23.

The CLT implies that $\bar{X}_n = O_P(1/\sqrt{n})$ due to Prokhorov's theorem.

Example 2.28 (Sample variance) *Consider univariate random variables $X_1, \cdots, X_n \sim F$ for some unknown distribution with mean $\mu = \mathbb{E}(X_1)$ and variance $\sigma^2 = \mathsf{Var}(X_1) < \infty$. We further assume that $\mathbb{E}|X_1|^4 < \infty$.*

Let $S_n^2 \equiv \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ be the sample variance.

Let $M_n = \frac{1}{n}\sum_{i=1}^n X_i^2$ be the empirical second moment.

Clearly, we have

$$S_n^2 = \frac{n}{n-1}\left[M_n - \bar{X}_n^2\right]. \tag{2.6}$$

Since both M_n and $\bar{X}_n$ are average, we can apply CLT to them, which leads to

$$M_n = \mathbb{E}(X_1^2) + O_P(n^{-1/2}), \qquad \bar{X}_n = \mathbb{E}(X_1) + O_P(n^{-1/2}).$$

Thus,

$$\bar{X}_n^2 = \left(\mathbb{E}(X_1) + O_P(n^{-1/2})\right)^2 = \mathbb{E}(X_1)^2 + O_P(n^{-1/2}) + O_P(n^{-1}) = \mathbb{E}(X_1)^2 + O_P(n^{-1/2}).$$

Putting these back to equation (2.6), we conclude that

$$\begin{aligned}
S_n^2 &= \frac{n}{n-1}\left[M_n - \bar{X}_n^2\right] \\
&\quad \frac{n}{n-1}\left[\mathbb{E}(X_1^2) + O_P(n^{-1/2}) - \mathbb{E}(X_1)^2 + O_P(n^{-1/2})\right] \\
&\quad \left(1 + \frac{1}{n-1}\right)\left[\mathbb{E}(X_1^2) + O_P(n^{-1/2}) - \mathbb{E}(X_1)^2 + O_P(n^{-1/2})\right] \\
&= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 + O_P(n^{-1/2}) + \underbrace{\frac{1}{n-1}\left[\mathbb{E}(X_1^2) + O_P(n^{-1/2}) - \mathbb{E}(X_1)^2 + O_P(n^{-1/2})\right]}_{=O_P(n^{-1})} \\
&= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 + O_P(n^{-1/2}) \\
&= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 + o_P(1).
\end{aligned}$$

So $S_n^2 \xrightarrow{P} \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \mathsf{Var}(X_1)$ is a consistent estimator of the population variance.

In the above example, you see the beauty of the O_P and o_P notations–we can simply a lot of terms via dropping the constant in front of each quantity and keeping only the dominating term. Therefore, they have became a daily routine for asymptotic analysis.

2.6 Delta method

In Statistics, we often encounter scenarios where we have shown that for an statistic $W_n \in \mathbb{R}^d$, it has an asymptotic normality toward a fixed point $\omega_0 \in \mathbb{R}^d$

$$\sqrt{n}(W_n - \omega_0) \xrightarrow{d} N(0, \Sigma)$$

or more generally,

$$r_n(W_n - \omega_0) \xrightarrow{d} Z$$

for some random variable Z and a decreasing sequence r_n.

But what we want is not exactly W_n but some smooth transformation of it, i.e., our estimator is $f(W_n)$ for some smooth function f. For instance, we know that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ for general case. But if we are estimating the rate parameter λ of an exponential distribution, our maximum likelihood estimator will be $\hat{\lambda}_n = 1/\bar{X}_n$. Do we still have asymptotic normality of $1/\bar{X}_n$?

The delta method offers a solution to this problem.

For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ at a point ω_0, we require that there is a gradient function $g : \mathbb{R}^d \to \mathbb{R}^d$, denoted as $g = \nabla f$, at ω_0 such that

$$\lim_{\epsilon \to 0} \sup_{h \in \mathbb{R}^d : \|h\| = 1} \frac{|f(\omega_0 + \epsilon h) - f(\omega_0) - \epsilon h^T g(\omega_0)|}{\epsilon} = 0. \tag{2.7}$$

A sufficient condition to equation (2.7) is that f is partially differentiable in the neighborhood of ω_0 and all partial derivatives are continuous at ω_0.

Theorem 2.29 (Delta method) *If f is differentiable at ω_0 and equation (2.7) holds at ω_0, and we have $r_n(W_n - \omega_0) \xrightarrow{d} Z$, then*

- $f(W_n) - f(\omega_0) - (W_n - \omega_0)^T \nabla f(\omega_0) = o_P(r_n^{-1})$,
- $f(W_n) - f(\omega_0) \xrightarrow{d} Z^T \nabla f(\omega_0)$.

Example 2.30 (Asymptotic linearity) *Suppose we have an asymptotic linear estimator*

$$W_n = \frac{1}{n} \sum_{i=1}^{n} w(X_i)$$

such that $\omega_0 = \mathbb{E}(w(X_1))$. Under suitable conditions, $\sqrt{n}(W_n - \omega_0) \xrightarrow{d} N(0, \mathbb{E}(w(X_1)w(X_1)^T))$.

Consider a transformation $f(W_n)$. Then this transformed quantity satisfies

$$f(W_n) - f(\omega_0) = (W_n - \omega_0)^T \nabla f(\omega_0) + o_P(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(w(X_i) - \omega_0)^T \nabla f(\omega_0)}_{=\psi(W_i)} + o_P(n^{-1/2}).$$

Estimators with the property

$$\hat{\theta}_n - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} f(X_i) + o_P(n^{-1/2})$$

is called asymptotic linear estimator. So $f(W_n)$ is an asymptotic linear estimator of $f(\omega_0)$.

Using the Cramer-Wald device, we can easily generalize the delta method to smooth vector-valued function. In this case, the gradient $\nabla f(\omega)$ will be a Jacobian matrix $J_f(\omega) \in \mathbb{R}^{k \times d}$ and equation (2.7) is replaced by

$$\lim_{\epsilon \to 0} \sup_{h \in \mathbb{R}^d : \|h\| = 1} \frac{\|f(\omega_0 + \epsilon h) - f(\omega_0) - \epsilon J_f(\omega_0) h\|}{\epsilon} = 0.$$

Theorem 2.31 (Vector-valued Delta method) *If $f : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at ω_0 and equation (2.7) holds at ω_0, and we have $r_n(W_n - \omega_0) \xrightarrow{d} Z$, then*

- $f(W_n) - f(\omega_0) - J_f(\omega_0)(W_n - \omega_0) = o_P(r_n^{-1})$,
- $f(W_n) - f(\omega_0) \xrightarrow{d} J_f(\omega_0)Z$.

Example 2.32 (Relative risk) *Suppose we observed IID random vectors*

$$(T_1, Y_1), \cdots, (T_n, Y_n) \in \{0, 1\}^2$$

such that $P(T = 1) = \frac{1}{2}$. We want to estimate the relative risk

$$\theta \equiv \frac{P(T = 1|Y = 1)}{P(T = 0|Y = 1)} = \frac{P(T = 1, Y = 1)}{P(T = 0, Y = 1)} = \frac{\mathbb{E}(TY)}{\mathbb{E}((1 - T)Y)}.$$

A natural estimator is to use the empirical proportion ratio:

$$\hat{\theta}_n = \frac{\sum_{i=1}^n T_i Y_i}{\sum_{j=1}^n (1 - T_j)Y_j} = \frac{\hat{\zeta}_{11}}{\hat{\zeta}_{01}},$$

where

$$\hat{\zeta}_{11} = \frac{1}{n}\sum_{i=1}^n T_i Y_i, \qquad \hat{\zeta}_{01} = \frac{1}{n}\sum_{i=1}^n (1 - T_i)Y_i.$$

By the multivariate Central Limit Theorem, we have:

$$\sqrt{n}\left(\begin{pmatrix}\hat{\zeta}_{11}\\ \hat{\zeta}_{01}\end{pmatrix} - \begin{pmatrix}\zeta_{11}\\ \zeta_{01}\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \Sigma\right),$$

where $\zeta_{11} = P(T = 1, Y = 1)$ and $\zeta_{01} = P(T = 0, Y = 1)$.

To find the covariance matrix $\Sigma = Var(W_i)$, where $W_i = (T_i Y_i, (1 - T_i)Y_i)^T$, we calculate its components. Since $T_i, Y_i \in \{0, 1\}$, they act as indicator variables, so $X^2 = X$.

Variance of $T_i Y_i$ is $\mathsf{Var}(T_i Y_i) = \mathbb{E}[(T_i Y_i)^2] - (\mathbb{E}[T_i Y_i])^2 = \zeta_{11} - \zeta_{11}^2 = \zeta_{11}(1 - \zeta_{11})$.

Variance of $(1 - T_i)Y_i$ is $\mathsf{Var}((1 - T_i)Y_i) = \zeta_{01} - \zeta_{01}^2 = \zeta_{01}(1 - \zeta_{01})$.

For the covariance, notice that T_i and $(1-T_i)$ are mutually exclusive. You cannot simultaneously have $T_i = 1$ and $T_i = 0$. Thus, their product is always 0. Therefore,

$$\mathsf{Cov}(T_i Y_i, (1 - T_i)Y_i) = \mathbb{E}[T_i(1 - T_i)Y_i^2] - \mathbb{E}[T_i Y_i]\mathbb{E}[(1 - T_i)Y_i] = 0 - \zeta_{11}\zeta_{01} = -\zeta_{11}\zeta_{01}$$

Thus, our covariance matrix is:

$$\Sigma = \begin{pmatrix}\zeta_{11}(1 - \zeta_{11}) & -\zeta_{11}\zeta_{01}\\ -\zeta_{11}\zeta_{01} & \zeta_{01}(1 - \zeta_{01})\end{pmatrix}$$

Our estimator is a function of the sample means: $\hat{\theta}_n = g(\hat{\zeta}_{11}, \hat{\zeta}_{01})$, where $f(x, y) = \frac{x}{y}$. To apply the Delta method, we need the gradient of f evaluated at the true parameters:

$$\nabla f(x, y) = \begin{pmatrix}\frac{\partial f}{\partial x}\\ \frac{\partial f}{\partial y}\end{pmatrix} = \begin{pmatrix}\frac{1}{y}\\ -\frac{x}{y^2}\end{pmatrix}$$

Evaluated at (ζ_{11}, ζ_{01})*, this yields:*

$$\nabla f(\zeta_{11}, \zeta_{01}) = \begin{pmatrix} \frac{1}{\zeta_{01}} \\ -\frac{\zeta_{11}}{\zeta_{01}^2} \end{pmatrix} = \frac{1}{\zeta_{01}} \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

(Here we cleverly factored out $1/\zeta_{01}$ *and substituted* $\theta = \zeta_{11}/\zeta_{01}$ *to simplify the upcoming algebra).*

By the Delta method, the asymptotic distribution of our estimator is:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, V),$$

where the asymptotic variance is

$$\begin{aligned} V &= (\nabla f)^T \Sigma (\nabla f) \\ &= \frac{1}{\zeta_{01}^2} \begin{pmatrix} 1 & -\theta \end{pmatrix} \begin{pmatrix} \zeta_{11}(1-\zeta_{11}) & -\zeta_{11}\zeta_{01} \\ -\zeta_{11}\zeta_{01} & \zeta_{01}(1-\zeta_{01}) \end{pmatrix} \begin{pmatrix} 1 \\ -\theta \end{pmatrix} \\ &= \frac{\zeta_{11}}{\zeta_{01}^2}(1+\theta) = \frac{\theta(1+\theta)}{\zeta_{01}} \end{aligned}$$

A Consistency of Bootstrap

The bootstrap is a common approach to numerically approximate the limiting distribution of an estimator. Suppose our estimator of a parameter θ is $\hat{\theta}_n = f(X_1, \cdots, X_n)$. We generate the *bootstrap sample* by sampling with replacement from $X_1, \cdots, X_n$, leading to a new bootstrap sample

$$X_1^*, \cdots, X_n^*.$$

Given the original data, you can show that the bootstrap sample are IID from the empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$. We then use the bootstrap sample to compute the bootstrap estimator $\hat{\theta}_n^*$. In mild conditions, we have $\hat{\theta}_n^* - \hat{\theta}_n \approx \hat{\theta}_n - \theta$. So we can repeat the bootstrap procedure multiple times, leading to many numerical realizations of $\hat{\theta}_n^* - \hat{\theta}_n$, and use these values to derive the distribution of $\hat{\theta}_n - \theta$. The confidence interval via this approach is called the bootstrap confidence interval.

In this section, we show how the bootstrap work under a simple scenario: estimating the population mean via a sample mean. While this is a simple problem, it can be easily generalized to asymptotic linear estimators. Suppose we observe univariate $X_1, \cdots, X_n$ and we are interested in estimating the population mean, i.e., $\theta = \mathbb{E}(X_1)$, using the sample mean $\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n$.

Let $Z_n = \sqrt{n}(\hat{\theta}_n - \theta)$ and $Z_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$. To formally prove the validity bootstrap, we need to prove that

$$\sup_t \left| P(Z_n^* \le t | \hat{F}_n) - P(Z_n \le t) \right| \xrightarrow{P} 0. \tag{2.8}$$

The above bound is also known as the *Kolomogrov distance* between two random variables.

Although this seems to be hard to prove, there are two popular approaches to derive equation (2.8). The first approach is to show that Z_n has an asymptotic linear form and then apply Lindeberg-Feller central limit theorem (triangular arrays) to Z_n^* since Z_n^* is sampled from a 'random distribution function' $\hat{F}_n$. The second approach is via the Berry-Esseen bound of the sample mean, which is our preferred route.

A.1 Bootstrap consistency: Lindeberg-Feller's CLT

The conventional central limit theorem (CLT) shows that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = \mathsf{Var}(X_1)$. This result is NOT enough for bootstrap consistency because it assumes that observations $X_1, \cdots, X_n$ are sampled from a fixed CDF F, not a distribution function that can change with respect to the sample size n. In the bootstrap case, the bootstrap sample $X_1^*, \cdots, X_n^*$ are IID from the EDF $\hat{F}_n$ given $X_1, \cdots, X_n$. Thus, the 'population' of the bootstrap sample changes with respect to n, so conventional CLT is not applicable.

To resolve this issue, we use the Lindeberg-Feller's CLT in the triangular array setting, which can be derived from Theorem 2.17.

Theorem 2.33 (Triangular array version of Lindeberg-Feller CLT) *For each $n = 1, 2, 3, \cdots$, let $W_n = (W_{n,1}, \cdots, W_{n,k_n})$ be a vector of independent elements with finite variance, i.e., $W_{n,1}, \cdots, W_{n,k_n}$ are independent from each other. Assume that*

- **Uniform integrability.** $\sum_{i=1}^{k_n} \mathbb{E}[W_{n,i}^2 I(|W_{n,i}| > \epsilon)] \to 0$ *for every* $\epsilon > 0$.
- **Finite variance.** $\sum_{i=1}^{k_n} \mathsf{Var}(W_{n,i}) \to \sigma^2$.

Then

$$\sum_{i=1}^{k_n} (W_{n,i} - \mathbb{E}(W_{n,i})) \xrightarrow{d} N(0, \sigma^2).$$

Compared to Theorem 2.17, we relax two conditions. First, for each n, we may have $1, \cdots, k_n$ independent observations, rather than $k_n = n$. Second, the total variance only need to converge to σ^2 rather than being exactly σ.

We consider the scenario where the original data $X_1, \cdots, X_n$ are fixed so that the bootstrap sample $X_1^*, \cdots, X_n^*$ are IID from $\hat{F}_n$.

To use Theorem 2.33 in the bootstrap setting, each $W_{n,i} = \frac{1}{\sqrt{n}} X_i^*$ is sampled from $\hat{F}_n$, so $\mathbb{E}_{\hat{F}_n}[W_{n,i}] = \frac{1}{\sqrt{n}} \hat{\theta}_n$. Note that the expectation $\mathbb{E}_{\hat{F}_n}(\cdot)$ is with respect to the distribution $\hat{F}_n$. Also, $k_n = n$. Under this setting, $\sum_{i=1}^{k_n} W_{n,i} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i^* = \sqrt{n}\hat{\theta}_n^*$ and

$$\sum_{i=1}^{k_n} (W_{n,i} - \mathbb{E}_{\hat{F}_n}(W_{n,i})) = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n).$$

Thus, the conclusion of Theorem 2.33 is applicable to the setting of the bootstrap.

Now we investigate the two conditions in Theorem 2.33. The first uniform integrability condition

$$\sum_{i=1}^{k_n} \mathbb{E}_{\hat{F}_n}[W_{n,i}^2 I(|W_{n,i}| > \epsilon)] \to 0$$

becomes

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 I(|X_i| > \sqrt{n}\epsilon) \to 0,$$

When the true distribution F has a finite second moment, i.e., $\mathbb{E}(X_i^2) < \infty$, strong law of large numbers implies $\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{a.s.} \mathbb{E}(X_i^2) < \infty$, so $\frac{1}{n}\sum_{i=1}^n X_i^2 I(|X_i| > \sqrt{n}\epsilon) \xrightarrow{a.s.} 0$.

The finite variance condition becomes

$$\sum_{i=1}^{k_n} \mathsf{Var}(W_{n,i}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}_{\hat{F}_n}(X_i^*) = \hat{\sigma}_n^* \xrightarrow{P} \sigma^2,$$

where $\hat{\sigma}_n^* = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\sigma^2 = \mathsf{Var}(X_1)$.

As a result, we conclude that

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{d} N(0, \sigma^2);$$

namely, it converges to the same limit as the original estimator $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. The above two convergences in distribution implies equation (2.8), so we have the consistency of the bootstrap.

A more general form of this approach can be found in Theorem 23.4 of [Van der Vaart].

A.2 Bootstrap consistency: Berry-Esseen bound

The Berry-Esseen bound offers a finite sample bounds on how fast the asymptotic normality of a sample average converges to the actual normal distribution.

Theorem 2.34 (Berry-Esseen bound) *Assume that* $\mathbb{E}(|X_1|^3) < \infty$. *Let* $Z \sim N(0,1)$ *and* $\theta = \mathbb{E}(X_1)$ *and* $\sigma^2 = \mathsf{Var}(X_1)$. *Then for any* n, *we have*

$$\sup_t \left| P\left(\sqrt{n}\left(\frac{\bar{X}_n - \theta}{\sigma}\right) < t\right) - P(Z < t) \right| \le C\frac{\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}},$$

for a constant $C \ge \frac{\sqrt{10}+3}{6\sqrt{2\pi}}$.

It is important to note that the Berry-Esseen bound is a *finite sample* bound, meaning that its result holds *for any* n (some finite sample bound holds when n is larger than some constant). So it is a much stronger result than the conventional central limit theorem. The finite sample bound is important in deriving the validity of the bootstrap (see the proof below).

The Berry-Esseen bound can be used to derive bounds like equation (2.8). Now consider very simple scenario that we are interested in estimating the population mean $\theta = \mathbb{E}(X_1)$ and we use the sample mean as the estimator $\hat{\theta}_n$.

Theorem 2.35 *Suppose that we are considering the sample mean problem, i.e.,* $\theta = \mathbb{E}(X_1)$ *and* $\hat{\theta}_n = \bar{X}_n$ *is the original sample mean estimator and* $\hat{\theta}_n^* = \bar{X}_n^*$ *is the sample mean of the bootstrap sample. Assume that* $\mathbb{E}(|X_1|^3) < \infty$. *Let*

$$Z_n = \sqrt{n}(\hat{\theta}_n - \theta), \qquad Z_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n).$$

Then

$$\sup_t \left| P(Z_n^* \le t | \hat{F}_n) - P(Z_n \le t) \right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof:

Let $\Psi_\sigma(t)$ be the CDF of $N(0,\sigma^2)$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. We bound the difference using

$$\sup_t \left| P(Z_n^* \le t|\hat{F}_n) - P(Z_n \le t) \right| \le \sup_t \left| P(Z_n^* \le t|\hat{F}_n) - \Psi_{\hat{\sigma}}(t) \right| + \sup_t |\Psi_{\hat{\sigma}}(t)) - \Psi_\sigma(t)| + \sup_t |P(Z_n \le t) - \Psi_\sigma(t)|.$$

The Berry Esseen theorem implies that

$$\sup_t |P(Z_n \le t) - \Psi_\sigma(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

so the third quantity is bounded. Similarly, we can apply the Berry-Esseen bound to the first quantity by replacing $\mathbb{E}(\cdot)$ with the empirical version of it (sample average operation), which implies

$$\sup_t \left| P(Z_n^* \le t|\hat{F}_n) - P(Z_n \le t) \right| \le C\frac{\frac{1}{n}\sum_{i=1}^n X_i^3}{\sigma^3\sqrt{n}}.$$

Note that we can apply the Berry-Esseen theory to the bootstrap because this theory holds in *finite sample*! In the bootstrap world, the EDF is the population distribution generating our data, and that is why we replace the expectation $\mathbb{E}$ by the empirical version of it.

By strong law of large number, the probability that the right hand side is less than $2C\frac{\mathbb{E}|X_1|^3}{\sigma^3\sqrt{n}}$ is 1. Thus, we conclude that

$$\sup_t \left| P(Z_n^* \le t|\hat{F}_n) - P(Z_n \le t) \right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

For the second term, $\sup_t |\Psi_{\hat{\sigma}}(t)) - \Psi_\sigma(t)|$, because $|\hat{\sigma} - \sigma| = O_P\left(\frac{1}{\sqrt{n}}\right)$ so differentiating the CDF with respect to σ and take a uniform bound leads to

$$\sup_t |\Psi_{\hat{\sigma}}(t)) - \Psi_\sigma(t)| = O_P\left(\frac{1}{\sqrt{n}}\right),$$

which completes the proof.

■

The Lindeberg-Feller central limit theorem approach requires a slightly less condition than the Berry-Esseen bound (we do not need third-moment but just need a bounded second moment). However, the Lindeberg-Feller approach will not give us a convergence rate while the Berry-Esseen approach gives us a convergence rate.