# STAT575
# Lrage Sample Theory

Pingbang Hu

April 25, 2024

**Abstract**

This is a graduate-level theoretical statistics course taught by Georgios Fellouris at University of Illinois Urbana-Champaign, aiming to provide an introduction to asymptotic analysis of various statistical methods, including weak convergence, Lindeberg-Feller CLT, asymptotic relative efficiency, etc.

We list some references of this course, although we will not follow any particular book page by page: *Asymnptotic Statistics* [Vaa98], *Asymptotic Theory of Statistics and Probability* [Das08], *A course in Large Sample Theory* [Fer17], *Approximation Theorems of Mathematical Statistics* [Ser09], and *Elements of Large-Sample Theory* [Leh04].

This course is taken in Spring 2024, and the date on the cover page is the last updated time.

# Contents

# Chapter 1

# Introduction

## Lecture 1: Introduction to Large Sample Theory

16 Jan. 9:30

Say we first collect $n$ data points $x_1, \dots, x_n \in \mathbb{R}^d$, where we may treat $x_i$ as a realization of a random vector $X_i$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this course, we will primarily consider the case that $X_i$'s are i.i.d., i.e., independent and identically distributed from a distribution function, or the *cumulative density function* (cdf) $F$ such that

$$X = (X^1, \dots, X^d) \sim F(x_1, \dots, x_d) \equiv \mathbb{P}(X^1 \le x_1, \dots, X^d \le x_d)$$

for all $x_i \in \mathbb{R}$. If we have access to $F$, we can compute the corresponding *probability density function* (pdf) $f$, and then have access to $\mathbb{P}(X \in A)$ for all (measurable) $A \subseteq \mathbb{R}^d$ of interest.

> **Notation.** In the measure-theoretic sense, the measure $\mathbb{P}$ in $(\Omega, \mathscr{F}, \mathbb{P})$ is the Lebesgue-Stieltjes measure $\mu_F$ induced by the distribution function $F$. When doing integration, we will often denote
>
> $$\mathrm{d}\mu_F(x) = \mathrm{d}\mathbb{P}(x) =: F(\mathrm{d}x) =: \mathrm{d}F(x) =: f(x)\mathrm{d}x$$

> **Remark.** If we know any of the above, we know every thing about the population.

Hence, the goal is to compute this by collecting data $x_i$'s, which is a statistical inference problem. Notably, *large sample theory* concerns with the limiting theory as $n \to \infty$.

## 1.1 Parametrized Approach

There are various ways of doing this task, one way is the so-called parametrized approach. By postulating a family of cdfs $\{F_\theta, \theta \in \Theta\}$ where $\Theta$ is often a subset of $\mathbb{R}^m$ for some $m$ (generally $\neq n$), the goal is to select a member of this family that is the "closet", or the "best fit" to the truth, i.e., $F$, based on the data.

> **Note.** To emphasize that this depends on the data, we sometimes write the function we found as $\hat{\theta}_n(x_1, \dots, x_n)$ so that $F_{\hat{\theta}_n(x_1,\dots,x_n)}$ is our proxy for $F$.

Now, assume that the family is initially given, the problem is then how to select $\hat{\theta}_n$.

> **Example.** Fisher suggested that we should look at the maximum likelihood estimator (MLE).

The justification for MLE is not about finite $n$, but about its asymptotic behavior when $n \to \infty$. Specifically, we have the following theorem due to Fisher (informally stated).

> **Theorem 1.1.1** (Fisher). If $F \in \{F_\theta : \theta \in \Theta\}$, i.e., if $F = F_{\theta^*}$ for some $\theta^* \in \Theta$, then under certain conditions, $\hat{\theta}_n$ will be "close" to $\theta^*$ as $n \to \infty$. Under some other conditions, $\sqrt{n}(\hat{\theta}_n - \theta)$ is approximately Gaussian with variance being the "best possible" in some sense.

On the other hand, in the misspecified case, i.e., $F \notin \{F_\theta, \theta \in \Theta\}$, we can still compute the MLE, which leads to another justification for MLE since even in this case, $\hat{\theta}_n$ will still be "close" to $\theta^*$ such that $F_{\theta^*}$ is, in some sense, the "closest" to $F$ among all possible $F_\theta$ (minimizing divergence, to be precise).

## 1.2 Hypothesis Testing

We will also develop theory for hypothesis testing for some hypothesis we're interested in, e.g., whether the data we collect is really i.i.d., or whether our proposed family is reasonable enough. Say now $X_i$'s are scalar random variable with $\mathbb{E}[X] = \mu$, and we want to test the null hypothesis $H_0 \colon \mu = 0$.

**Example.** Consider a controlled group $Z$ and a treatment group $Y$, and we observe $Z_1, \dots, Z_n$, and $Y_1, \dots, Y_n$, respectively, and compute $X_i = Z_i - Y_i$ for all $i$. Testing $H_0$ on the distribution of $X$ will show the effect of the treatment.

To do this, a well-known, elementary, and fundamental method is the so-called $t$-test.

**Definition 1.2.1** ($t$-statistic). Given a sample $X_1, \dots, X_n$, the *$t$-statistic* is defined as

$$T_n = \frac{\overline{X}_n}{s_n/\sqrt{n}},$$

where $s_n$ is the sample standard derivation.

**Note.** As long as $X$ is Gaussian, $T_n \sim t_{n-1}$, i.e., the $t$-distribution with $n-1$ degrees of freedom.

Hence, one can reject $H_0$ when $T_n$ is too large (or small) when $X \sim \mathcal{N}$ as we know the exact distribution $T_n$ will follow. What if $X$ is not an Gaussian? We will show that even if $X$ is not Gaussian, this result is "approximately valid" when $n$ is "large enough" as long as $\mathrm{Var}[X] < \infty$.

**Remark** (Sample Size). When we say $n$ is "large enough", what we mean really depends on how fast the underlying distribution will approach Gaussian as $n$ grows. Hence, if we can say more about the underlying population, we can say more about when does $n$ is "large enough"; otherwise such a limiting theory might be completely useless in practice.

What if $\mathrm{Var}[X]$ doesn't exit for some heavy tailed distribution like the Cauchy?

**Example** (Cauchy distribution). Cauchy distribution doesn't have finite moment of order greater than 1.

In this case, other tests are needed. A simple test would be looking at the sign of $X_i$.

**Example** (Sign test). We might reject $H_0$ if $\sum_{i=1}^n \mathbb{1}_{X_i>0}$ is large. Note that under $H_0$, $\sum_{i=1}^n \mathbb{1}_{X_i>0} \sim \mathrm{Bin}(n, 1/2)$, and this test is valid even if expectation doesn't exist.

We see that without saying anything about $F$, the sign test is valid even for $n = 3$ or 5 as the sum is exactly binomial distribution under $H_0$. Although simple and have good property, only looking at the sign of $X_i$ might be too weak. A natural idea is to look at the absolute value of $X_i$.

**Example** (Wilcoxon's rank-sum test). Let $R_{i,n}$ to be the rank of $|X_i|$, then consider the so-called *Wilcoxon's rank-sum test*

$$\sum_{i=1}^n \mathbb{1}_{X_i>0} R_{i,n}.$$

As one can imagine, the closed form of the above sum will be complicated; however, asymptotically, the above statics will follow Gaussian again, such that the rate of convergence doesn't depend on the underlying population.

Finally, we also ask how can we compare these different tests? This will also be addressed in this course.

# Chapter 2

# Modes of Convergence

## Lecture 2: Modes of Convergence

18 Jan. 9:30

### 2.1 Different Modes of Convergence

Given a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, consider a sequence of $d$-dimensional random vectors $(X_n)$ and a random vector $X$, i.e., $X_n, X \colon \Omega \to \mathbb{R}^d$. We now discuss different modes of convergence for $(X_n)$.

**Definition 2.1.1** (Point-wise converge). $(X_n)$ *point-wise converges* to $X$, denoted as $X_n \to X$, if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.[a]

[a] I.e., for every $\epsilon > 0$, there exists $n_0(\omega) \in \mathbb{N}$ such that for every $n \geq n_0$, $\|X_n(\omega) - X(\omega)\|_2 < \epsilon$.

Since we don't care about measure zero sets, we may instead consider the following.

**Definition 2.1.2** (Converge almost-surely). $(X_n)$ *converges almost-surely* to $X$, denoted as $X_n \overset{\text{a.s.}}{\to} X$, if $\mathbb{P}(X_n \to X) = 1$.[a]

[a] I.e., $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega \setminus N$ where $\mathbb{P}(N) = 0$.

However, this might still be too strong.

**Definition 2.1.3** (Converge in probability). $(X_n)$ *converges in probability* to $X$, denoted as $X_n \overset{p}{\to} X$, if for every $\epsilon > 0$, $\mathbb{P}(\|X_n - X\| > \epsilon) \to 0$ as $n \to \infty$.

**Remark.** $X_n \to X$ if and only if $\|X_n - X\| \to 0$. The same also holds for $\overset{p}{\to}$ and $\overset{\text{a.s.}}{\to}$.

A related notion is the following, where we now sum over $n$.

**Definition 2.1.4** (Converge completely). $(X_n)$ *converges completely* to $X$, denoted as $X_n \overset{\text{comp}}{\to} X$, if for every $\epsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(\|X_n - X\| > \epsilon) < \infty$.

Finally, we have the following.

**Definition 2.1.5** (Converge in $L^p$). $(X_n)$ *converges in $L^p$* to $X$ for some $p > 0$, denoted as $X_n \overset{L^p}{\to} X$, if $\mathbb{E}\left[\|X_n - X\|^p\right] \to 0$ as $n \to \infty$.

#### 2.1.1 Connection Between Modes of Convergence

We have the following connections between different modes of convergence.

$$\text{completely} \implies \text{almost-surely} \implies \text{in probability} \impliedby \text{in } L^p$$

To show the above, the following characterization for almost-surely convergence is useful.

**Proposition 2.1.1.** For a sequence of random vectors $(X_n)$ and a random vector $X$, we have

$$\begin{aligned} X_n \overset{\text{a.s.}}{\to} X &\Leftrightarrow \mathbb{P}(\|X_k - X\| > \epsilon \text{ for some } k \geq n) \overset{n\to\infty}{\to} 0 \\ &\Leftrightarrow \mathbb{P}(\|X_n - X\| > \epsilon \text{ for infinitely many } n\text{'s}) = 0 \\ &\Leftrightarrow \mathbb{P}(\limsup_{n\to\infty} \|X_n - X\| > \epsilon) = 0, \end{aligned}$$

where the above holds for every $\epsilon > 0$.

From Proposition 2.1.1, it's clear that $\overset{\text{a.s.}}{\to}$ implies $\overset{p}{\to}$ since

$$\mathbb{P}(\|X_k - X\| > \epsilon \text{ for some } k \geq n\ ) \geq \mathbb{P}(\|X_n - X\| > \epsilon),$$

hence if the former goes to 0, so does the latter. On the other hand, $\overset{\text{comp}}{\to}$ implies $\overset{\text{a.s.}}{\to}$ follows from the third equivalence. Lastly, the convergence in $L^p$ implies the convergence in probability since

$$\mathbb{P}(\|X_n - X\| > \epsilon) \leq \frac{1}{\epsilon^p}\mathbb{E}\left[\|X_n - X\|^p\right]$$

from Markov's inequality. However, the converse is not always true.

**Theorem 2.1.1** (Dominated convergence theorem). If $X_n \overset{p}{\to} X$ and $\|X_n - X\| \leq Z$ for all $n \geq 1$ where $\mathbb{E}\left[\|Z\|^p\right] < \infty$, then $X_n \overset{L^p}{\to} X$.

**Theorem 2.1.2** (Scheffé's theorem). Let $X_n \overset{p}{\to} X$. If $\mathbb{E}[\|X_n\|^p] \to \mathbb{E}[\|X\|^p] < \infty$, then $X_n \overset{L^p}{\to} X$. In particular, with Fatou's lemma, $\limsup_{n\to\infty} \mathbb{E}\left[\|X_n\|^p\right] \leq \mathbb{E}\left[\|X\|^p\right] < \infty$ suffices.

### 2.1.2 Consistent Estimator

Let $(X_n) \overset{\text{i.i.d.}}{\sim} F$ where $F$ is a distribution function. Say we're interested in some aspect of $F$, for example, some parameter $\theta = T(F) \in \mathbb{R}^m$. By collecting data $X_1, \ldots, X_n$, we estimate $\theta$ by computing an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ of $\theta$. There are some properties we might want for $\hat{\theta}_n$.

**Definition 2.1.6** (Consistent). $\hat{\theta}_n$ is *consistent* of $\theta$ if $\hat{\theta}_n \overset{p}{\to} \theta$.

**Definition 2.1.7** (Strongly consistent). $\hat{\theta}_n$ is *strongly consistent* of $\theta$ if $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta$.

**Definition 2.1.8** (Converge in mean squared error). $\hat{\theta}_n$ converges to $\theta$ *in mean squared error* if $\hat{\theta}_n \overset{L^2}{\to} \theta$.

**Remark.** When $d = 1$, $\mathbb{E}[(\hat{\theta}_n - \theta)^2] = \mathrm{Var}[\hat{\theta}_n] + (\mathbb{E}[\hat{\theta}_n - \theta])^2$. Therefore, $\hat{\theta}_n$ converges in mean squared error to $\theta$ if and only if $\mathbb{E}[\hat{\theta}_n] \to \theta$ and $\mathrm{Var}[\hat{\theta}_n] \to 0$.

Let's first see the most well-known estimation problem, the mean estimation.

**Example** (Mean esimation). Suppose $d = 1$, and let $X$ be non-negative. Say we're interested in $\theta = \mathbb{E}[X]$. It's standard that in this case, we can compute $\mathbb{E}[X]$ by

$$\theta = \mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t = \int_0^\infty (1 - F(t))\,\mathrm{d}t.$$

If $X$ has a pmf $f$, then $\mathbb{E}[X] = \sum_x xf(x) = \sum_x x\Delta F(x)$ where $f(x) = \Delta F(x) \equiv F(x) - F(x^-)$; if

$X$ has a pdf $f$, then

$$\mathbb{E}[X] = \int_0^\infty x f(x)\,\mathrm{d}x = \int_0^\infty x F(\mathrm{d}x).$$

Now, let $\hat{\theta}_n$ to be the sample mean, i.e., $\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. From the strong law of large number, $\overline{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$, which implies that $\hat{\theta}_n$ is a strongly consistent estimator of $\theta$.

On the other hand, if $\mathrm{Var}[X] < \infty$, then $\overline{X}_n \xrightarrow{L^2} \mathbb{E}[X]$, which further implies $\overline{X}_n \xrightarrow{p} \mathbb{E}[X]$, hence $\hat{\theta}_n$ is consistent.[a]

[a] The latter is true even when $\mathrm{Var}[X] = \infty$ as we expect.

**Proof.** We show the last statement. Since $\mathrm{Var}[X] < \infty$, then

$$\frac{\mathrm{Var}[X]}{n} = \mathrm{Var}\left[\overline{X}_n\right] = \mathbb{E}\left[\left(\overline{X} - \mathbb{E}[X]\right)^2\right] \to 0$$

as $n \to \infty$, which implies $\overline{X}_n \xrightarrow{p} \mathbb{E}[X]$. ⊛

Another interesting problem is the supremum estimation.

**Example** (Supremum estimation). Suppose $d = 1$ and there is a $\theta \in \mathbb{R}$ and a distribution function $F$ such that $F(\theta - \epsilon) < 1 = F(\theta)$ for all $\epsilon > 0$, i.e., $\theta = \sup_\omega X(\omega)$ since $\mathbb{P}(X \le \theta - \epsilon) = F(\theta - \epsilon)$ and $F(\theta) = \mathbb{P}(X \le \theta)$.[a] Then $\hat{\theta}_n = \max_{1 \le i \le n} X_i$ is indeed a strongly consistent estimator of $\theta$.

[a] Such a distribution exists, for example, $\mathcal{U}(0, \theta)$.

**Proof.** We see that for any $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) &= \mathbb{P}(\hat{\theta}_n > \theta + \epsilon) + \mathbb{P}(\hat{\theta}_n < \theta - \epsilon) \\
&= \mathbb{P}\left(\bigcup_{i=1}^n \{X_i > \theta + \epsilon\}\right) + \mathbb{P}\left(\bigcap_{i=1}^n \{X_i < \theta - \epsilon\}\right) \\
&\le \sum_{i=1}^n \underbrace{\mathbb{P}(X_i > \theta + \epsilon)}_{0} + \prod_{i=1}^n \mathbb{P}(X_i < \theta - \epsilon) = \left(\mathbb{P}(X_1 < \theta - \epsilon)\right)^n \le \left(F(\theta - \epsilon)\right)^n \to 0
\end{aligned}$$

as $n \to \infty$ since $F(\theta - \epsilon) < 1$. This shows that $\hat{\theta}_n$ is indeed consistent. Moreover, since $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon)$ decays exponentially, so this is absolutely summable, hence it's also strongly consistency. ⊛

Proving convergence of $\hat{\theta}_n$ is useful, but this might not be enough.

**Example.** Consider any deterministic sequence $(a_n)$ in $\mathbb{R}$ which converges to 0. Adding $a_n$ to $\hat{\theta}_n$ will not change the convergence of $\hat{\theta}_n$.

The above suggests that we should look at the *distribution* of $\hat{\theta}_n - \theta$ in order to say how does $\hat{\theta}_n \to \theta$.

**Example** (Mean estimation for Gaussian). Suppose $X \sim \mathcal{N}(\theta, 1)$. Then $\hat{\theta}_n = \overline{X}_n \sim \mathcal{N}(\theta, 1/n)$, i.e., $\sqrt{n}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$, i.e., we can write down a confidence interval such as $\hat{\theta}_n \pm 1.96/\sqrt{n}$ with 95% confidence level for $\theta$.

Doing this for other kind of estimators and $F$ is not that straightforward and will be challenging.

**Remark.** Let $(X_n)$ and $X$ be $d$-dimensional random vectors, $h\colon \mathbb{R}^d \to \mathbb{R}^m$, and $c \in \mathbb{R}^d$ constant.

(a) If $X_n \to c$, then $h(X_n) \to h(c)$ if $h$ is continuous at $c$.[a] This also holds for $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{p}$.

(b) If $X_n \to X$, then $h(X_n) \to h(X)$ if $h$ is continuous. This also holds for $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{p}$.

[a] This is an if and only if condition if this holds for any $h$.

Let's see some examples.

**Example.** If $d = 1$, and $X_n \to \theta \neq 0$. Then $1/X_n \to 1/\theta$ where

$$h(x) = \begin{cases} \frac{1}{x}, & \text{if } x \neq 0; \\ c, & \text{if } x = 0 \end{cases}$$

for any $c \in \mathbb{R}$. The same holds for $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{p}$.

**Example.** If $X_n \to X$ and $Y_n \to Y$, then $(X_n, Y_n) \to (X, Y)$.[a] The same holds for $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{p}$.

[a] The converse is also true since projections are continuous.

**Proof.** $\|(X_n, Y_n) - (X, Y)\| \to 0$ since $\|(X_n, Y_n) - (X, Y)\| \leq \|X_n - X\| + \|Y_n - Y\|$ for all $n \geq 1$.[a] The latter two terms go to 0 (in whatever sense) by assumption. ⊛

[a] This can be seen from $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$.

# Lecture 3: Weak Convergence Portmanteau Theorem

25 Jan. 9:30

## 2.2 Weak Convergence

The convergences we have seen are not "distribution-wise" since to evaluate $\|X_n - X\|$, $X_n$ and $X$ need to be defined on the same probability space. If all we care about is distribution, consider probability spaces $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ (and $(\Omega, \mathscr{F}, \mathbb{P})$) for which $X_n$ (and $X$) is defined on.

### 2.2.1 Convergence in Total Variation

**Definition 2.2.1** (Total variation). The *total variation* distance between $X$ and $Y$ on $\Omega$ is defined as

$$\mathrm{TV}(X, Y) = \sup_{B \in \mathscr{F}} |\mathbb{P}(X \in B) - \mathbb{P}(Y \in B)|$$

The above makes sense even if $X$ and $Y$ are defined on different probability spaces, e.g., in our situation, consider a sequence or random variables $(X_n)$ and a random variable $X$.

**Definition 2.2.2** (Converge in total variation). $(X_n)$ *converges in total variation* to $X$, denoted as $X_n \xrightarrow{\text{TV}} X$, if $\mathrm{TV}(X_n, X) \to 0$ as $n \to \infty$.

**Note.** Specifically, $X_n \xrightarrow{\text{TV}} X$ if $\mathbb{P}_n(X_n \in B) \to \mathbb{P}(X \in B)$ for all $B \in \mathcal{B}(\mathbb{R}^d)$.

**Remark.** If $X_n$ has density $f_n$ and $X$ has density $f$, then $\mathrm{TV}(X_n, X) = \frac{1}{2} \int |f_n - f|$, hence $f_n \to f$ implies $X_n \xrightarrow{\text{TV}} X$ from Scheffé's theorem.

**Example.** If $X_n \sim \mathrm{Bin}(n, p_n)$ such that $np_n \to \lambda \in \mathbb{R}$, then $X_n \sim \mathrm{Bin}(n, p_n) \xrightarrow{\text{TV}} X \sim \mathrm{Pois}(\lambda)$.

**Example.** Let $X_n \sim f_{\theta_n}$ where $f_\theta(x) = f(x)e^{\theta x - \psi(\theta)}$ for some $\theta \in \Theta$. If $\theta_n \to \theta$, then $X_n \xrightarrow{\text{TV}} X \sim f_\theta$. For example, if $X_n \sim \mathrm{Pois}(\theta_n)$ and $\theta_n \to \theta$, then $X_n \xrightarrow{\text{TV}} X \sim \mathrm{Pois}(\theta)$.

### 2.2.2 Weak Convergence

However, convergence in total variation might be too strong to work with.

**Example.** Let $X_n \sim \mathcal{U}\{0, 1/n, \dots, (n-1)/n\}$, which should be converging to $X \sim \mathcal{U}(0,1)$. However, this doesn't happen in total variation distance as we can take $B$ to be $\mathbb{Q}$.

This suggests that we should look at something weaker.

**Definition 2.2.3** (Converge weakly). $(X_n)$ *converges weakly* to $X$, denoted as $X_n \xrightarrow{\text{w}} X$, if for all bounded continuous $g\colon \mathbb{R}^d \to \mathbb{R}$, $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$.

To see how is weak convergence compared to convergence in total variation, we revisit the above.

**Example.** Let $X_n \sim \mathcal{U}\{0, 1/n, \dots, (n-1)/n\}$, which should be converging to $X \sim \mathcal{U}(0,1)$. We have

$$\mathbb{E}_n[g(X_n)] = \sum_{k=0}^{n-1} g(k/n)\left(\frac{k+1}{n} - \frac{k}{n}\right) \to \int_0^1 g(x)\,\mathrm{d}x = \mathbb{E}[g(X)]$$

as $g$ is bounded and continuous on $[0,1]$, hence Riemann integrable.

### 2.2.3 Portmanteau Theorem

The following is our main tool of proving weak convergence.

**Theorem 2.2.1** (Portmanteau theorem). The following are equivalent.

(a) $X_n \xrightarrow{\text{w}} X$.

(b) $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$ for all bounded Lipschitz $g\colon \mathbb{R}^d \to \mathbb{R}$.

(c) $\mathbb{P}(X \in A) \le \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A)$ for all $A \subseteq \mathbb{R}^d$ open.

(d) $\mathbb{P}(X \in A) \ge \limsup_{n\to\infty} \mathbb{P}_n(X_n \in A)$ for all $A \subseteq \mathbb{R}^d$ closed.

(e) $\mathbb{P}_n(X_n \in A) \to \mathbb{P}(X \in A)$ for all $A \in \mathscr{F}$ such that $\mathbb{P}(X \in \partial A) = 0$.

Before we prove Portmanteau theorem, we should note that all our discussion can be extended to metric spaces from Euclidean spaces. Let's first recall some basic results for metric spaces.

**Claim.** Given a metric space $(S, \rho)$, $\rho(\cdot, A)$ is Lipschitz for any $A \subseteq S$, i.e., for any $x, y \in S$,

$$|\rho(x, A) - \rho(y, A)| \le \rho(x, y).$$

**Proof.** Since for any $z \in S$, $\rho(x,z) \le \rho(x,y) + \rho(y,z)$, hence $\rho(x,A) - \rho(y,A) \le \rho(x,y)$ by taking the infimum over $z \in A$. Interchanging $x$ and $y$ gives another inequality. ⊛

**Claim.** Given a metric space $(S, \rho)$, for any $A \subseteq S$, $x \in \overline{A} \Leftrightarrow \rho(x, A) = 0$.

**Proof.** If $x \in \overline{A}$, there exists $(x_n)$ in $A$ such that $\rho(x_n, x) \to 0$. Then for any $z \in A$, $\rho(x,z) \le \rho(x, x_n) + \rho(x_n, z)$, implying

$$\rho(x, A) \le \rho(x, x_n) + \rho(x_n, A) \to 0,$$

hence $\rho(x, A) = 0$. On the other hand, suppose $\rho(x, A) = 0$. As $\rho(x, A) = \inf_{y\in A} \rho(x, y)$, there exists $(y_n)$ in $A$ such that $\rho(x, y_n) \to \rho(x, A) = 0$, i.e., $x \in \overline{A}$. ⊛

The crucial lemma we're going to use to prove Portmanteau theorem is the following.

**Lemma 2.2.1.** Given a metric space $(S, \rho)$ and let $A \subseteq S$ be a closed subset. Then there exists bounded Lipschitz $g_k\colon S \to \mathbb{R}$, decreasing in $k$ such that $g_k(x) \searrow \mathbb{1}_A(x)$.

**Proof.** To motivate, since $A$ is closed, $A = \overline{A}$ and

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{if } x \in A \Leftrightarrow \rho(x, A) = 0; \\ 0, & \text{if } x \notin A \Leftrightarrow \rho(x, A) > 0. \end{cases}$$

Then, consider

$$g_k(x) = \begin{cases} 0, & \text{if } \rho(x, A) > \dfrac{1}{k}; \\ 1 - k\rho(x, A), & \text{otherwise;} \end{cases} = 1 - (k\rho(x, A) \wedge 1)\,.$$

We see that

- if $x \in A$: $\mathbb{1}_A(x) = 1$, and $g_k(x) = 1$ since $\rho(x, A) = 0$;
- if $x \notin A$: $\mathbb{1}_A(x) = 0$, and $\rho(x, A) > 0$ since $A$ closed, and $g_k(x) = 0$ for all large enough $k$.

Finally, it's clear that $g_k(x)$ takes values in $[0, 1]$, and we now show it's Lipschitz. We have

$$|g_k(x) - g_k(y)| = |(k\rho(x, A) \wedge 1) - (k\rho(y, A) \wedge 1)| \leq k\rho(x, y)$$

for all $x, y \in S$. ■

Then we can prove the Portmanteau theorem.

**Proof of Theorem 2.2.1.** (a) $\Rightarrow$ (b) is clear, and we start by proving (c) $\Leftrightarrow$ (d).

**Claim.** (c) $\Leftrightarrow$ (d).

**Proof.** We first prove that (d) $\Rightarrow$ (c). Since when $A$ is open,

$$\begin{aligned} \mathbb{P}(X \in A) = 1 - \mathbb{P}(X \in A^c) &\leq 1 - \limsup_{n\to\infty} \mathbb{P}_n(X_n \in A^c) && \text{(d)} \\ &= 1 - \limsup_{n\to\infty} (1 - \mathbb{P}_n(X_n \in A)) = \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A). \end{aligned}$$

(c) $\Rightarrow$ (d) is exactly the same, hence (c) $\Leftrightarrow$ (d). ⊛

Next, we prove (b) $\Rightarrow$ (d), which gives us (a) $\Rightarrow$ (b) $\Rightarrow$ (d) $\Leftrightarrow$ (c).

**Claim.** (b) $\Rightarrow$ (d).

**Proof.** From Lemma 2.2.1, there exists bounded Lipschitz $g_k \searrow \mathbb{1}_A$ such that for all closed $A$,

$$\mathbb{P}_n(X_n \in A) = \mathbb{E}_n\left[\mathbb{1}_A(X_n)\right] \leq \mathbb{E}_n\left[g_k(X_n)\right].$$

This is true for every $k$ and $n$ since $g_k \geq \mathbb{1}_A$, and by taking the limit as $n \to \infty$,

$$\limsup_{n\to\infty} \mathbb{P}_n(X_n \in A) \leq \limsup_{n\to\infty} \mathbb{E}_n\left[g_k(X_n)\right] = \mathbb{E}\left[g_k(X)\right]$$

from our assumption (b). Finally, as $k \to \infty$, it goes to $\mathbb{E}\left[\mathbb{1}_A(X)\right] = \mathbb{P}(X \in A)$ as desired. ⊛

*The proof will be continued...*

# Lecture 4: Continuous Mapping Theorem

Before finishing the proof of Portmanteau theorem, we need one additional tool. 30 Jan. 9:30

**Lemma 2.2.2.** If $\{A_i\}_{i\in I}$ are pairwise disjoint events, then $\{i \in I : \mathbb{P}(A_i) > 0\}$ is countable.[a]

[a] Note that $I$ can be uncountable.

**Proof.** It suffices to show $|I_k| < \infty$ where $I_k \coloneqq \{i \in I : \mathbb{P}(A_i) \geq 1/k\}$ for any $k \geq 1$ since

$$\{i \in I : \mathbb{P}(A_i) > 0\} = \bigcup_{k=1}^{\infty} \left\{ i \in I : \mathbb{P}(A_i) \geq \frac{1}{k} \right\} =: \bigcup_{k=1}^{\infty} I_k.$$

We show $|I_k| \leq k$ for any $k$. Suppose not, then there exists a countable $J_k \subseteq I_k$ such that $|J_k| > k$,

$$\mathbb{P}\left( \bigcup_{i \in J_k} A_i \right) = \sum_{i \in J_k} \mathbb{P}(A_i) \geq \frac{|J_k|}{k} > 1,$$

which is a contradiction. ■

We now finish the proof of Portmanteau theorem.

**Proof of Theorem 2.2.1 (cont.)** We already proved (a) ⇒ (b) ⇒ (d) ⇔ (c).

**Claim.** (c) + (d) ⇒ (e).

**Proof.** We see that for any $A$, $A^o \subseteq A \subseteq \overline{A}$, and from (c),

$$\begin{aligned} \mathbb{P}(X \in A^o) &\leq \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A^o) \leq \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A) \\ &\leq \limsup_{n\to\infty} \mathbb{P}_n(X_n \in A) \leq \limsup_{n\to\infty} \mathbb{P}_n(X_n \in \overline{A}) \leq \mathbb{P}(X \in \overline{A}) \end{aligned}$$

where the last step follows from (d). Finally, since

$$\mathbb{P}(X \in \overline{A}) - \mathbb{P}(X \in A^o) = \mathbb{P}(\{X \in \overline{A}\} \setminus \{X \in A^o\}) = \mathbb{P}(X \in (\overline{A} \setminus A^o)) = \mathbb{P}(X \in \partial A),$$

which is 0 by our assumption, i.e., inequalities above are all equalities. In particular, since

$$\liminf_{n\to\infty} \mathbb{P}_n(X_n \in A) \leq \lim_{n\to\infty} \mathbb{P}_n(X_n \in A) \leq \limsup_{n\to\infty} \mathbb{P}_n(X_n \in A)$$

and $\mathbb{P}(X \in A^o) \leq \mathbb{P}(X \in A) \leq \mathbb{P}(X \in \overline{A})$, $\mathbb{P}(X \in A) = \lim_{n\to\infty} \mathbb{P}_n(X_n \in A)$. ⊛

Finally, we prove the following.

**Claim.** (e) ⇒ (a).

**Proof.** For every $g : \mathbb{R}^d \to \mathbb{R}$ bounded and continuous, we want to show $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$. Suppose $g \geq 0$,[a] and let $K \geq g(x)$ for every $x \in \mathbb{R}^d$ (which exists since $g$ is bounded), then

$$\mathbb{E}_n[g(X_n)] = \int_0^K \mathbb{P}_n(g(X_n) > t)\,\mathrm{d}t, \quad \mathbb{E}[g(X)] = \int_0^K \mathbb{P}(g(X) > t)\,\mathrm{d}t,$$

so we just need to prove the convergence of the above two integrals. From bounded convergence theorem, it suffices to show that for almost every $t \in [0, K]$,

$$\mathbb{P}_n(g(X_n) > t) \to \mathbb{P}(g(X) > t).$$

Observe that $\mathbb{P}_n(g(X_n) > t) = \mathbb{P}_n(X_n \in \{g > t\})$ and $\mathbb{P}(g(X) > t) = \mathbb{P}(X \in \{g > t\})$, so from (e) with $A \coloneqq \{g > t\}$, it suffices to show $\mathbb{P}(X \in \partial\{g > t\}) = 0$ for almost all $t$. Firstly,

$$\mathbb{P}(X \in \partial\{g > t\}) = \mathbb{P}(X \in \overline{\{g > t\}} \setminus \{g > t\}^o) = \mathbb{P}(X \in \overline{\{g \geq t\}} \setminus \{g > t\}) = \mathbb{P}(g(X) = t).$$

Moreover, consider the events $\{g(X) = t\}_{t \in [0,K]}$, which are pairwise disjoint, hence Lemma 2.2.2 implies $\mathbb{P}(g(X) = t) = 0$ for all but countably many $t$'s, exactly what we want to show. ⊛

[a] Otherwise, we consider $g = g^+ - g^-$ where $g^+ = \max(g, 0)$ and $g^- = \max(-g, 0)$, and everything follows.

This finishes the proof. ■

### 2.2.4 Continuous Mapping Theorem

A common scenario is that given a nice function $h$ (in terms of continuity), if $X_n \xrightarrow{\text{w}} X$, we want to know when will $h(X_n) \xrightarrow{\text{w}} h(X)$. To develop the theorem of this, we need some more facts about metric spaces.

> **As previously seen.** Given two metric spaces $(S, \rho)$, $(S', \rho')$, $g\colon S \to S'$ is continuous if $x_n \xrightarrow{\rho} x$ implies $g(x_n) \xrightarrow{\rho'} g(x)$, or for open $A \subseteq S'$, $g^{-1}(A) \subseteq S$ is open.

> **Notation.** We sometimes write $g^{-1}(A) =: \{g \in A\}$.

It's clear that the following holds.

> **Note.** If $g\colon S \to S'$ is continuous and $A \subseteq S'$ is closed, then $\overline{\{g \in A\}} = \{g \in \overline{A}\}$.

However, when $g$ is not continuous and $A$ is not closed, the situation is a bit more complicated. But at least we can first look at the set where $g$ is continuous.

> **Notation** (Continuous set). For any $g\colon S \to S'$, we denote the *continuous set* as $C_g := \{x \in S\colon g \text{ is continuous at } x\}$.

Then we have the following.

> **Proposition 2.2.1.** Given $g\colon S \to S'$ between metric spaces and $A \subseteq S'$,
>
> $$C_g \cap \overline{\{g \in A\}} \subseteq \{g \in \overline{A}\}.$$
>
> **Proof.** Let $x \in C_g \cap \overline{\{g \in A\}}$. Since $x \in \overline{\{g \in A\}}$, there exists $(x_n) \in \{g \in A\}$ such that $x_n \xrightarrow{\rho} x$. Moreover, $x \in C_g$ implies $g$ is continuous at $x$, hence $g(x_n) \xrightarrow{\rho'} g(x)$, i.e., $g(x) \in \overline{A}$. ■

This allows us to prove the following theorem, which answers our main question in this section.

> **Theorem 2.2.2** (Continuous mapping theorem). Consider $X_n \xrightarrow{\text{w}} X$ and $h\colon \mathbb{R}^d \to \mathbb{R}^m$. If $\mathbb{P}(X \in C_h) = 1$, then $h(X_n) \xrightarrow{\text{w}} h(X)$.
>
> **Proof.** Let $A \subseteq \mathbb{R}^m$ be a closed set. Then from Portmanteau theorem (d), we need to show
>
> $$\limsup_{n\to\infty} \mathbb{P}_n(h(X_n) \in A) \le \mathbb{P}(h(X) \in A).$$
>
> Since $\limsup_{n\to\infty} \mathbb{P}_n(h(X_n) \in A) = \limsup_{n\to\infty} \mathbb{P}_n(X_n \in \{h \in A\})$, implying
>
> $$\limsup_{n\to\infty} \mathbb{P}_n(h(X_n) \in A) \le \limsup_{n\to\infty} \mathbb{P}_n(X_n \in \overline{\{h \in A\}}) \le \mathbb{P}(X \in \overline{\{h \in A\}}),$$
>
> where the last inequality follows again from Portmanteau theorem (d) since $\overline{\{h \in A\}}$ is clearly closed and $X_n \xrightarrow{\text{w}} X$. Finally, as $\mathbb{P}(X \in C_h) = 1$,
>
> $$\mathbb{P}(X \in \overline{\{h \in A\}}) = \mathbb{P}(X \in \overline{\{h \in A\}} \cap C_h) \le \mathbb{P}(X \in \{h \in \overline{A}\})$$
>
> from Proposition 2.2.1, i.e.,
>
> $$\limsup_{n\to\infty} \mathbb{P}_n(h(X_n) \in A) \le \mathbb{P}(X \in \{h \in \overline{A}\}) = \mathbb{P}(X \in \{h \in A\}) = \mathbb{P}(h(X) \in A)$$
>
> since $A$ is closed, hence we're done. ■

> **Example.** Let $d = 1$ and $X_n \xrightarrow{\text{w}} X$ where $X$ is continuous. Then $1/X_n \xrightarrow{\text{w}} 1/X$ and $X_n^2 \xrightarrow{\text{w}} X^2$.

**Proof.** For $X_n^2 \xrightarrow{w} X^2$, continuous mapping theorem applies with $h(x) = x^2$. For $1/X_n \xrightarrow{w} 1/X$,

$$h(x) = \begin{cases} \frac{1}{x}, & \text{if } x \neq 0; \\ 0, & \text{if } x = 0 \end{cases}$$

is suitable with $C_h = \mathbb{R} \setminus \{0\}$. To apply continuous mapping theorem, we show $\mathbb{P}(X \in C_h) = 1$. Observe that this is the same as asking $\mathbb{P}(X = 0) = 0$, which is true when $X$ is continuous.[a] ⊛

[a] Even if $X$ is not continuous, as long as this is true we can conclude the same thing.

### 2.2.5 Slutsky's Theorem

Another useful theorem for proving weak convergence is the following.

**Theorem 2.2.3** (Converging together). Let $X_n \xrightarrow{w} X$, and if $Y_n$ on the same probability space as $X_n$ such that $\|X_n - Y_n\| \xrightarrow{p} 0$, i.e., for all $\epsilon > 0$, $\mathbb{P}_n(\|X_n - Y_n\| > \epsilon) \to 0$ as $n \to \infty$. Then, $Y_n \xrightarrow{w} X$.

The following corollary draws connections between weak convergence and convergence in probability.

**Corollary 2.2.1.** If $Y_n \xrightarrow{p} X$, then $Y_n \xrightarrow{w} X$. The converse holds if $\mathbb{P}(X = c) = 1$ for a constant $c$.

**Proof.** By considering $X_n = X$ for all $n$, converging together implies that if $Y_n \xrightarrow{p} X$, $Y_n \xrightarrow{w} X$. Conversely, if $Y_n \xrightarrow{w} c$, from Portmanteau theorem (c), for any fixed $\epsilon > 0$,[a]

$$1 = \mathbb{P}(c \in B(c, \epsilon)) \leq \liminf_{n \to \infty} \mathbb{P}_n(Y_n \in B(c, \epsilon)),$$

implying $\mathbb{P}_n(Y_n \in B(c, \epsilon)) \to 1$, i.e., $\mathbb{P}_n(\|Y_n - c\| < \epsilon) \to 1$. ■

[a] Recall that $B(c, \epsilon)$ is the open ball centered at $c$ with radius $\epsilon$.

**Remark.** Weak convergence doesn't give convergence in probability even if $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n) = (\Omega, \mathscr{F}, \mathbb{P})$.

**Example.** Let $X \sim \mathcal{N}(0, 1)$, $Y_n = -X$ for all $n \geq 1$. Then, $Y_n \xrightarrow{w} X$, but clearly not in probability.

## Lecture 5: Convergence in Distribution and Weak Convergence

Now we prove converging together. 1 Feb. 9:30

**Proof of Theorem 2.2.3.** From Portmanteau theorem (b), we want to prove that $\mathbb{E}_n[g(Y_n)] \to \mathbb{E}[g(X)]$ for all bounded and Lipschitz $g \colon \mathbb{R}^d \to \mathbb{R}$. Specifically, let $|g(x)| \leq C$ for all $x \in \mathbb{R}^d$ and $|g(x) - g(y)| \leq K\|x - y\|$ for all $x, y \in \mathbb{R}^d$. From triangle inequality,

$$|\mathbb{E}_n[g(Y_n)] - \mathbb{E}[g(X)]| \leq |\mathbb{E}_n[g(Y_n)] - \mathbb{E}_n[g(X_n)]| + |\mathbb{E}_n[g(X_n)] - \mathbb{E}[g(X)]|.$$

Since $X_n \xrightarrow{w} X$, the second term goes to 0. As for the first term, we see that

$$\begin{aligned} |\mathbb{E}_n[g(Y_n)] - \mathbb{E}_n[g(X_n)]| &= |\mathbb{E}_n[g(Y_n) - g(X_n)]| \\ &\leq \mathbb{E}_n[|g(Y_n) - g(X_n)|] \\ &= \mathbb{E}_n\left[|g(Y_n) - g(X_n)| \cdot \mathbb{1}_{\|X_n - Y_n\| > \epsilon}\right] + \mathbb{E}_n\left[|g(Y_n) - g(X_n)| \cdot \mathbb{1}_{\|X_n - Y_n\| \leq \epsilon}\right] \\ &\leq 2C\mathbb{P}_n(\|X_n - Y_n\| > \epsilon) + K\epsilon\mathbb{P}_n(\|X_n - Y_n\| \leq \epsilon) \\ &\leq 2C\mathbb{P}_n(\|X_n - Y_n\| > \epsilon) + K\epsilon. \end{aligned}$$

As $n \to \infty$, $\limsup_{n \to \infty} |\mathbb{E}_n[g(Y_n)] - \mathbb{E}[g(X)]| \leq K\epsilon$ for all $\epsilon > 0$, by letting $\epsilon \to 0$, we're done. ■

Another characterization regards the difference between marginal and joint weak convergence.

**As previously seen.** $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ if and only if $(X_n, Y_n) \xrightarrow{p} (X, Y)$. Same for $\xrightarrow{\text{a.s.}}$.

However, even if $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n) = (\Omega, \mathscr{F}, \mathbb{P})$, the marginal and joint weak convergences are not equivalent. Specifically, in the case of weak convergence, from continuous mapping theorem, if $(X_n, Y_n) \xrightarrow{\text{w}} (X, Y)$, then $X_n \xrightarrow{\text{w}} X$ and $Y_n \xrightarrow{\text{w}} Y$. However, the converse needs not be true.

**Example.** Let $X_n = X, Y_n = -X$ for all $n \geq 1$. If $X \sim \mathcal{N}(0, 1)$, we see that $\mathbb{P}(X \in A) = \mathbb{P}(-X \in A)$ for all $A \subseteq \mathbb{R}^d$, implying $X_n \xrightarrow{\text{w}} X$ and $Y_n \xrightarrow{\text{w}} X$.

However, this does not imply $(X_n, Y_n) \xrightarrow{\text{w}} (X, X)$ since otherwise, by continuous mapping theorem, $X_n + Y_n \xrightarrow{\text{w}} X + X = 2X$, which is not true since $X_n + Y_n = 0$.

But in the case of $Y$ is a constant, the converse is actually true, and the result is quite useful.

**Theorem 2.2.4** (Slutsky's theorem). If $X_n \xrightarrow{\text{w}} X$ in $\mathbb{R}^d$ and $Y_n \xrightarrow{p} c$ in $\mathbb{R}^m$,[a] then $(X_n, Y_n) \xrightarrow{\text{w}} (X, c)$.

[a] Recall that from Corollary 2.2.1, for a constant $c$, weak convergence is equivalent to convergence in probability.

**Proof.** Firstly, we show that $(X_n, c) \xrightarrow{\text{w}} (X, c)$. Indeed, since for every continuous and bounded $g\colon \mathbb{R}^{d+m} \to \mathbb{R}$, from $X_n \xrightarrow{\text{w}} X$ with $g(\cdot, c)$ being continuous and bounded, $\mathbb{E}_n[g(X_n, c)] \to \mathbb{E}[g(X, c)]$.

Secondly, we show that $\|(X_n, Y_n) - (X_n, c)\| \xrightarrow{p} 0$. This is easy since

$$\|(X_n, Y_n) - (X_n, c)\| \leq \|X_n - X_n\| + \|Y_n - c\| = \|Y_n - c\|,$$

which goes to 0 in probability. Combining the above with converging together gives the result. ■

Revisiting the counter-example, we see that now it's not the case when $Y$ is a constant.

**Corollary 2.2.2.** If $X_n \xrightarrow{\text{w}} X$ and $Y_n \xrightarrow{p} c$ in $\mathbb{R}^d$, $X_n \pm Y_n \xrightarrow{\text{w}} X \pm c$, $X_n \cdot Y_n \xrightarrow{\text{w}} X \cdot c$. If $d = 1$ and $c \neq 0$, then $X_n / Y_n \xrightarrow{\text{w}} X/c$.

**Proof.** This follows directly from Slutsky's theorem and continuous mapping theorem. ■

## 2.3 Convergence in Distribution

The convergences we have been talking about applies to general probability space, not necessarily $\mathbb{R}^d$. However, compared to weak convergence, $\mathbb{R}^d$ is considered first in terms of distributional convergence.

**Intuition.** There's a conical ordering available in $\mathbb{R}^d$ to define $F_X$ and $F_{X_n}$.

**Definition 2.3.1** (Converge in distribution). Let $(X_n)$ and $X$ be random vectors in $\mathbb{R}^d$. Then $(X_n)$ *converges in distribution* to $X$, denoted as $X_n \xrightarrow{D} X$, if for all $(t_1, \ldots, t_d) \in C_{F_X}$,

$$F_{X_n}(t_1, \ldots, t_d) \to F_X(t_1, \ldots, t_d).$$

Specifically, to see how this relates to what we have seen, recall that

$$F_{X_n}(t_1, \ldots, t_d) = \mathbb{P}_n(X_n^i \leq t_i, \forall 1 \leq i \leq d) = \mathbb{P}_n(X_n \in (-\infty, t_1] \times \cdots \times (-\infty, t_d]),$$

same for $F_X$. So this reduces to the form we're familiar with, i.e., $\mathbb{P}_n(X_n \in A)$ for some $A$. Let's make some remarks for this new notion of convergence.

**Remark.** $X_n \xrightarrow{\text{TV}} X$ implies $X_n \xrightarrow{D} X$.

**Proof.** Since $X_n \xrightarrow{\text{TV}} X$ means $\mathbb{P}_n(X_n \in A) \to \mathbb{P}(X \in A)$ uniformly in $A$, but $X_n \xrightarrow{D} X$ only requires the above holds for $A$ in the form of $(-\infty, t_1] \times \cdots \times (-\infty, t_d]$, which is weaker. ⊛

There are more classical results that are worth mentioning.

**Remark** (De Moivre's central limit theorem). Let $X_n \sim \text{Bin}(n,p)$, then for every $t \in \mathbb{R}$, as $n \to \infty$,

$$\mathbb{P}\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq t\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2}\, \mathrm{d}u = \Phi(t).$$

**Proposition 2.3.1.** Let $X_n$ and $X$ be in $\mathbb{Z}$ such that $f_n$ and $f$ are their corresponding pmf's, then

$$f_n \to f \Leftrightarrow X_n \overset{\text{TV}}{\to} X \Leftrightarrow X_n \overset{D}{\to} X.$$

**Proof.** The forward implications are clear, so we just need to show $X_n \overset{D}{\to} X$ implies $f_n \to f$. Since for every $t \in \mathbb{Z}$, since $X_n$ and $X$ are discrete in $\mathbb{Z}$, for some $\epsilon > 0$ small enough,

$$f_n(t) = \mathbb{P}_n(X_n = t) = \mathbb{P}_n(X_n \leq t + \epsilon) - \mathbb{P}_n(X_n \leq t - \epsilon).$$

Since $t \pm \epsilon \in C_X$, $X_n \overset{D}{\to} X$ implies $\mathbb{P}_n(X_n \leq t+\epsilon) \to \mathbb{P}(X \leq t+\epsilon)$. The same holds for $t-\epsilon$, hence

$$\begin{aligned} f_n(t) = \mathbb{P}_n(X_n = t) &= \mathbb{P}_n(X_n \leq t + \epsilon) - \mathbb{P}_n(X_n \leq t - \epsilon) \\ &\to \mathbb{P}(X \leq t + \epsilon) - \mathbb{P}(X \leq t - \epsilon) = \mathbb{P}(X = t) = f(t). \end{aligned}$$

As this holds for every $t \in \mathbb{Z}$, we're done. ■

One important remark is the following.

**Remark.** It's necessary to not require the condition for all $t \in \mathbb{R}^d$, but only $t \in C_{F_X}$.

**Proof.** Consider for $d = 1$ with $X = c \in \mathbb{R}$, i.e., $F_X$ is the step function at $c$. To show $X_n \overset{D}{\to} c$, we don't have to show $\mathbb{P}_n(X_n \leq c) \to \mathbb{P}(X \leq c) = 1$. Otherwise, if we need to show this for all $t$, in particular, $c$, $X_n = c + 1/n$ would not satisfy this. ⊛

In terms of continuity, if $X_n \overset{D}{\to} X$ and $X$ is continuous, then $F_{X_n}$ converges to $F_X$ not only point-wise, but uniformly. Specifically, we have the following.

**Remark** (Pólya's theorem). If $F_X$ is continuous, $X_n \overset{D}{\to} X$ is equivalent as

$$\sup_{t \in \mathbb{R}^d} |F_{X_n}(t) - F_X(t)| \to 0.$$

### 2.3.1 Equivalency of Convergence in Distribution and Weak Convergence

Surprisingly, convergence in distribution is actually just a renaming of weak convergence in $\mathbb{R}^d$.

**Theorem 2.3.1.** Given $(X_n)$ and $X$ in $\mathbb{R}^d$, $X_n \overset{\text{w}}{\to} X$ if and only if $X_n \overset{D}{\to} X$.

**Proof.** We prove for the case of $d = 1$, then it's easy to see the same holds for $d \geq 1$. For the forward direction, we want to show that for all $t \in C_{F_X}$, $\mathbb{P}_n(X_n \leq t) \to \mathbb{P}(X \leq t)$. Note that

$$\mathbb{P}(X \leq t) = \mathbb{P}(X \in (-\infty, t]), \text{ and } \mathbb{P}_n(X_n \leq t) = \mathbb{P}_n(X_n \in (-\infty, t]),$$

hence, from Portmanteau theorem (e) with $A = (-\infty, t]$, $X_n \overset{\text{w}}{\to} X$ is equivalently to $\mathbb{P}_n(X_n \leq t) \to \mathbb{P}(X \leq t)$ if $\mathbb{P}(X \in \partial A) = 0$, i.e.,

$$\mathbb{P}(X \in \partial(-\infty, t]) = \mathbb{P}(X \in \{t\}) = \mathbb{P}(X = t) = 0,$$

which is true since $t \in C_{F_X}$.

To show the backward direction, we need the following lemma.

**Lemma 2.3.1.** $X_n \xrightarrow{D} X$ if and only if for all $x \in \mathbb{R}^d$,

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

**Proof.** The backward direction is clear, so we prove the forward direction. When $x \in C_{F_X}$, we're clearly done, so consider $x \notin C_{F_X}$. Firstly, note that $|C_{F_X}^c|$ is countable, so there exists $(x_k) \nearrow x$ and $(y_k) \searrow x$, both in $C_{F_X}$. Hence, for all $n \ge 1$ and $k \ge 1$,

$$F_{X_n}(x_k) \le F_{X_n}(x) \le F_{X_n}(y_k)$$

as $F_{X_n}$ is increasing. We now have for every $k \ge 1$,

$$\begin{aligned}
F_X(x_k) &= \lim_{n\to\infty} F_{X_n}(x_k) && x_k \in C_{F_X} \\
&\le \liminf_{n\to\infty} F_{X_n}(x^-) \\
&\le \liminf_{n\to\infty} F_{X_n}(x) && F_{X_n} \text{ is increasing} \\
&\le \limsup_{n\to\infty} F_{X_n}(x) \\
&\le \limsup_{n\to\infty} F_{X_n}(y_k) = F_X(y_k). && y_k \in C_{F_X}
\end{aligned}$$

By taking $k \to \infty$, $F_X(x_k) \to F_X(x^-)$, while $F_X(y_k) \to F_X(x)$,[a] and we're done.

[a] Recall that the distribution function is always right-continuous.

*The proof will be continued...*

# Lecture 6: Stochastic Boundedness and Delta Theorem

Before we finish the proof of Theorem 2.3.1, we recall one important characterization of lim inf. 2 Feb. 17:30

**As previously seen.** Given two real sequence $x_n$ and $y_n$,

$$\liminf_{n\to\infty}(x_n + y_n) \ge \liminf_{n\to\infty} x_n + \liminf_{n\to\infty} y_n,$$

where the equality holds when either $x_n$ or $y_n$ converges (not if and only if).

We can then finish the proof of Theorem 2.3.1.

**Proof of Theorem 2.3.1 (cont.)** Now we can prove the backward direction. Form Portmanteau theorem (c), it suffices to show that for every open $A \subseteq \mathbb{R}$, we have

$$\mathbb{P}(X \in A) \le \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A).$$

From the elementary analysis, we see that it suffices to show when $A = (a, b)$ since when $A \subseteq \mathbb{R}$ is open, one can write $A = \bigcup_{k=1}^{\infty}(a_k, b_k)$ where $(a_k, b_k)$'s disjoint, and have

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{k=1}^{\infty} \mathbb{P}(X \in (a_k, b_k)) \\
&\le \sum_{k=1}^{\infty} \liminf_{n\to\infty} \mathbb{P}_n(X_n \in (a_k, b_k)) && \text{assume true for each } (a_k, b_k) \\
&\le \liminf_{n\to\infty} \sum_{k=1}^{\infty} \mathbb{P}_n(X_n \in (a_k, b_k)) = \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A),
\end{aligned}$$

where the last inequality follows from an induction on $\liminf_{n\to\infty}(x_n + y_n) \ge \liminf_{n\to\infty} x_n + \liminf_{n\to\infty} y_n$. Now, we show that $\mathbb{P}(X \in A) \le \liminf_{n\to\infty} \mathbb{P}_n(X_n \in A)$ when $A = (a, b)$.

**Claim.** $\mathbb{P}(X \in (a,b)) \leq \liminf_{n\to\infty} \mathbb{P}_n(X_n \in (a,b))$.

**Proof.** Observe that $\mathbb{P}(X \in (a,b)) = F_X(b^-) - F_X(a)$, with Lemma 2.3.1, we further have

$$\begin{aligned}
\mathbb{P}(X \in (a,b)) &= F_X(b^-) - F_X(a) \\
&\leq \liminf_{n\to\infty} F_{X_n}(b^-) - \left(\limsup_{n\to\infty} F_{X_n}(a)\right) \\
&\leq \liminf_{n\to\infty} F_{X_n}(b^-) + \liminf_{n\to\infty}(-F_{X_n}(a)) \\
&\leq \liminf_{n\to\infty} \left(F_{X_n}(b^-) - F_{X_n}(a)\right) = \liminf_{n\to\infty} \mathbb{P}_n(X_n \in (a,b)),
\end{aligned}$$

which proves the claim. ⊛

This proves the case of $d = 1$. ■

Theorem 2.3.1 means that when talking about random vectors, we can use every result we have proved for the case of weak convergence. Let's see one application.

**Proposition 2.3.2.** If $X_n \xrightarrow{D} X$ and $t_n \to t \in C_{F_X}$, then $\mathbb{P}_n(X_n \leq t_n) \to \mathbb{P}(X \leq t)$.

**Proof.** We see that from Corollary 2.2.2, $X_n - t_n \xrightarrow{\text{w}} X - t$, i.e., $X_n - t_n \xrightarrow{D} X - t$. Hence,

$$\mathbb{P}_n(X_n \leq t_n) = \mathbb{P}_n(X_n - t_n \leq 0) = F_{X_n - t_n}(0) \to F_{X-t}(0) = \mathbb{P}(X - t \leq 0)$$

as long as $0 \in C_{F_{X-t}}$, i.e., $\mathbb{P}(X - t = 0) = \mathbb{P}(X = t) = 0$, which is just $t \in C_{F_X}$ as we assumed. ■

## 2.4 Stochastic Boundedness

So far we have been talking about the notion of convergence, now we switch the gear a bit and consider boundedness. In this section, let $(X_i)_{i\in I}$ be a family of $d$-dimensional random vectors defined on probability spaces $(\Omega_i, \mathscr{F}_i, \mathbb{P}_i)$, with the non-empty index set $I$, which can be either finite or infinite.

**Definition 2.4.1** (Bounded in probability). $(X_i)_{i\in I}$ is said to be *bounded in probability* if for every $\epsilon > 0$, there exists an $M > 0$ such that for every $i \in I$,

$$\mathbb{P}_i(\|X_i\| \geq M) < \epsilon.$$

In other words, for every $\epsilon > 0$, there is an $M > 0$ such that $\mathbb{P}_i(\|X_i\| < M) \geq 1 - \epsilon$ for every $i \in I$.

**Intuition.** For any arbitrary large probability close to 1 we want, one can find an upper-bound $M$ on $\|X_i\|$ uniformly for all $i \in I$.

**Note.** When $X_i = X$ on $(\Omega, \mathscr{F}, \mathbb{P})$ for every $i \in I$, $(X_i)_{i\in I}$ is trivially bounded in probability.

**Proof.** Since if not, there exists $\epsilon > 0$, for every $M > 0$, $\mathbb{P}(\|X\| \geq M) \geq \epsilon$. Then as $M \to \infty$, $\mathbb{P}(\|X\| = \infty) \geq \epsilon$, which is a contradiction since $\|X\| = \infty$. ⊛

**Remark.** When $I$ is finite, $(X_i)_{i\in I}$ is also trivially bounded in probability. On the other hand, when $I$ is infinite, by considering $X_n = n$ (deterministic), which is not bounded in probability anymore.

### 2.4.1 Sufficient Conditions for Stochastic Boundedness

We now provide some sufficient conditions for being bounded in probability.

**Proposition 2.4.1.** If $(X_i)_{i\in I}$ is bounded in $L^p$ for some $p > 0$, i.e., $\sup_{i\in I} \mathbb{E}_i [\|X_i\|^p] < \infty$, then $(X_i)_{i\in I}$ is bounded in probability.

**Proof.** Denote $K := \sup_{i\in I} \mathbb{E}_i [\|X_i\|^p] < \infty$. Since for any $\epsilon > 0$, from Markov's inequality,

$$\mathbb{P}_i(\|X_i\| > M) \leq \frac{\mathbb{E}_i [\|X_i\|^p]}{M^p} \leq \frac{K}{M^p} =: \epsilon$$

for $M := \sqrt[p]{K/\epsilon}$. Hence, we're done. ■

We can generalize some relations between convergence and boundedness from the elementary analysis.

**As previously seen.** If a deterministic sequence in $\mathbb{R}$ converges, then it's bounded.

In our context, we might expect something like "if $X_n \xrightarrow{p} X$, then $(X_n)$ is bounded in probability." In fact, we have the following "stronger" result where we only require convergence in distribution.

**Proposition 2.4.2.** If $X_n \xrightarrow{D} X$, then $(X_n)$ is bounded in probability.

**Proof.** Fix an $\epsilon > 0$. There is an $M > 0$ such that $\mathbb{P}(\|X\| \geq M) < \epsilon$ since this is a single random vector. To relate this back to $X_n$, from Portmanteau theorem (d),

$$\epsilon > \mathbb{P}(\|X\| \geq M) = \mathbb{P}(X \in B^c(0, M)) \geq \limsup_{n\to\infty} \mathbb{P}_n(X_n \in B^c(0, M)) = \limsup_{n\to\infty} \mathbb{P}_n(\|X_n\| \geq M).$$

In other words, $\liminf_{n\to\infty} \mathbb{P}_n(\|X_n\| < M) > 1 - \epsilon$, hence there exists an $n_0$ such that for every $n \geq n_0$, $\mathbb{P}_n(\|X_n\| < M) \geq 1 - \epsilon$. As for those $n < n_0$, since $\{X_n : n < n_0\}$ is a finite family, we can find $M' > 0$ such that $\mathbb{P}_n(\|X_n\| < M') > 1 - \epsilon$ for every $n < n_0$. Finally, by considering $M'' := \max(M, M')$, we have $\mathbb{P}_n(\|X_n\| < M'') > 1 - \epsilon$, i.e., $\mathbb{P}_n(\|X_n\| \geq M'') < \epsilon$ as desired. ■

A kind of converse theorem is called Prokhorov's theorem, but we won't prove it here right now. We now see another useful characterization that generalizes our intuition in $\mathbb{R}$. Recall the following.

**As previously seen.** In $\mathbb{R}$, if $a_n \to 0$ and $b_n$ is bounded, $a_n b_n \to 0$.

The generalization is the following.

**Proposition 2.4.3.** Let $d = 1$ such that $(X_n)$ and $(Y_n)$ are defined on the same probability space. If $X_n \xrightarrow{p} 0$ and $Y_n$ is bounded in probability, then $X_n Y_n \xrightarrow{p} 0$.

**Proof.** Fix an $\epsilon > 0$. We want to show that $\mathbb{P}_n(|X_n Y_n| > \epsilon) \to 0$. This is because

$$\begin{aligned}\mathbb{P}_n(|X_n Y_n| > \epsilon) &= \mathbb{P}_n(|X_n Y_n| > \epsilon, |Y_n| > M) + \mathbb{P}_n(|X_n Y_n| > \epsilon, |Y_n| \leq M) \\ &\leq \mathbb{P}_n(|Y_n| > M) + \mathbb{P}_n(|X_n Y_n| > \epsilon, |Y_n| \leq M) \leq \mathbb{P}_n(|Y_n| > M) + \mathbb{P}_n(|X_n| > \epsilon/M)\end{aligned}$$

for any $M$. Now, we see that

- since $Y_n$ is bounded in probability, there's an $M > 0$ such that $\mathbb{P}_n(|Y_n| > M) < \epsilon$ for all $n$;
- since $X_n \xrightarrow{p} 0$, for the $M$ (depends on the fixed $\epsilon$) above, $\mathbb{P}_n(|X_n| > \epsilon/M) \to 0$ as $n \to \infty$.

We see that the second term always goes to 0, while the first term can always be upper-bounded by $\epsilon$. Hence, by letting $\epsilon \to 0$, we're done. ■

We often write the following.

**Notation.** We write $X_n = o_p(1)$ for $X_n \xrightarrow{p} 0$, and $X_n = O_p(1)$ when $(X_n)$ is bounded in probability.

**Remark.** Proposition 2.4.3 means $o_p(1) \times O_p(1) = o_p(1)$.

### 2.4.2 Delta Method

Let's see one important application which combines the above. Consider an estimator $T_n$ of $\theta$, and a deterministic sequence $b_n$ which goes to $\infty$. In this case, we often have

$$b_n(T_n - \theta) \xrightarrow{D} Y.$$

**Example.** When $X_n \sim \text{Bin}(n, p)$, then for $b_n = \sqrt{n/p(1-p)} \to \infty$, $T_n = X_n/n$, and $\theta = p$, we have

$$\frac{X_n - np}{\sqrt{np(1-p)}} = \sqrt{\frac{n}{p(1-p)}}\left(\frac{X_n}{n} - p\right) = b_n(T_n - \theta) \to Y \sim \mathcal{N}(0,1).$$

This allows us to compute the rate of convergence and the limiting distribution. But what can we say when we care about $g(T_n)$ for a function $g$?

**Theorem 2.4.1** (Delta method). Let $\theta \in \mathbb{R}^d$, $(T_n)$ and $Y$ be random vectors in $\mathbb{R}^d$, and $b_n \to \infty$ be a positive deterministic sequence. If $b_n(T_n - \theta) \xrightarrow{D} Y$, then $T_n \xrightarrow{p} \theta$. Moreover, if $g\colon \mathbb{R}^d \to \mathbb{R}^m$ is differentiable at $\theta$, $b_n(g(T_n) - g(\theta)) \xrightarrow{D} \nabla g(\theta) Y$, where $\nabla g \in \mathbb{R}^{d \times m}$ is the Jacobian of $g$.

**Proof.** We first observe that $\|b_n(T_n - \theta)\| \in O_p(1)$ since $b_n(T_n - \theta) \xrightarrow{D} Y$, with continuous mapping theorem and the fact that $\|\cdot\|$ is continuous, $\|b_n(T_n - \theta)\| \xrightarrow{p} \|Y\|$, so $\|b_n(T_n - \theta)\| \in O_p(1)$ by Proposition 2.4.2. With this, as $b_n \to \infty$, $T_n \xrightarrow{p} \theta$ since

$$\|T_n - \theta\| = \frac{1}{b_n}\|b_n(T_n - \theta)\| = o(1)O_p(1) \xrightarrow{p} 0$$

as $o(1)O_p(1) = o_p(1)$ from Proposition 2.4.3. For the second claim, since $g$ is differentiable at $\theta$,

$$\frac{g(x) - g(\theta) - \nabla g(\theta)(x - \theta)}{\|x - \theta\|} \to 0$$

when $x \to \theta$. Let $r(x) \coloneqq g(x) - g(\theta) - \nabla g(\theta)(x - \theta)$ for $x \in \mathbb{R}^d$ be the remainder, and consider

$$h(x) = \begin{cases} 0, & \text{if } x = \theta; \\ \dfrac{r(x)}{\|x - \theta\|}, & \text{if } x \neq \theta, \end{cases}$$

which is continuous at $\theta$. Rewriting everything, we have

$$r(x) = g(x) - g(\theta) - \nabla g(\theta)(x - \theta) = h(x)\|x - \theta\|$$

for every $x \in \mathbb{R}^d$. Now, let $x = T_n$, multiply both sides by $b_n$, and take the norm, we see that

$$\|b_n\left(g(T_n) - g(\theta)\right) - \nabla g(\theta) b_n(T_n - \theta)\| = \|h(T_n)\|\|b_n(T_n - \theta)\|.$$

We observe the following.

**Claim.** It suffices to show that the right-hand sides goes to 0 in probability.

**Proof.** Since it implies that $b_n(g(T_n) - g(\theta))$ has the same weak limit as $\nabla g(\theta) b_n(T_n - \theta)$ from converging together, i.e., $\nabla g(\theta) Y$ from our assumption with continuous mapping theorem. ⊛

It's enough to show $\|h(T_n)\| = o_p(1)$ since we know that $\|b_n(T_n - \theta)\| = O_p(1)$ and $o_p(1)O_p(1) = o_p(1)$ from Proposition 2.4.3. Indeed, as $T_n \xrightarrow{p} \theta$, $h(T_n) \xrightarrow{p} h(\theta) = 0$ again by continuous mapping theorem with $h$ being continuous at $\theta$. This further implies $\|h(T_n)\| \xrightarrow{p} 0$ as we desired.[a] Combining the above, the result follows. ■

[a] This involves continuous mapping theorem and Corollary 2.2.1 since $h(\theta) = 0$, a constant (so does its norm).

Hence, we see that the answer to our original question is rather simple: as $b_n(T_n - \theta) \xrightarrow{D} Y$,

$$b_n(g(T_n) - g(\theta)) \xrightarrow{D} \nabla g(\theta) \cdot Y$$

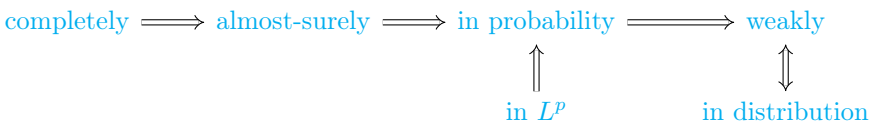
for any differentiable $g$ at $\theta$.

# Lecture 7: Skorohod's Representation Theorem

## 2.5 Skorohod's Representation Theorem

6 Feb. 9:30

So far, we have seen the following.




Now, we show an interesting result that one might not expect.

**Theorem 2.5.1** (Skorohod's representation theorem)**.** If $X_n \xrightarrow{D} X$, there exists $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$ on which we can define random vectors $(Y_n)$ and $Y$ such that $Y_n \overset{D}{=} X_n$ for all $n$ and $Y \overset{D}{=} X$, and $\widetilde{\mathbb{P}}(Y_n \to Y) = 1$.

**Intuition.** We have convergence in distribution "implies" almost surely convergence.

### 2.5.1 Quantile Function

We want to prove Skorohod's representation theorem for $d = 1$. To start, say $X \sim F$ on $(\Omega, \mathscr{F}, \mathbb{P})$. We will consider $F^{-1}(p)$, which exists if there exists a unique $t \in \mathbb{R}$ such that $F(t) = p$, then $F^{-1}(p) = t$. However, this is not really practical since in the discrete case, the preimage might not exist; and even if in the continuous $F$, when $F$ flats out (at $p = 1$), the preimage is not unique.

**Definition 2.5.1** (Quantile)**.** A $p^{th}$ *quantile* of $X$ is defined as any $t \in \mathbb{R}$ such that

$$\mathbb{P}(X \le t) \ge p \ge \mathbb{P}(X < t).$$

Now, we can define $F^{-1}(p)$ as the smallest quantile.

**Definition 2.5.2** (Quantile function)**.** The *quantile function* of $X \sim F$ is defined as

$$F^{-1}(p) = \inf\{t \in \mathbb{R} \colon F(t) \ge p\}.$$

We sometimes also call $F^{-1}$ as the *generalized inverse* of $F$.

**Remark.** $t \ge F^{-1}(p)$ if and only if $F(t) \ge p$; in other words, $t < F^{-1}(p)$ if and only if $F(t) < p$.

One application of $F^{-1}$ is that given any cdf $F$, we can construct a corresponding random variable.

**Remark** (Construction of random variable)**.** Let $U \sim \mathcal{U}(0,1)$ be a uniform random variable on $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$. Then, $F^{-1}(U) =: Y$ is a random variable with cdf $F$.

**Proof.** Since for any $t \in \mathbb{R}$,

$$\widetilde{\mathbb{P}}(Y \le t) = \widetilde{\mathbb{P}}(F^{-1}(U) \le t) = \mathbb{P}(U \le F(t)) = F(t).$$

⊛

### 2.5.2 Proof of Skorohod's representation theorem

Now we can prove Skorohod's representation theorem.

**Proof of Theorem 2.5.1.** Consider $\widetilde{\Omega} = (0,1)$, and $\widetilde{\mathbb{P}}((a,b)) = b - a$ for all $a < b$. Then, we can define $U(p) = p$ for all $p \in \widetilde{\Omega}$, i.e., $U \sim \mathcal{U}(0,1)$. Define $Y_n = F_{X_n}^{-1}(U)$ and $Y = F_X^{-1}(U)$ from the quantile functions. Denote $\Phi$ be the cdf of $\mathcal{N}(0,1)$, and let $Z = \Phi^{-1}(U)$.

It's clear that $Y_n \overset{D}{=} X_n$ and $Y \overset{D}{=} X$, so we just need to show $\widetilde{\mathbb{P}}(Y_n \to Y) = 1$.

**Claim.** It's equivalent to $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$ for almost all $p$'s.

**Proof.** Observe further that $Y_n(p) = F_{X_n}^{-1}(p)$, $Y(p) = F_X^{-1}(p)$, and $Z(p) = \Phi^{-1}(p)$ for all $p \in (0,1)$. Since for almost all $p$'s, $Y_n(p) \to Y(p)$ if and only if $\Phi(Y_n(p)) \to \Phi(Y(p))$ as $\Phi$ is strictly increasing and continuous, or equivalently,

$$\Phi(Y_n(p)) = \widetilde{\mathbb{P}}(Z \le Y_n(p)) \to \widetilde{\mathbb{P}}(Z \le Y(p)) = \Phi(Y(p)).$$

As $Z$ is continuous, this is equivalent to $\widetilde{\mathbb{P}}(Z < Y_n(p)) \to \widetilde{\mathbb{P}}(Z < Y(p))$, i.e.,

$$\widetilde{\mathbb{P}}(Z < F_{X_n}^{-1}(p)) \to \widetilde{\mathbb{P}}(Z < F_X^{-1}(p)),$$

which holds if and only if $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$.[a] ⊛

[a] Follows from the reamrk. Explicitly, firstly, it's equivalent to $\widetilde{\mathbb{P}}(Z \ge F_{X_n}^{-1}(p)) \to \widetilde{\mathbb{P}}(Z \ge F_X^{-1}(p))$, and with $\widetilde{\mathbb{P}}(Z \ge F_{X_n}^{-1}(p)) = \widetilde{\mathbb{P}}(F_{X_n}(Z) \ge p)$ and $\widetilde{\mathbb{P}}(Z \ge F_X^{-1}(p)) = \widetilde{\mathbb{P}}(F_X(Z) \ge p)$, the result follows.

Now we show $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$ for almost all $p$'s. Since $X_n \overset{D}{\to} X$ means $F_{X_n}(t) \to F_X(t)$, from Lemma 2.3.1, it further implies $F_{X_n}(t^-) \to F_X(t^-)$ for all $t \in C_{F_X}$. Note that $\widetilde{\mathbb{P}}(Z \in C_{F_X}) = 1$ since there can be only countably many discontinuities of $F_X$. Hence,

$$\widetilde{\mathbb{P}}(F_{X_n}(Z) \to F_X(Z)) = 1,$$

i.e., converges almost surely, which implies $F_{X_n}(Z) \overset{D}{\to} F_X(Z)$, i.e., for all $p \in C_{F_X(Z)}$

$$\widetilde{\mathbb{P}}(F_{X_n}(Z) \le p) \to \widetilde{\mathbb{P}}(F_X(Z) \le p),$$

and also $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$ from Lemma 2.3.1. Again, as $F_X$ can have only countably many discontinuities, this holds for almost all $p$'s, which is what we want to show. ■

We now see some applications of Skorohod's representation theorem, where we can obtain relatively simple proofs for several theorems, such as Theorem 2.3.1.

**Remark.** Theorem 2.3.1 can be proved from Skorohod's representation theorem.

**Proof.** If $X_n \overset{D}{\to} X$, from Skorohod's representation theorem, we can obtain $Y_n \overset{\text{a.s.}}{\to} Y$ on $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$ such that $X_n \overset{D}{=} Y_n$ and $X \overset{D}{=} Y$. Then for any bounded and continuous $g$,

$$\mathbb{E}[g(X_n)] = \widetilde{\mathbb{E}}[g(Y_n)] \to \widetilde{\mathbb{E}}[g(Y)] = \mathbb{E}[g(X)]$$

by the bounded convergence theorem, which proves $X_n \overset{\text{w}}{\to} X$. ⊛

Another application is to generalize Fatou's lemma.

**Proposition 2.5.1** (Fatou's lemma). Let $X_n \overset{D}{\to} X$[a] and $g\colon \mathbb{R}^d \to [0,\infty)$ continuous. Then

$$\mathbb{E}[g(X)] \le \liminf_{n\to\infty} \mathbb{E}_n[g(X_n)].$$

[a] Can be on different probability spaces.

**Proof.** Let $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$, from Skorohod's representation theorem, we can construct $Y_n \overset{D}{=} X_n$, $Y \overset{D}{=} X$, and $Y_n \overset{\text{a.s.}}{\to} Y$, which implies $g(Y_n) \overset{\text{a.s.}}{\to} g(Y)$. From Fatou's lemma in $d = 1$, $\widetilde{\mathbb{E}}[g(Y)] \le \liminf_{n\to\infty} \widetilde{\mathbb{E}}[g(Y_n)]$. The result then follows directly from

$$\mathbb{E}[g(X)] = \widetilde{\mathbb{E}}[g(Y)] \le \liminf_{n\to\infty} \widetilde{\mathbb{E}}[g(Y_n)] = \liminf_{n\to\infty} \mathbb{E}_n[g(X_n)].$$

■

The following is well-known from real analysis dominated convergence theorem.

**Theorem 2.5.2.** If $X_n \overset{\text{a.s.}}{\to} X$, $g\colon \mathbb{R}^d \to \mathbb{R}$ is continuous and $(g(X_n))$ is uniformly integrable[a] if and only if $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$.

in $L_1$?

[a] I.e., $\lim_{t\to\infty} \sup_{n\ge 1} \mathbb{E}[|g(X_n)| \mathbb{1}_{g(X_n)\ge t}] = 0$.

If $X_n \overset{\text{w}}{\to} X$, then from the definition, we will have $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$ if $g$ is continuous and bounded. We can indeed relax both continuity and boundedness as follows.

**Proposition 2.5.2.** If $X_n \overset{\text{w}}{\to} X$ and $\mathbb{P}(X \in C_g) = 1$ where $g\colon \mathbb{R}^d \to \mathbb{R}$ such that $(g(X_n))$ is uniformly integrable, then $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$.

Seems no need of $(g(X_n))$ being u.i.

**Proof.** From $\mathbb{P}(X \in C_g) = 1$ and $X_n \overset{\text{w}}{\to} X$, from continuous mapping theorem, $g(X_n) \overset{\text{w}}{\to} g(X)$, hence $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$. ■

**Remark.** Proposition 2.5.2 can be proved with Skorohod's representation theorem also.

the in $L_1$ version?

## 2.6 Characteristic Function

It turns out that convergence in distribution has a very neat characterization. To motivate the idea, consider the problem of proving $X_n \overset{D}{\to} X$, which is usually inefficient if we start from the definition. To get some intuition for potential proof strategies, consider a deterministic sequence $(x_n)$ in a metric space $(S, \rho)$.

**Theorem 2.6.1.** $(x_n) \to x$ if and only if every subsequence of $(x_n)$ has a subsequence that converges to the same limit $x$.

**Proof.** The forward direction is clear. For the backward direction, if not, there exists $(x_{n_k})$ and $\epsilon > 0$ such that $\rho(x_{n_k}, x) \ge \epsilon$ for every $k \ge 1$. But if there exists a subsubsequence $(x_{n_{k_\ell}})$ that converges to $x$, this is clearly a contradiction. ■

In the same vein, with the same argument, we have the following.

**Theorem 2.6.2.** $X_n \overset{\text{w}}{\to} X$ if and only if every subsequence of $(X_n)$ has a subsequence that converges weakly, and all weakly convergent subsequences have the same limit $X$.

**Proof.** Mimicking the proof as in Theorem 2.6.1. ■

# Lecture 8: Characteristic Functions

We see other similar theorems apart from Theorem 2.6.2.

8 Feb. 9:30

**Theorem 2.6.3.** If $X_n \overset{\text{w}}{\to} X$ and $X_n \overset{\text{w}}{\to} Y$, then $X \overset{D}{=} Y$. More generally, if $X_n \overset{\text{w}}{\to} X$ and $Y_n \overset{\text{w}}{\to} Y$, with $X_n \overset{D}{=} Y_n$ for all $n \ge 1$, $X \overset{D}{=} Y$.

**Proof.** For every $n \geq 1$, $\mathbb{E}_n[g(X_n)] = \mathbb{E}_n[g(Y_n)]$ for all $g\colon \mathbb{R}^d \to \mathbb{R}$. If $g$ is bounded and continuous, $\mathbb{E}_n[g(X_n)] \to \mathbb{E}[g(X)]$ and $\mathbb{E}_n[g(Y_n)] \to \mathbb{E}[g(Y)]$. To show that $X \overset{D}{=} Y$, we want to show $F_X = F_Y$, or $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ for all $B \in \mathscr{F} = \mathcal{B}(\mathbb{R}^d)$. In fact, it's enough to show this for closed $B$. With Lemma 2.2.1, there exists $(g_k) \searrow \mathbb{1}_B$ for closed $B$ and bounded, Lipschitz $g_k$, i.e.,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_B(X)] = \lim_{k\to\infty} \mathbb{E}[g_k(X)] &= \lim_{k\to\infty} \lim_{n\to\infty} \mathbb{E}_n[g_k(X_n)] \\ &= \lim_{k\to\infty} \lim_{n\to\infty} \mathbb{E}_n[g_k(Y_n)] = \lim_{k\to\infty} \mathbb{E}[g_k(Y)] = \mathbb{E}[\mathbb{1}_B(Y)],\end{aligned}$$

where the third equality follows from the fact that $X_n \overset{D}{=} Y_n$. ■

One question is that, if we don't have things like weak convergent but just some moment information (i.e., when $g(x) = x^k$ when computing $\mathbb{E}[g(X)]$), can we conclude the same thing?

**Problem** (Method of Moments). If $\mathbb{E}[X^k] = \mathbb{E}[Y^k] < \infty$ for all $k \geq 1$, does $X \overset{D}{=} Y$?

**Answer.** Not in general. We will discuss this more in the assignment. ⊛

### 2.6.1 Characteristic Function

To answer the question left above, we will see that it actually suffices to show only for $g(x) = \cos(t \cdot x)$ or $\sin(t \cdot x)$ for $t, x \in \mathbb{R}^d$. This leads to the so-called characteristic functions.

**Definition 2.6.1** (Characteristic function). The *characteristic function* of a $d$-dimensional random vector $X$ is defined as $\phi_X\colon \mathbb{R}^d \to \mathbb{C}$ where $t \in \mathbb{R}^d$ such that

$$\phi_X(t) = \mathbb{E}[\cos(t \cdot X)] + i\mathbb{E}[\sin(t \cdot X)] = \mathbb{E}[e^{i(t \cdot X)}].$$

**Notation.** We sometimes drop the inner product, i.e., write $t \cdot X =: tX$.

If we write $\phi_X$ explicitly, we have

$$\phi_X(t) = \mathbb{E}[e^{itX}] = \int e^{itx} f_X(x)\,\mathrm{d}x = \int e^{itx} F_X(\mathrm{d}x).$$

**Remark.** Characteristic functions are bounded.

**Proof.** Since

$$|\phi_X(t)| = \sqrt{(\mathbb{E}[\cos(tX)])^2 + (\mathbb{E}[\sin(tX)])^2} \leq \sqrt{\mathbb{E}[\cos^2(tX)] + \mathbb{E}[\sin^2(tX)]} = 1.$$

⊛

This implies that $\phi_X$ is meaningful for any random vector $X$, unlike the moment generating function.

**Remark.** If $X$ and $Y$ are independent, $\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$.

We make one more remark for future reference.

**Remark.** If $X, Y$ are discrete, $f_{X+Y}(x) = \sum_y f_Y(x-y) f_X(y)$. More generally, if $X, Y$ have pdfs,

$$f_{X+Y}(x) = \int f_Y(x-y) f_X(y)\,\mathrm{d}y = \int f_Y(x-y) F_X(\mathrm{d}y).$$

Furthermore, even if $X$ doesn't have pdf, as long as $Y$ does, the above still holds.

### 2.6.2 Uniqueness Theorem

Now we can prove the following uniqueness theorem, which states that indeed, it suffices to check only $\sin(tx)$ and $\cos(tx)$ when proving weak convergence.

**Theorem 2.6.4** (Uniqueness). If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^d$, then $X \overset{D}{=} Y$. The converse is trivial.

**Proof.** Consider $d = 1$. Observe that if we can write $F_X$ in terms of only $\phi_X$, then $\phi_X = \phi_Y$ implies $F_X = F_Y$. To do this, consider the following.

**Claim.** For $Z, Z' \sim \mathcal{N}(0,1)$ (independent of $X$ and $Y$), if one can write $F_{X+\sigma Z}$ for all $\sigma > 0$ in terms of only $\phi_X$, $\phi_X = \phi_Y$ implies $X \overset{D}{=} Y$.

**Proof.** Fix some $\sigma > 0$. In this case, if we can write $F_{X+\sigma Z}$ in terms of only $\phi_X$, $\phi_X = \phi_Y$ implies $F_{X+\sigma Z} = F_{Y+\sigma Z'}$. This implies $X + \sigma Z \overset{D}{=} Y + \sigma Z'$. Now, for $\sigma = 1/k$, $k \in \mathbb{N}$,

$$X + \frac{1}{k} Z \overset{D}{=} Y + \frac{1}{k} Z'.$$

With Corollary 2.2.2, since $Z/k \xrightarrow{p} 0$ (and also $Z'/k \xrightarrow{p} 0$), we have $X + Z/k \xrightarrow{D} X$ and $Y + Z'/k \xrightarrow{D} Y$, which implies $X \overset{D}{=} Y$ from Theorem 2.6.3. ⊛

Hence, our goal now is to write $F_{X+\sigma Z}$ in terms of $\phi_X$. Firstly, for all $t \in \mathbb{R}$,

$$\phi_Z(t) = \int e^{itz} F_Z(\mathrm{d}z) = \int e^{itz} f_Z(z)\,\mathrm{d}z = \int e^{itz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,\mathrm{d}z = e^{-t^2/2}. \tag{2.1}$$

Now, consider $f_{X+\sigma Z}(x)$ instead, which exists since $Z$ has a pdf from the remark. We see that

$$\begin{aligned} f_{X+\sigma Z}(x) &= \int f_{\sigma Z}(x-y) F_X(\mathrm{d}y) \\ &= \int \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-y)^2/2\sigma^2} F_X(\mathrm{d}y), \end{aligned}$$

by replacing $e^{-(x-y)^2/2\sigma^2}$ from Equation 2.1 with $t = (x-y)/\sigma$,

$$\begin{aligned} &= \int \frac{1}{\sigma\sqrt{2\pi}} \int e^{i\frac{y-x}{\sigma}z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,\mathrm{d}z F_X(\mathrm{d}y). \\ &= \frac{1}{2\pi} \iint e^{i(y-x)u} e^{-\sigma^2 u^2/2}\,\mathrm{d}u F_X(\mathrm{d}y), && z/\sigma =: u \\ &= \frac{1}{2\pi} \int e^{-ixu-\sigma^2u^2/2} \underbrace{\int e^{iyu} F_X(\mathrm{d}y)}_{\phi_X(u)}\,\mathrm{d}u, \end{aligned}$$

where we interchange the order of integrals with Fubini's theorem (justified by Tonelli's theorem) when integrands are absolute integrable. This implies that $F_{X+\sigma Z}(\mathrm{d}x)$ can be written in terms of $\phi_X$ where with no other dependencies, hence we're done. ■

**Note.** Now showing $X \overset{D}{=} Y$ reduces to calculus.

### 2.6.3 Continuity Theorem

One immediate consequence of the uniqueness theorem is that it's enough to have the characteristic functions converging to some function (not necessarily a characteristic functions of some $X$) for us to conclude that the subsequences of $(X_n)$ have the same weak limit. To do this, we need to prove Prokhorov's theorem.

**Theorem 2.6.5** (Prokhorov's theorem). If $(X_n) = O_p(1)$, then there exists a weakly convergent subsequence of $(X_n)$.

**Proof.** Based on Helly's selection theorem, $F_{X_n}(t) \to F(t)$ for all $t \in C_F$, there exists an increasing $F$, right continuous, $F(+\infty) \leq 1$ and $F(-\infty) \geq 0$ (called the *defective cdf*). Consider $d = 1$, we show that this $F$ is indeed a cdf when $X_n = O_p(1)$.

Fix $\epsilon > 0$, then there exists $M_\epsilon > 0$ in $C_F$ such that

$$F_{X_n}(M_\epsilon) = \mathbb{P}_n(X_n \leq M_\epsilon) \geq \mathbb{P}_n(|X_n| \leq M_\epsilon) \geq 1 - \epsilon$$

for all $n \geq 1$. Since $M_\epsilon \in C_F$, $F_{X_n}(M_\epsilon) \to F(M_\epsilon)$. We then see that for all $\epsilon > 0$, there exists $M_\epsilon > 0$ such that $F(+\infty) \geq F(M_\epsilon) \geq 1 - \epsilon$. As $\epsilon \to 0$, $F(+\infty) = 1$. Similarly, $F(-\infty) = 0$. ■

We now state the theorem.

**Theorem 2.6.6** (Lévy-Cramer continuity theorem). If $\phi_{X_n}(t) \to \phi(t)$ for all $t \in \mathbb{R}^d$, then all weakly convergent subsequences of $(X_n)$ have the same weak limit. Furthermore, if also $\phi$ is continuous at 0, then there exists $X$ such that $\phi = \phi_X$ and $X_n \xrightarrow{D} X$.

**Proof.** For the first claim, suppose $Y_n \xrightarrow{\text{w}} Y$ and $Z_n \xrightarrow{\text{w}} Z$ are two subsequences of $X_n$ such that $Y \neq Z$. But since $\phi_{Y_n}(t) \to \phi_Y(t)$ and $\phi_{Z_n}(t) \to \phi_Z(t)$, with the fact that $(\phi_{Y_n}(t))$ and $(\phi_{Z_n}(t))$ are subsequences of $(\phi_{X_n}(t))$ for every $t$, as $\phi_{X_n}(t) \to \phi(t)$, both subsequences need to converge to the same limit, i.e., $\phi_Y(t) = \phi(t) = \phi_Z(t)$ for all $t \in \mathbb{R}^d$. From the uniqueness theorem, $Y \overset{D}{=} Z$.

For the second claim, we just need to prove the following.

**Claim.** It's enough to show that if $\phi$ is continuous at 0, $(X_n) = O_p(1)$.

**Proof.** Since if $(X_n) = O_p(1)$, Prokhorov's theorem implies there exists a weakly convergent subsequence of $(X_n)$. With the first claim, we can find the weak limit $X$. ⊛

*The proof will be continued...*

# Lecture 9: Proof of Lévy-Cramer Continuity Theorem

13 Feb. 9:30

We now finish the proof of Lévy-Cramer continuity theorem.

**Proof of Theorem 2.6.6 (cont.)** Fix $\epsilon > 0$. Then there exists $\delta > 0$ such that for all $|t| < \delta$,

$$|\phi(t) - \phi(0)| = |\phi(t) - 1| < \frac{\epsilon}{4}$$

since for any $n \geq 1$, $\phi_{X_n}(0) = 1$, so is $\phi(0)$. Hence, we have

$$\frac{\epsilon}{2} = \frac{1}{\delta}\int_{-\delta}^{\delta} \frac{\epsilon}{4}\,\mathrm{d}t > \frac{1}{\delta}\int_{-\delta}^{\delta} |\phi(t) - 1|\,\mathrm{d}t.$$

We claim that we can find an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $\mathbb{P}_n(|X_n| \geq 2/\delta) < \epsilon$.[a] To bound $|X_n|$ with $\phi_{X_n}$, firstly, for all $x$, $|\sin x| \leq |x|$. This bound is good only when $x$ is close to 0. If it's not the case, then we can use $|\sin x/x| \leq 1/|x| \leq 1/2$ if $|x| \geq 2$. Hence, in general, for $x \neq 0$,

$$\frac{\sin x}{x} \leq \left|\frac{\sin x}{x}\right| \leq \frac{1}{2} \cdot \mathbb{1}_{|x|\geq 2} + 1 \cdot \mathbb{1}_{|x|<2} = 1 - \frac{1}{2}\mathbb{1}_{|x|\geq 2} \Rightarrow \mathbb{1}_{|x|\geq 2} \leq 2\left(1 - \frac{\sin x}{x}\right)$$

as $\mathbb{1}_{|x|<2} = 1 - \mathbb{1}_{|x|\geq 2}$. Plug in $\delta x$, for any $x \neq 0$, we have

$$\mathbb{1}_{|\delta x|\geq 2} \leq 2\left(1 - \frac{\sin(\delta x)}{\delta x}\right) = \frac{1}{\delta}\left(2\delta - 2\frac{\sin(\delta x)}{x}\right) = \frac{1}{\delta}\int_{-\delta}^{\delta} 1 - \cos(tx)\,\mathrm{d}t.$$

Indeed, the above is true for all $x \in \mathbb{R}$ by manually checking. Finally, by replacing $x$ by $X_n$ and

take the expectation on the both sides,

$$\mathbb{P}_n(|\delta X_n| \geq 2) \leq \frac{1}{\delta}\int_{-\delta}^{\delta} 1 - \mathbb{E}_n[\cos(tX_n)]\,\mathrm{d}t = \frac{1}{\delta}\int_{-\delta}^{\delta} \mathrm{Re}(1-\phi_{X_n}(t))\,\mathrm{d}t \leq \frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi_{X_n}(t)|\,\mathrm{d}t,$$

where we pass the expectation (i.e., limit) inside the integral from Fubini's theorem since $\cos(tX_n)$ is bounded. It remains to show that there is some $\delta > 0$ such that the right-hand side is less than $\epsilon$ for all $n \geq n_0$. As $\phi_{X_n}(t) \to \phi(t)$ for all $t$, we have $|1-\phi_{X_n}(t)| \to |1-\phi(t)|$ point-wise, hence by the bounded convergence theorem,

$$\frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi_{X_n}(t)|\,\mathrm{d}t \to \frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi(t)|\,\mathrm{d}t < \frac{\epsilon}{2}$$

from our assumption. Putting everything together, there is an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\mathbb{P}(|\delta X_n| \geq 2) = \mathbb{P}(|X_n| \geq 2/\delta) \leq \frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi_{X_n}(t)|\,\mathrm{d}t < \frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi(t)|\,\mathrm{d}t + \frac{\epsilon}{2} < \epsilon,$$

where the second-last inequality follows from the point-wise convergence of $\frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi_{X_n}(t)|\,\mathrm{d}t$ to $\frac{1}{\delta}\int_{-\delta}^{\delta} |1-\phi(t)|\,\mathrm{d}t$ being $\epsilon/2$-close for $n$ large enough, i.e., when $n \geq n_0$ for some $n_0$. ■

[a] If this is the case, then we can handle the $n < n_0$ case easily as usual by taking the maximum over all $n < n_0$.

### 2.6.4 Inversion Theorem

On the other hand, another way to prove Lévy-Cramer continuity theorem is to directly calculate the pdf of $X$, given $\phi_X$. It's follows the same vein of the proof of uniqueness theorem.

**Intuition.** In the proof of uniqueness theorem, we only obtain a pdf for $X+\sigma Z$. Imposing constraints on $\phi_X$ and calculate $\mathbb{E}[g(X)]$ in terms of $\phi_X$ will tell us which condition should we add.

**Theorem 2.6.7** (Feller's inversion formula)**.** Let $X$ be a $d$-dimensional random vector with the characteristic function $\phi_X$.

(a) If $g$ has a bounded support and $\mathbb{P}(X \in C_g) = 1$, then

$$\mathbb{E}[g(X)] = \lim_{\sigma \searrow 0} \frac{1}{2\pi}\iint g(x)e^{-iux-\sigma^2u^2/2}\,\mathrm{d}u\,\mathrm{d}x.$$

(b) For any $a, b \in C_{F_X}$,

$$F_X(b) - F_X(a) = \lim_{\sigma \searrow 0} \frac{1}{2\pi}\int_a^b \int e^{-iux-\sigma^2u^2/2}\phi_X(u)\,\mathrm{d}u\,\mathrm{d}x.$$

(c) If further, $\phi_X$ is absolute integrable, then $X$ has a pdf

$$f_X(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-iux}\phi_X(u)\,\mathrm{d}u.$$

**Proof.** The proof is based on uniqueness theorem.

(a) In the uniqueness theorem, $\sigma \searrow 0$ such that $X + \sigma Z \xrightarrow{D} X$, which implies $g(X+\sigma Z) \xrightarrow{D} g(X)$ when $\mathbb{P}(X \in C_g) = 1$. Since now $g$ is also bounded, by the bounded convergence theorem,

$$\mathbb{E}[g(X)] = \lim_{\sigma \searrow 0} \mathbb{E}[g(X+\sigma Z)].$$

We now calculate $\mathbb{E}[g(X+\sigma Z)]$. Since $g\colon \mathbb{R} \to \mathbb{R}$ has bounded support, the same calculation

from the proof of uniqueness theorem gives

$$\mathbb{E}[g(X+\sigma Z)] = \lim_{\sigma \searrow 0} \frac{1}{2\pi} \int g(x) \int e^{-ixu-\sigma^2 u^2/2} \phi_X(u)\,\mathrm{d}u\,\mathrm{d}x.$$

It remains to change the order of integration, which is justified by Tonelli's theorem as $\mathbb{E}[|g(X+\sigma Z)|] < \infty$ for all $\sigma > 0$, hence we obtain the result for the first part.

(b) Given $a, b \in C_{F_X}$, consider $g(x) = \mathbb{1}_{(a,b)}(x)$, which implies $\mathbb{P}(X \in C_g) = 1$ (and trivially $g$ has a bounded support), hence the result above applies.

(c) Finally, if $\phi_X$ is absolute integrable, our goal now is to pass the limit $\sigma \searrow 0$ inside the integral for $F_X(b) - F_X(a)$ given $a, b \in C_{F_X}$, i.e., to get

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \int_a^b \int \lim_{\sigma \searrow 0} e^{-iux-\sigma^2 u^2/2} \phi_X(u)\,\mathrm{d}u\,\mathrm{d}x = \frac{1}{2\pi} \int_a^b \int e^{-iux} \phi_X(u)\,\mathrm{d}u\,\mathrm{d}x.$$

Since cdfs are characterized by values in $C_{F_X}$, i.e., if the above holds for $a, b \in C_{F_X}$, the same holds for $a, b \in \mathbb{R}$, and we're done. To do so, dominated convergence theorem states that

$$\int_a^b \int \sup_{\sigma > 0} \left| e^{-ixu-\sigma^2 u^2/2} \phi_X(u) \right| \,\mathrm{d}u\,\mathrm{d}x < \infty$$

is the right condition. We see that the left-hand side is less than

$$\int_a^b \int_{\mathbb{R}} |\phi_X(u)| \sup_{\sigma > 0} |e^{-\sigma^2 u^2/2}|\,\mathrm{d}u\,\mathrm{d}x \le \int_a^b \int_{\mathbb{R}} |\phi_X(u)|\,\mathrm{d}u\,\mathrm{d}x$$

which is finite since $\int |\phi_X(u)|\,\mathrm{d}u < \infty$.

■

**Corollary 2.6.1.** Given $(X_n)$ and $X$ such that $\phi_X$ and $\phi_{X_n}$ for every $n$ are integrable. If $\phi_{X_n} \overset{L^1}{\to} \phi_X$, i.e., $\int_{\mathbb{R}} |\phi_{X_n}(t) - \phi_X(t)|\,\mathrm{d}t \to 0$, then $X_n \overset{\mathrm{TV}}{\to} X$.

**Proof.** It suffices to prove that $|f_{X_n}(x) - f_X(x)| \to 0$, where these pdfs exist due to Feller's inversion formula (c). We see that

$$|f_{X_n}(x) - f(x)| \le \frac{1}{2\pi} \int_{\mathbb{R}} |e^{-iux}| \cdot |\phi_{X_n}(u) - \phi_X(u)|\,\mathrm{d}u, \le \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_{X_n}(u) - \phi_X(u)|\,\mathrm{d}u$$

with the assumption the right-hand side goes to 0. ■

### 2.6.5 Properties of Characteristic Function

Finally, we see the following characterizations of $\phi_X$. The first one is that it's uniformly continuous.

**Proposition 2.6.1.** For any random vector $X$, $\phi_X$ is uniformly continuous, i.e.,

$$\lim_{h \to 0} \sup_{t \in \mathbb{R}^d} |\phi_X(t+h) - \phi_X(t)| = 0.$$

**Proof.** We see that for any $h$,

$$|\phi_X(t+h) - \phi_X(t)| = \left| \mathbb{E}[e^{i(t+h)X}] - \mathbb{E}[e^{itX}] \right| \le \mathbb{E}\left[ \left| e^{itX} \right| \left| e^{ihX} - 1 \right| \right] \le \mathbb{E}\left[ \left| e^{ihX} - 1 \right| \right],$$

which goes to 0 as $h \to 0$ since $|e^{ihX} - 1| \le 2$ with bounded convergence theorem. ■

The next theorem gives us a way to calculate the derivatives of $\phi_X$ and its connection to moments.

**Theorem 2.6.8.** If $X \in L^p$ for any $p \in \mathbb{N}$, then the $p^{\text{th}}$ derivative of $\phi_X(t)$ is given by

$$\phi_X^{(p)}(t) = \mathbb{E}[(iX)^p e^{itX}]$$

for every $t$. In particular, $\phi_X^{(p)}(0) = i^p \mathbb{E}[X^p]$ and $\sup_t |\phi_X^{(p)}(t)| \leq \mathbb{E}[|X|^p] < \infty$.

**Proof.** Consider $p = 1$ since for $p > 1$, it can be shown by induction. It's enough to prove

$$\lim_{h \to 0} \left| \frac{\phi_X(t+h) - \phi_X(t)}{h} - \mathbb{E}\left[iXe^{itX}\right] \right| = 0$$

Writing the $\phi_X$ explicitly, by Jensen's inequality, for any $h \neq 0$, the left-hand side is

$$\begin{aligned} \left| \frac{\mathbb{E}\left[e^{i(t+h)X}\right] - \mathbb{E}\left[e^{itX}\right] - \mathbb{E}\left[ihXe^{itX}\right]}{h} \right| &\leq \frac{\mathbb{E}\left[\left|e^{i(t+h)X} - e^{itX} - ihXe^{itX}\right|\right]}{|h|} \\ &= \frac{\mathbb{E}\left[\left|e^{itX}\right| \left|e^{ihX} - 1 - ihX\right|\right]}{|h|} \leq \frac{\mathbb{E}\left[\left|e^{ihX} - 1 - ihX\right|\right]}{|h|} \end{aligned}$$

Let $G(h) = e^{ihX}$, then $G'(h) = iXe^{ihX}$, and the right-hand side is equal to

$$\frac{\mathbb{E}\left[|G(h) - G(0) - G'(0)h|\right]}{|h|}.$$

Since $G$ is differentiable, $G(h) - G(0) = \int_0^h G'(y)\,\mathrm{d}y$, hence

$$G(h) - G(0) - G'(0)h = \int_0^h G'(y) - G'(0)\,\mathrm{d}y = h\int_0^1 G'(uh) - G'(0)\,\mathrm{d}u = h\int_0^1 iXe^{iuhX} - iX\,\mathrm{d}u$$

where we let $y = uh$. Plugging in, we have

$$\begin{aligned} \mathbb{E}\left[\frac{\left|e^{ihX} - 1 - ihX\right|}{|h|}\right] &\leq \mathbb{E}\left[\int_0^1 |G'(uh) - G'(0)|\,\mathrm{d}u\right] \\ &= \mathbb{E}\left[\int_0^1 |iXe^{iuhX} - iX|\,\mathrm{d}u\right] \leq \mathbb{E}\left[|X|\int_0^1 |e^{iuhX} - 1|\,\mathrm{d}u\right]. \end{aligned}$$

Finally, taking the limit as $h \to 0$, with the fact that $\mathbb{E}[|X|] < \infty$ and $\int_0^1 |e^{ihuX} - 1|\,\mathrm{d}u \leq 2$, we see that $|X|\int_0^1 |e^{ihuX} - 1|\,\mathrm{d}u \leq 2|X|$, and the latter is integrable since $\mathbb{E}[X] < \infty$, hence dominated convergence theorem applies, i.e., we can pass the limit into the expectation,

$$\lim_{h \to 0} \mathbb{E}\left[|X|\int_0^1 |e^{ihuX} - 1|\,\mathrm{d}u\right] = \mathbb{E}\left[|X| \lim_{h \to 0}\int_0^1 |e^{ihuX} - 1|\,\mathrm{d}u\right] = 0$$

since $\lim_{h \to 0}\int_0^1 |e^{iuhX} - 1|\,\mathrm{d}u = 0$, again from the bounded convergence theorem. ■

**Corollary 2.6.2.** If $X \in L^p$ for some $p \in \mathbb{N}$, then $\phi_X^{(p)}$ is uniformly continuous.[a]

[a] This is a generalization of Proposition 2.6.1.

**Proof.** To show that $\phi_X^{(p)}$ is uniformly continuous, we show that $\sup_{t \in \mathbb{R}} |\phi^{(p)}(t+h) - \phi_X^{(p)}(t)| \to 0$ as $h \to 0$. But this is clear since for any $h \in \mathbb{R}$, with Theorem 2.6.8,

$$\sup_{t \in \mathbb{R}} |\phi_X^{(p)}(t+h) - \phi_X^{(p)}(t)| \leq \mathbb{E}\left[|X|^p |e^{ihX} - 1|\right],$$

which goes to 0 as $h \to 0$ from the dominated convergence theorem. ■

# Chapter 3

# Fundamental Theorems of Probability

## Lecture 10: WLLN and CLT, and Applications to Inferences

With the tools we developed, we can now prove the fundamental theorems of probability and see some applications to inferences. 15 Feb. 9:30

## 3.1 Law of Large Number and Central Limit Theorem

In this section, we will study the weak law of large number and the central limit theorem.

### 3.1.1 Weak Law of Large Number

The first result, the weak law of large number, states that the sample mean converges to the mean.

**Theorem 3.1.1** (Khintchin's weak law of large number)**.** Let $X$ and $(X_n)$ be i.i.d. random vectors with $X \in L^1$, i.e., $\mathbb{E}[|X|] < \infty$. Then $\overline{X}_n \xrightarrow{p} \mathbb{E}[X]$.

**Proof.** Since $c \coloneqq \mathbb{E}[X]$ is a constant, it suffices to show that $\phi_{\overline{X}_n}(t) \to \phi_c(t) = e^{itc}$ for all $t$ from Corollary 2.2.1. Firstly, let $\overline{X}_n = S_n/n$, we have

$$\phi_{\overline{X}_n}(t) = \mathbb{E}[e^{itS_n/n}] = \phi_{S_n}(t/n) = \prod_{i=1}^{n} \phi_{X_i}(t/n) =: (\phi(t/n))^n$$

where we let $\phi_{X_i} =: \phi$ since $(X_n)$ are i.i.d. From the fundamental theorem of calculus, with the fact that the first moment of $X$ exists, $\phi$ is differentiable such that

$$(\phi(t/n))^n = \left(1 + \frac{t}{n}\int_0^1 \phi'(ut/n)\,\mathrm{d}u\right)^n.$$

Since $(1 + a_n)^n \to e^c$ if $na_n \to c$, it remains to show $\int_0^1 \phi'(ut/n)\,\mathrm{d}u \to ic$. First, $\phi'(t)$ is continuous at 0 from Corollary 2.6.2,[a] as $n \to \infty$

$$\phi'(ut/n) \to \phi'(0) = i\mathbb{E}[X] = ic.$$

With the fact that $\sup_t |\phi'(t)| \le \mathbb{E}[|X|]$, the bounded convergence theorem implies

$$\int_0^1 \phi'(ut/n)\,\mathrm{d}u \to \int_0^1 ic\,\mathrm{d}u = ic$$

since we can now pass the limit inside the integral. ■

[a] We see that assuming $\phi$ is differentiable at 0 such that $\phi'(0) = ic$ is enough.

Although we will not show, but the stronger version holds, i.e., it converges almost surely.

**Theorem 3.1.2** (Strong law of large number)**.** Let $X$ and $(X_n)$ be i.i.d. random vectors with $X \in L^1$. Then $\overline{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$.

### 3.1.2 Central Limit Theorem

In terms of the distributional result, we need higher-order moments. In particular, if the second moment exists, then we can generalize we have done as in the proof of Theorem 2.6.8.

**As previously seen.** If $g$ is continuously differentiable at 0, then for $x$ around 0,

$$g(x) = g(0) + g'(0)x + x\int_0^1 g'(ux) - g'(0)\,\mathrm{d}u.$$

**Note.** If in addition, $g'$ is also continuously differentiable at 0, then for $x$ around 0,

$$\begin{aligned} g(x) &= g(0) + g'(0)x + x\int_0^1\int_0^{ux} g''(y)\,\mathrm{d}y\,\mathrm{d}u \\ &= g(0) + g'(0)x + x^2\int_0^1\int_0^1 g''(xuv)u\,\mathrm{d}v\,\mathrm{d}u. \qquad\qquad y = uxv,\ \mathrm{d}y = ux\mathrm{d}v \end{aligned}$$

We now state the theorem.

**Theorem 3.1.3** (Lindeberg-Lévy central limit theorem)**.** Let $(X_n)$ be i.i.d. random variables (i.e., $d = 1$) with $\mathbb{E}[X_i] =: \mu$, $\mathrm{Var}[X_i] =: \sigma^2 < \infty$ for all $1 \le i \le n$. Then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

**Proof.** Without loss of generality, let $\mu = 0$, $\sigma = 1$. Since $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, it's enough to show that $\phi_{S_n/\sqrt{n}}(t) \to e^{-t^2/2}$ for any $t \in \mathbb{R}$ from Lévy-Cramer continuity theorem and Equation 2.1. Firstly,

$$\phi_{S_n/\sqrt{n}}(t) = \mathbb{E}[e^{itS_n/\sqrt{n}}] = \phi_{S_n}(t/\sqrt{n}) = \big(\phi(t/\sqrt{n})\big)^n$$

where we let $\phi_{X_n} =: \phi$ since $(X_n)$ are i.i.d. By applying the above note, we further have

$$\begin{aligned} \big(\phi(t/\sqrt{n})\big)^n &= \left(\phi(0) + \phi'(0)\frac{t}{\sqrt{n}} + \frac{t^2}{n}\int_0^1\int_0^1 u\phi''(uvt/\sqrt{n})\,\mathrm{d}u\,\mathrm{d}v\right)^n \\ &= \left(1 + \frac{t^2}{n}\int_0^1\int_0^1 u\phi''(uvt/\sqrt{n})\,\mathrm{d}u\,\mathrm{d}v\right)^n \end{aligned}$$

since $\phi(0) = 1$ and $\phi'(0) = i\mu = 0$. It remains to show that the double integral converges to $-1/2$ since it'll imply $(\phi(t/\sqrt{n}))^n \to e^{-t^2/2}$. We see that as $n \to \infty$, the integrand

$$u\phi''(uvt/\sqrt{n}) \to u\phi''(0) = u(i^2\mathbb{E}[X^2]) = -u(\mathrm{Var}[X] + (\mathbb{E}[X])^2) = -u(1+0) = -u.$$

Hence, from the bounded convergence theorem,

$$\int_0^1\int_0^1 u\phi''(ut/\sqrt{n})\,\mathrm{d}u\,\mathrm{d}v \to \int_0^1\int_0^1 -u\,\mathrm{d}u\,\mathrm{d}v = -\frac{1}{2},$$

which shows the result. ∎

**Remark.** From the central limit theorem, we can indeed deduce the weak law of large number. But since the former requires more conditions, hence weak law of large number still has its own merit.

## 3.2 Inference for Population Mean and Variance

We now apply what we have proved to one of the most basic problems, inference for mean and variance.

### 3.2.1 Population Mean

Firstly, let's consider the applications for mean estimation. Let $X, X_1, \dots, X_n$ be i.i.d. samples such that $\mathbb{E}[X] = \mu < \infty$, $\mathrm{Var}[X] = \sigma^2$. If, also, $X_i$'s are Gaussian, $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, i.e.,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

In this case, a natural estimator of $\mu$ is $\overline{X}_n$, and we have the distribution of $\overline{X}_n - \mu$, i.e., we know how our estimator perform in terms of distribution, which can in turn provides a confidence interval.

> **Intuition.** We want to make the distribution, specifically, its variance (denominator at the left-hand side) independent of parameters to get a corresponding confidence interval.

Right now, our confidence interval depends on $\sigma$. To solve this, consider replacing $\sigma$ by the sample standard deviation $s_n$, then

$$T_n \coloneqq \frac{\overline{X}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1} \xrightarrow{\text{TV}} \mathcal{N}(0,1)$$

as $n \to \infty$, where $T_n$ follows $t$-distribution with $n-1$ degrees of freedom.

> **Notation.** We let $s_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ and $\hat{\sigma}_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

We see that when $X$ is Gaussian, an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\mu$ is

$$\overline{X} \pm Z_{\alpha/2}\frac{s_n}{\sqrt{n}}.$$

> **Notation.** For any $\alpha > 0$, we define $Z_\alpha$ by $\alpha = \mathbb{P}(Z > Z_\alpha)$, **not** $\alpha = \mathbb{P}(Z < Z_\alpha)$.

The first question we will address is "what if $X_i$'s are not Gaussian, and can we replace $s_n$ by $\hat{\sigma}_n$."

**Proposition 3.2.1.** If $X \in L^2$, then both $\hat{\sigma}_n^2$ and $s_n^2$ are consistent estimators of $\sigma^2$. Furthermore, $T_n \xrightarrow{D} \mathcal{N}(0,1)$, and the same holds if $s_n$ is replaced by $\hat{\sigma}_n$ in the definition of $T_n$.

**Proof.** Indeed, by letting $Y_i \coloneqq X_i - \mu$ for all $i$ (and also $Y = X - \mu$), as $n \to \infty$,

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(Y_i - \overline{Y}_n\right)^2 = \frac{1}{n}\sum_{i=1}^n Y_i^2 - (\overline{Y}_n)^2 \xrightarrow{p} \sigma^2 + 0$$

since $\frac{1}{n}\sum_{i=1}^n Y_i^2 \xrightarrow{p} \mathbb{E}[Y^2] = \mathrm{Var}[X] = \sigma^2 < \infty$ as $X \in L^2$, and $(\overline{Y}_n)^2 \xrightarrow{p} (\mathbb{E}[Y])^2 = 0$, both from weak law of large number. This implies that $s_n^2$ is also a consistent estimator of $\sigma^2$ since

$$s_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2 \xrightarrow{p} 1 \cdot \sigma^2 = \sigma^2,$$

again from Slutsky's theorem. The distributional result follows directly from central limit theorem for $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1)$ and Slutsky's theorem. ■

Proposition 3.2.1 says that for mean estimation, even if the data is not Gaussian, we're fine.

**Corollary 3.2.1.** If $X \in L^2$, then $\overline{X}_n \pm Z_{\alpha/2}s_n/\sqrt{n}$ and $\overline{X}_n \pm Z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ are both asymptotically valid $100(1-\alpha)\%$ confidence intervals for $\mu$.

### 3.2.2 Population Variance

Next, let's consider variance estimation and further assume that $\sigma^2 < \infty$. Again, let $X, X_1, \ldots, X_n$ be i.i.d. samples. If they are Gaussian,

$$(n-1)\frac{s_n^2}{\sigma^2} \overset{D}{=} \sum_{i=1}^{n-1} Z_i^2 \sim \chi^2_{n-1}$$

where $(Z_{n-1}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Firstly, since $\mathbb{E}[Z_i^2] = \text{Var}[Z_i] + (\mathbb{E}[Z_i])^2 = 1$, and $\text{Var}[Z_i^2] = \mathbb{E}[Z_i^4] - (\mathbb{E}[Z_i^2])^2 = 3 - 1 = 2$, standardizing, from the normal approximation to the chi-square distribution,

$$\frac{(n-1)\frac{s_n^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} \overset{D}{=} \frac{\sum_{i=1}^{n-1} Z_i^2 - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{D} \mathcal{N}(0,1),$$

i.e., as $n \to \infty$,

$$\sqrt{n-1}\left(\frac{s_n^2}{\sigma^2} - 1\right) \xrightarrow{D} \mathcal{N}(0,2) \Leftrightarrow \sqrt{n}\left(\frac{s_n^2}{\sigma^2} - 1\right) \xrightarrow{D} \mathcal{N}(0,2) \Leftrightarrow \sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0, 2\sigma^4),$$

and an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\sigma^2$ is

$$\frac{s_n^2}{1 \pm Z_{\alpha/2}\sqrt{2/n}}.$$

Let's again ask what will happen when $X_i$'s are not Gaussian anymore.

**Proposition 3.2.2.** If $X \in L^2$, then the following hold when $\hat{\sigma}_n^2$ is replaced by $s_n^2$. Firstly,

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i^2 - \sigma^2) + o_p(1).$$

Moreover, if $X \in L^4$ and $\mathbb{E}[((X-\mu)/\sigma)^4] > 1$, then $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0, \mathbb{E}[(X-\mu)^4] - \sigma^4)$.

**Proof.** We see that from the same calculation as above, with $Y_i := X_i - \mu$ (and also $Y = X - \mu$),

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}_n^2 \Rightarrow \hat{\sigma}_n^2 - \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - \sigma^2) - \overline{Y}_n^2$$
$$\Rightarrow \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i^2 - \sigma^2) - \frac{(\sqrt{n}\overline{Y}_n)^2}{\sqrt{n}}.$$

As $n \to \infty$, since $(\sqrt{n}\overline{Y}_n)^2$ converges in distribution from the central limit theorem for $\sqrt{n}\overline{Y}_n$ (as $X \in L^2$) and continuous mapping theorem, $(\sqrt{n}\overline{Y}_n)^2 = O_p(1)$ from Proposition 2.4.2, hence

$$\frac{(\sqrt{n}\overline{Y}_n)^2}{\sqrt{n}} = o(1)O_p(1) = o_p(1),$$

proving the first claim. Now, if further $\text{Var}[Y_i^2] < \infty$ from $X \in L^4$, central limit theorem gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i^2 - \mathbb{E}[Y_i^2]) \xrightarrow{D} \mathcal{N}(0, \text{Var}[Y_i^2]),$$

implying $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0, \text{Var}[Y^2])$ from the first claim and Slutsky's theorem, where

$$\text{Var}[Y^2] = \mathbb{E}[(X-\mu)^4] - \left(\mathbb{E}[(X-\mu)^2]\right)^2 = \sigma^4\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - \sigma^4 = \sigma^4\left(\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 1\right),$$

which proves the second claim. Finally, we note that

$$\sqrt{n}(\hat{\sigma}_n^2 - s_n^2) = \frac{\sqrt{n}}{n-1}\hat{\sigma}_n^2 \xrightarrow{p} 0 \cdot \sigma^2 = 0,$$

hence the same results above hold for replacing $\hat{\sigma}_n^2$ by $s_n^2$ from Slutsky's theorem. ■

The quantity (and a related one) in our assumption deserves a special name.

**Definition 3.2.1** (Kurtosis). The *Kurtosis* of a random variable $X$ is defined as $\mathbb{E}[((X-\mu)/\sigma)^4]$.

**Definition 3.2.2** (Skewness). The *skewness* of a random variable $X$ is defined as $\mathbb{E}[((X-\mu)/\sigma)^3]$.

**Example** (Kurtosis for Gaussian). The Kurtosis of the standard Gaussian is 3.

Let $Z = (X-\mu)/\sigma$, we note that Proposition 3.2.2 requires $\mathbb{E}[Z^4] > 1$. However, from Jensen's inequality, $\mathbb{E}[Z^4] \geq \left(\mathbb{E}[Z^2]\right)^2 \geq 1$, hence indeed, the assumption might not be true in general.

**Example.** If $\mathbb{E}[Z^4] = 1$,

$$\mathrm{Var}[Y^2] = 0 \Leftrightarrow \mathbb{P}(Y^2 = \mathbb{E}[Y^2]) = 1 \Leftrightarrow \mathbb{P}(Y = \pm\sigma) = 1 \Leftrightarrow \mathbb{P}(X = \mu \pm \sigma) = 1,$$

i.e., the violation might happen for $X$ being concentrated on two points.

The takeaway is when $X$ is not a normal (or when the Kurtosis of $X$ is different from 3), then the distribution of $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2)$ is different. Specifically, if the Kurtosis exists and is not equal to 1, then an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\sigma^2$ is

$$\frac{\hat{\sigma}_n^2}{1 \pm Z_{\alpha/2}\sqrt{(\mathbb{E}[((X-\mu)/\sigma)^4] - 1)/n}}.$$

However, if we don't know the Kurtosis of $X$, we can't say anything about the confidence interval.

**Intuition.** By Slutsky's theorem, if we have a consistent estimator of the Kurtosis, we can then use it instead and get a desired asymptotic confidence interval.

# Lecture 11: Sample Standardized Central Moments

20 Feb. 9:30

Following the intuition, let's find such consistent estimators. Let $Y \coloneqq X - \mu = X - \mathbb{E}[X]$ (and also $Y_i = X_i - \mu$ as usual), $\mu_k \coloneqq \mathbb{E}[Y^k] = \mathbb{E}[(X-\mu)^k]$ for all $k \geq 2$, and finally $\widetilde{\mu}_k = \mu_k/\sigma^k = \mathbb{E}\left[(X-\mu)^k/\sigma^k\right]$.

**As previously seen.** In this notation, Proposition 3.2.2 gives $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \to \mathcal{N}(0, (\widetilde{\mu}_4 - 1)\sigma^4)$, i.e.,

$$\frac{\sqrt{n}}{\sqrt{\widetilde{\mu}_4 - 1}}\left(\frac{\hat{\sigma}_n^2}{\sigma^2} - 1\right) \to \mathcal{N}(0,1).$$

The task is then the following.

**Problem.** How to estimate $\widetilde{\mu}_4$, or more generally, how to estimate $\widetilde{\mu}_k$ consistently?

**Answer.** Consider the $k^{th}$ *sample central moment*

$$M_k \coloneqq \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k.$$

Let's also define the $k^{th}$ *sample standardized central moment* as $\widetilde{M}_k \coloneqq M_k/\hat{\sigma}_n^k$. ⊛

The above essentially is motivated from the following observation.

**Intuition.** If we know $\mu$, then $\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^k \xrightarrow{p} \mu_k$ by the weak law of large number. However, since we don't know $\mu$, we need to use $\overline{X}_n$.

We now show that this still yields a consistent estimator.

**Proposition 3.2.3.** If $X \in L^k$ for $k>2$, then $M_k \xrightarrow{p} \mathbb{E}[Y^k] = \mu_k$. Same for $\widetilde{M}_k$ and $\widetilde{\mu}_k$.

**Proof.** Let's denote $\overline{X}_n =: \overline{X}$ and $\overline{Y}_n =: \overline{Y}$. Then

$$M_k = \frac{1}{n}\sum_{i=1}^n (X_i-\overline{X})^k = \frac{1}{n}\sum_{i=1}^n (Y_i-\overline{Y})^k = \frac{1}{n}\sum_{i=1}^n\sum_{\ell=0}^k \binom{k}{\ell} Y_i^\ell (-\overline{Y})^{k-\ell} = \sum_{\ell=0}^k \binom{k}{\ell}(-\overline{Y})^{k-\ell}\frac{1}{n}\sum_{i=1}^n Y_i^\ell.$$

Let $\frac{1}{n}\sum_{i=1}^n Y_i^\ell =: \overline{Y^\ell}$, then we further get

$$M_k = \sum_{\ell=0}^k \binom{k}{\ell}(-\overline{Y})^{k-\ell}\overline{Y^\ell} = \overline{Y^k} + \sum_{\ell=0}^{k-1}\binom{k}{\ell}(-\overline{Y})^{k-\ell}\overline{Y^\ell}. \tag{3.1}$$

By the weak law of large number, $\overline{Y^k} \xrightarrow{p} \mathbb{E}[Y^k] = \mu_k$ and $(-\overline{Y})^{k-\ell} \xrightarrow{p} 0$ for $\ell < k$ from $-\overline{Y} \xrightarrow{p} 0$ (by weak law of large number) and continuous mapping theorem, hence $M_k \xrightarrow{p} \mu_k$ by Slutsky's theorem. The consistency of $\hat{\sigma}_n$ implies $\widetilde{M}_k \xrightarrow{p} \widetilde{\mu}_k$ clearly. ■

This gives the following useful confidence interval for $\sigma^2$.

**Corollary 3.2.2.** If the Kurtosis of $X$ exists and is not equal to 1, then an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\sigma^2$ is

$$\frac{\hat{\sigma}_n^2}{1 \pm Z_{\alpha/2}\sqrt{(\widetilde{M}_4-1)/n}}.$$

**Proof.** This directly follows from Proposition 3.2.2 and Proposition 3.2.3. ■

## 3.3 Testing for Normality

As we will soon see, it's natural to extend what we just discussed to the problem of hypothesis testing, and specifically, testing for normality.

### 3.3.1 Asymptotic Distribution of Sample Central Moments

It turns out that asking for the asymptotic distribution of $M_k$, i.e., $\sqrt{n}(M_k-\mu_k)$ is quite valuable, although the motivation is not so clear right now. Anyway, we have the following.

**Theorem 3.3.1.** If $X \in L^k$ for some $k>2$, then

$$\sqrt{n}(M_k-\mu_k) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i^k - \mu_k - k\mu_{k-1}Y_i) + o_p(1).$$

Moreover, if $X \in L^{2k}$ and $v_k > 0$ where

$$v_k := \mathrm{Var}[Y^k - k\mu_{k-1}Y] = \mu_{2k} - \mu_k^2 + k^2\mu_{k-1}^2\sigma^2 - 2k\mu_{k-1}\mu_{k+1},$$

then $\sqrt{n}(M_k-\mu_k) \xrightarrow{D} \mathcal{N}(0, v_k)$.

**Proof.** Firstly, if $X \in L^k$, from Equation 3.1,

$$\sqrt{n}(M_k - \mu_k) = \sqrt{n}(\overline{Y^k} - \mu_k) + \sum_{\ell=0}^{k-1} \binom{k}{\ell} (-\overline{Y})^{k-\ell} \overline{Y^\ell} \sqrt{n} = \sqrt{n}(\overline{Y^k} - \mu_k) + \sum_{\ell=0}^{k-1} \binom{k}{\ell} \frac{(-\overline{Y}\sqrt{n})^{k-\ell}}{\sqrt{n}^{k-\ell-1}} \overline{Y^\ell}.$$

We see that from Proposition 2.4.2, for $\ell < k-1$,

- $(-\overline{Y}\sqrt{n})^{k-\ell} = O_p(1)$ from central limit theorem and continuous mapping theorem;
- $\overline{Y^\ell} = O_p(1)$ since $\overline{Y^\ell} \xrightarrow{p} \mathbb{E}[Y^\ell]$ from weak law of large number;
- $1/\sqrt{n}^{k-\ell-1} = o(1)$.

Combining, every term in the summation is $O_p(1)O_p(1)o(1) = o_p(1)$ except for $\ell = k-1$, hence

$$\begin{aligned}
\sqrt{n}(M_k - \mu_k) &= \sqrt{n}(\overline{Y^k} - \mu_k) - \binom{k}{k-1} \overline{Y^{k-1}} \sqrt{n\overline{Y}} + \sum_{\ell=0}^{k-2} \binom{k}{\ell} o_p(1) \\
&= \sqrt{n}(\overline{Y^k} - \mu_k) - k\overline{Y^{k-1}} \sqrt{n\overline{Y}} + o_p(1)
\end{aligned}$$

while $\sqrt{n\overline{Y}} = O_p(1)$, $\overline{Y^{k-1}}$ is not $o_p(1)$. By replacing $\overline{Y^{k-1}}$ by $\overline{Y^{k-1}} - \mu_{k-1} + \mu_{k-1}$,

$$\begin{aligned}
&= \sqrt{n}(\overline{Y^k} - \mu_k) - k\left(\overline{Y^{k-1}} - \mu_{k-1}\right)\sqrt{n\overline{Y}} - k\mu_{k-1}\sqrt{n\overline{Y}} + o_p(1) \\
&= \sqrt{n}(\overline{Y^k} - \mu_k) - k\mu_{k-1}\sqrt{n\overline{Y}} + o_p(1)
\end{aligned}$$

since $\overline{Y^{k-1}} - \mu_{k-1} \xrightarrow{p} 0$ from the weak law of large number, finally,

$$\begin{aligned}
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}(Y_i^k - \mu_k) - k\mu_{k-1}\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} Y_i + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i^k - \mu_k - k\mu_{k-1}Y_i\right) + o_p(1),
\end{aligned}$$

proving the first claim. Moreover, since $Y_i^k - \mu_k - k\mu_{k-1}Y_i$'s are i.i.d., it converges in distribution to $\mathcal{N}(0, v_k) = \mathcal{N}(0, \operatorname{Var}\left[Y^k - \mu_k - k\mu_k Y\right])$ by central limit theorem and Slutsky's theorem, where

$$\begin{aligned}
v_k \coloneqq \operatorname{Var}\left[Y^k - \mu_k - k\mu_{k-1}Y\right] &= \operatorname{Var}\left[Y^k - k\mu_{k-1}Y\right] \\
&= \operatorname{Var}[Y^k] + k^2\mu_{k-1}^2 \operatorname{Var}[Y] - 2k\mu_{k-1}\operatorname{Cov}[Y, Y^k] \\
&= \mu_{2k} - \mu_k^2 + k^2\mu_{k-1}^2\sigma^2 - 2k\mu_{k-1}\mu_{k+1}
\end{aligned}$$

since $\operatorname{Cov}[Y, Y^k] = \mathbb{E}[Y \cdot Y^k] - \mathbb{E}[Y]\mathbb{E}[Y^k] = \mathbb{E}[Y^{k+1}] = \mu_{k+1}$ and $\mu_{2k} < \infty$ from $X \in L^{2k}$. ■

**Note.** Theorem 3.3.1 doesn't give an asymptotic distribution of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$ since it requires the joint distribution of $\hat{\sigma}_n^k$ and $M_k$.

However, it turns out that when $k$ is odd and the distribution is symmetric, Theorem 3.3.1 does give an asymptotic distribution for $\widetilde{M}_k$.

### 3.3.2 Testing Normality with Odd Moments

To motivate why we want to have an asymptotic distribution for $\widetilde{M}_k$, consider the problem of testing normality, i.e., let $H_0 \colon X \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu, \sigma^2$.

**Intuition.** The idea is that to reject $H_0$ if $|\widetilde{M}_k| = |M_k/\hat{\sigma}_n^k|$ deviates significantly.

In this regard, Theorem 3.3.1 is not enough since it's only for $M_k$, but we really need $\widetilde{M}_k$.

**Problem.** What is the asymptotic distribution of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$?

First observe that if $X_i$'s are Gaussian, as Gaussian is symmetric, $\mu_k = 0$ (and hence $\widetilde{\mu}_k = 0$) for all odd $k$. It turns out that this property allows us to bypass the joint if we focus on odd $k$. Formally, suppose $k$ is odd, and $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mu_k = 0$, hence Theorem 3.3.1 gives

$$\sqrt{n}(M_k - \mu_k) = \sqrt{n}M_k \xrightarrow{D} \mathcal{N}(0, \operatorname{Var}\left[Y^k - k\mu_{k-1}Y\right]) \Rightarrow \sqrt{n}\frac{M_k}{\sigma^k} \xrightarrow{D} \mathcal{N}(0, \sigma^{-2k}\operatorname{Var}\left[Y^k - k\mu_{k-1}Y\right]).$$

Then, by Slutsky's theorem, $\sqrt{n}M_k/\hat{\sigma}_n^k$ also converges to this normal. Since all we use is the fact that $\mu_k = 0$ for odd $k$ and Theorem 3.3.1, let's write this general result as a corollary.

**Corollary 3.3.1.** If $X \in L^{2k}$ for some odd $k > 2$ such that $\mu_k = 0$ and $\widetilde{v}_k := v_k/\sigma^{2k} > 0$, then $\sqrt{n}M_k/\hat{\sigma}_n^k = \sqrt{n}\widetilde{M}_k \xrightarrow{D} \mathcal{N}(0, \widetilde{v}_k)$.

**Remark.** We get the asymptotic distribution of $M_k/\hat{\sigma}_n^k$ without computing the joint of $M_k$ and $\hat{\sigma}_n^k$.

**Example.** Consider $k = 3$, under $H_0 \colon X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\sqrt{\frac{n}{6}}\frac{M_3}{\hat{\sigma}_n^3} \xrightarrow{D} \mathcal{N}(0, 1).$$

**Proof.** From the symmetry of normal distribution, Corollary 3.3.1 gives

$$\sqrt{n}\frac{M_3}{\hat{\sigma}_n^3} \xrightarrow{D} \mathcal{N}(0, \sigma^{-6}\operatorname{Var}[Y^3 - 3\sigma^2 Y]) = \mathcal{N}(0, \sigma^{-6}\left(\operatorname{Var}[Y^3] + 9\sigma^4\sigma^2 - 6\sigma^2\mathbb{E}[Y^4]\right))$$

where $\mu_2 = \sigma^2$ and $\operatorname{Cov}\left[Y^3, Y\right] = \mathbb{E}[Y^4] - \mathbb{E}[Y]\mathbb{E}[Y^3] = \mathbb{E}[Y^4]$. Hence, by plugging $\operatorname{Var}[Y^3] = \mu_{2\times 3} = \mu_6$,[a] the variance of the normal is further equal to

$$\frac{\mu_6 + 9\sigma^6 - 6\sigma^2\mu_4}{\sigma^6} = \widetilde{\mu}_6 + 9 - 6\widetilde{\mu}_4 = 15 + 9 - 6 \times 3 = 6,$$

which provides the result. ⊛

[a] More generally, $\operatorname{Var}[Y^k] = \mathbb{E}[Y^{2k}] - (\mathbb{E}[Y^k])^2 = \mathbb{E}[Y^{2k}] = \mu_{2k}$ since $(\mathbb{E}[Y^k])^2 = \mu_k^2 = 0$.

For even $k$ or odd $k$ but $\mu_k \neq 0$, we really need to work out the joint. Since we know the asymptotic distribution of both $M_k$ and $\hat{\sigma}_n^2$, the joint can be obtained by the delta method with $g(M_k, \hat{\sigma}_n^2) = M_k/\hat{\sigma}^k = \widetilde{M}_k$ and the "multivariate" version of central limit theorem.

### 3.3.3 Multivariate Central Limit Theorem

As mentioned above, we now prove the multivariate central limit theorem, i.e., the high dimensional generalization of central limit theorem. We first need the following tool.

**Theorem 3.3.2** (Cramér-Wold device). Let $(X_n)$ be a sequence of random vectors and $X$ be a random vector in $\mathbb{R}^d$. Then $X_n \xrightarrow{D} X$ if and only if $t \cdot X_n \xrightarrow{D} t \cdot X$ for every $t \in \mathbb{R}^d$.

**Proof.** The forward direction is clear from continuous mapping theorem for the linear functional induced from $t$. For the backward direction, assume that $t \cdot X_n \xrightarrow{D} t \cdot X$. Then

$$\phi_{X_n}(t) = \mathbb{E}[e^{it\cdot X_n}] = \phi_{t\cdot X_n}(1) \to \phi_{t\cdot X}(1) = \mathbb{E}[e^{it\cdot X}] = \phi_X(t),$$

which implies $X_n \xrightarrow{D} X$ by the Lévy-Cramer continuity theorem. ■

**Remark.** Proving $X_n \xrightarrow{D} X$ reduces to proving something in the scalar case.

# Lecture 12: Asymptotic Joint Distribution by Multivariate CLT

22 Feb. 9:30

**Theorem 3.3.3** (Multivariate central limit theorem). Let $(X_n)$ be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = \mu \in \mathbb{R}^d$, $\mathrm{Var}[X_i] = \Sigma \in \mathbb{R}^{d\times d}$ for all $1 \le i \le n$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

**Proof.** Set $\mu = 0$, and from Cramér-Wold device, it suffices to show that for any $t \in \mathbb{R}^d$,

$$t \cdot \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right) \xrightarrow{D} t \cdot Z \sim \mathcal{N}(0, t^\top \Sigma t)$$

where $Z \sim \mathcal{N}(0, \Sigma)$. Indeed, since from the univariate central limit theorem, the left-hand side converges to $\mathcal{N}(0, \mathrm{Var}[t \cdot X_i])$, with $\mathrm{Var}[t \cdot X] = t^\top \mathrm{Var}[X_i] t = t^\top \Sigma t = \mathrm{Var}[t \cdot Z]$, we're done. ■

### 3.3.4 Testing Normality with General Moments

With multivariate central limit theorem, we can now generalize Corollary 3.3.1, i.e., finding the asymptotic distribution of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$ for general $k$. Recall the setup, where we let $(X_n)$ and $X$ be i.i.d. random variable, $Y_i = X_i - \mu$ (and $Y = X - \mu$), $\sigma^2 = \mathrm{Var}[X]$, $\mu_k = \mathbb{E}[Y^k]$, and $\widetilde{\mu}_k = \mu_k/\sigma^k$. Let's start with $k = 1$, i.e., compute the asymptotic law of $\overline{X}_n/\hat{\sigma}_n$. In this case, we have proved the following.

**As previously seen.** From Proposition 3.2.1 and Proposition 3.2.2,

- $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ from $\sqrt{n}(\overline{X}_n - \mu) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$, assuming $X \in L^2$;
- $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0, \mu_4 - \sigma^4)$ from $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i^2 - \sigma^2) + o_p(1)$, assuming $X \in L^4$ and $\widetilde{\mu}_4 > 1$.[a]

[a] The latter representation result needs only the assumption of $X \in L^2$.

This together with multivariate central limit theorem and Slutsky's theorem give the following.

**Proposition 3.3.1.** If $X \in L^2$,

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n \\ \hat{\sigma}_n^2\end{pmatrix} - \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}Y_i \\ Y_i^2 - \sigma^2\end{pmatrix} + o_p(1).$$

Moreover, if $X \in L^4$ and $\widetilde{\mu}_4 = \mu_4/\sigma^4 > 1$, then the above converge in distribution to $\mathcal{N}(0, \Sigma)$ where

$$\Sigma = \mathrm{Var}\left[\begin{pmatrix}Y \\ Y^2\end{pmatrix}\right] = \begin{pmatrix}\mathrm{Var}[Y] & \mathrm{Cov}[Y, Y^2] \\ \mathrm{Cov}[Y, Y^2] & \mathrm{Var}[Y^2]\end{pmatrix} = \begin{pmatrix}\sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4\end{pmatrix}.$$

**Remark** (Asymptotically independent). We know that when $X$ is Gaussian, $\overline{X}_n$ and $s_n^2$ are independent. Related back to Corollary 3.3.1, when their skewness is 0, $\overline{X}_n$ and $\hat{\sigma}_n^2$ (or $s_n^2$) are asymptotically independent, which is again confirmed by Proposition 3.3.1 here.

Proposition 3.3.1 gives an asymptotic distribution of $\overline{X}_n$ and $\hat{\sigma}_n^2$, but not $\hat{\sigma}_n$. This is fine since we can further apply the delta method with $g(\overline{X}_n, \hat{\sigma}_n^2) \coloneqq \overline{X}_n/\hat{\sigma}_n$ to get the distribution of $\overline{X}_n/\hat{\sigma}_n$. However, let's leave the application of the delta method to the general $k$. We note the following.

**Note.** The actual characterization of $\overline{X}_n$ and $\hat{\sigma}_n^2$ right before applying central limit theorem is much more useful than the final asymptotic distributions.

Next, we compute the asymptotic law of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$ for general $k > 2$. Following a similar calculation, for $\hat{\sigma}_n^k$, we can again use the result from Proposition 3.2.2 for $\hat{\sigma}_n^2$.

**As previously seen.** From Theorem 3.3.1, if $X \in L^k$,

$$\sqrt{n}(M_k - \mu_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i^k - \mu_k - k\mu_{k-1} Y_i) + o_p(1),$$

and $\sqrt{n}(M_k - \mu_k) \to \mathcal{N}(0, \mathrm{Var}[Y^k - k\mu_{k-1}Y])$ if $X \in L^{2k}$ and the variance is strictly positive.

This implies that for $X \in L^k$ for any $k > 2$,[1]

$$Y \coloneqq \sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_n^2 \\ M_k\end{pmatrix} - \begin{pmatrix}\sigma^2 \\ \mu_k\end{pmatrix}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} Y_i^2 - \sigma^2 \\ Y_i^k - \mu_k - k\mu_{k-1} Y_i \end{pmatrix} + o_p(1), \tag{3.2}$$

which converges to $\mathcal{N}(0, \Sigma)$ from multivariate central limit theorem when $X \in L^{2k}$, where

$$\Sigma = \begin{pmatrix} \mathrm{Var}[Y^2] & \mathrm{Cov}[Y^2, Y^k - k\mu_{k-1}Y] \\ \mathrm{Cov}[Y^2, Y^k - k\mu_{k-1}Y] & \mathrm{Var}[Y^k - k\mu_{k-1}Y] \end{pmatrix}.$$

**Remark.** In general $k$, if $\mu_\ell = 0$ for all odd $\ell$, then $M_k$ and $\hat{\sigma}_n^2$ are asymptotically independent. This is why we get a simplification for odd case in Corollary 3.3.1.

Putting everything together formally, we have the following result for general $k$.

**Theorem 3.3.4.** Let $X \in L^k$ for some $k > 2$. Then for $Z = (X - \mu)/\sigma = Y/\sigma$,

$$\sqrt{n}(\widetilde{M}_k - \widetilde{\mu}_k) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( -\frac{k}{2}\widetilde{\mu}_k (Z_i^2 - 1) + (Z_i^k - \widetilde{\mu}_k - k\widetilde{\mu}_{k-1} Z_i) \right) + o_p(1).$$

Moreover, if $X \in L^{2k}$ and $\widetilde{v}_k \coloneqq \mathrm{Var}\left[-\frac{k}{2}\widetilde{\mu}_k Z^2 + Z^k - k\widetilde{\mu}_{k-1} Z\right] > 0$, then $\sqrt{n}(\widetilde{M}_k - \widetilde{\mu}_k) \xrightarrow{D} \mathcal{N}(0, \widetilde{v}_k)$.

**Proof.** Since Proposition 3.2.2 is for $\hat{\sigma}_n^2$ but not $\hat{\sigma}_n^k$, we need to use delta method by considering $\widetilde{M}_k = M_k / \hat{\sigma}_n^k = g(\hat{\sigma}_n^2, M_k)$ where $g(x, y) \coloneqq y/x^{k/2}$ for $x > 0$, $y \in \mathbb{R}$. We see that

$$\nabla g(\sigma^2, \mu_k) = \begin{pmatrix} -\frac{k}{2}\mu_k \sigma^{-k-2} & \sigma^{-k} \end{pmatrix} = \begin{pmatrix} -\frac{k}{2}\widetilde{\mu}_k \sigma^{-2} & \sigma^{-k} \end{pmatrix}$$

since $\widetilde{\mu}_k = \mu_k/\sigma^k$, $\partial g/\partial x = -kyx^{-k/2-1}/2$, and $\partial g/\partial y = x^{-k/2}$. From delta method and Equation 3.2 with $X \in L^k$, with $\widetilde{\mu}_k = g(\sigma^2, \mu_k)$, we get $\sqrt{n}(g(\hat{\sigma}_n^2, M_k) - g(\sigma^2, \mu_k)) \xrightarrow{D} \nabla g Y$, i.e.,

$$\begin{aligned}
\sqrt{n}(\widetilde{M}_k - \widetilde{\mu}_k) &= \nabla g(\sigma^2, \mu_k) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} Y_i^2 - \sigma^2 \\ Y_i^k - \mu_k - k\mu_{k-1} Y_i \end{pmatrix} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( -\frac{k}{2}\widetilde{\mu}_k \frac{1}{\sigma^2}(Y_i^2 - \sigma^2) + \frac{1}{\sigma^k}(Y_i^k - \mu_k - k\mu_{k-1} Y_i) \right) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( -\frac{k}{2}\widetilde{\mu}_k (Z_i^2 - 1) + (Z_i^k - \widetilde{\mu}_k - k\widetilde{\mu}_{k-1} Z_i) \right) + o_p(1)
\end{aligned}$$

by letting $Z_i \coloneqq (X_i - \mu)/\sigma = Y_i/\sigma$, proving the first claim. Then by the multivariate central limit theorem and Slutsky's theorem, the above further converges in distribution to $\mathcal{N}(0, \widetilde{v}_k)$ when

$$\widetilde{v}_k \coloneqq \mathrm{Var}\left[ -\frac{k}{2}\widetilde{\mu}_k (Z^2 - 1) + (Z^k - \widetilde{\mu}_k - k\widetilde{\mu}_{k-1} Z) \right] = \mathrm{Var}\left[ -\frac{k}{2}\widetilde{\mu}_k Z^2 + Z^k - k\widetilde{\mu}_{k-1} Z \right] > 0,$$

as we assumed. ∎

Compared to Corollary 3.3.1 for odd $k$ and $\mu_k = 0$, there we only get an asymptotic distribution, not an explicit decomposition. With this explicit formula, we can do more. Consider the following example.

[1] This "$Y$" will be used in the delta method later, although this is not exact since $Y$ should be the random vector corresponding the asymptotic distribution. But this is fine in the end from Slutsky's theorem.

**Example.** Consider using both $\widetilde{M}_3$ and $\widetilde{M}_4$ to test $H_0 \colon X \sim \mathcal{N}$. We see that under $H_0$,

$$\left(\sqrt{\frac{n}{\widetilde{v}_3}}\widetilde{M}_3\right)^2 + \left(\sqrt{\frac{n}{\widetilde{v}_4}}(\widetilde{M}_4 - \widetilde{\mu}_4)\right)^2 \xrightarrow{D} \chi^2_2.$$

**Proof.** One can write down $\sqrt{n}(\widetilde{M}_\ell - \widetilde{\mu}_\ell)$ for even $\ell$, and also $\sqrt{n}(\widetilde{M}_k - \widetilde{\mu}_k) = \sqrt{n}\widetilde{M}_k$ for odd $k$, and see that while they both converge to $\mathcal{N}(0,1)$, their covariance is 0, i.e., asymptotically independent, so the square of them add up to $\chi^2_2$. ⊛

Generalizing the above example, for any $X$ with $k > 1$ odd and $\ell > 2$ even, such that every odd central moments vanish with $\widetilde{v}_k, \widetilde{v}_\ell < \infty$,

$$\frac{n}{\widetilde{v}_k}\widetilde{M}_k^2 + \frac{n}{\widetilde{v}_\ell}(\widetilde{M}_\ell - \widetilde{\mu}_\ell)^2 \xrightarrow{D} \chi^2_2.$$

## 3.4 A Quick Detour

We take a slight detour discussing how to asymptotically compare two estimators and how to make the confidence interval (when it depends on too many estimators) more stable.

### 3.4.1 Asymptotic Relative Efficiency

First, consider the following illustrative example.

**Example.** Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$. To estimate $\theta$, as $\theta = \mathbb{E}[X] = \text{Var}[X]$, two natural estimators are $\overline{X}_n$ and $\hat{\sigma}^2_n$. To compare them, we see that

- $\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$;
- $\sqrt{n}(\hat{\sigma}^2_n - \theta) \xrightarrow{D} \mathcal{N}(0, \mu_4 - \sigma^4)$.

As $\sigma^2 = \theta$ and $\mu_4 = 3\theta^2 + \theta$, we see that $\overline{X}_n$ is better since its variance is smaller.

To further quantify how much better is it, we ask how many data we need to we get a similar precision: consider the problem of estimating a scalar parameter $\theta$ such that for two estimators $T^1_n$ and $T^2_n$,

$$\sqrt{n}(T^i_n - \theta) \xrightarrow{D} \sigma_i(\theta) Z \sim \mathcal{N}(0, \sigma^2_i(\theta))$$

Our goal is to find a single number which compares these two estimators. Firstly, for $n$ large enough,

$$\mathbb{P}(\theta \in I^i_n) \coloneqq \mathbb{P}\left(\theta \in T^i_n \pm Z_{\alpha/2}\frac{\sigma_i(\theta)}{\sqrt{n}}\right) \cong 1 - \alpha$$

where $I^i_n \coloneqq T^i_n \pm Z_{\alpha/2}\sigma_i(\theta)/\sqrt{n}$. Let $n_i(\gamma)$ be the value of $n$ such that $|I^i_n| = \gamma$, for $\gamma$ small enough,

$$\gamma \cong 2Z_{\alpha/2}\frac{\sigma_i(\theta)}{\sqrt{n_i(\gamma)}} \Rightarrow n_i(\gamma) \cong \left(\frac{2Z_{\alpha/2}}{\gamma}\sigma_i(\theta)\right)^2,$$

i.e., $n_1(\gamma)/n_2(\gamma) \cong \sigma^2_1(\theta)/\sigma^2_2(\theta)$. We called this the asymptotic relative efficiency $\text{ARE}_\theta(T^1, T^2)$.

**Definition 3.4.1** (Asymptotic relative efficiency for estimator)**.** The *asymptotic relative efficiency* between two estimators $T^1_n$ and $T^2_n$ for $\theta$ such that $\sqrt{n}(T^i_n - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma^2_i(\theta))$ is defined as

$$\text{ARE}_\theta(T^1, T^2) = \frac{\sigma_1(\theta)^2}{\sigma_2(\theta)^2}.$$

**Intuition.** We can read $\text{ARE}_\theta(T^1, T^2) < 1$ as $n_1 < n_2$ and infer $T^1$ is better than $T^2$.

**Note.** Definition 3.4.1 is different from the convention, where we usually define the *asymptotic relative efficiency of* $T^1$ *w.r.t.* $T^2$ as $\text{ARE}_\theta(T^1, T^2) = (\sigma_2(\theta)/\sigma_1(\theta))^2$. But it's just the convention.

### 3.4.2 Variance Stabilizing Transformation

Continuing on the previous example, say we use $\overline{X}_n$ as the estimator of $\theta$. We have

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{D} \sqrt{\theta} Z \sim \sqrt{\theta}\mathcal{N}(0,1) = \mathcal{N}(0,\theta).$$

As the asymptotic distribution depends on $\theta$, we don't directly get a confidence interval.

**As previously seen.** We will usually write $\sqrt{n}/\sqrt{\theta}(\overline{X}_n - \theta) \xrightarrow{D} Z \sim \mathcal{N}(0,1)$, replace $\sqrt{\theta}$ by $\sqrt{\overline{X}_n}$, and apply continuous mapping theorem and Slutsky's theorem to get a confidence interval.

We see that our usual approach relies on (consistently) estimating the variance of the asymptotic distribution, which is potentially "unstable" for small $n$. To get around this, observe that from the delta method with some $g\colon \mathbb{R} \to \mathbb{R}$ differentiable at $\theta$ and $g'(\theta) \neq 0$,

$$\sqrt{n}(g(\overline{X}_n) - g(\theta)) \xrightarrow{D} g'(\theta)\sqrt{\theta} Z.$$

This suggests that if we can select $g$ such that $g'(\theta)\sqrt{\theta} = c > 0$ is some constant for every $\theta > 0$, our goal is achieved since now we have

$$\frac{\sqrt{n}}{c}(g(\overline{X}_n) - g(\theta)) \xrightarrow{D} \mathcal{N}(0,1).$$

In this case, we get an asymptotic confidence interval for $g(\theta)$ with confidence level $1-\alpha$ as

$$\left(g(\overline{X}_n) - Z_{\alpha/2}\frac{c}{\sqrt{n}}, g(\overline{X}_n) + Z_{\alpha/2}\frac{c}{\sqrt{n}}\right),$$

and hence an asymptotic confidence interval for $\theta$ with confidence level $1-\alpha$ is just

$$\left(g^{-1}\left(g(\overline{X}_n) - Z_{\alpha/2}\frac{c}{\sqrt{n}}\right), g^{-1}\left(g(\overline{X}_n) + Z_{\alpha/2}\frac{c}{\sqrt{n}}\right)\right),$$

This is the so-called *variance stabilizing transformation.*

**Claim.** For $c = 1/2$, $g(\theta) = \sqrt{\theta}$ suffices. Hence, in this case, $g^{-1}(u) = u^2$.

**Proof.** Since for $g'(\theta) = \frac{1}{2\sqrt{\theta}}$, we have $g(\theta) = \sqrt{\theta}$. ⊛

Lastly, we note that the above can be easily generalized.

**Remark.** Consider estimating a scalar parameter $\theta$ in some open interval $\Theta$, where we replace:

- $\sqrt{\theta}$ by $h(\theta)$, a positive function;[a]
- $\sqrt{n}$ by $b_n$, a positive divergent strictly increasing sequence;
- $\overline{X}_n$ by $T_n$, a consistent estimator of $\theta$.

In this way, letting $g'(\theta)h(\theta) = c > 0$ for all $\theta \in \Theta$ asserts $g'(\theta) > 0$ for all $\theta \in \Theta$, hence $g$ is strictly increasing and its usual inverse $g^{-1}$ is well-defined.

[a] We don't need continuity since we don't need $h(T_n)$ when doing the variance stabilizing transformation.

**Note.** Variance stabilizing transformation doesn't have any theoretical guarantees. Rather, it's just our guess that by making the variance independent of $n$, it'll become better in terms of stability.

# Lecture 13: Bahadur's Representation for Quantiles

## 3.5 Inference for Population Quantiles

27 Feb. 9:30

Let $X, X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ for some distribution function $F$, and let $\theta_p$ for some $p \in (0,1)$ be the $p^{\text{th}}$ quantile, which we recall is defined as $F^{-1}(p) = \inf\{t \in \mathbb{R} \colon F(t) \geq p\}$.

> **Intuition.** Since $F^{-1}(p)$ depends on $F$, if we have an estimation of $F$ itself, then we can have an estimation of $F^{-1}(p)$.

Specifically, to estimate $F$, consider the empirical cdf $\hat{F}_n(t)$ such that for all $t \in \mathbb{R}$,

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$$

Now, from $\hat{F}_n(t)$, we estimate $\theta_p = F^{-1}(p)$ by the *$p^{th}$-sample quantile*

$$\hat{\theta}_p \coloneqq \hat{F}_n^{-1}(p) \coloneqq \inf\{t \in \mathbb{R} \colon \hat{F}_n(t) \geq p\}.$$

> **Remark.** If $F$ is continuous, then apart from a null set we have
>
> $$\hat{\theta}_p = \inf\left\{X_{(i)} \colon \hat{F}_n(X_{(i)}) = i/n \geq p\right\} = \inf\left\{t \in \mathbb{R} \colon \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} \geq \lceil np \rceil\right\} = X_{(\lceil np \rceil)}.$$
>
> **Proof.** Since $F$ is continuous, with probability 1 there are no ties among $X_i$'s, hence $\hat{F}_n(t)$ has jumps of size $1/n$ at every order statistic $X_{(i)}$. Finally, the ceiling can be taken since $\sum_{i=1}^{n} \mathbb{1}_{X_i \geq t} \in \mathbb{N}$. ⊛

### 3.5.1 Consistency

Firstly, $\hat{F}_n$ is a consistent estimator of $F$ since by weak law of large number, $\hat{F}_n(t) \xrightarrow{p} \mathbb{P}(X \leq t) = F(t)$. In fact, the convergence is exponentially fast in $n$ by observing the following.

> **Note.** By fixing $t$, $\mathbb{1}_{X \leq t}$ is $\text{Ber}(F(t))$, hence $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{D} \mathcal{N}(0, F(t)(1-F(t)))$.

This implies that $\hat{F}_n(t)$ is an average of i.i.d. Bernoulli random variables, hence Hoeffding's inequality implies that the convergence is exponentially fast, i.e., for all $n \in \mathbb{N}$, $t \in \mathbb{R}$, and $\epsilon > 0$,

$$\mathbb{P}(|\hat{F}_n(t) - F(t)| > \epsilon) \leq 2\exp\left(-n\epsilon^2/2\right).$$

We now show the consistency of $\hat{\theta}_p$ when the corresponding $\theta_p$ is unique. Recall the following.

> **As previously seen.** $t \geq F^{-1}(p) \Leftrightarrow F(t) \geq p$ and $t < F^{-1}(p) \Leftrightarrow F(t) < p$. This is also true for $\hat{F}_n$.

> **Theorem 3.5.1.** If $F(\theta_p + \epsilon) > F(\theta_p) \geq p$ for any $\epsilon > 0$, then $\hat{\theta}_p \xrightarrow{p} \theta_p$. More generally, if $p_n \to p$, then $\hat{\theta}_{p_n} \xrightarrow{p} \theta_p$.
>
> **Proof.** We want to show that for any $\epsilon > 0$, $\mathbb{P}(|\hat{\theta}_{p_n} - \theta_p| > \epsilon) \to 0$. We see that
>
> $$\mathbb{P}(|\hat{\theta}_{p_n} - \theta_p| > \epsilon) = \mathbb{P}(\hat{\theta}_{p_n} > \theta_p + \epsilon) + \mathbb{P}(\hat{\theta}_{p_n} < \theta_p - \epsilon).$$
>
> For the first term, $\hat{\theta}_{p_n} = \hat{F}_n^{-1}(p_n) > \theta + \epsilon$, hence $p_n > \hat{F}_n(\theta_p + \epsilon)$, which gives
>
> $$p_n - p + p - F(\theta_p + \epsilon) > \hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon).$$

Since $p < F(\theta_p + \epsilon)$, let $-\delta := p - F(\theta_p + \epsilon)$ for some $\delta > 0$, then

$$\hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon) < p_n - p - \delta < \frac{\delta}{2} - \delta = -\frac{\delta}{2}$$

for large enough $n$ such that $|p_n - p| < \delta/2$, which implies $|\hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon)| > \delta/2$, i.e.,

$$\mathbb{P}(\hat{\theta}_{p_n} > \theta + \epsilon) \leq \mathbb{P}(|\hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon)| > \delta/2),$$

which goes to 0 as $n \to \infty$ from the consistency of $\hat{F}_n$. The second term can be proved similarly. ■

**Note.** The convergence in Theorem 3.5.1 is also exponentially fast in $n$.

### 3.5.2 Bahadur's Representation Theorem

If $F$ is differentiable, we can establish the asymptotic normality of $\hat{\theta}_{p_n}$.

**Theorem 3.5.2** (Bahadur's representation). If $F'(\theta_p) =: f(\theta_p) > 0$ and $\sqrt{n}(p_n - p) = O(1)$, then

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{p_n - \mathbb{1}_{X_i \leq \theta_p}}{f(\theta_p)} + o_p(1).$$

Let's postpone the proof and discuss its implication first.

**Corollary 3.5.1.** If $F'(\theta_p) =: f(\theta_p) > 0$ and $\sqrt{n}(p_n - p) \to c \in [0, \infty)$, then

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) \xrightarrow{p} \frac{c}{f(\theta_p)}$$

and

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) \xrightarrow{D} \mathcal{N}\left(\frac{c}{f(\theta_p)}, \frac{p(1-p)}{f^2(\theta_p)}\right).$$

**Proof.** From Bahadur's representation shows

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{p - \mathbb{1}_{X_i \leq \theta_p}}{f(\theta_p)} + \frac{\sqrt{n}(p_n - p)}{f(\theta_p)} + o_p(1),$$

implying the first claim. For the second claim, firstly, if $\sqrt{n}(p_n - p) \to 0$, from central limit theorem,

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) \xrightarrow{D} \mathcal{N}\left(0, \frac{F(\theta_p)(1 - F(\theta_p))}{f^2(\theta_p)}\right) = \mathcal{N}\left(0, \frac{p(1-p)}{f^2(\theta_p)}\right).$$

Now for $\sqrt{n}(p_n - p) \to c$, we first look at $\hat{\theta}_{p_n}$ and $\hat{\theta}_p$ instead, which gives

$$\sqrt{n}(\hat{\theta}_{p_n} - \hat{\theta}_p) = \sqrt{n}\left((\hat{\theta}_{p_n} - \theta_p) - (\hat{\theta}_p - \theta_p)\right) = \sqrt{n}\frac{p_n - p}{f(\theta_p)} + o_p(1) \xrightarrow{p} \frac{c}{f(\theta_p)}.$$

Moreover, from central limit theorem and Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) = \sqrt{n}(\hat{\theta}_{p_n} - \hat{\theta}_p) + \sqrt{n}(\hat{\theta}_p - \theta_p) \xrightarrow{D} \mathcal{N}\left(\frac{c}{f(\theta_p)}, \frac{p(1-p)}{f^2(\theta_p)}\right),$$

where the variance calculation is the same as the case of $c = 0$ above. ■

**Intuition.** This is expected since if the density is low, then we don't have many data to evaluate $\theta_p$ in the first place, hence the precision will be low (large variance).

### 3.5.3 Confidence Intervals

When $c = 0$, Corollary 3.5.1 gives an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\theta_p$ as

$$\hat{\theta}_{p_n} \pm Z_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n} f(\theta_p)}.$$

However, to implement this confidence interval, we need to estimate $f(\theta_p)$ consistently. To avoid this, consider a sequence of intervals $(\hat{\theta}_{\ell_n}, \hat{\theta}_{u_n})$ for some $\ell_n < p_n < u_n$ such that

$$\hat{\theta}_{\ell_n} \xrightarrow{p} \hat{\theta}_p - Z_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n} f(\theta_p)} \text{ and } \hat{\theta}_{u_n} \xrightarrow{p} \hat{\theta}_p + Z_{\alpha/2} \frac{\sqrt{p(1-p)}}{\sqrt{n} f(\theta_p)}.$$

This will also give us an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\theta_p$. The upshot is that, this is easy to construct without estimating $f(\theta_p)$ explicitly.

**Example.** Consider $\ell_n = p - Z_{\alpha/2}\sqrt{p(1-p)}/\sqrt{n}$, and similarly, $u_n = p + Z_{\alpha/2}\sqrt{p(1-p)}/\sqrt{n}$.

The above construction works due to the following.

**Proposition 3.5.1.** Let $c = Z_{\alpha/2}\sqrt{p(1-p)}$, and let $\ell_n$ and $u_n$ such that $\sqrt{n}(\ell_n - p) \to -c$ and $\sqrt{n}(u_n - p) \to c$. If $F'(\theta_p) =: f(\theta_p) > 0$, then $\mathbb{P}(\hat{\theta}_{\ell_n} \le \theta_p \le \hat{\theta}_{u_n}) \to 1 - \alpha$.

**Proof.** First, consider $\ell_n$. Since $\sqrt{n}(\ell_n - p) \to -c$, then $\hat{\theta}_{\ell_n}$ defined above is guaranteed from Corollary 3.5.1 since it's equivalent to

$$\sqrt{n}(\hat{\theta}_{\ell_n} - \hat{\theta}_p) \xrightarrow{p} \frac{-c}{f(\theta_p)} = -Z_{\alpha/2} \frac{\sqrt{p(1-p)}}{f(\theta_p)}.$$

The same holds for $u_n$, hence we're done. ■

**Remark.** We can construct $(\hat{\theta}_{\ell_n}, \hat{\theta}_{u_n})$ without assuming knowledge or having to estimate $f(\theta_p)$.

### 3.5.4 Estimating the Center of a Distribution

Another implication is comparing the sample mean and the sample median as estimators of the center of a symmetric distribution.

**Definition 3.5.1** (Median). When $p = 1/2$, $\theta_{1/2}$ is called the *median*.

Firstly, for $p = 1/2$, if $F'(\theta_{1/2}) =: f(\theta_{1/2}) > 0$, from Corollary 3.5.1 we have

$$\sqrt{n}(\hat{\theta}_{1/2} - \theta_{1/2}) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{4f^2(\theta_{1/2})}\right).$$

Suppose further, $\theta_{1/2} = \mu$ and $\mathrm{Var}[X] = \sigma^2 < \infty$. Then both $\hat{\theta}_{1/2}$ and $\overline{X}_n$ are two possible estimators of $\mu$, and in this case, we might want to look at the asymptotic relative efficiency. Specifically,

$$\mathrm{ARE}(\overline{X}_n, \hat{\theta}_{1/2}) = \frac{\sigma^2}{\frac{1}{4f^2(\theta_{1/2})}} = 4\sigma^2 f^2(\theta_{1/2}).$$

Let's summarize the above in the following.

**Proposition 3.5.2.** Suppose $\mu = \mathbb{E}[X]$ exists and $\sigma^2 = \mathrm{Var}[X] < \infty$ such that $\mu = \theta_{1/2}$. If $F'(\mu) =: f(\mu) > 0$, then $\mathrm{ARE}(\overline{X}, \hat{\theta}_{1/2}) = (2\sigma f(\mu))^2$.

The following two examples suggest that the sample median is asymptotically better than the sample mean when $X$ has heavy tails.

**Example.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\overline{X}_n$ is a better estimator of $\mu$ than $\hat{\theta}_{1/2}$.

**Proof.** Since $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, hence $f(\mu) = 1/\sigma\sqrt{2\pi}$, i.e.,

$$\mathrm{ARE}(\overline{X}_n, \hat{\theta}_{1/2}) = 4\sigma^2 \frac{1}{\sigma^2 2\pi} = \frac{2}{\pi} < 1,$$

which means $\overline{X}_n$ is a better estimator of $\mu$ than $\hat{\theta}_{1/2}$. ⊛

**Example.** If $X \sim \mathrm{Laplace}(\mu, b)$ where $\sigma^2 = 2b^2$, then $\hat{\theta}_{1/2}$ is a better estimator of $\mu$ than $\overline{X}_n$.

**Proof.** Since $f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} = \frac{1}{\sigma\sqrt{2}} e^{-\frac{|x-\mu|}{\sigma/\sqrt{2}}}$, hence $f(\mu) = 1/\sqrt{2}\sigma$, i.e.,

$$\mathrm{ARE}(\overline{X}_n, \hat{\theta}_{1/2}) = 4\sigma^2 \frac{1}{2\sigma^2} = 2 > 1,$$

which means $\hat{\theta}_{1/2}$ is a better estimator of $\mu$ than $\overline{X}_n$. ⊛

One might want to consider $c\overline{X} + (1-c)\hat{\theta}_{1/2}$ for any $c \in [0,1]$. In this case, by Bahadur's representation and delta method, one can have

$$\sqrt{n}\left((c\overline{X} + (1-c)\hat{\theta}_{1/2}) - \mu\right) \xrightarrow{D} \mathcal{N}(0, V)$$

where

$$V = c^2 \operatorname{Var}[X] + (1-c)^2 \frac{1}{4f^2(\mu)} + 2c(1-c) \operatorname{Cov}\left[X - \mu, \frac{1/2 - \mathbb{1}_{X \le \mu}}{f(\mu)}\right].$$

# Lecture 14: Proof of Bahadur's Representation Theorem

## 3.5.5 Proof of Bahadur's Representation Theorem

29 Feb. 9:30

Now we prove the Bahadur's representation theorem. Recall the statement.

**As previously seen.** Given $F'(\theta_p) =: f(\theta_p) > 0$ and $\sqrt{n}(p_n - p) = O(1)$, we want to prove that

$$\sqrt{n}(\hat{\theta}_{p_n} - \theta_p) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{p - \mathbb{1}_{X_i \le \theta_p}}{f(\theta_p)} - \sqrt{n}\frac{p_n - p}{f(\theta_p)} = o_p(1).$$

We now start the proof.

**Proof of Theorem 3.5.2.** Firstly, we write

$$W_n := \sqrt{n}(\hat{\theta}_{p_n} - \theta_p) - \sqrt{n}\frac{p_n - p}{f(\theta_p)},$$

and from $p = F(\theta_p)$,

$$U_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{p - \mathbb{1}_{X_i \le \theta_p}}{f(\theta_p)} = \frac{\sqrt{n}(p - \hat{F}_n(\theta_p))}{f(\theta_p)} = \frac{\sqrt{n}(F(\theta_p) - \hat{F}_n(\theta_p))}{f(\theta_p)},$$

so we want to show $W_n - U_n = o_p(1)$. Consider the following lemma.

**Lemma 3.5.1.** Given two sequences of random variable $(W_n), (U_n)$ such that one of them is $O_p(1)$ and for every $\epsilon > 0$ and every $t \in \mathbb{R}$,

$$\mathbb{P}(W_n \le t, U_n \ge t + \epsilon) + \mathbb{P}(U_n \le t, W_n \ge t + \epsilon) \to 0,$$

then $U_n - W_n \xrightarrow{p} 0$.

**Proof.** Without loss of generality, suppose $W_n = O_p(1)$, and we show that for every $\epsilon > 0$, $\mathbb{P}(|W_n - U_n| > \epsilon) \to 0$. Firstly, observe that for every fixed $\epsilon > 0$, if $b - a < \epsilon/2$,

$$\mathbb{P}(a \le W_n \le b, |W_n - U_n| > \epsilon) \to 0$$

since the left-hand side is equal to

$$\begin{aligned}
&\mathbb{P}(a \le W_n \le b, U_n > W_n + \epsilon) + \mathbb{P}(a \le W_n \le b, U_n < W_n - \epsilon) \\
&\le \mathbb{P}(W_n \le b, U_n > a + \epsilon) + \mathbb{P}(a \le W_n, U_n < b - \epsilon) \\
&\le \mathbb{P}(W_n \le b, U_n > a + (2b - 2a)) + \mathbb{P}(a \le W_n, U_n < b - (2b - 2a)) \\
&= \mathbb{P}(W_n \le b, U_n > b + (b - a)) + \mathbb{P}(a \le W_n, U_n < a + (a - b)),
\end{aligned}$$

which goes to 0 from our assumption. Furthermore, fix any $\delta > 0$, since $W_n = O_p(1)$, there exists $M > 0$ such that $\mathbb{P}(|W_n| \le M) \ge 1 - \delta$ for every $n \ge 1$. Then,

$$\begin{aligned}
\mathbb{P}(|U_n - W_n| > \epsilon) &\le \mathbb{P}(|W_n| > M) + \mathbb{P}(|W_n| \le M, |U_n - W_n| > \epsilon) \\
&\le \delta + \mathbb{P}(-M \le W_n \le M, |U_n - W_n| > \epsilon).
\end{aligned}$$

The second term is like the first observation, but now we have a larger interval $[-M, M]$ rather than some $[a, b]$ with $b - a < \epsilon/2$. To compensate this, consider pair-wise disjoint intervals $(a_i, b_i)$ for $i \in I$ with $|I| < \infty$ such that $b_i - a_i < \epsilon/2$ for all $i \in I$ and $\bigcup_{i \in I} [a_i, b_i] \supseteq [-M, M]$,

$$\mathbb{P}(-M \le W_n \le M, |U_n - W_n| > \epsilon) \le \sum_{i \in I} \mathbb{P}(a_i \le W_n \le b_i, |U_n - W_n| > \epsilon).$$

Since $I$ is finite, together with the first observation, implies $\limsup_{n \to \infty} \mathbb{P}(|U_n - W_n|) \le \delta$. As $\delta$ is arbitrary, letting $\delta \to 0$ completes the proof. ■

Clearly, $U_n = O_p(1)$ since it converges in distribution, so we can try to apply Lemma 3.5.1. First, we study the numerator of $U_n$, i.e., $Z_n(t) \coloneqq \sqrt{n}(F(t) - \hat{F}_n(t))$. We have seen that $\mathbb{E}[Z_n(t)] = 0$ and $\mathrm{Var}[Z_n(t)] = F(t)(1 - F(t))$, and $Z_n(t) \xrightarrow{D} \mathcal{N}(0, F(t)(1 - F(t)))$ by central limit theorem.

**Claim.** For any $t, s \in \mathbb{R}$, $\mathrm{Var}[Z_n(t) - Z_n(s)] = \mathbb{E}[(Z_n(t) - Z_n(s))^2] \le |F(t) - F(s)|$. Hence, if $s_n \to s$ and $F$ is continuous at $s$, $Z_n(s_n) - Z_n(s) \xrightarrow{L^2} 0$, hence $Z_n(s_n) - Z_n(s) \xrightarrow{p} 0$.

**Proof.** Observe that $\mathrm{Var}[Z_n(t) - Z_n(s)] = \mathrm{Var}[\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}] \le \mathbb{E}[|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}|^2]$ where

$$|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}| = \begin{cases} 1, & \text{if } s < X \le t \text{ or } t < X \le s; \\ 0, & \text{otherwise.} \end{cases}$$

Hence, as $|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}| = |\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}|^2$,

$$\begin{aligned}
\mathbb{E}[|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}|^2] &= \mathbb{P}(s < X \le t) + \mathbb{P}(t < X \le s) \\
&= (F(t) - F(s))^+ + (F(s) - F(t))^+ = |F(t) - F(s)|,
\end{aligned}$$

i.e., $|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}| \sim \mathrm{Ber}(|F(t) - F(s)|)$. ⊛

From Lemma 3.5.1, it suffices to show $\mathbb{P}(W_n \le t, U_n \ge t + \epsilon) \to 0$ and $\mathbb{P}(U_n \le t, W_n \ge t + \epsilon) \to 0$ for every $t \in \mathbb{R}$ and $\epsilon > 0$. Let's show the first one only. Fix $t \in \mathbb{R}$ and $\epsilon > 0$, then

$$\begin{aligned}
W_n \le t &\Leftrightarrow \sqrt{n}(\hat{\theta}_{p_n} - \theta_p) - \sqrt{n}\frac{p_n - p}{f(\theta_p)} \le t \\
&\Leftrightarrow \hat{\theta}_{p_n} = \hat{F}_n^{-1}(p_n) \le \theta_p + \frac{t}{\sqrt{n}} + \frac{p_n - p}{f(\theta_p)} =: \theta_p + \delta_n \qquad \delta_n \coloneqq \frac{t}{\sqrt{n}} + \frac{p_n - p}{f(\theta_p)}
\end{aligned}$$

From the property of $\hat{F}_n^{-1}$, $p_n \le \hat{F}_n(\theta_p + \delta_n)$,

$$\Leftrightarrow \sqrt{n}(p_n - F(\theta_p + \delta_n)) \le \sqrt{n}(\hat{F}_n(\theta_p + \delta_n) - F(\theta_p + \delta_n)) = -Z_n(\theta_p + \delta_n),$$

which can be written as

$$Z_n(\theta_p+\delta_n)\le\sqrt{n}(F(\theta_p+\delta_n)-p_n)\Leftrightarrow\frac{Z_n(\theta_p+\delta_n)}{f(\theta_p)}\le\frac{\sqrt{n}(F(\theta_p+\delta_n)-p_n)}{f(\theta_p)}=:t_n.$$

Putting everything together, with $U_n=Z_n(\theta_p)/f(\theta_p)$, we have

$$\begin{aligned}\mathbb{P}(W_n\le t,U_n\ge t+\epsilon)&=\mathbb{P}(Z_n(\theta_p+\delta_n)\le t_nf(\theta_p),Z_n(\theta_p)\ge f(\theta_p)(t+\epsilon))\\&\le\mathbb{P}(Z_n(\theta_p+\delta_n)-Z_n(\theta_p)\le(t_n-t-\epsilon)f(\theta_p))\\&=\mathbb{P}\left(\frac{Z_n(\theta_p+\delta_n)-Z_n(\theta_p)}{f(\theta_p)}-(t_n-t)\le-\epsilon\right),\end{aligned}$$

which goes to 0 as $n\to\infty$ if $t_n\to t$ since from the previous claim:

- let $s_n:=\theta_p+\delta_n$, $s:=\theta_p$, with $F$ being continuous at $s$ and $\delta_n\to0$, $Z_n(\theta_p+\delta_n)-Z_n(\theta_p)\xrightarrow{p}0$;
- if further, $t_n\to t$, the left-hand side goes to 0, and the inequality tends to be vacuous.

**Claim.** Indeed, $t_n\to t$.

**Proof.** We want to show that

$$t_n=\frac{F(\theta_p+\delta_n)-p_n}{f(\theta_p)/\sqrt{n}}\to t.$$

By assumption, as $\delta_n\to0$ and $F'(\theta_p)=f(\theta_p)$,

$$\frac{F(\theta_p+\delta_n)-F(\theta_p)}{\delta_n}\to f(\theta_p)\Leftrightarrow\frac{F(\theta_p+\delta_n)-F(\theta_p)-\delta_nf(\theta_p)}{\delta_n}\to0,$$

i.e., $F(\theta_p+\delta_n)=F(\theta_p)+\delta_nf(\theta_p)+o(\delta_n)$. Since $F(\theta_p)=p$ and $\delta_n=t/\sqrt{n}+(p_n-p)/f(\theta_p)$,

$$F(\theta_p+\delta_n)=p+\left(\frac{t}{\sqrt{n}}+\frac{p_n-p}{f(\theta_p)}\right)f(\theta_p)+o(\delta_n)=p+\frac{t}{\sqrt{n}}f(\theta_p)+(p_n-p)+o(\delta_n).$$

Rearranging, with $o(\delta_n)\cdot\sqrt{n}/f(\theta_p)=\sqrt{n}o(\delta_n)$ from $f(\theta_p)>0$, we have

$$t_n=\frac{F(\theta_p+\delta_n)-p_n}{f(\theta_p)/\sqrt{n}}=t+\sqrt{n}o(\delta_n).$$

Finally, since $o(\delta_n)=\delta_no(1)$, with

$$\sqrt{n}\delta_n=\sqrt{n}\left(\frac{t}{\sqrt{n}}+\frac{p_n-p}{f(\theta_p)}\right)=t+\frac{\sqrt{n}(p_n-p)}{f(\theta_p)}=O(1)$$

from our assumption, we have $\sqrt{n}o(\delta_n)=O(1)o(1)=o(1)$, hence $t_n=t+o(1)\to t$. ⊛

The second claim $\mathbb{P}(U_n\le t,W_n\ge t+\epsilon)\to0$ can be proved similarly, hence we're done. ■

## 3.6 Inference for Distribution Function

Since we estimate $F$ by $\hat{F}_n$ when estimating $\theta_p$ by $\hat{\theta}_p$, one might just as well focus on the former task.

### 3.6.1 Consistency

Since $\sqrt{n}(\hat{F}_n(t)-F(t))\xrightarrow{D}\mathcal{N}(0,F(t)(1-F(t)))$ for any fixed $t$, so given $t_1,\dots,t_m\in\mathbb{R}$, we have

$$\begin{pmatrix}\sqrt{n}(\hat{F}_n(t_1)-F(t_1))\\\vdots\\\sqrt{n}(\hat{F}_n(t_m)-F(t_m))\end{pmatrix}=\begin{pmatrix}Z_n(t_1)\\\vdots\\Z_n(t_m)\end{pmatrix}=\frac{1}{\sqrt{n}}\sum_{i=1}^n\begin{pmatrix}\mathbb{1}_{X_i\le t_1}-F(t_1)\\\vdots\\\mathbb{1}_{X_i\le t_m}-F(t_m)\end{pmatrix}\xrightarrow{D}\mathcal{N}(0,\Sigma)$$

from multivariate central limit theorem where

$$\Sigma_{ij} = \mathrm{Cov}[\mathbb{1}_{X\le t_i}, \mathbb{1}_{X\le t_j}] = \mathbb{P}(X \le t_i \wedge X \le t_j) - \mathbb{P}(X \le t_i)\mathbb{P}(X \le t_j).$$

# Lecture 15: Inference for Cumulative Density Function

Surprisingly, this consistency property holds uniformly over $t$. 5 Mar. 9:30

**Theorem 3.6.1** (Glivenko-Cantelli). Given a cdf $F$, $\|\hat{F}_n - F\|_\infty \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$.

**Proof.** Fix some $\epsilon > 0$, and let $\epsilon > 2/k$ for some $k \in \mathbb{N}$. Then, for finitely many $t_1, \dots, t_{k-1}$, $\hat{F}_n(t_i) \overset{\text{a.s.}}{\to} F(t_i)$ and $\hat{F}_n(t_i^-) \overset{\text{a.s.}}{\to} F(t_i^-)$ for every $1 \le i \le k-1$ from the strong law of large number. This means that there exists $n_0$ such that for $n \ge n_0$, for every $\omega \notin N$ such that $\mathbb{P}(N) = 0$,

$$|\hat{F}_n(t_i, \omega) - F(t_i, \omega)| < \frac{1}{k}, \text{ and } |\hat{F}_n(t_i^-, \omega) - F(t_i^-, \omega)| < \frac{1}{k},$$

so when $t \in \{t_i\}_{i=1}^{k-1}$ for some finite $k$, the desired bound is established. In particular, consider $t_i = \inf\{t \in \mathbb{R} : F(t) > i/k\}$ for $1 \le i \le k-1$, and $t_0 := -\infty$, $t_k = \infty$. Then for any $t \in \mathbb{R} \setminus \{t_i\}_{i=1}^{k-1}$, there exists a unique $i$ such that $t \in (t_{i-1}, t_i)$, and furthermore, for all $n \ge n_0$,

$$\hat{F}_n(t) - F(t) \le \hat{F}_n(t_i^-) - F(t_{i-1}) = \hat{F}_n(t_i^-) - F(t_i^-) + F(t_i^-) - F(t_{i-1}) \le \frac{1}{k} + \frac{i}{k} - \frac{i-1}{k} = \frac{2}{k} < \epsilon$$

Similarly, we can show $\hat{F}_n(t) - F(t) > -\epsilon$ for all $t \in \mathbb{R} \setminus \{t_i\}_{i=1}^{k-1}$, which completes the proof. ■

## 3.6.2 Donsker's Theorem

On the other hand, for distributional result, first recall the *empirical process*

$$Z_n(t) := \sqrt{n}(F(t) - \hat{F}(t))$$

for $t \in \mathbb{R}$ introduced in the proof of Bahadur representation theorem. Recall the following.

**As previously seen.** We have seen that

$$Z_n(t) := \sqrt{n}(F(t) - \hat{F}_n(t)) \overset{D}{\to} B_F(t) := \mathcal{N}(0, F(t)(1 - F(t))).$$

Furthermore, for any $t_1, \dots, t_m \in \mathbb{R}$,

$$(Z_n(t_1), \dots, Z_n(t_m)) \overset{D}{\to} (B_F(t_1), \dots, B_F(t_m)) \sim \mathcal{N}(0, \Sigma_F(t_1, \dots, t_m))$$

where for $1 \le i \le j \le m$,

$$\mathrm{Cov}[B_F(t_i), B_F(t_j)] = \mathrm{Cov}[\mathbb{1}_{X\le t_i}, \mathbb{1}_{X\le t_j}] = F(t_i \wedge t_j) - F(t_i)F(t_j).$$

We now ask the same question, i.e., does the convergence hold uniformly over $t$?

**Intuition.** As the theory of weak convergence applies to sequences of random elements that take values in general metric spaces, it's reasonable to conjecture that $(Z_n)$ converges weakly to some random process $B_F$ with index set $\mathbb{R}$, i.e., $B_F = \{B_F(t)\}_{t\in\mathbb{R}}$, such that for every $t, s \in \mathbb{R}$,

$$\mathbb{E}[B_F(t)] = 0 \text{ and } \mathrm{Cov}[B_F(t), B_F(s)] = F(t \wedge s) - F(t)F(s).$$

The conjecture is indeed correct, and it's an extension of Donsker's theorem.

**Note.** The formal setup is to view each $Z_n$ as a random element that takes values on the space $\mathcal{D}$ of right continuous functions with left limits with the norm $\|\cdot\|_\infty$.

**Example** (Brownian bridge). A *Brownian bridge* is $B \coloneqq B_F$ when $F(t) = t$, i.e., $F \sim \mathcal{U}([0,1])$.

**Note.** For any cdf $F$, $B_F(t)$ is just $B$ index at $F(t)$ for any $t \in \mathbb{R}$, i.e., $B_F(t) = B(F(t))$.

Taking this convergence as granted (and work with continuous $F$), we have

$$(Z_n) \coloneqq (t \mapsto Z_n(t))_{n\geq 1} \xrightarrow{\text{w}} B_F \coloneqq \{t \mapsto B(t)\}_{t\in\mathbb{R}}.$$

One immediate consequence is the following.

**Proposition 3.6.1.** If $T\colon \mathcal{D} \to \mathbb{R}$ is continuous,[a] then $T(Z_n) \xrightarrow{D} T(B_F)$.

[a] I.e., $T(G_n) \to T(F)$ if $(G_n), F \in \mathcal{D}$ such that $\|G_n - F\|_\infty \to 0$.

**Proof.** Since $Z_n \xrightarrow{\text{w}} B_F$, continuous mapping theorem proves the result. ■

### 3.6.3 Confidence Bands

One immediate application of Proposition 3.6.1 is the following.

**Corollary 3.6.1.** If $F$ is continuous, then $\sqrt{n}\|\hat{F}_n - F\|_\infty \xrightarrow{D} \|B_F\|_\infty$.

**Proof.** From Proposition 3.6.1, we just note that $\|\cdot\|_\infty$ is continuous since it's a norm. ■

In particular, Corollary 3.6.1 allows us to do inference.

**Example** (Confidence band). For any $\alpha \in (0,1)$, let $d_\alpha$ to be defined as $\alpha =: \mathbb{P}(\|B_F\|_\infty > d_\alpha)$, then if $F$ is continuous, $\mathbb{P}(\sqrt{n}\|F - \hat{F}_n\|_\infty \geq d_\alpha) \to \alpha$, i.e.,

$$\mathbb{P}\left(\forall t \in \mathbb{R}\colon \hat{F}_n(t) - \frac{d_\alpha}{\sqrt{n}} \leq F(t) \leq \hat{F}_n(t) + \frac{d_\alpha}{\sqrt{n}}\right) \to 1 - \alpha.$$

Another application of Proposition 3.6.1 is the following.

**Corollary 3.6.2.** If $F$ is continuous, then

$$n\int_{\mathbb{R}} \left(\hat{F}_n(t) - F(t)\right)^2 F(\mathrm{d}t) \xrightarrow{D} \int_{\mathbb{R}} B_F^2(t) F(\mathrm{d}t).$$

**Proof.** From Proposition 3.6.1, it suffices to show that $G \mapsto \int_{\mathbb{R}} G^2\,\mathrm{d}F$ is continuous for $G \in \mathcal{D}$. Firstly, let $(G_n), G \in \mathcal{D}$ such that $\|G_n - G\|_\infty \to 0$. Then

$$\begin{aligned}|T(G_n) - T(G)| &= \left|\int_{\mathbb{R}} G_n^2 - G^2\,\mathrm{d}F\right| \\ &\leq \int_{\mathbb{R}} |G_n^2 - G^2|\,\mathrm{d}F \leq \|G_n - G\|_\infty \int_{\mathbb{R}} |G_n(t) + G(t)| F(\mathrm{d}t) \leq 2\|G_n - G\|_\infty\end{aligned}$$

since $\|G_n - G\|_\infty = \sup_t |G_n(t) - G(t)|$. As $\|G_n - G\|_\infty \to 0$, we're done. ■

**Note.** We can also obtain a confidence band from Corollary 3.6.2 as in the previous example.

### 3.6.4 Goodness of Fit Tests

We now consider the problem of testing the null hypothesis $H_0\colon F = F_0$ given $F_0$, i.e., the *goodness of fit test*. Suppose $F$ is continuous, then Corollary 3.6.1 and Corollary 3.6.2 suggest we can test $H_0$ using the following two statistics.

**Example** (Kolmogorov-Smirnov statistic). Consider the *Kolmogorov-Smirnov statistic*

$$K_n := \|\hat{F}_n - F_0\|_\infty.$$

For any $\alpha \in (0,1)$, we reject $H_0$ when $\sqrt{n}K_n > d_\alpha$, where $d_\alpha$ is defined as $\alpha =: \mathbb{P}(\|B_F\|_\infty > d_\alpha)$.

**Example** (Cramér-von Mises statistic). Consider the *Cramér-von Mises statistic*

$$C_n := \int_{\mathbb{R}} \left(\hat{F}_n(t) - F_0(t)\right)^2 F_0(\mathrm{d}t).$$

For any $\alpha \in (0,1)$, we reject $H_0$ when $nC_n > c_\alpha^2$, where $c_\alpha$ is defined as $\alpha =: \mathbb{P}(\int_0^1 B_F^2(t)\,\mathrm{d}t \geq c_\alpha^2)$.

In particular, Corollary 3.6.1 and Corollary 3.6.2 shows the following.

**Remark.** Under $H_0$, the above two tests will only reject with probability approaching $\alpha$.

On the other hand, when $F \neq F_0$, we will reject with probability approaching 1:

**Proposition 3.6.2.** Suppose $F$ is continuous. Consider testing $H_0 \colon F = F_0$ using the Kolmogorov-Smirnov statistic, then for any $F \neq F_0$, $\mathbb{P}_F(\text{reject}) = \mathbb{P}(\sqrt{n}K_n > d_\alpha) \to 1$ as $n \to \infty$.

**Proof.** For any metric $d$, we have

$$\mathbb{P}_F(\sqrt{n}d(\hat{F}_n, F_0) \geq d_\alpha) \geq \mathbb{P}_F(\sqrt{n}(d(\hat{F}_n, F) - d(F, F_0)) \geq d_\alpha) = \mathbb{P}_F(\sqrt{n}d(\hat{F}_n, F) \geq d_\alpha + \sqrt{n}d(F, F_0)).$$

As $F \neq F_0$, $d(F, F_0) > 0$ is a fixed number, so $\sqrt{n}d(F, F_0) \to \infty$. Hence, from Corollary 3.6.1, $\sqrt{n}\|\hat{F}_n - F\|_\infty = O_p(1)$, with $d(x,y) := \|x - y\|_\infty$, the right-hand side goes to 1. ■

The same result can be obtained when the Cramér-von Mises statistic is used as follows.

**Proposition 3.6.3.** Suppose $F$ is continuous. Consider testing $H_0 \colon F = F_0$ using the Cramér-von Mises statistic, then for any $F \neq F_0$, $\mathbb{P}_F(\text{reject}) = \mathbb{P}(nC_n > c_\alpha^2) \to 1$ as $n \to \infty$.

We omit the proof of Proposition 3.6.3, instead, we focus on a related result regarding the representation of $C_n$, and the proof of Proposition 3.6.3 can be done similarly. First, recall the following.

**As previously seen.** From Corollary 3.6.2, under $H_0 \colon F = F_0$,

$$nC_n = n\int_{\mathbb{R}} (\hat{F}_n - F_0)^2\,\mathrm{d}F_0 \xrightarrow{D} \int_{\mathbb{R}} B_F^2\,\mathrm{d}F_0.$$

**Problem.** What is the distribution in the case of $F \neq F_0$?

**Answer.** Intuitively, consider

$$h(F) := \int_{\mathbb{R}} (F - F_0)^2\,\mathrm{d}F_0.$$

Since $h$ is continuous, $C_n = h(\hat{F}_n) \to h(F)$. ⊛

Formally, for a distributional result, we have the following.

**Proposition 3.6.4.** There is a function $g$ so that $\mathbb{E}[g(X)] = 0$ and

$$\sqrt{n}\left(C_n - \int_{\mathbb{R}} (F - F_0)^2\,\mathrm{d}F_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i) + o_p(1).$$

If we further have $F \neq F_0$, then the above converges to $\mathcal{N}(0, \mathrm{Var}[g(X)])$.

**Proof.** We first note that the left-hand side is just $\sqrt{n}(h(\hat{F}_n) - h(F))$. Now, since

$$\begin{aligned} h(\hat{F}_n) = C_n &= \int_{\mathbb{R}} (\hat{F}_n - F_0)^2 \, \mathrm{d}F_0 \\ &= \int_{\mathbb{R}} (\hat{F}_n - F + F - F_0)^2 \, \mathrm{d}F_0 \\ &= \int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 + \underbrace{\int_{\mathbb{R}} (F - F_0)^2 \, \mathrm{d}F_0}_{h(F)} + 2 \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) \, \mathrm{d}F_0, \end{aligned}$$

we have

$$\sqrt{n} \left( h(\hat{F}_n) - h(F) \right) = \sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 + 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) \, \mathrm{d}F_0.$$

As $n \int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 \xrightarrow{\mathrm{w}} \int_{\mathbb{R}} B_F^2 \, \mathrm{d}F_0$, Proposition 2.4.2 implies

$$\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 = \frac{n}{\sqrt{n}} \int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 = \frac{O_p(1)}{\sqrt{n}} = o_p(1),$$

which gives

$$\sqrt{n} \left( h(\hat{F}_n) - h(F) \right) = 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) \, \mathrm{d}F_0 + o_p(1) =: \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) + o_p(1)$$

where we define the function $g \colon \mathbb{R} \to \mathbb{R}$ as

$$g(x) \coloneqq 2 \int_{\mathbb{R}} (\mathbb{1}_{x \le t} - F(t))(F(t) - F_0(t)) F_0(\mathrm{d}t)$$

since

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) &= \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \int_{\mathbb{R}} (\mathbb{1}_{X_i \le t} - F(t))(F(t) - F_0(t)) F_0(\mathrm{d}t) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} \sum_{i=1}^{n} (\mathbb{1}_{X_i \le t} - F(t))(F(t) - F_0(t)) F_0(\mathrm{d}t) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}(F(t) - F_0(t)) - nF(t)(F(t) - F_0(t)) F_0(\mathrm{d}t) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} n \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}(F(t) - F_0(t)) - F(t)(F(t) - F_0(t)) \right) F_0(\mathrm{d}t) \\ &= 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) \, \mathrm{d}F_0. \end{aligned}$$

To show $\mathbb{E}[g(X)] = 0$, as $F(t), F_0(t), \mathbb{1}_{x \le t}$ are all bounded by 1, Fubini's theorem gives

$$\mathbb{E}[g(X)] = 2 \int_{\mathbb{R}} (\mathbb{P}(X \le t) - F(t))(F(t) - F_0(t)) F_0(\mathrm{d}t) = 0$$

since $\mathbb{P}(X \le t) = F(t)$. Finally, when $F \ne F_0$, $0 < \mathbb{E}[g^2(X)] < \infty$ follows from the same calculation, hence central limit theorem gives the distributional result. ∎

# Chapter 4

# Lindeberg Central Limit Theorem

## Lecture 16: Lindeberg Central Limit Theorem

19 Mar. 9:30

We extend the central limit theorem to the case that $(X_n)$ are only independent but not identically distributed. In particular, consider the case that $S_n = X_1 + \cdots + X_n$'s are a sum of independent, but not necessarily identically distributed, random variables whose distribution may vary with $n$.

The following examples show that in this more general case, assuming finite variance doesn't suffice.

**Example.** If $X_i \sim \text{Pois}(1/i^2)$ for all $i \geq 1$ and are independent to each other, then

$$S_n \sim \text{Pois}\left(\sum_{i=1}^{n} \frac{1}{i^2}\right) \overset{\text{TV}}{\to} \text{Pois}\left(\sum_{i=1}^{\infty} \frac{1}{i^2}\right) = \text{Pois}\left(\frac{\pi^2}{6}\right),$$

which does not go to normal as we expected since $X_i$ are not identically distributed.

On the other hand, something tricker can happen when $X_i$ are "seemingly" identically distributed.

**Example.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(1/n)$ for every $n \geq 1$. But since $S_n \sim \text{Pois}(1)$ for all $n \geq 1$,

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}[S_n]}} \overset{D}{\not\to} \mathcal{N}(0,1).$$

This does not contradict to central limit theorem since $\text{Pois}(1/n)$ depends on $n$.

In general, for any $n \geq 1$, let $k_n \nearrow \infty$ be the number of independent random variables in the sequence $(X_{nk_n}) = (X_{n1}, \ldots, X_{nk_n})$ with $\text{Var}[X_{nj}] < \infty$ for all $1 \leq j \leq k_n$ and $n$. Again, we define $S_n = X_{n1} + \cdots + X_{nk_n}$. In picture, we have something like a *triangular array* of random variables:

$$\begin{aligned}
&n = 1: \quad && X_{11}, \ldots, X_{1k_1};\\
&n = 2: \quad && X_{21}, X_{22}, \ldots, X_{1k_2};\\
&\vdots\\
&n: && X_{n1}, X_{n2}, X_{n3}, \ldots, X_{nk_n}.
\end{aligned}$$

**Example.** As a special case, we previously have $X_{nj} = X_j$ for all $1 \leq j \leq n$, i.e., $k_n = n$.

**Remark.** For different $n$, $(X_{nk_n})$ can be defined on different probability space.

## 4.1 Lindeberg Central Limit Theorem

The goal of this section is to establish the following.

**Theorem 4.1.1** (Lindeberg central limit theorem). For every $n \geq 1$, let $(X_{nk_n})$ be a sequence of independent variables with $k_n \nearrow \infty$ and let $Y_{nj} \coloneqq (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\mathrm{Var}[S_n]}$ for every $1 \leq j \leq k_n$. If the Lindeberg condition holds, then

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}[S_n]}} = \sum_{j=1}^{k_n} \frac{X_{nj} - \mathbb{E}[X_{nj}]}{\sqrt{\mathrm{Var}[S_n]}} =: \sum_{j=1}^{k_n} Y_{nj} \xrightarrow{D} \mathcal{N}(0,1).$$

**Note.** In the above notation, for all $n \geq 1$, $\mathbb{E}[Y_{nj}] = 0$ for all $1 \leq j \leq k_n$ and $\sum_{j=1}^{k_n} \mathrm{Var}[Y_{nj}] = 1$.

### 4.1.1 Lindeberg Condition

We first explain the sufficient condition stated in the Lindeberg central limit theorem, i.e., the Lindeberg condition. Firstly, a weaker but more intuitive notion one might consider is the following.

**Definition 4.1.1** (Uniform asymptotic negligibility). Given a (family of) sequence $(X_{nk_n})$, we say it satisfies the *uniform asymptotic negligibility* (*UAN*), if as $n \to \infty$,

$$\frac{\max_{1 \leq j \leq k_n} \mathrm{Var}[X_{nj}]}{\mathrm{Var}[S_n]} \to 0.$$

However, as we have seen in the second examples, UAN doesn't suffice for the Lindeberg central limit theorem to hold since in this case, $\max_{i \leq j \leq n} \mathrm{Var}[X_{nj}] = 1/n \to 0$, but we know that Lindeberg central limit theorem fail. Hence, we consider the following stronger notion.

**Definition 4.1.2** (Lindeberg condition). Given a (family of) sequence $(X_{nk_n})$, let $Y_{nj} \coloneqq (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\mathrm{Var}[S_n]}$ for every $1 \leq j \leq k_n$ and every $n \geq 1$. Then we say $(X_{nk_n})$ satisfies the *Lindeberg condition* if for every $\epsilon > 0$, as $n \to \infty$,

$$\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] \to 0.$$

Indeed, Lindeberg condition is stronger than uniform asymptotic negligibility.

**Proposition 4.1.1.** The Lindeberg condition implies uniform asymptotic negligibility.

**Proof.** We want to prove that $\max_{1 \leq j \leq k_n} \mathrm{Var}[Y_n] \to 0$ as $n \to \infty$. Firstly, for any $n$ and every $1 \leq j \leq k_n$, by splitting up the expectation, for every $\epsilon > 0$, we have $\mathrm{Var}[Y_{nj}] \leq \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] + \epsilon^2$. Then with the Lindeberg condition, we have

$$\max_{i \leq j \leq k_n} \mathrm{Var}[Y_{nj}] \leq \sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] + \epsilon^2 \to \epsilon^2,$$

i.e., $\limsup_{n \to \infty} \max_{1 \leq j \leq k_n} \mathrm{Var}[Y_{nj}] \leq \epsilon^2$. By letting $\epsilon \to 0$, we complete the proof. ∎

While the UAN is insufficient, it's also not necessary.

**Remark.** Let $(X_n)$ be a sequence of independent Gaussian variables with $\mathrm{Var}[X_i] = 1/i^2$ for all $i \geq 1$. Then indeed, $S_n \sim \mathcal{N}$. However, the uniform asymptotic negligibility does not hold since

$$\frac{\max_{1 \leq i \leq n} \mathrm{Var}[X_i]}{\mathrm{Var}[S_n]} = \frac{1}{\sum_{i=1}^{n} 1/i^2} \to \frac{6}{\pi^2} > 0.$$

### 4.1.2 Proof of Lindeberg Central Limit Theorem

Now, to prove the Lindeberg central limit theorem, we again turn to the characteristic function and use the uniqueness theorem, as what we have done for the usual central limit theorem. Since this turns the problem into calculus, we will need a series of lemmas that provide some bounds.

**Lemma 4.1.1.** For any $w_1, \dots, w_n, z_1, \dots, z_n \in \mathbb{C}$ such that $|w_i|, |z_i| \le 1$ for all $1 \le i \le n$, we have

$$\left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| \le \sum_{i=1}^n |w_i - z_i|.$$

It turns out that we will also need to bound $e^{ix} - (1 + ix)$, i.e., the remainder of the second order Taylor expansion of $e^{ix}$. Let's first see two uniform bounds for this.

**Lemma 4.1.2.** For any $x \in \mathbb{R}$, $|e^{ix} - (1 + ix)| \le x^2/2$ and $|e^{ix} - 1 - ix - (ix)^2/2| \le x^2$.

**Proof.** Recall the specific form of Taylor expansion we used before, which gives

$$e^{ix} = 1 + ix + (ix)^2 \int_0^1 \int_0^1 e^{iuvx} u \, du \, dv = 1 + ix + \frac{(ix)^2}{2} + (ix)^2 \int_0^1 \int_0^1 (e^{iuvx} - 1) u \, du \, dv,$$

which gives both inequalities by bounding the two integrals differently. ■

On the other hand, when $|z|$ is small enough, we have the following tighter bounds.

**Lemma 4.1.3.** For any $z \in \mathbb{C}$ such that $|z| \le \epsilon < 1$, $|e^z - 1 - z - z^2/2| \le |z|^3/(1 - \epsilon)$.

**Proof.** Since

$$\left| e^z - 1 - z - \frac{z^2}{2} \right| \le \left| \sum_{n=3}^\infty z^n \right| \le |z|^3 \sum_{n=0}^\infty |z|^n = |z|^3 \cdot \frac{1}{1 - |z|} \le \frac{|z|}{1 - \epsilon},$$

where the series converges from the fact that $|z| < 1$. ■

**Lemma 4.1.4.** For any $z \in \mathbb{C}$ such that $|z| < \delta/2$ where $\delta \in (0, 1)$, $|e^{iz} - (1 + iz)| \le \delta |z|$.

**Proof.** Since

$$|e^{iz} - 1 - iz| = \left| \sum_{n=2}^\infty \frac{(iz)^n}{n!} \right| \le \sum_{n=2}^\infty |z|^n = |z|^2 \sum_{n=0}^\infty |z|^n = \frac{|z|^2}{1 - |z|} = \frac{|z|}{1 - |z|} |z| < \frac{\delta/2}{2 - \delta} |z| \le \delta |z|,$$

where the series converges from the fact that $|z| < \delta/2 < 1$. ■

Finally, recall the following.

**As previously seen.** From Equation 2.1, for $Z \sim \mathcal{N}(0, 1)$, $\phi_Z(t) = e^{-t^2/2}$.

We can now prove the Lindeberg central limit theorem.

**Proof of Theorem 4.1.1.** Let $\phi_{nj}(t) \coloneqq \mathbb{E}[e^{itX_{nj}}]$ for $t \in \mathbb{R}$. We want to show that

$$\sum_{j=1}^{k_n} Y_{nj} \xrightarrow{D} \mathcal{N}(0, 1) \Leftrightarrow \prod_{j=1}^{k_n} \phi_{nj}(t) \to e^{-t^2/2}$$

for every $t \in \mathbb{R}$ from the uniqueness theorem. Fix $t \in \mathbb{R}$, from triangle inequality, it suffices to show

$$\left| \prod_{j=1}^{k_n} \phi_{nj}(t) - \prod_{j=1}^{k_n} e^{\phi_{nj}(t) - 1} \right| + \left| \prod_{j=1}^{k_n} e^{\phi_{nj}(t) - 1} - e^{-t^2/2} \right| \to 0.$$

Firstly, consider the first term, and recall what we have shown in the homework.

> **As previously seen.** If $\phi$ is a characteristic function, so is $e^{\lambda(\phi-1)}$ for any $\lambda > 0$.

Hence, $e^{\phi_{nj}(t)-1}$ is a characteristic function, so both $\phi_{nj}(t)$ and $e^{\phi_{nj}(t)-1}$ are bounded by 1. This means we can apply Lemma 4.1.1 and get

$$\left|\prod_{j=1}^{k_n}\phi_{nj}(t)-\prod_{j=1}^{k_n}e^{\phi_{nj}(t)-1}\right|\le\sum_{j=1}^{k_n}\left|\phi_{nj}(t)-e^{\phi_{nj}(t)-1}\right|=\sum_{j=1}^{k_n}\left|e^{\phi_{nj}(t)-1}-(\phi_{nj}(t)-1)-1\right|.$$

Let $z_j \coloneqq \phi_{nj}(t)-1$, then the above is just $\sum_{j=1}^{k_n}|e^{z_j}-(z_j+1)|$, suggesting Lemma 4.1.4. Fixing some $\delta\in(0,1)$, we show that $\max_{1\le j\le k_n}|z_j|<\delta/2$ for large enough $n$.

> **Claim.** For any $\delta\in(0,1)$, $\max_{1\le j\le k_n}|\phi_{nj}(t)-1|\le\delta/2$ for $n$ large enough.
>
> **Proof.** As $\mathbb{E}[Y_{nj}]=0$ for all $1\le j\le k_n$, by using Lemma 4.1.2, we have
>
> $$\begin{aligned}\max_{1\le j\le k_n}|\phi_{nj}(t)-1| &= \max_{1\le j\le k_n}\left|\mathbb{E}[e^{itY_{nj}}-1-itY_{nj}]\right| \\ &\le \max_{1\le j\le k_n}\mathbb{E}\left[\left|e^{itY_{nj}}-(1+itY_{nj})\right|\right]\le\frac{t^2}{2}\max_{1\le j\le k_n}\mathbb{E}[Y_{nj}^2]\end{aligned}$$
>
> From the Lindeberg condition, $\max_{1\le j\le k_n}\mathbb{E}[Y_{nj}^2]\to 0$, hence we're done. ⊛

Hence, for any $\delta\in(0,1)$, when $n$ is large enough, Lemma 4.1.4 and the above calculation gives,

$$\left|\prod_{j=1}^{k_n}\phi_{nj}(t)-\prod_{j=1}^{k_n}e^{\phi_{nj}(t)-1}\right|\le\delta\sum_{j=1}^{k_n}|\phi_{nj}(t)-1|\le\delta\cdot\frac{t^2}{2}\sum_{j=1}^{k_n}\mathbb{E}[Y_{nj}^2]=\frac{\delta t^2}{2}$$

since $\sum_{j=1}^{k_n}\mathbb{E}[Y_{nj}^2]=\sum_{j=1}^{k_n}\operatorname{Var}[Y_{nj}]=1$. By letting $n\to\infty$, and $\delta\to 0$, we see that the first term indeed goes to 0 as $n\to\infty$. As for the second term, it suffices to show that for every $t\in\mathbb{R}$,

$$\sum_{j=1}^{k_n}(\phi_{nj}(t)-1)\to-\frac{t^2}{2}\Leftrightarrow\sum_{j=1}^{k_n}(\phi_{nj}(t)-1)+\frac{t^2}{2}\to 0\Leftrightarrow\sum_{j=1}^{k_n}\left[(\phi_{nj}(t)-1)+\frac{t^2}{2}\operatorname{Var}[Y_{nj}]\right]\to 0$$

as $\sum_{j=1}^{k_n}\operatorname{Var}[Y_{nj}]=1$. Since $Y_{nj}$ is centered, we have $\mathbb{E}[Y_{nj}]=0$ and $\operatorname{Var}[Y_{nj}]=\mathbb{E}[Y_{nj}^2]$, hence

$$\sum_{j=1}^{k_n}\left[(\phi_{nj}(t)-1)+\frac{t^2}{2}\operatorname{Var}[Y_{nj}]\right]=\sum_{j=1}^{k_n}\mathbb{E}\left[e^{itY_{nj}}-1-itY_{nj}-\frac{(itY_n)^2}{2}\right].$$

To bound this, we decompose it via the event $|Y_{nj}|>\epsilon$ (for some $\epsilon>0$ to be determined later) as

$$\sum_{j=1}^{k_n}\mathbb{E}\left[\left(e^{itY_{nj}}-1-itY_{nj}-\frac{(itY_n)^2}{2}\right)\mathbb{1}_{|Y_{nj}|>\epsilon}\right]+\sum_{j=1}^{k_n}\mathbb{E}\left[\left(e^{itY_{nj}}-1-itY_{nj}-\frac{(itY_n)^2}{2}\right)\mathbb{1}_{|Y_{nj}|\le\epsilon}\right].$$

We then see that

- from Lemma 4.1.2, with $x\coloneqq tY_{nj}$, we can bound the first term as

  $$\sum_{j=1}^{k_n}\mathbb{E}\left[\left(e^{itY_{nj}}-1-itY_{nj}-\frac{(itY_{nj})^2}{2}\right)\mathbb{1}_{|Y_{nj}|>\epsilon}\right]\le t^2\sum_{j=1}^{k_n}\mathbb{E}\left[Y_{nj}^2\mathbb{1}_{|Y_{nj}|>\epsilon}\right]\to 0$$

  as $n\to\infty$ by the Lindeberg condition;

- from Lemma 4.1.3, with $z \coloneqq itY_{nj}$ such that $|z| = |tY_{nj}| \le |t|\epsilon$ under the event. Let $\epsilon$ be defined such that $|t|\epsilon < 1$, then we can bound the second term by

$$\frac{1}{1-\epsilon}\sum_{j=1}^{k_n}\mathbb{E}[|tY_{nj}|^3 \cdot \mathbb{1}_{|Y_{nj}|\le\epsilon}] \le \frac{|t|^3}{1-\epsilon}\epsilon\sum_{j=1}^{k_n}\mathbb{E}[|Y_{nj}|^2 \cdot \mathbb{1}_{|Y_{nj}|\le\epsilon}] \le \frac{|t|^3}{1-\epsilon}\epsilon\sum_{j=1}^{k_n}\mathbb{E}[|Y_{nj}|^2] = \frac{|t|^3}{1-\epsilon}\epsilon$$

since again $\sum_{j=1}^{k_n}\mathbb{E}[Y_{nj}^2] = \sum_{j=1}^{k_n}\operatorname{Var}[Y_{nj}] = 1$. By letting $\epsilon \to 0$, $|t|^3\epsilon/(1-\epsilon) \to 0$ as desired.

Hence, we see that both terms go to 0 when $n \to \infty$, so the second term indeed go to 0 ■

### 4.1.3 Sufficiency of Lindeberg Condition

To apply Lindeberg central limit theorem, checking the Lindeberg condition might not be the most efficient way. We now study several sufficient conditions for the Lindeberg condition to hold.

**Corollary 4.1.1** (Hájek-Sidak central limit theorem). If $X_{ni} = c_{ni}Z_i$[a] for all $1 \le i \le k_n$ where $(Z_{k_n})$ are i.i.d. with $\mathbb{E}[Z_i] = \mu$ and $\operatorname{Var}[Z_i] = \sigma^2$. If $\max_{1\le i\le k_n} c_{ni}^2 / \sum_{i=1}^{k_n} c_{ni}^2 \to 0$ as $n \to \infty$, then the Lindeberg condition holds.

[a] Note that in this notation, we implicitly assume that when $n$ varies, only $c_{ni}$ varies, but not $Z_i$.

# Lecture 17: Rank Test and Two-Sample Problem

21 Mar. 9:30

**Proof.** Firstly, we see that $\operatorname{Var}[S_n] = \sum_{j=1}^{k_n} c_{nj}^2\sigma^2$, hence with our usual notation, we define

$$Y_{nj} \coloneqq \frac{c_{nj}(Z_j - \mu)}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}\sigma} =: \frac{c_{nj}}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}}W_j,$$

where $W_j \coloneqq (Z_j - \mu)/\sigma$. Then, for any $\epsilon > 0$, we can check that Lindeberg condition as

$$\sum_{j=1}^{k_n}\mathbb{E}[Y_{nj}^2\mathbb{1}_{|Y_{nj}|>\epsilon}] = \sum_{j=1}^{k_n}\frac{c_{nj}^2}{\sum_{i=1}^{k_n} c_{ni}^2}\mathbb{E}\left[W_j^2\mathbb{1}_{\frac{|c_{nj}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}}|W_j|>\epsilon}\right]$$

since the only dependence of $j$ in the expectation is $|c_{nj}|$ in the indicator,

$$\le \sum_{j=1}^{k_n}\frac{c_{nj}^2}{\sum_{i=1}^{k_n} c_{ni}^2}\mathbb{E}\left[W_j^2\mathbb{1}_{\max_{1\le k\le k_n}\frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}}|W_k|>\epsilon}\right]$$

which makes the term in the expectation i.i.d., so we can replace both $W_j$ and $W_k$ by $W \coloneqq W_1$,

$$= \sum_{j=1}^{k_n}\frac{c_{nj}^2}{\sum_{i=1}^{k_n} c_{ni}^2}\mathbb{E}\left[W^2\mathbb{1}_{\max_{1\le k\le k_n}\frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}}|W|>\epsilon}\right]$$

$$= \mathbb{E}\left[W^2\mathbb{1}_{\max_{1\le k\le k_n}\frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}}|W|>\epsilon}\right].$$

Hence, it reduces to show $\mathbb{E}[W^2\mathbb{1}_{|W|>x}] \to 0$ as $n \to \infty$ for $x \coloneqq \epsilon\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}/\max_{1\le k\le k_n}|c_{nk}|$. From our assumption, for any $\epsilon > 0$, $x \to \infty$ as $n \to \infty$, hence the expectation indeed goes to 0 as long as $W$ has finite second moment, which is indeed the case by our assumption. ■

We see that Hájek-Sidak central limit theorem is very common in practice.

**Intuition.** Often time for every $n$, $c_{ni} = c_n$, the same for every $1 \le i \le k_n$. In this case, we may write $X_{ni} = c_n \cdot X_{ni}/c_n =: c_nZ_i$ such that $Z_i \coloneqq X_{ni}/c_n$ is i.i.d. distributed, ready for checking the Hájek-Sidak condition.

Let's see three examples.

**Example** (Uniform distribution). For every $n \geq 1$, let $(X_{nk_n}) \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-c_n, c_n)$ for some $c_n > 0$. Then we see that $Z_i \coloneqq X_{ni}/c_n \sim \mathcal{U}(-1, 1)$ is now i.i.d. distributed.

**Example** (Rademacher distribution). For every $n \geq 1$, let $(X_{nk_n})$ be i.i.d. such that $\mathbb{P}(X_{ni}/c_n = \pm 1) = 1/2$. Then $Z_i \coloneqq X_{ni}/c_n$ is now i.i.d. distributed.

**Example** (Exponential distribution). For every $n \geq 1$, let $(X_{nk_n}) \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1/c_n)$ for some $c_n > 0$, hence $Z_i \coloneqq X_{ni}/c_n \sim \mathrm{Exp}(1)$ is now i.i.d. distributed.

On the other hand, if we insist the same setup as in the Lindeberg central limit theorem, it suffices to check a slightly higher moment rather than the truncated one used in the Lindeberg condition.

**Corollary 4.1.2** (Lyapunov's central limit theorem). Consider the setup as in the Lindeberg central limit theorem. If $\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^{2+\delta}] \to 0$ for some $\delta > 0$, then the Lindeberg condition holds.

**Proof.** Fix some $\delta > 0$ such that the assumption holds. Then for any $\epsilon > 0$, we have

$$\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^2 \mathbb{1}_{|Y_{nj}|>\epsilon}] \leq \sum_{j=1}^{k_n} \mathbb{E}\left[\left(\frac{|Y_{nj}|}{\epsilon}\right)^{\delta} |Y_{nj}|^2 \mathbb{1}_{|Y_{nj}|>\epsilon}\right] \leq \sum_{j=1}^{k_n} \mathbb{E}\left[\left(\frac{|Y_{nj}|}{\epsilon}\right)^{\delta} |Y_{nj}|^2\right] \to 0$$

by taking $\epsilon^{-\delta}$ out and then the result follows from the assumption. ■

For bounded random variable, we have a simpler form.

**Corollary 4.1.3.** Let $|X_{ni}| \leq C_n$ for all $1 \leq i \leq k_n$. Then if $C_n/\sqrt{\mathrm{Var}[S_n]} \to 0$, the Lindeberg condition holds. In particular, when $C_n =: C$, it suffices to check $\mathrm{Var}[S_n] \to \infty$.

**Proof.** We see that for every $1 \leq j \leq k_n$,

$$|Y_{nj}| = \frac{|X_{nj} - \mathbb{E}[X_{nj}]|}{\sqrt{\mathrm{Var}[S_n]}} \leq \frac{2C_n}{\sqrt{\mathrm{Var}[S_n]}}.$$

In this case, for any $\delta > 0$, recall that $\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = 1$, hence

$$\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^{2+\delta}] \leq \left(\frac{2C_n}{\sqrt{\mathrm{Var}[S_n]}}\right)^{\delta} \sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = \left(\frac{2C_n}{\sqrt{\mathrm{Var}[S_n]}}\right)^{\delta} \to 0,$$

hence the Lyapunov's condition holds. ■

**Example** (Bernoulli distribution). For every $n \geq 1$, let $X_{ni} \sim \mathrm{Ber}(p_{ni})$ for all $1 \leq i \leq k_n$. Since $X_{ni} \leq 1$, from Corollary 4.1.3, it suffices to check

$$\mathrm{Var}[S_n] = \sum_{i=1}^{k_n} p_{ni}(1 - p_{ni}) \to \infty.$$

- If $p_{ni} = 1/i$: then $\mathrm{Var}[S_n] = \sum_{i=1}^{k_n} 1/i - \sum_{i=1}^{k_n} 1/i^2 \to \infty$.
- If $p_{ni} = p_n$: then $\mathrm{Var}[S_n] = k_n p_n (1 - p_n)$. In particular, for $p_n = 1/n$ and $k_n = n$,

$$\mathrm{Var}[S_n] = n \cdot \frac{1}{n} \cdot \frac{n-1}{n} \to 1 \neq \infty.$$

In general, if $np_n \to \lambda > 0$, then the sum $S_n$ converges to $\mathrm{Pois}(\lambda)$.

### 4.1.4 Testing the I.I.D. Assumption

With this new tool in hand, i.e., the Lindeberg central limit theorem, let's see one illustrative application, i.e., *testing the i.i.d. assumption*.

**Problem** (Testing the i.i.d. assumption)**.** Consider collecting a sequence of data $X_1, \dots, X_n \sim F$ where $F$ is continuous. We want to test whether $X_i$'s are i.i.d. from $F$.

**Answer.** To do this, consider that for all $i \geq 1$, define the *rank* to be[a]

$$R_i = \sum_{j=1}^{i} \mathbb{1}_{X_j < X_i}.$$

In particular, if $R_i = i$ ($R_i = 1$), then $X_i$ is the largest (smallest) of $X_1, \dots, X_i$. ⊛

[a] We don't need to worry about the equality in the indicator since $F$ is continuous.

To see how $R_i$'s help us to decide whether $X_i$'s are i.i.d., under this hypothesis, we have the following.

**Theorem 4.1.2.** Let $(X_n) \overset{\text{i.i.d.}}{\sim} F$ where $F$ is continuous. Then $(R_n)$ are independent and $\mathbb{P}(R_i = r) = 1/i$ for all $1 \leq r \leq i$, i.e., $R_i \sim \mathcal{U}(\{1, \dots, i\})$. Moreover,

$$\frac{6}{\sqrt{n^3}} \left( \sum_{i=1}^{n} R_i - \frac{n(n+1)}{4} \right) \xrightarrow{D} \mathcal{N}(0, 1).$$

**Proof.** Since $X_i$'s are i.i.d. and $F$ it continues (hence no ties),

$$\mathbb{P}(X_{\sigma(1)} < X_{\sigma(2)} < \cdots < X_{\sigma(i)}) = \frac{1}{i!}$$

for any permutation $\sigma$ of $\{1, \dots, i\}$. This implies $\mathbb{P}(R_1 = r_1, \dots, R_i = r_i) = 1/i!$, hence

$$\mathbb{P}(R_i = r) = \sum_{r_1, \dots, r_{i-1}} \mathbb{P}(R_1 = r_1, \dots, R_{i-1} = r_{i-1}, R_i = r) = (i-1)! \cdot \frac{1}{i!} = \frac{1}{i}.$$

This proves the first part. Now, observe that $\mathbb{E}\left[\sum_{i=1}^n R_i\right] = \sum_{i=1}^n \frac{i}{2} = n(n+1)/4$ and

$$\operatorname{Var}\left[\sum_{i=1}^{n} R_i\right] = \sum_{i=1}^{n} \frac{i^2 - 1}{12} = \frac{n(2n^2 + 3n - 5)}{72} \sim \frac{n^3}{36},$$

hence if the Lindebert central limit theorem holds, then we're done. We check Corollary 4.1.3: by noting that $|R_i| \leq n$, we indeed have $n / \sqrt{\operatorname{Var}[\sum_{i=1}^n R_i]} \to 0$. ■

**Remark** (Record)**.** We may instead consider the *record* $Z_i = \mathbb{1}_{R_i = i}$.[a] Since $R_i \leq i$, $\mathbb{P}(Z_i = 1) = 1/i$, i.e., $Z_i \sim \text{Ber}(1/i)$. Then the previous example gives asymptotic normality of $\sum_{i=1}^n Z_i$.

[a] Intuitive, it indicates when $X_i$ is the largest among $X_1, \dots, X_{i-1}$.

In either case ($R_i$ or $Z_i$), the above gives us some hints about how to deal with this kind of problems.

**Intuition.** Find some statistics whose distribution is *independent* of the underlying $F$ under $H_0$.

## 4.2 Testing for Symmetry

Our next goal is to apply the intuition we get from the i.i.d. problem, and apply it to the problem of testing for symmetry.

**Problem 4.2.1** (Testing for symmetry). Let $X, X_1, \dots X_n \overset{\text{i.i.d.}}{\sim} F$ where $F$ is a continuous distribution function. The problem of *testing symmetry* is to test the null hypothesis, $H_0$, that $F$ is symmetric about 0, i.e., whether $X \overset{D}{=} -X$.

### 4.2.1 Motivation: Two-Sample Problem with no I.I.D. Assumption

Symmetry testing is important since it justifies various tests we use in practice for the classical two-sample problem, even if the i.i.d. assumption is violated.

**Problem 4.2.2** (Two-sample problem). Given the data of two random samples obtained from a different given population. The *two-sample problem* considers the task of determining whether the difference between these two populations is statistically significant.

One classical example of a two-sample problem is the *treatment effect testing*.

**Example** (Treatment effect). Consider a treatment group and a control group (the two populations), and we observe $Y_i^{\mathrm{T}}$'s and $Y_i^{\mathrm{C}}$'s from the treatment and the control group, respectively. The goal is to test whether the treatment takes affect. It's common to pair the data together and get $(Y_1^{\mathrm{T}}, Y_1^{\mathrm{C}}), \dots, (Y_n^{\mathrm{T}}, Y_n^{\mathrm{C}})$, and let $X_i := Y_i^{\mathrm{T}} - Y_i^{\mathrm{C}}$ for all $1 \le i \le n$.

Usually, we assume $X_1, \dots, X_n$ are i.i.d. with mean $\mu$ and finite variance, and the null hypothesis we want to test is $H_0 \colon \mu = 0$. As $X_i$'s are i.i.d., we may use the $t$-test to the treatment effect problem, namely, we reject $H_0$ if the $t$-statistic is too large, i.e.,

$$T_n = \sqrt{n}\frac{\overline{X}_n}{\hat{\sigma}_n} = \frac{\sum_{i=1}^n X_i}{\sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2}} = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}} > Z_\alpha$$

for some $\alpha > 0$. Since $\mathrm{Var}[X_i] < \infty$, $T_n \overset{D}{\to} \mathcal{N}(0,1)$ under $H_0$.

**Problem.** In reality, $X_i$'s are never i.i.d., but people still use the $t$-test. Why?

**Answer.** This is because we usually give the treatment in a "randomized" way. In addition, we can "condition" on the observed data $|X_i|$'s. In this case, we can design a hypothesis testing such that it only depends on statistics that are now i.i.d., and by applying the central limit theorem, we will get something like $\sum_{i=1}^n X_i / \sqrt{\sum_{i=1}^n X_i^2}$, pretty much like what we have in the $t$-test. ⊛

## Lecture 18: Wilcoxon Signed-Rank Test

26 Mar. 9:30

Before we explain the above answer, let's point out other potential problems: firstly, since even for i.i.d. $X_i$'s, the results for $T_n$ is asymptotic, and we might wonder what's the rate of convergence.

**Problem.** How fast does $T_n \overset{D}{\to} \mathcal{N}(0,1)$ under $H_0$?

**Answer.** It depends heavily on $F$, so it's unlikely to get a universal answer. ⊛

On the other hand, what if the underlying distribution has heavy tails?

**Problem.** What if $\mathrm{Var}[X_i]$, or even $\mathbb{E}[X_i]$, doesn't exist?

**Answer.** This is something we will solve along the way when we deal with the i.i.d. problem. ⊛

Now, we focus on the non-i.i.d. problem. It turns out that if each unit in each pair is equally likely to be in the treatment and the control group, independently of what happens in the other pairs, then $H_0 \colon \mu = 0$ can be expressed as for all $1 \le i \le n$, $\mathbb{P}(X_i = \pm|X_i| \mid |X_j| = |x_j|, 1 \le j \le n) = 1/2$, i.e.,

$$H_0 \colon \mathbb{P}(\mathrm{sgn}(X_i) = \pm 1 \mid |X_j| = |x_j|, 1 \le j \le n) = \frac{1}{2} \text{ for all } 1 \le i \le n.$$

Hence, we can test this "conditionally on the observed absolute values of $X_1, \dots, X_n$," i.e., given $|X_j| = |x_j|$ for $1 \le j \le n$. This is exactly the problem of testing for symmetry.

**Intuition.** We see that now under $H_0$, $\operatorname{sgn}(X_i)$'s are i.i.d., overcoming the problem of non-i.i.d.

**Note.** In practice, when doing inference, we usually implicitly condition on $|X_j|$'s too: by doing inference on the selected "design points", we're essentially condition on those.

**Example** ($t$-statistic for two-sample problem). Consider writing $\sum_{i=1}^n X_i = \sum_{i=1}^n |X_i| \operatorname{sgn}(X_i) = \sum_{i=1}^n |x_i| \operatorname{sgn}(X_i)$. Then by treating $|x_i|$'s as constants, under $H_0$,

$$T_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n x_i^2}} \sqrt{\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}} = \frac{\sum_{i=1}^n |X_i| \operatorname{sgn}(X_i)}{\sqrt{\sum_{i=1}^n x_i^2}} \sqrt{\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}} \xrightarrow{D} \mathcal{N}(0,1) \tag{4.1}$$

by the Lindeberg central limit theorem as long as $\max_{1 \le i \le n} x_i^2 / \sqrt{\sum_{j=1}^n x_j^2} \to 0$. We can then use the left-hand side for out test as usual.

Let's use this example as our running example for the problem of testing for symmetry.

### 4.2.2 Student's $t$-Test, Sign Test, and Wilcoxon Signed-Rank Test

Returning to the general problem of testing for symmetry, let $X, X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$.

**As previously seen.** Now $H_0$ is defined to be whether $F$ is symmetric about 0, or $X \overset{D}{=} -X$.

As suggested by the running example, one standard way to solve this is still considering the $t$-statistic,

$$T_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2}},$$

and reject $H_0$ when $T_n$ is larger than some critical value.

**As previously seen.** If $X$ has positive and finite variance, $\mathbb{P}_{H_0}(T_n > Z_\alpha) \to \alpha$ as $n \to \infty$. However, as we have seen, the rate of such an approximation depends on the unknown $F$.

An alternative and even more classical way to solve this, which works even if $X$ does not have any moments, is to consider the sign statistic.

**Definition 4.2.1** (Sign statistic). Given a sample $X_1, \dots, X_n$, the *sign statistic* is defined as

$$\operatorname{sign}_n := \sum_{i=1}^n \operatorname{sgn}(X_i) = 2 \sum_{i=1}^n \mathbb{1}_{X_i > 0} - n.$$

Observe that under $H_0$, the distribution of $\operatorname{sign}_n$ is independent of $F$ since $\sum_{i=1}^n \mathbb{1}_{X_i > 0} \overset{H_0}{\sim} \operatorname{Bin}(n, 1/2)$, i.e., we can conduct the hypothesis test non-asymptotically. On the other hand, under $H_0$, even if we consider its asymptotic behavior, i.e., from the usual central limit theorem,

$$\frac{1}{\sqrt{n}} \operatorname{sign}_n = \frac{\sum_{i=1}^n \operatorname{sgn}(X_i)}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1),$$

the quality of the approximation doesn't depend on $F$, and we reject $H_0$ when $\sum_{i=1}^n \operatorname{sgn}(x_i)/\sqrt{n} \ge Z_\alpha$.

**Example** (Sign-statistic for two-sample problem). One can also motivate the sign statistic from the use of $t$-statistic $T_n$ in our running example, i.e., by simply ignoring $|X_i|$'s used in $T_n$ to avoid conditioned on them. Hence, we reject $H_0$ if $\sum_{i=1}^n \operatorname{sgn}(X_i) = \operatorname{sign}_n$ is large.

The sign test with the sign statistic is simple, but it's questionable that how powerful it is as we're not using $|X_i|$'s at all. In view of the above intuition, let's utilize $|X_i|$'s in other ways.

**Intuition.** We can utilize $|X_i|$'s by replacing $|X_i|$'s in Equation 4.1 by their "rank" information.

In particular, consider the *rank* $R_i$ for $1 \le i \le n$

$$R_i \coloneqq \sum_{j=1}^{n} \mathbb{1}_{|X_j| \le |X_i|}.$$

It's clear that $R_i$'s are dependent to each other.

**As previously seen.** This is different from the previous definition $R_i = \sum_{j=1}^{i} \mathbb{1}_{X_j < X_i}$.

Following the intuition, consider the so-called Wilcoxon signed-rank statistic.

**Definition 4.2.2** (Wilcoxon signed-rank statistic)**.** Given a sample $X_1, \dots, X_n$, the *Wilcoxon signed-rank statistic* is defined as

$$W_n = \sum_{i=1}^{n} R_i \operatorname{sgn}(X_i).$$

Then, the corresponding Wilcoxon signed-rank test is to reject $H_0$ if $W_n$ is "large" as usual.

**Note.** Under $H_0$, $W_n$ is again independent of $F$.

However, it turns out that it's quite hard to get the exact critical value for a general problem of testing for symmetry.

**Example** (Wilcoxon signed-rank statistic for two-sample probelm)**.** If we condition on $|X_j|$'s as in the running example, without loss of generality, $R_i = i$ for every $1 \le i \le n$.[a] In this case, $W_n = \sum_{i=1}^{n} i \operatorname{sgn}(X_i)$, hence under $H_0$,

$$\frac{\sum_{i=1}^{n} i \operatorname{sgn}(X_i)}{\sqrt{\sum_{i=1}^{n} i^2}} \xrightarrow{D} \mathcal{N}(0,1) \Rightarrow \frac{W_n}{\sqrt{n^3}} \xrightarrow{D} \mathcal{N}(0, 1/3)$$

by the Lindeberg central limit theorem. We can then use this to test $H_0$.

[a] This is doable since $|X_j|$'s are treated as constants.

Without conditioning on $|X_j|$'s, the above doesn't hold. However, some sign tests are still possible.

**Intuition** (Another sign test)**.** Consider $H_0 \colon \operatorname{sgn}(X_i) = \epsilon_i$ where $\epsilon_i$ is a Rademacher random variable, i.e., $\mathbb{P}(X > 0) = \mathbb{P}(X < 0) = 1/2$ if $X$ is symmetric. This doesn't need to assume $\mathbb{E}[X_i] < \infty$.

If we further assume $\operatorname{Var}[X_i] < \infty$, then it's possible to say something about $W_n$ without conditioning on $|X_j|$'s. Let's first see a useful characterization of $W_n$ to get started.

**Proposition 4.2.1.** Let $h(x_1, x_2) \coloneqq \mathbb{1}_{x_1 + x_2 \ge 0}$. Then we have

$$W_n = \sum_{i=1}^{n} \operatorname{sgn}(X_i) + \sum_{i \ne j} \left( h(X_i, X_j) - \frac{1}{2} \right).$$

**Proof.** Consider $W_n =: W_n^+ - W_n^-$ where

$$W_n^+ = \sum_{i=1}^{n} R_i \mathbb{1}_{X_i > 0}, \text{ and } W_n^- = \sum_{i=1}^{n} R_i \mathbb{1}_{X_i < 0}.$$

Some calculation gives the following.

**Claim.** $W_n^+ - \sum_{i=1}^n \mathbb{1}_{X_i>0} = \frac{1}{2}\sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}$ and $W_n^- - \sum_{i=1}^n \mathbb{1}_{X_i<0} = \frac{1}{2}\sum_{i\neq j} \mathbb{1}_{X_i+X_j<0}$.

**Proof.** Let's first focus on $W_n^+$. We see that

$$W_n^+ = \sum_{i=1}^n \left( \sum_{j=1}^n \mathbb{1}_{|X_j|\leq |X_i|} \right) \mathbb{1}_{X_i>0} = \sum_{i=1}^n \mathbb{1}_{X_i>0} + \sum_{\substack{1\leq i,j\leq n \\ i\neq j}} \mathbb{1}_{|X_j|\leq |X_i|, X_i>0},$$

Let's abbreviate the argument $1 \leq i, j \leq n \colon i \neq j$ in the double summation by $i \neq j$, we have

$$\begin{aligned} W_n^+ - \sum_{i=1}^n \mathbb{1}_{X_i>0} &= \sum_{i\neq j} \mathbb{1}_{|X_j|\leq X_i} \\ &= \sum_{i\neq j} \mathbb{1}_{-X_i\leq X_j\leq X_i} \\ &= \underbrace{\sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}\mathbb{1}_{X_j\leq X_i}}_{S_1} = \sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0} - \underbrace{\sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}\mathbb{1}_{X_j>X_i}}_{S_2}, \end{aligned}$$

where the last equality can be justified by the fact that $1 = \mathbb{1}_{X_j\leq X_i} + \mathbb{1}_{X_j>X_i}$. Observe that

$$S_1 = \sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}\mathbb{1}_{X_j\leq X_i} = \sum_{i\neq j} \mathbb{1}_{X_j+X_i\geq 0}\mathbb{1}_{X_i\leq X_j} = \sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}\mathbb{1}_{X_j>X_i} = S_2,$$

since $F$ is continuous. Hence, by letting $S \coloneqq S_1 = S_2$, the above calculation gives

$$W_n^+ - \sum_{i=1}^n \mathbb{1}_{X_i>0} = S = \sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0} - S \Rightarrow S = W_n^+ - \sum_{i=1}^n \mathbb{1}_{X_i>0} = \frac{1}{2}\sum_{i\neq j} \mathbb{1}_{X_i+X_j\geq 0}.$$

The results for $W_n^-$ follows similarly. ⊛

By subtracting the above, since $\mathbb{1}_{X_i>0} - \mathbb{1}_{X_i<0} = \operatorname{sgn}(X_i)$, we have

$$W_n - \sum_{i=1}^n \operatorname{sgn}(X_i) = \frac{1}{2}\sum_{i\neq j} \left( \mathbb{1}_{X_i+X_j\geq 0} - \mathbb{1}_{X_i+X_j<0} \right) = \frac{1}{2}\sum_{i\neq j} \left( 2\mathbb{1}_{X_i+X_j\geq 0} - 1 \right).$$

Plugging the definition of $h$ yields the result. ■

Hence, from Proposition 4.2.1, we should first study $\sum_{i\neq j} h(X_i, X_j)$. First, consider testing

$$H_0 \colon X, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F \text{ where } F \text{ is continuous such that } X \overset{D}{=} -X,$$

i.e., $\mathbb{P}(X \geq x) = \mathbb{P}(X \leq -x)$ for all $x \in \mathbb{R}$. Note that we're not assuming $\mathbb{E}[X_i]$ to exists.

**Note.** Since $X_i$'s are i.i.d. under $H_0$, $h(X_i, X_j)$'s are identically distributed under $H_0$ for $i \neq j$.

However, it's clear that $h(X_i, X_j)$'s are not independent. Anyway, recall what we're trying to study.

**As previously seen.** From Proposition 4.2.1, under $H_0$ (in particular, $F$ being continuous),

$$W_n - \sum_{i=1}^n \operatorname{sgn}(X_i) = \sum_{i\neq j} \left( h(X_i, X_j) - \frac{1}{2} \right)$$

This is a sum of identically distributed random variables subtracting something.

**Intuition.** If $1/2$ is the expectation of $h(X_i, X_j)$ for $i \neq j$, we're getting closer to the familiar form.

Indeed, under $H_0$, $\mathbb{E}[h(X_i, X_j)] = 1/2$ for $i \neq j$ as

$$\begin{aligned}\mathbb{E}_{H_0}[h(X_i, X_j)] = \mathbb{E}_{H_0}[h(X_1, X_2)] &= \mathbb{P}(X_1 + X_2 \geq 0) \\ &= \mathbb{P}(X_1 \geq -X_2) \\ &= \int_{\mathbb{R}} \mathbb{P}(X_1 \geq -x) F(\mathrm{d}x) = \int_{\mathbb{R}} \mathbb{P}(X \leq x) F(\mathrm{d}x) = \mathbb{E}[F(X)] = \frac{1}{2}\end{aligned}$$

since $F$ is continuous, and $F(X) \sim \mathcal{U}(0,1)$ as we have shown in the homework. Hence, under $H_0$,

$$\sum_{i \neq j} \left( h(X_i, X_j) - \frac{1}{2} \right) = 2 \sum_{i<j} \left( h(X_i, X_j) - \frac{1}{2} \right) = 2 \sum_{i<j} \left( h(X_i, X_j) - \mathbb{E}[h(X_i, X_j)] \right).$$

By rescaling by the number of terms in the summation, we're looking at

$$\frac{1}{\binom{n}{2}} \left( W_n - \sum_{i=1}^n \operatorname{sgn}(X_i) \right) = \frac{2}{\binom{n}{2}} \sum_{i<j} \left( h(X_i, X_j) - \mathbb{E}[h(X_i, X_j)] \right) =: 2 \left( U_n - \frac{1}{2} \right)$$

where we define $U_n$ for some permutation symmetric function $h$[1] as

$$U_n := \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j).$$

This is an $U$-statistic, where $U$ stands for unbiased. We will formally define it later, and we will show that by multiplying $\sqrt{n}$ on the both sides, this converges to a standard normal, i.e.,

$$\frac{\sqrt{n}}{2\binom{n}{2}} \left( W_n - \sum_{i=1}^n \operatorname{sgn}(X_i) \right) = \sqrt{n} \left( U_n - \frac{1}{2} \right) \xrightarrow{D} \mathcal{N}(0, 1/3). \tag{4.2}$$

**Note.** This will imply an asymptotic distribution for $W_n$ exactly alone since under $H_0$,

$$\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^n \operatorname{sgn}(X_i) = \frac{1}{n-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \operatorname{sgn}(X_i) \right) \xrightarrow{p} 0.$$

## 4.3 $U$-Statistics

In this section, we develop the asymptotic theory for $U$-statistics. While postponing defining $U$-statistics, we first outline what we're going to do. The main tool is the following.

**Proposition 4.3.1.** Given two sequences $(S_n)$, $(\widetilde{S}_n)$ of random variables in $L^2$, if they satisfy $\operatorname{Var}[S_n]/\operatorname{Var}[\widetilde{S}_n] \to 1$ and $\operatorname{Corr}(S_n, \widetilde{S}_n) \to 1$, then

$$\frac{\operatorname{Var}[S_n - \widetilde{S}_n]}{\operatorname{Var}[S_n]} \to 0 \Leftrightarrow \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{Var}[S_n]}} - \frac{\widetilde{S}_n - \mathbb{E}[\widetilde{S}_n]}{\sqrt{\operatorname{Var}[S_n]}} \xrightarrow{L^2} 0.$$

**Proof.** Since for $n \in \mathbb{N}$, $\operatorname{Var}[S_n - \widetilde{S}_n] = \operatorname{Var}[S_n] - \operatorname{Var}[\widetilde{S}_n] - 2\sqrt{\operatorname{Var}[S_n] \operatorname{Var}[\widetilde{S}_n]} \operatorname{Corr}(S_n, \widetilde{S}_n)$, i.e.,

$$\frac{\operatorname{Var}[S_n - \widetilde{S}_n]}{\operatorname{Var}[S_n]} = 1 + \frac{\operatorname{Var}[\widetilde{S}_n]}{\operatorname{Var}[S_n]} - 2\sqrt{\frac{\operatorname{Var}[\widetilde{S}_n]}{\operatorname{Var}[S_n]}} \operatorname{Corr}(S_n, \widetilde{S}_n).$$

The right-hand side goes to 0 under the assumptions, proving the result. ∎

Hence, to show $(S_n - \mathbb{E}[S_n])/\sqrt{\operatorname{Var}[S_n]} \xrightarrow{D} Y$, we may find $(\widetilde{S}_n)$ such that $(\widetilde{S}_n - \mathbb{E}[\widetilde{S}_n])/\sqrt{\operatorname{Var}[S_n]} \xrightarrow{D} Y$ and apply Proposition 4.3.1 with Theorem 2.2.3. In particular, after defining what's an $U$-statistic, we will then apply the above strategy to which.

[1] Indeed, since $\mathbb{E}[h(X_1, X_2)] = \mathbb{P}(X_1 + X_2 \geq 0)$, so $h(X_1, X_2) = h(X_2, X_1)$.

# Lecture 19: Projection and $U$-Statistic

## 4.3.1 Projection

28 Mar. 9:30

A natural approach to find such $(\widetilde{S}_n)$ is to "project" $(S_n)$ onto some space, which potentially has nice properties for us to do the analysis. Consider the space $L^2 = \{Y : \|Y\| = \sqrt{\mathbb{E}[Y^2]} < \infty\}$, and let $K \subseteq L^2$ with $Y \in L^2$. The goal is to approximate $Y$ by a sequence in $K$.

**Definition 4.3.1** (Projection). Given a subspace $K \subseteq L^2$, $Y^* \in K$ is a *projection* of $Y \in L^2$ onto $K$ if $\|Y^* - Y\| \le \|Z - Y\|$ for all $Z \in K$.

The following characterization of the projection is useful in our analysis.

**Proposition 4.3.2.** Suppose $K \subseteq L^2$ is a linear subspace. If $Y^* \in K$ and $\mathbb{E}[Y^*Z] = \mathbb{E}[YZ]$ for every $Z \in K$, then $Y^*$ is a projection of $Y$ onto $K$. If in addition, $1 \in K$, then $\mathbb{E}[YY^*] = \mathbb{E}[(Y^*)^2]$ and $\operatorname{Cov}[Y, Y^*] = \operatorname{Var}[Y^*]$. In particular, $\operatorname{Corr}(Y, Y^*) = \sqrt{\operatorname{Var}[Y^*]/\operatorname{Var}[Y]}$.

**Proof.** We see that for all $Z \in K$,

$$\|Y - Z\|^2 = \|Y - Y^*\|^2 + \|Y^* - Z\|^2 + 2\mathbb{E}[(Y - Y^*)(Y^* - Z)]$$

from the assumption, for every $W \in K$, $\mathbb{E}[(Y^* - Y)W] = 0$, with $W := Y^* - Z \in K$,

$$= \|Y - Y^*\|^2 + \|Y^* - Z\|^2,$$

which is greater than $\|Y - Y^*\|^2$. Now, if $1 \in K$, by taking $Z = 1$, $\mathbb{E}[Y^*] = \mathbb{E}[Y]$. On the other hand, for $Z = Y^*$, we have $\mathbb{E}[YY^*] = \mathbb{E}[(Y^*)^2]$. Hence,

$$\operatorname{Cov}[Y, Y^*] = \mathbb{E}[YY^*] - \mathbb{E}[Y]\mathbb{E}[Y^*] = \mathbb{E}[(Y^*)^2] - \mathbb{E}[Y^*]^2 = \operatorname{Var}[Y^*].$$

The correlation is then $\operatorname{Corr}(Y, Y^*) = \operatorname{Var}[Y^*]/\sqrt{\operatorname{Var}[Y]\operatorname{Var}[Y^*]} = \sqrt{\operatorname{Var}[Y^*]/\operatorname{Var}[Y]}$. ■

If we combine the above, our plan is clear now. In particular, we have the following.

**Corollary 4.3.1.** For each $n \in \mathbb{N}$, let $\widetilde{S}_n$ be a projection of $S_n$ onto a linear subspace $K_n$ of $L^2$ with $1 \in K_n$. If $\operatorname{Var}[S_n]/\operatorname{Var}[\widetilde{S}_n] \to 1$, then

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{Var}[S_n]}} - \frac{\widetilde{S}_n - \mathbb{E}[\widetilde{S}_n]}{\sqrt{\operatorname{Var}[S_n]}} \xrightarrow{L^2} 0.$$

**Proof.** Since $\operatorname{Var}[S_n]/\operatorname{Var}[\widetilde{S}_n] \to 1$, from Proposition 4.3.2, $\operatorname{Corr}(Y, Y^*) \to 1$ is automatically satisfied. The result then follows from Proposition 4.3.1. ■

With Corollary 4.3.1, the goal is to find suitable $K_n$'s such that $\operatorname{Var}[S_n]/\operatorname{Var}[\widetilde{S}_n] \to 1$, and $\widetilde{S}_n - \mathbb{E}[\widetilde{S}_n]$ converges somewhere. However, we should first understand how to find projections in practice.

**Example.** Let $X$ be a random vector and $K = \{g(X) \in L^2\}$. Then $Y^* = \mathbb{E}[Y \mid X]$ is a projection of $Y \in L^2$ onto $K$. More generally, let $X_1, \dots, X_n$ be a sequence of random vectors, and let

$$K_n = \{g(X_1, \dots, X_n) \in L^2\}.$$

Then $Y^* = \mathbb{E}[Y \mid X_1, \dots, X_n]$ is a projection of $Y \in L^2$ onto $K_n$.

**Proof.** For all $g \in K$, from the basic properties for conditional expectation, we have

$$\mathbb{E}[\mathbb{E}[Y \mid X] \cdot g(X)] = \mathbb{E}[\mathbb{E}[g(X) \cdot Y \mid X]] = \mathbb{E}[g(X) \cdot Y].$$

Then from Proposition 4.3.2, the result follows. ⊛

A more elaborated example is the following, which turns out to be flexible enough for our purpose.

**Example.** Let $X_1, \dots, X_n$ be a sequence of independent random vectors, and let

$$K_n = \left\{ \sum_{i=1}^n g_i(X_i) \colon g_i(X_i) \in L^2 \text{ for all } 1 \le i \le n \right\}.$$

In this case, $Y^* - \mathbb{E}[Y] = \sum_{i=1}^n \mathbb{E}[Y - \mathbb{E}[Y] \mid X_i]$, where $Y^*$ is a projection of $Y \in L^2$ onto $K_n$.

**Proof.** We want to show that for all $\sum_{i=1}^n g_i \in K_n$, $\mathbb{E}[Y^* \sum_{i=1}^n g_i(X_i)] = \mathbb{E}[Y \sum_{i=1}^n g_i(X_i)]$, i.e.,

$$\mathbb{E}\left[ (Y^* - \mathbb{E}[Y^*]) \sum_{i=1}^n g_i(X_i) \right] = \mathbb{E}\left[ (Y - \mathbb{E}[Y]) \sum_{i=1}^n g_i(X_i) \right]$$

since $\mathbb{E}[Y^*] = \mathbb{E}[Y]$. Expanding the left-hand side, and using $\mathbb{E}[Y^*] = \mathbb{E}[Y]$ again,

$$\begin{aligned}
\mathbb{E}\left[ (Y^* - \mathbb{E}[Y^*]) \sum_{i=1}^n g_i(X_i) \right] &= \mathbb{E}\left[ \left( \sum_{i=1}^n \mathbb{E}[Y - \mathbb{E}[Y] \mid X_i] \right) \sum_{j=1}^n g_j(X_j) \right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[ \mathbb{E}[(Y - \mathbb{E}[Y]) \mid X_i] g_j(X_j) \right]
\end{aligned}$$

observe that for $i \neq j$, $\mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y]) \mid X_i]] = \mathbb{E}[Y - \mathbb{E}[Y]]$ is independent of $g_j(X_j)$, hence

$$\begin{aligned}
&= \sum_{i \neq j} \mathbb{E}[(Y - \mathbb{E}[Y])] \mathbb{E}[g_j(X_j)] + \sum_{i=1}^n \mathbb{E}\left[ \mathbb{E}[Y - \mathbb{E}[Y] \mid X_i] \cdot g_i(X_i) \right] \\
&= \sum_{i=1}^n \mathbb{E}\left[ \mathbb{E}[(Y - \mathbb{E}[Y]) \cdot g_i(X_i) \mid X_i] \right] \\
&= \mathbb{E}\left[ \sum_{i=1}^n (Y - \mathbb{E}[Y]) g_i(X_i) \right],
\end{aligned}$$

where we again use the property of conditional expectation. ⊛

We see that the projection in the above example is a sum of independent random variables, so we can apply the Lindeberg central limit theorem to establish asymptotic normality.

**Remark.** In particular, with the help of Corollary 4.3.1, we only need to make sure that $K_n$ is a linear subspace that contains constants and $\operatorname{Var}[S_n] / \operatorname{Var}[\widetilde{S}_n] \to 1$ as $n \to \infty$.

**Remark.** Since we're using Lindeberg central limit theorem, it's possible to generalize the above to the case of triangular array of random variables $X_{n1}, \dots, X_{nn}$ with what we're going to do next.

### 4.3.2 Asymptotic Distribution of $U$-Statistic

Now, let's start developing a theory for the $U$-statistics. First, consider the following.

**Definition 4.3.2** ($U$-statistic). Given a sequence of i.i.d. random vectors $(X_n)$ and a permutation symmetric function $h \colon \mathbb{R}^m \to \mathbb{R}$[a] with $n \ge m$, the *$U$-statistic* is defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{\substack{i_1, \dots, i_m \\ \{i_1, \dots, i_m\} \subseteq [n]}} h(X_{i_1}, \dots, X_{i_m}).$$

[a] We refer to $h$ as $U_n$'s *kernel*.

**Notation.** Let $[n] \coloneqq \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$.

**Remark.** The $U$-statistic is an **u**nbiased estimator of $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$, and in particular, an average of dependent, but identically distributed random variables.

**Note.** Since $h$ is permutation symmetric, the order of indices used in the argument doesn't matter.

**Example** (Wilcoxon signed-rank statistic). When studying the Wilcoxon signed-rank statistic $W_n$, we see that $\left(W_n - \sum_{i=1}^n \operatorname{sgn}(X_i)\right) / \binom{n}{2} = 2U_n - 1$ where $U_n$ is an $U$-statistic defined as[a]

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j).$$

[a] Note that $i < j$ is the same as $\{i, j\} \subseteq [n]$. Indeed, one can use $i_1 < \cdots < i_m$ without loss of generality.

From Corollary 4.3.1, given the projection $U_n^*$ of $U_n$ onto $K_n = \{\sum_{i=1}^n g_i(X_i)\}$, we want to show

$$\left| \frac{U_n - \mathbb{E}[U_n]}{\sqrt{\operatorname{Var}[U_n]}} - \frac{U_n^* - \mathbb{E}[U_n^*]}{\sqrt{\operatorname{Var}[U_n]}} \right| \xrightarrow{p} 0,$$

and establish the asymptotic normality of $U_n$. For this, assuming that $U_n \in L^2$ for each $n \geq m$, i.e., $\operatorname{Var}[h(X_1, \dots, X_m)] < \infty$. Let's first find $U_n^*$.

**Proposition 4.3.3.** For every $n \geq m$, let $K_n = \{\sum_{i=1}^n g_i(X_i) \colon g_i(X_i) \in L^2\}$. Furthermore, let $h^*(x) \coloneqq \mathbb{E}[h(x, X_2, \dots, X_m)]$, then the projection of $U_n$ onto $K_n$ is given by

$$U_n^* = \mathbb{E}[U_n] + \frac{m}{n} \sum_{k=1}^n (h^*(X_k) - \mathbb{E}[U_n]).$$

**Proof.** Denote $\theta \coloneqq \mathbb{E}[U_n] = \mathbb{E}[U_n^*]$ (recall Proposition 4.3.2), then from the last example,

$$\begin{aligned}
U_n^* - \theta &= \sum_{i=1}^n \mathbb{E}[U_n - \theta \mid X_i] \\
&= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_1, \dots, i_m\} \subseteq [n]} \mathbb{E}\left[h(X_{i_1}, \dots X_{i_m}) - \theta \mid X_k\right] \\
&= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_1, \dots, i_m\} \subseteq [n]} \left(\mathbb{E}\left[h(X_{i_1}, \dots X_{i_m}) \mid X_k\right] - \theta\right).
\end{aligned}$$

When $k \neq i_1, \dots, i_m$, the conditional expectation becomes an unconditional one, which is $\theta$, i.e., the only terms survive is when $i_j = k$ for some $1 \leq j \leq m$. Hence,

$$= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_2, \dots, i_m\} \subseteq [n] \setminus \{k\}} \left(\mathbb{E}\left[h(X_k, X_{i_2}, \dots X_{i_m}) \mid X_k\right] - \theta\right)$$

as $X_i$'s are i.i.d. and $h$ is permutation symmetry, with $h^*$, we have

$$\begin{aligned}
&= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_2, \dots, i_m\} \subseteq [n] \setminus \{k\}} (h^*(X_k) - \theta) \\
&= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \binom{n-1}{m-1} (h^*(X_k) - \theta) \\
&= \frac{m}{n} \sum_{k=1}^n (h^*(X_k) - \theta).
\end{aligned}$$

Rearranging the terms gives the result. ∎

As $h^*(X_k)$'s are i.i.d., we can apply the central limit theorem if $\operatorname{Var}[h^*(X)] \in (0, \infty)$ and get

$$\sqrt{n}(U_n^* - \theta) \xrightarrow{D} \mathcal{N}(0, m^2 \operatorname{Var}[h^*(X)]).$$

Hence, we just need to show that $\operatorname{Var}[U_n]/\operatorname{Var}[U_n^*] \to 1$. Firstly, since

$$\operatorname{Var}[U_n^*] = \frac{m^2}{n}\operatorname{Var}[h^*(X_i)],$$

it reduces to show $\operatorname{Var}[U_n] \sim m^2 \operatorname{Var}[h^*(X_i)]/n$. We have the following.

**Proposition 4.3.4.** For every $n \geq 2m-1$, we have

$$\binom{n}{m}\operatorname{Var}[U_n] = \sum_{r=1}^{m}\binom{m}{r}\binom{n-m}{m-r}\xi_{r,m},$$

where for every $r \in [m]$, let $\{i_{r+1},\dots,i_m\} \cap \{i'_{r+1},\dots,i'_m\}$ be disjoint subsets of $[n]$ and set

$$\xi_{r,m} \coloneqq \operatorname{Cov}[h(X_1,\dots,X_r,X_{i_{r+1}},\dots,X_{i_m}), h(X_1,\dots,X_r,X_{i'_{r+1}},\dots,X_{i'_m})].$$

**Proof.** By the definition of the variance, we have

$$\begin{aligned}\binom{n}{m}^2\operatorname{Var}[U_n] = &\sum_{\{i_1,\dots,i_m\}\subseteq[n]}\operatorname{Var}[h(X_{i_1},\dots,X_{i_m})]\\ &+\sum_{\substack{\{i_1,\dots,i_m\},\{i'_1,\dots,i'_m\}\subseteq[n]\\ \{i_1,\dots,i_m\}\neq\{i'_1,\dots,i'_m\}}}\operatorname{Cov}[h(X_{i_1},\dots,X_{i_m}), h(X_{i'_1},\dots,X_{i'_m})].\end{aligned}$$

The variance terms are clear since for every subset, they're all the same, hence

$$\sum_{\{i_1,\dots,i_m\}\subseteq[n]}\operatorname{Var}[h(X_{i_1},\dots,X_{i_m})] = \binom{n}{m}\operatorname{Var}[h(X_1,\dots,X_m)] = \binom{n}{m}\xi_{m,m}.$$

For the covariance terms, if $\{i_1,\dots,i_m\}\cap\{i'_1,\dots,i'_m\} = \varnothing$, the covariance vanishes from independence. Hence, consider iterate through subsets of $i$'s and $i'$'s with $r \geq 1$ common indices.

> **Intuition.** Let $1,\dots,r$ be the common indices between $\{i_1,\dots,i_m\}$ and $\{i'_1,\dots,i'_m\}$, i.e., $[r] = \{i_1,\dots,i_m\}\cap\{i'_1,\dots,i'_m\}$. In this case, the covariance term becomes
>
> $$\xi_{r,m} = \operatorname{Cov}[h(X_1,\dots,X_r,X_{i_{r+1}},\dots,X_{i_m}), h(X_1,\dots,X_r,X_{i'_{r+1}},\dots,X_{i'_m})]$$
>
> where $\{i_{r+1},\dots,i_m\}\cap\{i'_{r+1},\dots,i'_m\} = \varnothing$. Such a covariance is fixed for every $r$, regardless of other indices $i_j, i'_j$ for $r+1 \leq j \leq m$ since $X_{i_{r+1}},\dots,X_{i_m}$ and $X_{i'_{r+1}},\dots,X_{i'_m}$ are i.i.d.

With this, the second summation for covariance becomes

$$\sum_{\substack{\{i_1,\dots,i_m\},\{i'_1,\dots,i'_m\}\subseteq[n]\\ \{i_1,\dots,i_m\}\neq\{i'_1,\dots,i'_m\}}}\operatorname{Cov}[h(X_{i_1},\dots,X_{i_m}), h(X_{i'_1},\dots,X_{i'_m})] = \sum_{r=1}^{m-1}\binom{n}{m}\binom{m}{r}\binom{n-m}{m-r}\xi_{r,m},$$

where the summation is from $r=1$ to $m-1$ since we require $\{i_1,\dots,i_m\} \neq \{i'_1,\dots,i'_m\}$, i.e., $r \neq m$. We note that the counting is calculated as:

1. choose $m$ indices for the first $h$ arbitrarily from $n$ indices;
2. choose $r$ indices to be the common indices between the first and second $h$ from the $m$ indices chosen above;
3. choose $m-r$ indices for the second $h$ from indices different from those are chosen before. In total, $n-m$ of them.

Combining the sum of variances and covariances, dividing both sides by $\binom{n}{m}$ gives the result. ■

# Lecture 20: Comparing Different Tests for Symmetry

We can finally establish the asymptotic normality of $U_n$. 2 Apr. 9:30

**Theorem 4.3.1.** If for every $r \in [m]$, $\xi_{r,m}$ is bounded away from 0 and infinity as $n \to \infty$, then

$$\sqrt{n}(U_n - \mathbb{E}[U_n]) = \frac{m}{\sqrt{n}} \sum_{k=1}^{n} (h^*(X_k) - \mathbb{E}[U_n]) + o_p(1).$$

Furthermore, if $\operatorname{Var}[h^*(X)] > 0$, then $\sqrt{n}(U_n - \mathbb{E}[U_n]) \xrightarrow{D} \mathcal{N}(0, m^2 \operatorname{Var}[h^*(X)])$.

**Proof.** Recall that by Proposition 4.3.3 and Corollary 4.3.1, it suffices to show that $\operatorname{Var}[U_n] \sim \operatorname{Var}[\widetilde{U}_n] = m^2/n \cdot \operatorname{Var}[h^*(X_i)]$ for all $n \in \mathbb{N}$. By Proposition 4.3.4, and the Stirling's approximation,[a]

$$\operatorname{Var}[U_n] = \frac{1}{\binom{n}{m}} \sum_{r=1}^{m} \binom{m}{r}\binom{n-m}{m-r} \xi_{r,m} \sim \frac{\binom{m}{1}\binom{n-m}{m-1}}{\binom{n}{m}} \xi_{1,m} \sim \frac{m^2}{n} \xi_{1,m}$$

with some algebra (and our assumptions). It remains to show that $\operatorname{Var}[h^*(X)] = \xi_{1,m}$. Indeed,

$$\begin{aligned}
\xi_{1,m} &= \operatorname{Cov}[(X_1, X_2, \ldots, X_m), h(X_1, X_{m+1}, \ldots, X_{2m-1})] \\
&= \mathbb{E}[(h(X_1, X_2, \ldots, X_m) - \theta) \cdot (h(X_1, X_{m+1}, \ldots, X_{2m-1}) - \theta)] \\
&= \int \mathbb{E}[h(x, X_2, \ldots, X_m) - \theta] \cdot \mathbb{E}[h(x, X_{m+1}, \ldots, X_{2m-1})] F(\mathrm{d}x) \\
&= \int (h^*(x) - \theta)(h^*(x) - \theta) F(\mathrm{d}x) \\
&= \mathbb{E}[(h^*(X) - \theta)^2] \\
&= \operatorname{Var}[h^*(X)],
\end{aligned}$$

which is exactly what we want to show. ■

[a] Recall that as $n \to \infty$, for any fixed $m$, $\binom{n}{m} \sim n^m/m!$ by the Sterling's approximation.

Let's conclude the discussion by looking at our initial motivation.

**Example** (Wilcoxon signed-rank statistic). For $d = 1$, consider $h(x_1, x_2) = \mathbb{1}_{x_1+x_2>0}$. Then, $\theta = \mathbb{E}[h(X_1, X_2)] = \mathbb{P}(X_1 + X_2 > 0)$ and $h^*(x) = \mathbb{P}(x + X > 0) = 1 - F(-x)$. If $X$ is not trivial, then $\operatorname{Var}[h^*(X)] = \operatorname{Var}[F(-X)] > 0$. For example, under $H_0$: $X \overset{D}{=} -X$ and $X$ is continuous, $\operatorname{Var}[h^*(X)] = \operatorname{Var}[F(X)] = 1/12$ as $F(X) \sim \mathcal{U}(0,1)$ since $F$ is assumed to be continuous. Moreover,

$$\theta = \mathbb{P}(X_1 > -X_2) = \mathbb{E}[h^*(X)] = \mathbb{E}[1 - F(-X)] = 1 - \frac{1}{2} = \frac{1}{2}.$$

Therefore, $\sqrt{n}(U_n - 1/2) \xrightarrow{D} \mathcal{N}(0, 2^2/12) = \mathcal{N}(0, 1/3)$ from Theorem 4.3.1. Finally, recall the representation that the Wilcoxon signed-rank statistic $W_n$ admits under $H_0$, i.e.,

$$\frac{\sqrt{n}}{2 \cdot \binom{n}{2}} \left( W_n - \sum_{i=1}^{n} \operatorname{sgn}(X_i) \right) = \frac{\sqrt{n}}{n(n-1)} \left( W_n - \sum_{i=1}^{n} \operatorname{sgn}(X_i) \right) = \sqrt{n} \left( U_n - \frac{1}{2} \right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{3}\right),$$

where $U$ is a $U$-statistic with $h(x_1, x_2) = \mathbb{1}_{x_1+x_2 \geq 0}$, hence the above calculation[a] applies, giving the asymptotic normality. This further implies that under $H_0$, $W_n/n^{3/2} \xrightarrow{D} \mathcal{N}(0, 1/3)$.[b]

[a] Ignore the difference between $\mathbb{1}_{x_1+x_2 \geq 0}$ for $W_n$ and $\mathbb{1}_{x_1+x_2>0}$ in the above example as now $X$ is continuous.
[b] Since from the previous note, the sum of $\operatorname{sgn}(X_i)$'s will vanish.

**Note.** This is the same guarantee when we condition on $|X_j|$'s as seen in the previous example.

We make one last remark before we move on to the next topic.

**Remark.** $U_n$ is a function of $X_{(1)}, \dots, X_{(n)}$, i.e., the function of the order statistics, which is an UMVUE since in this non-parametric formulation,[a] the order statistic is complete and sufficient.

[a] Since we didn't assume anything about $F$ excepts for the continuity.

## 4.4 Asymptotic Relative Efficiency of Tests

For the testing for symmetry problem, we have three test statistics, i.e., $t$-statistic $T_n = \sqrt{n}\overline{X}_n/\hat{\sigma}_n$, the sign statistic $\text{sign}_n = \sum_{i=1}^n \text{sgn}(X_i)$, and the Wilcoxon signed-rank statistic $W_n = \sum_{i=1}^n R_i \,\text{sgn}(X_i)$.

**Problem.** How does $T_n$, $\text{sign}_n$, and $W_n$ compared in terms of their efficiency? Can we say something similar for the estimator's efficiency we have developed, i.e., the asymptotic relative efficiency?

To answer the above question, we'll look into their *powers*. Consider an alternative hypothesis testing, where we have $H_0 \colon \theta = \theta_0$ and $H_1 \colon \theta > \theta_0$.

**Example.** One example of $\theta$ is $\theta = \mathbb{P}(X_1 + X_2 > 0)$.

Let $T_n$ now denote a generic statistic (not necessarily the $t$-statistic) such that there exists an increasing $\mu(\theta)$ such that when the true parameter is $\theta$,

$$\frac{T_n - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1).$$

Then, we reject $H_0$ whenever $T_n$ is large.

**Example.** One can write all the $t$-statistic "$T_n$", the sign test statistic $\text{sign}_n$, and also the Wilcoxon signed-rank statistic $W_n$ in this form.

Specifically, under this setup, by treating the asymptotic normality as exact, we reject $H_0$ if

$$T_n \geq \mu(\theta_0) + Z_\alpha \frac{\sigma(\theta_0)}{\sqrt{n}},$$

which gives that $\mathbb{P}_{\theta_0}(\text{reject}) \to \alpha$ as we usually get.

**Note.** It's easy to see that we can always control the type-I error.

Since we can always control $\alpha$, the interesting question might be whether we can control the type-II error $\beta$. We see that under $H_1$,

$$\mathbb{P}_\theta(\text{reject}) = \mathbb{P}_\theta\left(\frac{T_n - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} \geq \frac{\mu(\theta_0) - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta)}\right) \overset{\theta \geq \theta_0}{\to} 1$$

as $n \to \infty$ since $\mu(\theta_0) - \mu(\theta) < 0$.

**Remark.** The power always approaches 1, not very interesting.

Hence, we might turn our focus to some non-asymptotic results. One way to look at this problem is by fixing the type-I and type-II error, and see how many samples we need to achieve them.

### 4.4.1 A Heuristic Approach

Given some fixed $\alpha$, $\beta$, and $\theta_0$, suppose we want $\mathbb{P}_{\theta^*}(\text{reject}) = 1 - \beta$ for some $\theta^*$ with some $n$. From the above calculation with asymptotic normality we assume for $T_n$, we have

$$1 - \beta = \mathbb{P}_{\theta^*}\left(\frac{T_n - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} \geq \frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)}\right) \to 1 - \Phi\left(\frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)}\right),$$

where the convergent is not rigorous since the right-hand side still depend on $n$. Anyway, this leads to[2]

$$Z_{1-\beta} = -Z_\beta = \frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)} \Rightarrow \sqrt{n}\frac{\mu(\theta^*) - \mu(\theta_0)}{\sigma(\theta^*)} = Z_\beta + \frac{\sigma(\theta_0)}{\sigma(\theta^*)} Z_\alpha.$$

Let $\theta_i$ be the $\theta^*$ for which the above holds when $n = i$, and denote the corresponding sequence as $(\theta_n)$.

**Note.** Obviously we then have $\theta_n \to \theta_0$ for $(\theta_n)$.

In this case, by replacing $\theta^*$ by $\theta_n$ for every $n \in \mathbb{N}$, we have

$$\sqrt{n}\frac{\mu(\theta_n) - \mu(\theta_0)}{\sigma(\theta_n)} = Z_\beta + \frac{\sigma(\theta_0)}{\sigma(\theta_n)} Z_\alpha.$$

If $\sigma$ is continuous at $\theta_0$, then as $\sigma(\theta_n) \to \sigma(\theta_0)$, the right-hand side becomes $Z_\alpha + Z_\beta$. If we further assume that $\mu$ is differentiable at $\theta_0$, with $\sqrt{n}(\mu(\theta_n) - \mu(\theta_0)) = \mu'(\theta_0)\sqrt{n}(\theta_n - \theta_0) + \sqrt{n} \cdot o(\theta_n - \theta_0)$,

$$\sqrt{n}(\theta_n - \theta_0) \to \frac{Z_\alpha + Z_\beta}{\mu'(\theta_0)/\sigma(\theta_0)}.$$

Let $n^*$ be the $n$ such that $\theta_{n^*} = \theta_n = \theta^*$. Then, we have

$$\sqrt{n^*} \to \frac{Z_\alpha + Z_\beta}{\frac{\mu'(\theta_0)}{\sigma(\theta_0)}(\theta_{n^*} - \theta_0)}.$$

**Note.** $n^*$ only depends on $\mu'(\theta_0)/\sigma(\theta_0)$.

**Proof.** Since $Z_\alpha + Z_\beta$ is fixed while $\theta_{n^*} - \theta_0$ is assumed to be independent of the statistic also since we treat their asymptotic normality as exact, so $\theta_{n^*}$ will be the same across different statistics. ⊛

**Definition 4.4.1** (Slope). For any statistics $T_n$ with $\mu$ and $\sigma$, its *slope* is defined as $\mu'(\theta_0)/\sigma(\theta_0)$.

We see that if the analysis can be made formal, then we can compare two statistics $T_n$ and $\widetilde{T}_n$ in terms of their required sample sizes to achieve $\alpha$ and $\beta$ for a fixed $\theta_0$.

**Remark.** This analysis relies on the fact that when $\sqrt{n}(\theta_n - \theta_0)$ converges, then

$$\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1).$$

# Lecture 21: Slope of a Statistics and Pitman (Local) Alternatives

## 4.4.2 Deriving the Slope

4 Apr. 9:30

Formally, let $\xi \geq 0$ such that $\sqrt{n}(\theta_n - \theta_0) \to \xi$, and suppose there exists $\mu(\theta)$ and $\sigma(\theta)$ such that

$$\sqrt{n}\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)} \xrightarrow{D} \mathcal{N}(0, 1) \Leftrightarrow \mathbb{P}_{\theta_n}\left(\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} \leq x\right) \to \Phi(x)$$

for all $x \in \mathbb{R}$. Firstly, when $\xi = 0$, then we know that

$$\sqrt{n}\frac{T_n - \mu(\theta_0)}{\sigma(\theta_0)} \xrightarrow{D} \mathcal{N}(0, 1).$$

Hence, we reject $H_0$ if $T_n > \mu(\theta_0) + \sigma(\theta_0)Z_\alpha/\sqrt{n}$. We see that this happens with probability

$$\mathbb{P}_{\theta_n}(\text{reject}) = \mathbb{P}_{\theta_n}\left(T_n > \mu(\theta_0) + \frac{\sigma(\theta_0)}{\sqrt{n}} Z_\alpha\right) = \mathbb{P}_{\theta_n}\left(\frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} > \frac{\mu(\theta_0) - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta_n)}\right).$$

[2] Recall that $Z_\alpha$ is defined for the right-tail.

If $\mu$ is differentiable at $\theta_0$ and $\sigma$ is continuous at $\theta_0$, then as $\sqrt{n}(\theta_n - \theta_0) \to \xi$, the above converges to

$$\Phi\left(-\left(-\frac{\mu'(\theta_0)}{\sigma(\theta_0)}\xi + Z_\alpha\right)\right) = \Phi\left(\frac{\mu'(\theta_0)}{\sigma(\theta_0)}\xi - Z_\alpha\right).$$

Let $\theta^*$ to be defined as $\mathbb{P}_{\theta^*}(\text{reject}) = 1 - \beta$ for some $\beta > 0$. Then, denote $n^*$ such that $\theta_{n^*} = \theta^*$, and define $\xi > 0$ such that $\sqrt{n^*}(\theta^* - \theta_0) = \xi$, i.e., $\theta^* = \theta_0 + \xi/\sqrt{n^*}$. Then $\mathbb{P}_{\theta^*}(\text{reject})$ will converge to

$$\Phi\left(\frac{\mu'(\theta_0)}{\sigma(\theta_0)}\xi - Z_\alpha\right) = 1 - \beta \Rightarrow \frac{\mu'(\theta_0)}{\sigma(\theta_0)}\xi - Z_\alpha = Z_\beta \Rightarrow \sqrt{n^*}(\theta^* - \theta_0) = \xi = \frac{Z_\alpha + Z_\beta}{\mu'(\theta_0)/\sigma(\theta_0)},$$

solving w.r.t. $\sqrt{n^*}$ gives

$$\sqrt{n^*} = \frac{Z_\alpha + Z_\beta}{\frac{\mu'(\theta_0)}{\sigma(\theta_0)}(\theta^* - \theta_0)},$$

which confirms our heuristic argument.

**Remark.** $n^*$ still only depends on $\mu'(\theta_0)/\sigma(\theta_0)$ from the same reason.

### 4.4.3 Asymptotic Relative Efficiency for Statistics

If we have another statistic $\widetilde{T}$ associates with $\widetilde{\mu}$, $\widetilde{\sigma}$, and $\widetilde{n}^*$, such that it also satisfies all the assumptions, i.e., asymptotic normality, differentiability for $\widetilde{\mu}$, and continuity for $\widetilde{\sigma}$, then from the same analysis, we can then compare how many samples we need to reach $\alpha$, $\beta$, given $\theta_0$.

**Definition 4.4.2** (Asymptotic relative efficiency for statistic)**.** Given $\theta_0$, $\alpha$, and $\beta$, the *asymptotic relative efficiency* between two statistics $T_n$ and $\widetilde{T}_n$ is defined as

$$\text{ARE}(T, \widetilde{T}) = \frac{n^*}{\widetilde{n}^*} = \left(\frac{\widetilde{\mu}'(\theta_0)/\widetilde{\sigma}(\theta_0)}{\mu'(\theta_0)/\sigma(\theta_0)}\right)^2.$$

**Note.** Same as Definition 3.4.1, Definition 4.4.2 is different from the convention, where we usually define the *asymptotic relative efficiency of* $T$ *w.r.t.* $\widetilde{T}$ as $\text{ARE}_\theta(T, \widetilde{T}) = \widetilde{n}^*/n^*$.

Let compare the $t$-test, sign test, and Wilcoxon signed-rank test on the problem of testing symmetry. In particular, consider the following variation of the problem.

**Problem.** Let $\epsilon, \epsilon_1, \dots, \epsilon_n \overset{\text{i.i.d.}}{\sim} F$ where $F$ is continuous and symmetric around 0, i.e., $\epsilon \overset{D}{=} -\epsilon$. Furthermore, assuming $X_{n1}, \dots, X_{nn} \overset{\text{i.i.d.}}{\sim} \theta_n + \epsilon_i$ such that $\sqrt{n}(\theta_n - \theta_0) = \xi$ for some fixed $\xi \geq 0$. We're interested in testing whether $H_0 \colon X_{n1} \overset{D}{=} -X_{n1}$. In other words, $H_0 \colon \theta_0 = 0$.

Let's try the simplest sign test and compute its slope.

**Example** (Sign test)**.** The slope of the averaged sign statistic $\overline{\text{sign}}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_{ni}>0}$ is $2f(0)$.

**Proof.** We first see that

$$\overline{\text{sign}}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_{ni}>0} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\epsilon_i > -\theta_n}.$$

The mean of $\mathbb{1}_{\epsilon_i > -\theta_n}$ is

$$\mathbb{P}(\epsilon > -\theta_n) = \mathbb{P}(\epsilon \leq \theta_n) = F(\theta_n)$$

since $\epsilon$ is symmetric and continuous. Hence, the Lindeberg central limit theorem gives

$$\frac{\sum_{i=1}^n (\mathbb{1}_{\epsilon_i > -\theta_n} - F(\theta_n))}{\sqrt{n}\sqrt{F(\theta_n)(1 - F(\theta_n))}} = \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\epsilon_i > -\theta_n} - F(\theta_n)}{\sqrt{F(\theta_n)(1 - F(\theta_n))}} = \sqrt{n}\frac{\overline{\text{sign}}_n - F(\theta_n)}{\sqrt{F(\theta_n)(1 - F(\theta_n))}} \xrightarrow{D} \mathcal{N}(0, 1)$$

by checking the Lyapunov condition, indeed, since we have

$$\operatorname{Var}\left[\sum_{i=1}^{n} \mathbb{1}_{\epsilon_i > -\theta_n}\right] = nF(\theta_n)(1-F(\theta_n)) \to nF(\theta_0)(1-F(\theta_0)) \to \infty.$$

We see that

- $\mu(\theta) = F(\theta)$, hence if $F$ is differentiable at 0 with $F'(0) =: f(0)$, then $\mu'(0) = f(0)$;
- $\sigma(\theta) = \sqrt{F(\theta)(1-F(\theta))}$, so $\sigma(0) = 1/2$ since $F(0) = 1/2$ by the symmetry of $F$.

We conclude that the slope of $\overline{\operatorname{sign}}_n$ is $\mu'(0)/\sigma(0) = 2f(0)$. ⊛

**Note.** For the sign test, we don't need any moment assumption. Additionally, it's expected to be a weak test since it seems only care about the density around 0, which is intuitive.

Now, let's try the $t$-test. This time, we will need the second moment to exist.

**Example** ($t$-test). Suppose $\operatorname{Var}[\epsilon] = \sigma^2 < \infty$, then the slope of the "normalized" $t$-statistic $\widetilde{T}_n := T_n/\sqrt{n}$ is $1/\sigma$.

**Proof.** Since $\widetilde{T}_n := T_n/\sqrt{n} = \overline{X}_n/\hat{\sigma}_n$, with $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n (\epsilon_i - \overline{\epsilon}_n)^2$, we have

$$\widetilde{T}_n = \frac{\overline{X}_n}{\hat{\sigma}_n} = \frac{\theta_n + \overline{\epsilon}_n}{\hat{\sigma}_n} = \left(\frac{\theta_n}{\hat{\sigma}_n} - \frac{\theta_n}{\sigma}\right) + \left(\frac{\overline{\epsilon}_n}{\hat{\sigma}_n} + \frac{\theta_n}{\sigma}\right) \Rightarrow \sqrt{n}\left(\widetilde{T}_n - \frac{\theta_n}{\sigma}\right) = \sqrt{n}\theta_n\left(\frac{1}{\hat{\sigma}_n} - \frac{1}{\sigma}\right) + \sqrt{n}\frac{\overline{\epsilon}_n}{\hat{\sigma}_n}.$$

Since $\theta_0 = 0$ and $\sqrt{n}(\theta_n - \theta_0) = \xi$, we have $\sqrt{n}\theta_n \to \xi \geq 0$, $1/\hat{\sigma}_n - 1/\sigma \xrightarrow{p} 0$, so the first term goes to 0. On the other hand, by the usual central limit theorem, $\sqrt{n}\overline{\epsilon}_n \xrightarrow{D} \mathcal{N}(0, \sigma^2)$, hence

$$\sqrt{n}\left(\widetilde{T}_n - \frac{\theta_n}{\sigma}\right) = \sqrt{n}\frac{\widetilde{T}_n - \frac{\theta_n}{\sigma}}{1} \xrightarrow{D} \mathcal{N}(0,1).$$

We see that

- $\mu(\theta) = \theta/\sigma$, which is clearly differentiable with $\mu'(\theta) = 1/\sigma$;
- $\sigma(\theta) = 1$, so $\sigma(0) = 1$.

We conclude that the slope of $\widetilde{T}_n$ is $\mu'(0)/\sigma(0) = 1/\sigma$. ⊛

This gives us a way to compare $\overline{\operatorname{sign}}_n$ and $\widetilde{T}_n$.

**Proposition 4.4.1.** Consider the problem of testing for symmetry, the asymptotic relative efficiency between the sign statistic and the $t$-statistic is

$$\operatorname{ARE}(\widetilde{T}_n, \overline{\operatorname{sign}}_n) = (2f(0)\sigma)^2.$$

**Remark.** From Proposition 3.5.2, $\operatorname{ARE}(\overline{X}_n, \hat{\theta}_{1/2}) = (2f(0)\sigma)^2$, exactly the same!

Let's see some actual example, where we can borrow from the previous calculation.[3]

**Example** (Gaussian). If $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$, then $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$, so $f(0) = 1/\sigma\sqrt{2\pi}$, hence

$$\operatorname{ARE}(\widetilde{T}_n, \overline{\operatorname{sign}}_n) = \left(\frac{2\sigma}{\sqrt{2\pi}\sigma}\right)^2 = \frac{2}{\pi} < 1$$

[3] The previous examples include normal and Laplace.

**Example** (Laplace). If $\epsilon \sim \text{Laplace}(\mu, b)$ with $\sigma^2 = 2b^2$, since $f(0) = 1/\sqrt{2}\sigma$, hence

$$\text{ARE}(\widetilde{T}_n, \overline{\text{sign}}_n) = 4\sigma^2 \frac{1}{2\sigma^2} = 2 > 1,$$

**Example** (Uniform). If $\epsilon \sim \mathcal{U}(-c, c)$ for some $c$ such that $\text{Var}[\epsilon] = \sigma^2$, we have

$$\text{ARE}(\widetilde{T}_n, \overline{\text{sign}}_n) = \frac{1}{3}.$$

**Proof.** We see that since $\sigma^2 = (2c)^2/12 = c^2/3$, $c$ should be $\sqrt{3}\sigma$. Hence, $f(0) = 1/2c = 1/2\sqrt{3}\sigma$. Plugging in $(2f(0)\sigma)^2$ gives $1/3$. ⊛

Finally, let's consider the Wilcoxon signed-rank test. Recall what we have shown.

**As previously seen.** Under $H_0$, i.e., $X$ is continuous and $X \overset{D}{=} -X$, with $U_n \coloneqq \binom{n}{2}^{-1} \sum_{i<j} h(X_i, X_j)$ where $h(x_1, x_2) = \mathbb{1}_{x_1+x_2 \geq 0}$, we have

$$\frac{\sqrt{n}}{2 \cdot \binom{n}{2}} \left( W_n - \sum_{i=1}^{n} \text{sgn}(X_i) \right) = \sqrt{n} \left( U_n - \frac{1}{2} \right) \xrightarrow{D} \mathcal{N}(0, 1/3).$$

In our case, since $X_i = \theta_n + \epsilon_i$, we have

$$U_n = \frac{1}{\binom{n}{2}} \sum_{\{i,j\} \subseteq [n]} \mathbb{1}_{X_i + X_j > 0} = \frac{1}{\binom{n}{2}} \sum_{\{i,j\} \subseteq [n]} \mathbb{1}_{\epsilon_i + \epsilon_j > -2\theta_n}.$$

**Example** (Wilcoxon signed-rank test). The slope of the corresponding $U$-statistic $U_n$ of the Wilcoxon signed-rank statistic $W_n$ is $2\sqrt{3} \cdot \int f^2(x)\,\mathrm{d}x$.

**Proof.** Let $h_n(\epsilon_1, \epsilon_2) = \mathbb{1}_{\epsilon_1+\epsilon_2 > -2\theta_n}$ (note that it depends on $n$), then from Theorem 4.3.1,

$$\sqrt{n}\,(U_n - \mathbb{E}[U_n]) = \frac{2}{\sqrt{n}} \sum_{k=1}^{n} (h_n^*(\epsilon_k) - \mathbb{E}[U_n]) + o_p(1)$$

where the $2 = m$ is the number of arguments of $h$. Here, $\mathbb{E}[h_n(\epsilon_1, \epsilon_2)] = \mathbb{E}[U_n] = \mathbb{E}[h_n^*(\epsilon)]$ where

$$h_n^*(x) = \mathbb{E}[h(x, \epsilon)] = \mathbb{P}(x + \epsilon > -2\theta_n) = \mathbb{P}(\epsilon > -x - 2\theta_n) = \mathbb{P}(\epsilon \leq x + 2\theta_n) = F(x + 2\theta_n),$$

with $\epsilon_i$'s being i.i.d., by dividing both sides by $2\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}$ and the central limit theorem,

$$\sqrt{n} \frac{U_n - \mathbb{E}[F(\epsilon + 2\theta_n)]}{2\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{h_n^*(\epsilon_k) - \mathbb{E}[F(\epsilon + 2\theta_n)]}{\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}} + o_p(1) \xrightarrow{D} \mathcal{N}(0, 1)$$

as long as $\text{Var}[F(\epsilon + 2\theta_n)] > 0$. We see that

- $\mu(\theta) = \mathbb{E}[F(\epsilon + 2\theta)]$, if we assume $f$ exists, it's just

  $$\mu(\theta) = \int_{\mathbb{R}} F(x + 2\theta) F(\mathrm{d}x) = \int_{\mathbb{R}} F(x + 2\theta) f(x)\,\mathrm{d}x.$$

  If we further assume that we can interchange the derivative and the integral, we then have $\mu'(\theta) = \int_{\mathbb{R}} 2f(x + 2\theta) f(x)\,\mathrm{d}x$, giving $\mu'(0) = 2\int_{\mathbb{R}} f^2(x)\,\mathrm{d}x$.
- $\sigma(\theta) = 2\sqrt{\text{Var}[F(\epsilon + 2\theta)]}$, hence $\sigma(0) = 2\sqrt{\text{Var}[F(\epsilon)]} = 2 \cdot \sqrt{1/12} = 1/\sqrt{3}$.

Hence, the slope of $U_n$ is $2\sqrt{3} \cdot \int_{\mathbb{R}} f^2(x)\,\mathrm{d}x$. ⊛

This gives us a way to compare $U_n$ and $\widetilde{T}_n$, i.e.,

$$\mathrm{ARE}(\widetilde{T}_n, U_n) = \left(\frac{2\sqrt{3}\int_{\mathbb{R}} f^2(x)\,\mathrm{d}x}{1/\sigma}\right)^2 = 12\sigma^2\left(\int_{\mathbb{R}} f^2(x)\,\mathrm{d}x\right)^2.$$

# Lecture 22: The Theory of Simple Linear Rank Statistics

## 4.5 Simple Linear Rank Statistics

9 Apr. 9:30

Consider $X_1, \dots, X_N, \dots \overset{\text{i.i.d.}}{\sim} F$ such that $F$ is continuous, and for $1 \le i \le N$, let the *rank* be

$$R_{Ni} \coloneqq \sum_{j=1}^{N} \mathbb{1}_{X_j \le X_i},$$

i.e., $R_{Ni}$ is the rank of the $i^{\text{th}}$ observation among the first $N$. Clearly, as $N$ varies, $R_{Ni}$ is distributed differently. In particular, we have the following relatively easy lemma where we omit the proof.

**Lemma 4.5.1.** The rank vector $\widetilde{R}_N \coloneqq (R_{N1}, \dots, R_{NN})$ is uniformly distributed on the permutations of $[N]$. Consequently,

- $R_{Ni}$ is uniformly distributed on $[N]$ for every $1 \le i \le N$.
- $(R_{Ni}, R_{Nj})$ is uniformly distributed on $\{(k,\ell)\colon 1 \le k \ne \ell \le N\}$ for every $1 \le i \ne j \le N$.

On the other hand, we also note the following.

**Remark.** Since $F$ is continuous, $U_i \coloneqq F(X_i) \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,1)$ for all $i \ge 1$, and we have $R_{Ni} = \sum_{j=1}^{N} \mathbb{1}_{U_j \le U_i}$ almost surely.

Now, let's revisit the two-sample problem.

**Example** (Two-sample problem). Consider two samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} G$, and we want to test whether $H_0\colon F = G$. Set $X_{n+i} = Y_i$ for $1 \le i \le m$, which gives

$$X_1, \dots, X_n, X_{n+1} = Y_1, \dots, X_{n+m} = Y_m$$

with $N \coloneqq n + m$. Under $H_0$, $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} F$. Naively, we will reject $H_0$ when $\overline{Y}_m - \overline{X}_n$ is "large" (or "small"). Equivalently, we may consider $\sum_{i=1}^{m} Y_i - \sum_{i=1}^{n} X_i = \sum_{i=1}^{N} c_{Ni} X_i$ where

$$c_{Ni} = \begin{cases} 1, & \text{if } n < 1 \le N; \\ -1, & \text{if } 1 \le i \le n. \end{cases}$$

From our experience, one might replace $X_i$ by $R_{Ni}$, i.e., consider

$$\sum_{i=1}^{N} c_{Ni} R_{Ni} = \sum_{i=1}^{N} (2\widetilde{c}_{Ni} - 1) R_{Ni} = 2\sum_{i=1}^{N} \widetilde{c}_{Ni} R_{Ni} - \sum_{i=1}^{N} R_{Ni} = 2\sum_{i=n+1}^{M} R_{Ni} - \frac{N(N+1)}{2}$$

since $\sum_{i=1}^{N} R_{Ni} = \sum_{i=1}^{N} i$, and we define

$$\widetilde{c}_{Ni} = \frac{c_{Ni} + 1}{2} = \begin{cases} 1, & \text{if } n < i \le N; \\ 0, & \text{if } 1 \le i \le n. \end{cases}$$

Observe that $\sum_{i=n+1}^{M} R_{Ni}$ is just a Wilcoxon two-sample rank statistic.

This suggests we look into the following.

**Definition 4.5.1** (Simple linear rank statistic)**.** Consider $X_1, \dots, X_N \overset{\text{i.i.d.}}{\sim} F$ where $F$ is continuous. The *simple linear rank statistic* is defined as

$$\sum_{i=1}^{N} c_N(i)\alpha_N(R_{Ni}),$$

where $c_N(i) =: c_{Ni}$ and $\alpha_N(i) =: \alpha_{Ni}$ for $1 \le i \le N$ are all constants.

**Remark.** The distribution of any simple linear rank statistics is independent of $F$!

In fact, simple linear rank statistic is very common in practice.

**Example** (Median statistic)**.** We can also consider $\sum_{i=n+1}^{M} \mathbb{1}_{R_{Ni} \ge (N+1)/2}$.

**Example** (Simple random sampling)**.** Given a finite population $\{x_1, \dots, x_N\}$, to estimate the population average $(x_1 + \cdots + x_N)/N$, we take a sample of size $n$ and evaluate the sample mean

$$\frac{1}{n}\sum_{i=1}^{N} x_i \mathbb{1}_{i^{\text{th}} \text{ population is in the sample}}.$$

Consider a *simple random sample*, i.e., all subset of size $n$ from the population is equally likely to be selected. In this case, $(R_{N1}, \dots, R_{NN})$ is equally likely to be any permutation of $[N]$. Hence, the above indicators are just $\mathbb{1}_{R_{Ni} \le n}$, which suggests $\alpha_N(R_{Ni}) \coloneqq \mathbb{1}_{R_{Ni} \le n}$ in the above notation.

All these examples motivates us to develop a general theory for the simple linear rank statistic in the form of $T_N \coloneqq \sum_{i=1}^{N} c_{Ni}\alpha_N(R_{Ni})$, in particular, to establish the asymptotic normality of them.

### 4.5.1 Moments of Linear Rank Statistics

To start, let's first compute the expectation and the variance. We will adopt the following notations.

**Notation.** We write $\overline{\alpha}_N$ to be the mean of $\alpha_{Ni}$ and $\overline{c}_N$ to be the average of $c_{Ni}$, i.e.,

$$\overline{\alpha}_N \coloneqq \frac{1}{N}\sum_{i=1}^{N} \alpha_N(i), \text{ and } \overline{c}_N \coloneqq \frac{1}{N}\sum_{i=1}^{N} c_{Ni}.$$

Moreover, let $\sigma^2_{N\alpha}$ to be the variance of $\alpha_{Ni}$, and similarly define $\sigma^2_{Nc}$, i.e.,

$$\sigma^2_{N\alpha} = \frac{1}{N}\sum_{i=1}^{N} (\alpha_N(i) - \overline{\alpha}_N)^2, \text{ and } \alpha^2_{Nc} = \frac{1}{N}\sum_{i=1}^{N} (c_{Ni} - \overline{c}_N)^2.$$

These make sense since for any individual $R_{Ni}$, marginally they're identically distributed from Lemma 4.5.1. On the other hand, there's no randomness in $c_i$'s.

With this notation, we can now compute the first and the second moments.

**Claim.** For any simple linear rank statistic $T_N \coloneqq \sum_{i=1}^{N} c_{Ni}\alpha_N(R_{Ni})$, $\mathbb{E}[T_N] = N\overline{\alpha}_N\overline{c}_N$.

**Proof.** Since marginally, $R_{Ni}$'s are just uniform over $[N]$, hence the expectation of $T_N$ is

$$\mathbb{E}[T_N] = \sum_{i=1}^{N} c_{Ni}\mathbb{E}[\alpha_N(R_{Ni})] = \sum_{i=1}^{N} c_{Ni}\mathbb{E}[\alpha_N(R_{N1})] = \sum_{i=1}^{N} c_{Ni}\sum_{j=1}^{N}\frac{\alpha_N(j)}{N} =: N\overline{\alpha}_N\overline{c}_N$$

as $\overline{\alpha}_N = \mathbb{E}[\alpha_N(R_{N1})] = \frac{1}{N}\sum_{i=1}^{N}\alpha_{Ni}$ and $\overline{c}_N = \frac{1}{N}\sum_{i=1}^{N} c_{Ni}$. ⊛

Computing the variance is a bit more challenging, but still doable.

**Claim.** For any simple linear rank statistic $T_N \coloneqq \sum_{i=1}^N c_{Ni}\alpha_N(R_{Ni})$, $\operatorname{Var}[T_N] = \frac{N^2}{N-1}\sigma_{Nc}^2\sigma_{N\alpha}^2$.

**Proof.** Let's first center $T_N$, which gives

$$T_N - \mathbb{E}[T_N] = \sum_{i=1}^N c_{Ni}\alpha_N(R_{Ni}) - \sum_{i=1}^N c_{Ni}\overline{\alpha}_N = \sum_{i=1}^N (c_{Ni}-\overline{c}_N)\alpha_N(R_{Ni}) + \overline{c}_N\sum_{i=1}^N \alpha_N(R_{Ni}) - \sum_{i=1}^N c_{Ni}\overline{\alpha}_N.$$

Now, observe that $\sum_{i=1}^N \alpha_N(R_{Ni}) = \sum_{i=1}^N \alpha_N(i)$, the last two terms will cancel out, giving $T_N - \mathbb{E}[T_N] = \sum_{i=1}^N (c_{Ni}-\overline{c}_N)\alpha_N(R_{Ni})$. Then, by the definition of variance,

$$\begin{aligned}\operatorname{Var}[T_N] &= \operatorname{Var}\left[\sum_{i=1}^N (c_{Ni}-\overline{c}_N)\alpha_N(R_{Ni})\right] \\ &= \sum_{i=1}^N (c_{Ni}-\overline{c}_N)^2 \operatorname{Var}[\alpha_N(R_{Ni})] + \sum_{i\neq j}(c_{Ni}-\overline{c}_N)(c_{Nj}-\overline{c}_N)\operatorname{Cov}[\alpha_N(R_{Ni}),\alpha_N(R_{Nj})].\end{aligned}$$

The first sum is just $N\sigma_{Nc}^2\sigma_{N\alpha}^2$, so we focus on the second sum.

**Intuition.** For $i \neq j$, $(R_{Ni}, R_{Nj})$ is equally likely to take any value in $\{(i,j)\colon 1 \le i \neq j \le N\}$.

Hence, we can replace $\operatorname{Cov}[\alpha_N(R_{Ni}),\alpha_N(R_{Nj})]$ by $\operatorname{Cov}[\alpha_N(R_{N1}),\alpha_N(R_{N2})]$, and focus only on $\sum_{i\neq j}(c_{Ni}-\overline{c}_N)(c_{Nj}-\overline{c}_j)$. In particular, we have the following.

**Note.** For any sequence $(x_N)$, we have $\sum_{i\neq j}(x_i-\overline{x}_N)(x_j-\overline{x}_N) = -\sum_{i=1}^N (x_i-\overline{x}_N)^2$.

**Proof.** From the identity $(\sum_{i=1}^N x_i)^2 = \sum_{i=1}^N x_i^2 + \sum_{i\neq j} x_i x_j$, hence

$$0 = \left(\sum_{i=1}^N (x_i-\overline{x}_N)\right)^2 = \sum_{i=1}^N (x_i-\overline{x}_N)^2 + \sum_{i\neq j}(x_i-\overline{x}_N)(x_j-\overline{x}_N).$$

Rearranging the terms gives the equality. ⊛

Hence, we see that by using the above identity and the fact that the joint distribution of $R_{N1}$ and $R_{N2}$ is the uniform, with the definition of the covariance,

$$\begin{aligned}&\sum_{i\neq j}(c_{Ni}-\overline{c}_N)(c_{Nj}-\overline{c}_N)\operatorname{Cov}[\alpha_N(R_{Ni}),\alpha_N(R_{Nj})] \\ &= \operatorname{Cov}[\alpha_N(R_{n1}),\alpha_N(R_{N2})]\cdot\sum_{i\neq j}(c_{Ni}-\overline{c}_N)(c_{Nj}-\overline{c}_N) \\ &= \left[\frac{1}{N(N-1)}\sum_{i\neq j}(\alpha_{Ni}-\overline{\alpha}_N)(\alpha_{Nj}-\overline{\alpha}_N)\right](-N\sigma_{Nc}^2) \\ &= \left[-\frac{1}{N(N-1)}\sum_{i=1}^N(\alpha_{Ni}-\overline{\alpha}_N)^2\right](-N\sigma_{Nc}^2) = \frac{N}{N-1}\sigma_{N\alpha}^2\sigma_{Nc}^2.\end{aligned}$$

Putting everything together, we have

$$\operatorname{Var}[T_N] = N\sigma_{Nc}^2\sigma_{N\alpha}^2 + \frac{N}{N-1}\sigma_{N\alpha}^2\sigma_{Nc}^2 = \frac{N^2}{N-1}\sigma_{Nc}^2\sigma_{N\alpha}^2,$$

which gives the desired result. ⊛

### 4.5.2 Asymptotic Normality of Linear Rank Statistics

With the above calculation, we now want to establish the asymptotic normality of the simple linear rank statistic. Specifically, we consider a special form of $\alpha_N(i)$ given some $\phi\colon [0,1] \to \mathbb{R}$, i.e., for $1 \le i \le N$,

$$\alpha_N(i) = \phi\left(\frac{i}{N+1}\right).$$

It might seem cryptic and mysterious at the first glance why we want to consider $\alpha_N(i)$ in this form.

**Intuition.** Consider order statistics $U_{N(1)} \le \cdots \le U_{N(N)}$ for the uniform $U_i$'s. Then, for $1 \le i \le N$, $\mathbb{E}[U_{N(i)}] = i/(N+1)$, implying $\alpha_N(i) = \phi(\mathbb{E}[U_{N(i)}])$.

**Example.** The simple linear rank statistic we have seen so far can be written in the above form.

To proceed, one might expect something like $\phi(\mathbb{E}[U_{N(i)}]) \approx \mathbb{E}[\phi(U_{N(i)})]$ to hold, and in fact, while both of them are of our interests, the latter is easier to analyze than $\phi(\mathbb{E}[U_{N(i)}])$.

**Intuition.** For $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$, we have $\alpha_N(R_{Ni}) = \mathbb{E}[\phi(U_i) \mid \widetilde{R}_N]$.

The above shows that $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$ is more convenient since we will have

$$T_N = \sum_{i=1}^N c_{Ni}\alpha_N(R_{Ni}) = \sum_{i=1}^N c_{Ni}\mathbb{E}[\phi(U_i) \mid \widetilde{R}_N],$$

which is a conditional expectation. We can then apply the theory of projection.

**Notation.** In what follows, we denote $\alpha_{Ni} \coloneqq \mathbb{E}[\phi(U_{N(i)})]$ and $\alpha'_{Ni} \coloneqq \phi(\mathbb{E}[U_{N(i)}])$ for $1 \le i \le N$, and refer to the corresponding simple linear rank statistic as $T_N$ and $T'_N$, respectively.

## Lecture 23: Asymptotically Normality of Linear Rank Statistics

We will first work with $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$, where we recall that $\alpha_N(R_{Ni}) = \mathbb{E}[\phi(U_i) \mid \widetilde{R}_N]$. 11 Apr. 9:30

**Proposition 4.5.1.** For every $N \ge 1$, we have $T_N - \mathbb{E}[T_N] = \mathbb{E}[\widetilde{T}_N \mid \widetilde{R}_N]$ where

$$\widetilde{T}_N \coloneqq \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\phi(U_i).$$

**Proof.** It'll be convenient to consider

$$T_N = \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\alpha_N(R_{Ni}) + \overline{c}_N \sum_{i=1}^N \alpha_N(R_{Ni}) = \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\alpha_N(R_{Ni}) + N\overline{c}_N\overline{\alpha}_N,$$

with the fact that $\mathbb{E}[T_N] = N\overline{c}_N\overline{\alpha}_N$, we see that

$$T_N - \mathbb{E}[T_N] = \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\alpha_N(R_{Ni})$$

and since $\alpha_N(R_{Ni}) = \mathbb{E}[\phi(U_i) \mid \widetilde{R}_N]$, we further have

$$= \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\mathbb{E}[\phi(U_i) \mid \widetilde{R}_N] = \mathbb{E}\left[\sum_{i=1}^N (c_{Ni} - \overline{c}_N)\phi(U_i) \mid \widetilde{R}_N\right] =: \mathbb{E}[\widetilde{T}_N \mid \widetilde{R}_N]$$

where we let $\widetilde{T}_N \coloneqq \sum_{i=1}^N (c_{Ni} - \overline{c}_N)\phi(U_i)$. ∎

To see how can we obtain asymptotic normality by the theory of projection, observe the following.

**Claim.** We can easily have asymptotically normality for $\widetilde{T}_N$.

**Proof.** We see that by the Hájek-Sidak central limit theorem, if $\phi(U_i)$'s are i.i.d. such that

$$0 < \mathbb{E}[\phi^2(U)] = \int_0^1 \phi^2(u)\,\mathrm{d}u < \infty, \text{ and } \max_{1\le i\le N} \frac{(c_{Ni}-\bar{c}_N)^2}{\sum_{j=1}^N (c_{Nj}-\bar{c}_N)^2} \to 0,$$

then $\widetilde{T}_N / \sqrt{\operatorname{Var}[\widetilde{T}_N]} \xrightarrow{D} \mathcal{N}(0,1)$ as $\mathbb{E}[\widetilde{T}_N] = 0$. ⊛

Applying the projection theory we have developed, we have the following.

**Theorem 4.5.1.** Suppose that $(c_{Ni})$ satisfy $\max_{1\le i\le N}(c_{Ni}-\bar{c}_N)^2 / \sum_{i=1}^N (c_{Ni}-\bar{c}_N)^2 \to 0$ as $N\to\infty$. If $\operatorname{Var}[\phi(U)] \in (0,\infty)$, then

$$\frac{T_N - \mathbb{E}[T_N]}{\sqrt{\operatorname{Var}[T_N]}} \xrightarrow{D} \mathcal{N}(0,1).$$

**Proof.** From Corollary 4.3.1 and the above claim, it suffices to show that $\operatorname{Var}[T_N]/\operatorname{Var}[\widetilde{T}_N] \to 1$ as $N\to\infty$. Recall that $\operatorname{Var}[T_N] = \frac{N^2}{N-1}\sigma_{N\alpha}^2\alpha_{Nc}^2$, furthermore, $\operatorname{Var}[\widetilde{T}_N] = N\sigma_{Nc}^2\operatorname{Var}[\phi(U_1)]$ since

$$\operatorname{Var}[\widetilde{T}_N] = \operatorname{Var}[\phi(U_1)]\cdot\sum_{i=1}^N (c_{Ni}-\bar{c}_N)^2 = N\sigma_{Nc}^2\operatorname{Var}[\phi(U)],$$

so it suffices to show $\sigma_{N\alpha}^2 = \operatorname{Var}[\alpha_N(R_{N1})] \to \operatorname{Var}[\phi(U_1)]$.

**Note.** To show $\operatorname{Var}[X]\to\operatorname{Var}[Y]$, it suffices to show $X \xrightarrow{L^2} Y$ since it implies convergence for both the first and second moments, hence the variance

In particular, it reduces to show $\alpha_N(R_{N1}) = \mathbb{E}[\phi(U_1)\mid \widetilde{R}_N] \xrightarrow{L^2} \phi(U_1)$. Firstly, we write

$$\mathbb{E}[\phi(U_1)\mid \widetilde{R}_N] = \mathbb{E}[\phi(U_1)\mid \widetilde{R}_1,\dots,\widetilde{R}_N]$$

as the condition on the right-hand side is equivalent to $\widetilde{R}_N, U_2,\dots,U_N$, and $U_1$ is independent of $U_2,\dots,U_N$. From the martingale limit theorem, we have

$$\mathbb{E}[\phi(U_1)\mid \widetilde{R}_1,\dots,\widetilde{R}_N] \xrightarrow{\text{a.s.}} \mathbb{E}[\phi(U_1)\mid \widetilde{R}_N, N\ge 1],$$

which will equal to $\phi(U_1)$ if $\phi(U_1)$ is a function of the conditions. Hence, it remains to show that $U_1$ is a measurable function of $\{\widetilde{R}_N\}_{N\ge1}$.

**Claim.** Knowing the first components of $\widetilde{R}_N$ for every $N\ge1$, i.e., $R_{N1}$, determines $U_1$ (in $L^2$).

**Proof.** We observe that $\mathbb{E}[U_1\mid R_{N1}] = R_{N1}/(N+1)$, i.e., the expectation of $U_{(R_{N1})}$. Hence, by Corollary 4.3.1, $U_1 - R_{N1}/(N+1) \xrightarrow{L^2} 0$ if the ratio between variances converges to 1. Indeed,

$$\frac{\operatorname{Var}[U_1]}{\operatorname{Var}\left[\dfrac{R_{N1}}{N+1}\right]} = \frac{1/12}{\dfrac{1}{(N+1)^2}\operatorname{Var}[R_{N1}]} = \frac{1/12}{\dfrac{1}{(N+1)^2}\dfrac{N^2-1}{12}} = \frac{N+1}{N-1} \to 1$$

since $R_{N1}\sim\mathcal{U}([N])$, hence we're done. ⊛

Finally, note that the limits in $L^2$ are unique up to a null set, hence we're done. ■

We now discuss the connection between $T_N$ and $T_N'$. Firstly, recall that

$$T_N = \sum_{i=1}^N c_{Ni}\alpha_N(R_{Ni}), \text{ and } T_N' = \sum_{i=1}^N c_{Ni}\alpha_N'(R_{Ni})$$

where $\alpha_{Ni} = \mathbb{E}[\phi(U_{N(i)})]$ and $\alpha'_{Ni} \coloneqq \phi(\mathbb{E}[U_{N(i)}]) = \phi(i/(N+1))$.

**Theorem 4.5.2.** Suppose that $(c_{Ni})$ satisfy $\max_{1\le i\le N}(c_{Ni}-\bar{c}_N)^2/\sum_{i=1}^N(c_{Ni}-\bar{c}_N)^2 \to 0$ as $N\to\infty$. If $\mathrm{Var}[\phi(U)] \in (0,\infty)$, and in addition, if $\phi$ is almost surely continuous and

$$\limsup_{n\to\infty}\frac{1}{N}\sum_{i=1}^N \phi^2\left(\frac{i}{N+1}\right) \le \int_0^1 \phi^2(u)\,\mathrm{d}u,$$

then

$$\frac{T'_N - \mathbb{E}[T'_N]}{\sqrt{\mathrm{Var}[T_N]}} \xrightarrow{D} \mathcal{N}(0,1).$$

**Proof.** From Theorem 4.5.1, to show that $T'_N$ is asymptotically normal as $T_N$, it suffices to show

$$\frac{T_N - \mathbb{E}[T_N]}{\sqrt{\mathrm{Var}[T_N]}} - \frac{T'_N - \mathbb{E}[T'_N]}{\sqrt{\mathrm{Var}[T_N]}} \xrightarrow{L^2} 0 \Leftrightarrow \frac{\mathrm{Var}[T_N - T'_N]}{\mathrm{Var}[T_N]} \to 0$$

Firstly, recall that $\mathrm{Var}[T_N] = \frac{N^2}{N-1}\sigma^2_{Nc}\sigma^2_{N\alpha}$, so $\mathrm{Var}[T_N - T'_N] = \frac{N^2}{N-1}\sigma^2_{Nc}\sigma^2_{N(\alpha-\alpha')}$ since we can write

$$T_N - T'_N = \sum_{i=1}^N c_{Ni}(\alpha_N - \alpha'_N)(R_{Ni}),$$

which is again a simple linear rank statistic, so the same calculation applies. Hence, it suffices to show $\sigma^2_{N(\alpha-\alpha')}/\sigma^2_{N\alpha} \to 0$. Recall what we have already shown earlier.

> **As previously seen.** In the proof of Theorem 4.5.1, we have $\sigma^2_{N\alpha} \to \mathrm{Var}[\phi(U_1)] > 0$.

Hence, we just need to show that $\sigma^2_{N(\alpha-\alpha')} = \mathrm{Var}[(\alpha_N - \alpha'_N)(R_{N1})] \to 0$. From the same reason as before, it suffices to show that $(\alpha_N - \alpha'_N)(R_{N1}) \xrightarrow{L^2} 0$, i.e.,

$$\mathbb{E}[\phi(U_1) \mid \widetilde{R}_N] - \phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{L^2} 0.$$

Let's again recall what we have shown earlier.

> **As previously seen.** In the proof of Theorem 4.5.1, we have $\mathbb{E}[\phi(U_1) \mid \widetilde{R}_N] \xrightarrow{L^2} \phi(U_1)$.

Hence, it reduces to show $\phi(R_{N1}/(N+1)) \xrightarrow{L^2} \phi(U_1)$.

> **Intuition.** If $\phi = \mathrm{id}$, then we have showed that $R_{N1}/(N+1) \xrightarrow{L^2} U_1$ in the previous claim.

Hence, $R_{N1}/(N+1) \xrightarrow{p} U_1$. As $\phi$ is assumed to be continuous almost surely, we know that

$$\phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{p} \phi(U_1) \overset{?}{\Rightarrow} \phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{L^2} \phi(U_1).$$

The implication can be provided by Scheffé's theorem, i.e., we only need to check

$$\limsup_{N\to\infty}\mathbb{E}\left[\phi^2\left(\frac{R_{N1}}{N+1}\right)\right] = \limsup_{N\to\infty}\frac{1}{N}\sum_{i=1}^N \phi^2\left(\frac{i}{N+1}\right) \le \mathbb{E}[\phi^2(U_1)] = \int_0^1 \phi^2(u)\,\mathrm{d}u$$

since $R_{N1} \sim \mathcal{U}([N])$ as we mentioned. This is what we assumed exactly. ■

We end this chapter by noting that the condition required in Theorem 4.5.2, in particular, the integral inequality, is not that hard to satisfy. The following is one example.

**Example.** The conditions required in Theorem 4.5.2 are satisfied when $\phi$ is increasing, or, more generally, the difference of two increasing functions.

**Proof.** Let's only illustrate the case of increasing functions. Observe that we can understand

$$\limsup_{n\to\infty} \frac{1}{N} \sum_{i=1}^{N} \phi^2(i/(N+1))$$

in the condition of Theorem 4.5.2 as a Riemann integral. In particular,

$$\int_0^1 \phi^2(u)\,\mathrm{d}u = \lim_{N\to\infty} \sum_{i=1}^{N+1} \int_{\frac{i-1}{N+1}}^{\frac{i}{N+1}} \phi^2(u)\,\mathrm{d}u$$

since $\phi$ is increasing, so is $\phi^2$, hence

$$\geq \lim_{N\to\infty} \sum_{i=1}^{N+1} \phi^2\left(\frac{i-1}{N+1}\right) \cdot \frac{1}{N+1} \geq \lim_{N\to\infty} \frac{N}{N+1} \frac{1}{N} \sum_{i=1}^{N} \phi^2\left(\frac{i}{N+1}\right),$$

which is what we want. ⊛

# Chapter 5

# $M$-Estimation

## Lecture 24: The Maximum Likelihood Estimator

Consider $X, X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$, and we would like to say something about $F$ or $T(F)$. To do this, one can we first obtain the empirical cdf $\hat{F}_n$, and plug in $T(\hat{F}_n)$. This is what we did previously. 16 Apr. 9:30

**Example** (Quantile). The $p^{\text{th}}$-quantile is defined as $T(F) = \theta_p = F^{-1}(p)$, which is estimated by the sample $p^{\text{th}}$-quantile $T(\hat{F}_n) = \hat{\theta}_p = \hat{F}_n^{-1}(p)$.

**Example** (Moment). The $k^{\text{th}}$-moment is defined as $T(F) = \mu_k = \int (x-\mu)^k F(\mathrm{d}x)$ where $\mu = \int xF(\mathrm{d}x)$, which is estimated by the sample $k^{\text{th}}$-central moment $T(\hat{F}_n) = M_k = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^k$.[a]

[a] Note that we're essentially interpreting $M_k = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^k$ as $\int (x-\overline{X}_n)\hat{F}_n(\mathrm{d}x)$ with $\overline{X}_n = \int x\hat{F}_n(\mathrm{d}x)$.

On the other hand, instead of directly estimating $F$, we can start by postulating a family of cdfs $\{G_\theta : \theta \in \Theta\}$ where $\Theta$ is a metric space, with the goal being to choose $\hat{\theta}_n$ among $\Theta$ such that $G_{\hat{\theta}_n}$ approximates $F$. Our choice, $\hat{\theta}_n$, should be a function of the data $X_1, \dots, X_n$.

## 5.1 Maximum Likelihood Estimator

If we assume $G_\theta$ has a corresponding density $g_\theta$ for all $\theta \in \Theta$, then a natural choice is to select $\hat{\theta}_n \in \Theta$ by maximizing the *likelihood function*, i.e., the well-known maximum likelihood estimator (MLE).

**Definition 5.1.1** (Maximum likelihood eestimator). Given $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ and a family of cdfs $\{G_\theta : \theta \in \Theta\}$ for some metric space $\Theta$. The *maximum likelihood estimator* $\hat{\theta}_n$ is the maximizer of the likelihood function, i.e., $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \prod_{i=1}^n g_\theta(X_i)$.

### 5.1.1 Divergence Minimizing

This is all good, since we're just proposing a method. The true motivating question is the following.

**Problem.** What does the maximum likelihood estimator $\hat{\theta}_n$ estimating?

Observe that $\hat{\theta}_n$ is also a maximizer of $\frac{1}{n}\sum_{i=1}^n \log(g_\theta(X_i))$.[1] As $n \to \infty$, from the strong law of large number, we know that for all $\theta \in \Theta$,

$$\frac{1}{n}\sum_{i=1}^n \log(g_\theta(X_i)) \overset{\text{a.s.}}{\to} \mathbb{E}[\log(g_\theta(X))] =: M(\theta) \in [-\infty, \infty].$$

[1] We adapt the convention that $\log(0) := -\infty$ since it happens when $\{G_\theta\}$ is misspecified.

**Note.** We will need to assume that $M(\theta) > -\infty$ for some $\theta$, otherwise there's no hope.

Since $\hat{\theta}_n$ is maximizing the left-hand side, it's natural to conjecture the following.

**Conjecture 5.1.1.** $\hat{\theta}_n$ should "converge" to the maximizer of $M(\theta)$.

To see why this might be the case, assuming $F$ has a density $f$, then we can write

$$M(\theta) = \int_{\mathbb{R}} \log g_\theta(x) f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \log \frac{g_\theta(x)}{f(x)} f(x)\,\mathrm{d}x + \int_{\mathbb{R}} \log f(x) f(x)\,\mathrm{d}x.$$

The second term is independent of $\theta$, and the first term is just the Kullback-Leibler divergence since

$$\int_{\mathbb{R}} \log \frac{g_\theta(x)}{f(x)} f(x)\,\mathrm{d}x = -\int_{\mathbb{R}} \log \frac{f(x)}{g_\theta(x)} f(x)\,\mathrm{d}x = -\operatorname{KL}(f \| g_\theta).$$

**Remark.** The maximizer of $M(\theta)$ is minimizing the KL divergence $\operatorname{KL}(f\|g_\theta)$.

Let $\theta^* = \arg\max_{\theta\in\Theta} M(\theta)$. Then if we can actually show Conjecture 5.1.1, then we can conclude that $\hat{\theta}_n$ is estimating $\theta^*$, i.e., $\hat{\theta}_n$ also tries to minimize the KL divergence between $f$ and $g_{\hat{\theta}_n}$.

**Example** (Supremum estimation). Suppose $X_1, \dots, X_n$ follows $\mathbb{P}(0 \le X \le u) = 1$ for some $u > 0$ such that for all $\epsilon > 0$, $F(u-\epsilon) < 1 = F(u)$. Consider the family of uniform distribution $\{G_\theta = \mathcal{U}(0,\theta) \colon \theta > 0\}$. The likelihood function is given by

$$\Theta \ni \theta \mapsto \prod_{i=1}^n \frac{1}{\theta^n} \mathbb{1}_{\max_{1\le i\le n} X_i \le \theta},$$

hence the maximizer of the likelihood function is given by $\hat{\theta}_n = \max_{1\le i\le n} X_i$.

On the other hand, $M(\theta) = \mathbb{E}[\log g_\theta(X)]$ for all $\theta > 0$, and we want to find $\sup_{\theta>0} M(\theta)$. But if $\theta < u$, $\mathbb{P}(g_\theta(x) = 0) > 0$, then $M(\theta) = -\infty$, hence

$$\sup_{\theta>0} M(\theta) = \sup_{\theta\ge u} M(\theta) = \sup_{\theta\ge u} \log \frac{1}{\theta},$$

which is attained by $u$. From the very first example, indeed, $\hat{\theta}_n \xrightarrow{\text{a.s.}} u$.

Next, for the following examples let's consider the following setup: consider a family $\{G_\theta(x) = G(x-\theta)\}$ for some cdf $G$ with a density $g$. In this case, $\hat{\theta}_n$ maximizes $\prod_{i=1}^n g(X_i - \theta)$, or equivalently,

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n \log(g(X_i - \theta)).$$

The following two examples also confirm Conjecture 5.1.1.

**Example** (Normal). If $G \sim \mathcal{N}(0,1)$, i.e., $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, then

- $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n -(X_i - \theta)^2 = \overline{X}_n$;
- $\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}[-(X-\theta)^2] = \mathbb{E}[X]$.

From strong law of large number, $\overline{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$ assuming it converges.

**Example** (Laplace). If $G \sim \text{Laplace}(\mu, b)$ where $\sigma^2 = 2b^2$, i.e., $g(x) = \frac{1}{2b} e^{-\frac{|x|}{b}} = \frac{1}{\sigma\sqrt{2}} e^{-\frac{|x-\mu|}{\sigma/\sqrt{2}}}$, then

- $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n -|X_i - \theta|$, which is the sample median;

- $\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}[-|X-\theta|]$, which is the population median.

Hence, from Corollary 3.5.1, $\hat{\theta}_n \xrightarrow{p} \theta^*$.

**Remark.** We see that the MLE will do the right thing: when estimating $\mu$ (hence $\theta_{1/2}$), as we have shown, for normal, $\overline{X}_n$ is better than $\hat{\theta}_{1/2}$, while for Laplace $\hat{\theta}_{1/2}$ is better.

However, sometimes we don't know $\hat{\theta}_n$ and $\theta^*$, hence we don't know whether Conjecture 5.1.1 is true.

**Example** (Cauchy). If $G \sim$ Cauchy with location 0 and scale 1, i.e., $g(x) = 1/\pi(1+x^2)$, then

- $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n -\log\big(1+(X_i-\theta)^2\big)$;
- $\theta^* = \arg\max_{\theta\in\Theta} \mathbb{E}[-\log\big(1+(X-\theta)^2\big)]$.

However, we don't know what are $\hat{\theta}_n$ and $\theta^*$ this time, hence we need a different technique.

### 5.1.2 $M$-Estimator

The upshot is that in the above examples, we're dealing with $\hat{\theta}_n = \arg\max_{\theta\in\Theta} \sum_{i=1}^n m(X_i-\theta)$ where $m(x) = -\log(g(x))$. But we're not limited to choosing this specific $m$.

**Example.** Consider $m(x) = x^2 \mathbb{1}_{|x|\le k} + (2k|x| - k^2)\mathbb{1}_{|x|>k}$ for some $k \in \mathbb{R}$.

This kind of general estimator in the form of maximizing some $m$ is called the $M$-estimator.

**Definition 5.1.2** ($M$-estimator). Given data $X_1, \dots, X_n$, not necessarily i.i.d., and a metric space $\Theta$, let $m_\theta(x)$ be a function for every $\theta \in \Theta$, and define a function $M_n(\theta)$ of $X_i$'s as $M_n(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$. An *$M$-estimator* $\hat{\theta}_n$ is a maximizer of $M_n$, i.e., $\hat{\theta}_n = \arg\max_{\theta\in\Theta} M_n(\theta)$.

We see that if $X_i$'s are i.i.d., then $M_n(\theta) \to M(\theta) \coloneqq \mathbb{E}[m_\theta(X)] \in [-\infty, \infty)$[2] as $n \to \infty$ for any fixed $\theta \in \Theta$, where the convergence can be either almost surely or just in probability.

**Notation.** We will use $\xrightarrow[p]{\text{a.s.}}$ from now on to indicate either $\xrightarrow{\text{a.s.}}$ or $\xrightarrow{p}$, depending on the assumption.

Hence, the goal of an $M$-estimator is to use $(\hat{\theta}_n)$ to approximate $\theta^* = \arg\max_{\theta\in\Theta} M(\theta)$, what we really care about. With $M_n(\theta) \xrightarrow[p]{\text{a.s.}} M(\theta)$, we can also approximate $\max_{\theta\in\Theta} M(\theta)$.

**Intuition.** We don't need to specify the family $\{G_\theta : \theta \in \Theta\}$ explicitly, allowing more flexibility.

## 5.2 Consistency

Our first goal is to establish $\hat{\theta}_n \xrightarrow[p]{\text{a.s.}} \theta^*$, i.e., the consistency of the $M$-estimator. However, we will see that the *approximate maximizer* $\hat{\theta}_n$ of $\theta \mapsto M_n(\theta)$ suffices for our purpose. In particular, we define $\hat{\theta}_n$ in the way that for a (positive) vanishing sequence $z_n \xrightarrow[p]{\text{a.s.}} 0$ as $n \to 0$, $\hat{\theta}_n$ satisfies

$$M_n(\hat{\theta}_n) \ge \sup_{\theta\in\Theta} M_n(\theta) - z_n. \tag{5.1}$$

**Intuition.** We don't need to solve the optimization problem $\arg\max_{\theta\in\Theta} M_n(\theta)$ exactly.

[2] Again, assume that $M(\theta) > -\infty$ for some $\theta \in \Theta$, otherwise such a optimization problem is meaningless.

### 5.2.1 Assumptions for Consistency

We first consider the most fundamental question:

**Problem.** Where does $\hat{\theta}_n$ converge, in what sense, if anywhere?

We will need some assumptions around $\theta^*$ to answer this. In particular, consider the following.

(A1) $M$ has a unique maximizer $\theta^*$, and is well-separated, i.e., for all $\epsilon > 0$, $M(\theta^*) > \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta)$.

(A2) For all $\delta > 0$ and $\epsilon > 0$, $\mathbb{P}(\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) - \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \geq \delta) \to 0$.

We can also consider a stronger version of (A2):

(A3) For all $\delta > 0$ and $\epsilon > 0$, $\mathbb{P}(\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \geq \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) + \delta \text{ i.o.}) = 0$.[3]

**Notation** (Open ball). Let $B(\theta^*, \epsilon) = \{\theta \in \Theta \colon d(\theta, \theta^*) < \epsilon\}$ be an open ball around $\theta^* \in \Theta$ with radii $\epsilon > 0$, where $d$ is the metric on $\Theta$.

Let's understand them one-by-one. Firstly, we see that if $\widetilde{\theta} \neq \theta^*$, then $\widetilde{\theta} \notin B(\theta^*, \epsilon)$ for some $\epsilon > 0$. Hence, (A1) implies $M(\theta^*) > \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \geq M(\widetilde{\theta})$. To satisfy (A1), consider the following.

**Definition 5.2.1** (Upper semi-continuous). A function $M \colon \Theta \to \mathbb{R}$ is *upper semi-continuous* if for all $\theta \in \Theta$ and $\theta_n \to \theta$, $M(\theta) \geq \limsup_{n \to \infty} M(\theta_n)$.

Upper semi-continuity is useful since we have the following result from analysis.

**Theorem 5.2.1.** If $\Theta$ is compact and $M(\theta)$ is upper semi-continuous, then $\sup_{\theta \in \Theta} M(\theta)$ is obtained, i.e., there exists a maximizer $\theta^*$.

A simple application of Theorem 5.2.1 allows us to prove (A1).

**Corollary 5.2.1.** If $\Theta$ is compact, $M$ is upper semi-continuous, and the maximizer $\theta^*$ is unique, then (A1) is satisfied.

**Proof.** Firstly, we can write $\sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) = \sup_{\theta \in (B(\theta^*, \epsilon))^c} M(\theta)$. Since $B(\theta^*, \epsilon)$ is open, $(B(\theta^*, \epsilon))^c$ is closed, hence compact since $\Theta$ is compact. Applying Theorem 5.2.1, we see that $\sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta)$ is obtained by a maximizer $\theta^*$, which is unique as assumed. ■

Hence, as long as $\Theta$ is compact and $M(\theta)$ is reasonable enough, (A1) can be satisfied. On the other hand, (A2) is a bit more involved. To understand it, observe the following.

**Claim.** $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ implies (A2).

**Proof.** By the triangle inequality, $M_n(\theta) \leq |M_n(\theta) - M(\theta)| + M(\theta)$, hence

$$\sup_{\theta \in \Theta} M_n(\theta) \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + \sup_{\theta \in \Theta} M(\theta) \Rightarrow \sup_{\theta \in \Theta} M_n(\theta) - \sup_{\theta \in \Theta} M(\theta) \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|.$$

Moreover, taking the supremum over a smaller subset, we have

$$\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) - \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|,$$

hence, $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} 0$ implies $\mathbb{P}(\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) - \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \geq \delta) \to 0$ for any $\delta > 0$, which is (A2). ⊛

[3] Here, i.o. means *infinitely often* (in terms of $n$).

**Intuition.** (A2) uniformly controls the convergence.

In this regard, by reformulating (A3) as

$$\mathbb{P}\left(\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \geq \sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)+\delta \text{ i.o.}\right)=0 \Leftrightarrow \mathbb{P}\left(\limsup_{n\to\infty}\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \leq \sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)\right)=1,$$

since the i.o. can be expressed as $\mathbb{P}(\limsup_{n\to\infty}\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta)\geq \sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)+\delta)=0$. Then, the same argument shows that $\sup_{\theta\in\Theta}|M_n(\theta)-M(\theta)|\xrightarrow{\text{a.s.}}0$ implies (A3).

**Theorem 5.2.2.** If (A1) and (A2) holds, then $\hat{\theta}_n\xrightarrow{p}\theta^*$. On the other hand, if we replace (A2) by (A3), then we have $\hat{\theta}_n\xrightarrow{\text{a.s.}}\theta^*$.

# Lecture 25: Understanding Assumptions for Consistency

18 Apr. 9:30

We first prove Theorem 5.2.2.

**Proof.** From the definition, we have

- $\hat{\theta}_n\xrightarrow{p}\theta^*\Leftrightarrow\mathbb{P}(d(\hat{\theta}_n,\theta^*)\geq\epsilon)\to 0$ for all $\epsilon>0$;
- $\hat{\theta}_n\xrightarrow{\text{a.s.}}\theta^*\Leftrightarrow\mathbb{P}(\limsup_{n\to\infty} d(\hat{\theta}_n,\theta^*)\geq\epsilon)=0$ for all $\epsilon>0$.

Hence, we need to show that for all $\epsilon>0$, $\mathbb{P}(d(\hat{\theta}_n,\theta^*)\geq\epsilon)\to 0$ or $\mathbb{P}(d(\hat{\theta}_n,\theta^*)>\epsilon \text{ i.o.})=0$. Firstly, suppose $d(\hat{\theta}_n,\theta^*)\geq\epsilon$, which implies $\hat{\theta}_n\notin B(\theta^*,\epsilon)$, and with Equation 5.1,

$$\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta)\geq M_n(\hat{\theta}_n)\geq \sup_{\theta\in\Theta} M_n(\theta)-z_n\geq M_n(\theta^*)-z_n.$$

This gives $\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta)-M_n(\theta^*)+z_n\geq 0$, i.e.,

$$\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta)-M_n(\theta^*)+z_n-\sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)+M(\theta^*)\geq M(\theta^*)-\sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)=:\delta_\epsilon>0$$

by (A1). Write $A_n:=\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta)-\sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)$, $C_n:=M(\theta^*)-M_n(\theta^*)$, we have

$$A_n+C_n+z_n\geq\delta_\epsilon.$$

This gives $\{d(\hat{\theta}_n,\theta^*)\geq\epsilon\}\subseteq\{A_n+C_n+z_n\geq\delta_\epsilon\}$, hence

$$\mathbb{P}(d(\hat{\theta}_n,\theta^*)\geq\epsilon)\leq\mathbb{P}\left(A_n\geq\frac{\delta_\epsilon}{3}\right)+\mathbb{P}\left(C_n\geq\frac{\delta_\epsilon}{3}\right)+\mathbb{P}\left(z_n\geq\frac{\delta_\epsilon}{3}\right).$$

Now, with (A2), $\mathbb{P}(A_n\geq\delta_\epsilon/3)$ vanishes. Moreover, as $M_n(\theta)\xrightarrow{p}M(\theta)$ for any fixed $\theta$, in this case, $\theta^*$ for $C_n$, $\mathbb{P}(C_n\geq\delta_\epsilon/3)$ also vanishes. Finally, $\mathbb{P}(z_n\geq\delta_\epsilon/3)$ vanishes if we require $z_n\xrightarrow{p}0$.

The case for (A3) is the same, but this time we need $M_n(\theta)\xrightarrow{\text{a.s.}}M(\theta)$ and $z_n\xrightarrow{\text{a.s.}}0$. ■

## 5.2.2 Proving the Assumptions

With Theorem 5.2.2, the only task left is to check (A1) and (A2) or (A3). Although we have briefly seen how to prove (A1), i.e., Corollary 5.2.1, but there are still lots to be discussed. Let's assume that $X_1,\ldots,X_n$ are i.i.d. and $M_n(\theta)=\frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ where $\theta\mapsto m_\theta(x)$ is upper semi-continuous for all $x$.

**Note.** Hence, $\theta\mapsto M_n(\theta)$ is also upper semi-continuous.

**Proof.** Since the sum of upper semi-continuous functions is still upper semi-continuous. ⊛

But what about $M$? I.e., can we say that the upper semi-continuity is preserved in the limit?

**Proposition 5.2.1.** If for all $\theta$, there exists $\rho_\theta > 0$ such that $\mathbb{E}[\sup_{u \in B(\theta,\rho_\theta)} m_u(X)] < \infty$, then $M$ is upper semi-continuous.

**Proof.** We want to show that as $\theta_n \to \theta$, $M(\theta) \geq \limsup_{n\to\infty} M(\theta_n)$. From the strong law of large number, $M_n(\theta) \overset{\text{a.s.}}{\to} M(\theta) = \mathbb{E}[m_\theta(X)] \in [-\infty, \infty]$. Since $\theta \mapsto m_\theta(x)$ is upper semi-continuous,

$$M(\theta) = \mathbb{E}[m_\theta(X)] \geq \mathbb{E}\left[\limsup_{n\to\infty} m_{\theta_n}(X)\right] \geq \limsup_{n\to\infty} \mathbb{E}[m_{\theta_n}(X)] = \limsup_{n\to\infty} M(\theta_n),$$

where the last inequality is provided by the reverse Fatou's lemma. In particular, it requires $\mathbb{E}[\sup_{n\geq n_0} m_{\theta_n}(X)] < \infty$ for some large enough $n_0$. But since for any $\epsilon > 0$, as $\theta_n \to \theta$, there exists $n_0$ such that for all $n \geq n_0$, $\theta_n \in B(\theta, \epsilon)$, such a requirement becomes what we assumed. ■

With Proposition 5.2.1, we now have a tool to check upper semi-continuous for $M$, and with Corollary 5.2.1, we only need to check that $M$ has a unique maximizer.

**Example** (MLE). For MLE, if the model class $\{g_\theta : \theta \in \Theta\}$ we postulated is identifiable is well-specified, i.e., $g_\theta = g_{\theta'}$ implies $\theta = \theta'$ and the true pdf is in $\{g_\theta\}$, then $M$ has a unique maximizer.

**Proof.** For MLE, recall that $m_\theta(x) = \log(g_\theta(x))$, and say $X, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f = g_{\theta^*}$. In this case,

$$M(\theta) = \mathbb{E}[\log g_\theta(X)] = -\operatorname{KL}(f \| g_\theta) + C \leq C$$

for some constant $C$ as we have seen. As $\theta = \theta^*$, the maximum is achieved. However, suppose there exists $\widetilde{\theta}$ such that $M(\widetilde{\theta}) = M(\theta^*)$, we have $\operatorname{KL}(g_{\theta^*} \| g_{\widetilde{\theta}}) = 0$ implies $g_{\theta^*} = g_{\widetilde{\theta}}$ almost surely. Hence, if $g_{\theta^*} = g_{\widetilde{\theta}}$ doesn't imply $\theta^* = \widetilde{\theta}$, i.e., not identifiable, the maximizer might not be unique. ⊛

In view of this, (A1) can be checked in general. Before we turn our focus to checking either (A2) or (A3), we see one more tool related to upper semi-continuity, which will be useful later.

**Proposition 5.2.2.** Let $M$ be an upper semi-continuous function on a compact $\Theta$, and let $\theta \in \Theta$ be fixed. Then, as $\rho \to 0$, $\sup_{u \in B(\theta,\rho)} M(u) \to M(\theta)$.

**Proof.** Suppose not, then we have the following.

- If $M(\theta) > -\infty$: there exists $\epsilon > 0$ and $(\rho_n) \to 0$ such that $\sup_{u \in B(\theta,\rho_n)} M(u) \geq M(\theta) + \epsilon$.
- If $M(\theta) = -\infty$: there exists $C_0 \in \mathbb{R}$ such that for all $C \leq C_0$, $\sup_{u \in B(\theta,\rho_n)} M(u) \geq C$.

The above holds for all $n \geq 1$. Combining the above, we see that

$$\sup_{u \in B(\theta,\rho_n)} M(u) \geq \max(M(\theta) + \epsilon, C) \Rightarrow \sup_{u \in \overline{B(\theta,\rho_n)}} M(u) \geq \max(M(\theta) + \epsilon, C).$$

Since $\overline{B(\theta, \rho_n)}$ is also compact, there exists $(\theta_n)$ in $B(\theta, \rho_n)$ such that $M(\theta_n) \geq \max(M(\theta) + \epsilon, C)$ for all $n \geq 1$, $C \leq C_0$, for some $\epsilon > 0$. Taking the limit, we see that

$$\liminf_{n\to\infty} M(\theta) \geq \max(M(\theta) + \epsilon, C) \geq \max\left(\limsup_{n\to\infty} M(\theta_n) + \epsilon, C\right)$$

since $M$ is upper semi-continuous and $\theta_n \to \theta$ as $\rho_n \to 0$. By letting $C \to -\infty$, we have

$$\liminf_{n\to\infty} M(\theta_n) \geq \limsup_{n\to\infty} M(\theta_n) + \epsilon > \limsup_{n\to\infty} M(\theta_n),$$

which is a contradiction. ■

This concludes the discussion on upper semi-continuity. Now, let's focus on the stronger assumption, i.e., (A3), since as we will soon see, the assumption used in Proposition 5.2.1 suffices to prove (A3).

Explicitly, for (A3), we want to show that for any $\epsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty} \sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \le \sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta)\right) = 1.$$

Replacing $\theta \notin B(\theta^*,\epsilon)$ by $\theta \in (B(\theta^*,\epsilon))^c$, where $(B(\theta^*,\epsilon))^c$ is compact as shown before. Now, consider a trivial open cover

$$(B(\theta^*,\epsilon))^c \subseteq \bigcup_{\theta\in(B(\theta^*,\epsilon))^c} B(\theta,\rho_\theta) \text{ for some } \rho_\theta > 0 \text{ for all } \theta \in (B(\theta^*,\epsilon))^c,$$

from the compactness of $(B(\theta^*,\epsilon))^c$, there exists a finite subset $K \subseteq (B(\theta^*,\epsilon))^c$ such that

$$(B(\theta^*,\epsilon))^c \subseteq \bigcup_{\theta\in K} B(\theta,\rho_\theta).$$

This implies that

$$\sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \le \max_{\theta\in K} \sup_{u\in B(\theta,\rho_\theta)} M_n(u) \le \max_{\theta\in K} \frac{1}{n}\sum_{i=1}^n \sup_{u\in B(\theta,\rho_\theta)} m_u(X_i) \overset{\text{a.s.}}{\to} \max_{\theta\in K} \mathbb{E}\left[\sup_{u\in B(\theta,\rho_\theta)} m_u(X)\right]$$

by the strong law of large number. However, we need to be careful here by combining $\overset{\text{a.s.}}{\to}$ and $\max_{\theta\in K}$.

**Claim.** We can indeed interchange the limit and the maximum.

**Proof.** Since the convergence happens for every $\theta \in K$ such that

$$\frac{1}{n}\sum_{i=1}^n \sup_{u\in B(\theta,\rho_\theta)} m_u(X_i(\omega)) \overset{\text{a.s.}}{\to} \mathbb{E}\left[\sup_{u\in B(\theta,\rho_\theta)} m_u(X(\omega))\right].$$

In other words, for all $\theta \in K$, there exists a null set $N_\theta$ such that the convergence holds for all $\omega \notin N_\theta$. With $|K| < \infty$, $\bigcup_{\theta\in K} N_\theta$ is a *countable* union of null sets, which is still a null set. Hence, $K$ might be countable if this is the only issue. However, this is not the only story.

Consider $|K| = 2$, or in particular, $K = \{\theta_1, \theta_2\}$. In this case, we can say that the maximum of the limits is the limits of the maximum since for these two sequences of convergence sequence, we're saying that for $\theta_i$, the limiting statement involves "there exists $n_i$ such that for all $n \ge n_{\theta_i}$, such that ...." Hence, if $K$ is finite, one can just take $n_0 \coloneqq \max_{\theta\in K} n_\theta$, which is guaranteed to be finite. If $K$ is countable, $n_0$ can be unbounded, causing problems. ⊛

Hence, with the above justification, taking limits on both sides we have

$$\limsup_{n\to\infty} \sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \le \max_{\theta\in K} \mathbb{E}\left[\sup_{u\in B(\theta,\rho_\theta)} m_u(X)\right].$$

With this, the upshot is that by using the same assumption used in Proposition 5.2.1, the above can be controlled carefully. Specifically, since $\theta \mapsto m_\theta(x)$ is upper semi-continuous for all $x$, we know that as $\rho \to 0$, $\sup_{u\in B(\theta,\rho)} m_u(X) \to m_\theta(X)$ for every $\theta \in \Theta$, as shown in Proposition 5.2.2. Hence,

$$M(\theta) = \mathbb{E}[m_\theta(X)] = \lim_{\rho\to 0} \mathbb{E}\left[\sup_{u\in B(\theta,\rho)} m_u(X)\right]$$

by the monotone convergence theorem as the sequence is the supremum over smaller and smaller sets. This implies that for all $\theta \in \Theta$ and $\delta > 0$, there exists $\rho_\theta > 0$ such that $\mathbb{E}[\sup_{u\in B(\theta,\rho_\theta)} m_u(X)] < M(\theta)+\delta$. Hence, combining this with the above, by choosing $\theta_\theta$ for every $\theta \in \Theta$ in this way, we have

$$\limsup_{n\to\infty} \sup_{\theta\notin B(\theta^*,\epsilon)} M_n(\theta) \le \max_{\theta\in K} \mathbb{E}\left[\sup_{u\in B(\theta,\rho_\theta)} m_u(X)\right] \le \max_{\theta\in K} M(\theta) + \delta \le \sup_{\theta\notin B(\theta^*,\epsilon)} M(\theta) + \delta$$

almost surely, which is exactly (A3).

**Intuition.** We can use a single assumption to prove both (A1) and (A3)!

# Lecture 26: Asymptotic Normality of $M$-Estimators

Let's summarize what we have done. 23 Apr. 9:30

**Theorem 5.2.3** (Wall). Given $X, (X_n) \overset{\text{i.i.d.}}{\sim} F$, a compact metric space $\Theta$, and $\theta \mapsto m_\theta(x)$ upper semi-continuous for all $x$. Define $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$, and let $\hat{\theta}_n$ to be the approximate maximizer of $M_n$.[a] If $M(\theta) = \mathbb{E}[m_\theta(X)]$ has a unique maximizer $\theta^*$ and $\mathbb{E}[\sup_{u\in B(\theta,\rho)} m_u(X)] < \infty$ for some small $\rho > 0$ and all $\theta \in \Theta$, then $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$.

[a] I.e., Equation 5.1: $M_n(\hat{\theta}_n) \geq M_n(\theta^*) - z_n$ for some $z_n \overset{\text{a.s.}}{\to} 0$.

**Proof.** From the strong law of large number, $M_n(\theta) \overset{\text{a.s.}}{\to} M(\theta) = \mathbb{E}[m_\theta(X)] \in [-\infty, \infty)$ for all $\theta \in \Theta$. Furthermore, since for some $\rho > 0$, $\mathbb{E}[\sup_{u\in B(\theta,\rho)} m_u(X)] < \infty$, Proposition 5.2.1 implies that $M$ is also upper semi-continuous as $m_\theta(x)$ is. Hence, as $\Theta$ is compact and the maximizer $\theta^*$ of $M$ is unique, Corollary 5.2.1 implies that (A1) is satisfied. Finally, as we discussed above, (A3) is also satisfied by the same set of assumptions, hence Theorem 5.2.2 proves that $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$. ■

Indeed, Theorem 5.2.3 helps us find the MLE of Cauchy that we previously don't know how to solve.

**Example** (Cauchy). Let $X \sim F = G_{\theta^*}$ with a family of cdfs $\{G_\theta \sim \text{Cauchy}(\theta, 1) : \theta \in \Theta = \overline{\mathbb{R}}\}$ where the pdf of $G_\theta$ is given by $g_\theta(x) = 1/\pi(1 + (x-\theta)^2)$. Consider the setup of MLE, i.e., $m_\theta(x) = \log(g_\theta(x))$, and let $\hat{\theta}_n := \arg\max_{\theta\in\Theta} M_n(\theta)$ and $\theta^* = \arg\max_{\theta\in\Theta} M(\theta)$, then $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$.

**Proof.** We first check the following.

- $\theta \mapsto m_\theta(x)$ is upper semi-continuous: clearly since $\theta \mapsto g_\theta(x)$ is continuous.
- $\Theta$ is compact: we consider the extended real line $\overline{\mathbb{R}} = [-\infty, \infty]$. However, we need to be careful to define $m_{\pm\infty}(x) = -\infty$ for all $x \in \mathbb{R}$, which indeed preserves the continuity of $m_\theta(x)$ since $g_\theta(x) \to 0$ as $\theta \to \pm\infty$.
- $\mathbb{E}[\sup_{u\in B(\theta,\rho)} m_u(X)] < \infty$ for small $\rho$: indeed, since $m_\theta(x) \leq -\log(\pi)$ for all $\theta \in \Theta$ and $x \in \mathbb{R}$.
- $M$ does have a unique maximizer: since Cauchy distribution is identifiable.

Hence, we can apply Theorem 5.2.3, and conclude the result. ⊛

Another example is the following.

**Example** (Uniform). Let $X \sim F = G_{\theta^*}$ with a family of cdfs $\{G_\theta \sim \mathcal{U}(0, \theta) : \theta > 0\}$. From the previous example, the MLE is given by $\hat{\theta}_n = \max_{1\leq i\leq n} X_i$. Applying Theorem 5.2.3 gives $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$.

## 5.3 Asymptotic Distributions

With Theorem 5.2.3, under reasonable assumptions, we can establish consistency for $M$-estimators. A further natural question is that whether we can obtain some asymptotic distribution for $M$-estimators as well. However, asymptotic normality is impossible in general.

**Example** (Uniform is not asymptotically normal). From the previous example, clearly we don't have $\sqrt{n}(\theta^* - \hat{\theta}_n) \overset{D}{\to} \mathcal{N}$ as we always have $\theta^* - \hat{\theta}_n \geq 0$.

More specifically, for the uniform example, the asymptotic distribution is the exponential.

**Proposition 5.3.1.** If $X_i \sim \mathcal{U}(0, \theta^*)$ with the family $\{\mathcal{U}(0,\theta) \colon \theta > 0\}$, then $n(\theta^* - \hat{\theta}_n) \xrightarrow{D} \text{Exp}(\theta^*)$.

**Proof.** We want to show that for all $x > 0$, $\mathbb{P}(n(\theta^* - \hat{\theta}_n) > x) \to e^{-x/\theta^*}$. As $\hat{\theta}_n = \max_{1 \le i \le n} X_i$,

$$\mathbb{P}(n(\theta^* - \hat{\theta}_n) > x) = \mathbb{P}\left(\hat{\theta}_n < \theta^* - \frac{x}{n}\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i < \theta^* - \frac{x}{n}\right) = \left(\mathbb{P}\left(X < \theta^* - \frac{x}{n}\right)\right)^n.$$

For $n$ large enough, $x/n < \theta^*$, then $\mathbb{P}(X < \theta^* - x/n) = (\theta^* - x/n)/\theta^*$. This further gives

$$\mathbb{P}(n(\theta^* - \hat{\theta}_n) > x) = \left(\frac{\theta^* - x/n}{\theta^*}\right)^n = \left(1 + \frac{-x/\theta^*}{n}\right)^n \to e^{-x/\theta^*}$$

as $n \to \infty$. ■

**Remark.** In general, we don't have asymptotic normality for $M$-estimators.

Hence, our next goal is to show that under what conditions we can establish the asymptotic normality. But firstly, we discuss the method of moments estimators, which turns out to be useful for later.

### 5.3.1 Method of Moments Estimators

Let $X, X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ and let $\{G_\theta \colon \theta \in \Theta \subseteq \mathbb{R}^m\}$ be a family of cdfs. Write $h(x) = (x, \ldots, x^m)$, then consider $T(\theta) = \mathbb{E}_\theta[h(X)]$ for $\theta \in \Theta \subseteq \mathbb{R}^m$, i.e.,

$$T(\theta) = (\mathbb{E}_\theta[X], \ldots, \mathbb{E}_\theta[X^m]).$$

If $T$ is one-to-one, we can then define the *method of moments (MOM) estimator* $\widetilde{\theta}_n \coloneqq T^{-1}(\overline{h(X)}_n)$ where $\overline{h(X)}_n = \frac{1}{n}\sum_{i=1}^n h(X_i)$.

**Note.** More generally, consider $h(x) = (h_1(x), \ldots, h_m(x))$ and $T(\theta) = (\mathbb{E}_\theta[h_1(X)], \ldots, \mathbb{E}_\theta[h_m(X)])$.

Clearly, from the strong law of large number, $\overline{h(X)}_n \xrightarrow{\text{a.s.}} \mathbb{E}_F[h(X)]$ as long as $\mathbb{E}_F[h(X)] < \infty$. Hence, if $T^{-1}$ is continuous at $\mathbb{E}_F[h(X)]$, continuous mapping theorem implies

$$\widetilde{\theta}_n = T^{-1}(\overline{h(X)}_n) \xrightarrow{p} T^{-1}(\mathbb{E}_F[h(X)]).$$

Moreover, since $\sqrt{n}(\overline{h(X)} - \mathbb{E}_F[h(X)]) \xrightarrow{D} \mathcal{N}(0, \text{Var}[h(X)])$ by the central limit theorem as long as the variance is well-defined. Hence, if $T^{-1}$ is differentiable at $\mathbb{E}_F[h(X)]$, by the Delta method,

$$\sqrt{n}(\widetilde{\theta}_n - T^{-1}(\mathbb{E}_F[h(X)])) \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

We see the following.

**Lemma 5.3.1.** If the model is well-specified, i.e., $F = G_{\theta^*}$ for some $\theta^* \in \Theta$, $\widetilde{\theta}_n \xrightarrow{p} \theta^*$.

**Proof.** Since $T^{-1}(\mathbb{E}_F[h(X)]) = T^{-1}(\mathbb{E}_{G_{\theta^*}}[h(X)]) = \theta^*$ from the definition of $T^{-1}$. ■

On the other hand, what if $F$ is not in $\{G_\theta \colon \theta \in \Theta\}$?

**Example** (Uniform). Let $F(u - \epsilon) < 1 = F(u)$ for all $\epsilon > 0$ such that $\mathbb{P}(0 \le X \le u) = 1$, i.e., $X$ has support $[0, u]$. Consider the family of cdfs $\{G_\theta \sim \mathcal{U}(0,\theta) \colon \theta > 0\}$, then $T(\theta) = \mathbb{E}_\theta[X] = \theta/2$ for $\theta > 0$, which is clearly invertible, and we can define $T^{-1}$ to get $\widetilde{\theta}_n = 2\overline{X}_n$.

Hence, from the strong law of large number, $\widetilde{\theta}_n \xrightarrow{\text{a.s.}} 2\mathbb{E}[X]$, so $\widetilde{\theta}_n$ is a strongly consistent estimator of $u$ as long as $\mathbb{E}[X] = u/2$. Furthermore, by the central limit theorem, we also have $\sqrt{n}(\widetilde{\theta}_n - 2\mathbb{E}[X]) \xrightarrow{D} \mathcal{N}(0, 4\,\text{Var}[X])$.

Comparing the above example and Proposition 5.3.1, we see that while the MOM estimator is asymptotically normal, the MLE is not. Moreover, they converge with different rates: specifically, $n(u - \hat{\theta}_n) \xrightarrow{D} \text{Exp}(u)$ for the MLE and $\sqrt{n}(\widetilde{\theta}_n - 2\mathbb{E}[X]) \xrightarrow{D} \mathcal{N}(0, 4\,\text{Var}[X])$ for the MOM estimator. We see that MLE is better than the MOM estimator since the rate is much faster ($n$ v.s. $\sqrt{n}$).

**Remark** (Consistent under symmetry). The above example for MOM holds whenever $X$ is symmetric about $u/2$ since in this case, $\mathbb{E}[X] = u/2$ also. That means, the MOM estimator is also consistent and asymptotically normal with a rate $\sqrt{n}$.

### 5.3.2 $M$-Estimators

Again, let $X, X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F$ such that $M_n(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i)$ where $\theta \in \Theta \subseteq \mathbb{R}^m$. There are two closely related problems we're going to address in this section.

**Problem.** How to find the approximate maximizer $\hat{\theta}_n$ of $M_n$ in practice?

**Answer.** If $\theta^* \in \text{int}(\Theta)$ and $\theta \mapsto M_n(\theta)$ is differentiable, then we can find $\hat{\theta}_n$ by searching the solution set of $\nabla M_n(\theta) =: \Psi_n(\theta) = 0$ via some numerical algorithms, e.g., the classical Newton-Raphson. ⊛

**As previously seen** (Newton-Raphson algorithm). Starting with some $\hat{\theta}_n^{(0)} \in \Theta$. Then for $k \geq 0$, consider updating $\hat{\theta}_n^{(k)}$ by

$$\hat{\theta}_n^{(k+1)} = \hat{\theta}_n^{(k)} - \left(\nabla \Psi_n(\hat{\theta}_n^{(k)})\right)^{-1} \Psi_n(\hat{\theta}_n^{(k)}).$$

However, we should note that typically, $\Psi_n(\theta)$ has many roots (sometimes even grows linearly with $n$) and in general, we don't know whether our numerical algorithm gives us the actual maximizer of $M_n(\theta)$. One might want to try different $\hat{\theta}_n^{(0)}$ and run the numerical algorithm multiple times.

**Intuition.** The up-shot is that we don't need to do this: if $\hat{\theta}_n^{(0)}$ is "good," then we're fine.

If we can find $\hat{\theta}_n$, with Theorem 5.2.3, we can have $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$. However, a follow-up question arises:

**Problem.** Under what condition can we establish asymptotic normality of $\hat{\theta}_n$?

**Answer.** In general, $\hat{\theta}_n$ is not asymptotically normal, e.g., the uniform example. On the other hand, we will show that if $\hat{\theta}_n^{(0)}$ is "good," or more specifically, if $\sqrt{n}(\hat{\theta}_n^{(0)} - \theta^*) = O_p(1)$, then by only one Newton-Raphson update, we can obtain a sequence of estimators $(\hat{\theta}_n^{(1)})$ that is asymptotically normal, hence also consistent. The above approach is the so-called *one-step MLE*. ⊛

# Lecture 27: Consistency with Computational Consideration

25 Apr. 9:30

To continue the discussion, we will need to establish a *uniform* version of the strong law of large number. In particular, say $X, (X_n) \overset{\text{i.i.d.}}{\sim} F$ and $h_\theta(\cdot) \in \mathbb{R}$ is given for every $\theta \in \Theta$. From the strong law of large number, $\frac{1}{n}\sum_{i=1}^n h_\theta(X_i) \overset{\text{a.s.}}{\to} \mathbb{E}[h_\theta(X)]$ for any fixed $\theta \in \Theta$. Now, we interested in whether

$$\frac{1}{n}\sum_{i=1}^n h_{\theta_n}(X_i) \overset{\text{a.s.}}{\to} \mathbb{E}[h_{\theta^*}(X)]$$

for some sequence $(\theta_n)$, either deterministic or random, such that $\theta_n \to \theta^* \in \Theta$ (or $\theta_n \overset{\text{a.s.}}{\to} \theta^*$).

**Theorem 5.3.1** (Uniform strong law of large number). Let $X, (X_n) \overset{\text{i.i.d.}}{\sim} F$ and $h_\theta(x) \in \mathbb{R}$ be given. Furthermore, let $(\theta_n)$ be a sequence such that $\theta_n \to \theta^*$. If $\theta \mapsto h_\theta(x)$ is continuous for all $x$ and $\mathbb{E}[\sup_{\theta \in \overline{B(\theta^*, \rho)}} |h_\theta(X)|] < \infty$ for some $\rho > 0$, then

$$\frac{1}{n}\sum_{i=1}^n h_{\theta_n}(X_i) \overset{\text{a.s.}}{\to} \mathbb{E}[h_{\theta^*}(X)].$$

**Proof.** The main idea is to apply Proposition 5.2.1 twice with $h$ and $-h$, and conclude

$$\sup_{\theta \in \overline{B(\theta^*, \rho)}} \left| \frac{1}{n} \sum_{i=1}^{n} h_\theta(X_i) - \mathbb{E}[h_\theta(X)] \right| \to 0.$$

This will conclude the proof since we can write

$$\left| \frac{1}{n} \sum_{i=1}^{n} h_{\theta_n}(X_i) - \mathbb{E}[h_{\theta^*}(X)] \right| \leq \left| \frac{1}{n} \sum_{i=1}^{n} h_{\theta_n}(X_i) - \mathbb{E}[h_{\theta_n}(X)] \right| + \left| \mathbb{E}[h_{\theta_n}(X)] - \mathbb{E}[h_{\theta^*}(X)] \right|.$$

Firstly, the second term goes to 0 as $n \to \infty$ if $\theta \mapsto \mathbb{E}[h_\theta(X)]$ is continuous at $\theta^*$. Indeed, since $\theta \mapsto h_\theta(x)$ is continuous for all $x$, $\theta \mapsto \mathbb{E}[h_\theta(X)]$ is also continuous at $\theta^*$. On the other hand,

$$\left| \frac{1}{n} \sum_{i=1}^{n} h_{\theta_n}(X_i) - \mathbb{E}[h_{\theta_n}(X)] \right| \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} h_\theta(X_i) - \mathbb{E}[h_\theta(X)] \right|,$$

and since we only care about $n \to \infty$, we may assume that for some $\rho > 0$,

$$\mathbb{P}(\theta_n \in B(\theta^*, \rho) \text{ for all } n \text{ large enough}) = 1.$$

Hence, we can replace $\Theta$ by $B(\theta^*, \rho)$, which gives

$$\left| \frac{1}{n} \sum_{i=1}^{n} h_{\theta_n}(X_i) - \mathbb{E}[h_{\theta_n}(X)] \right| \leq \sup_{\theta \in B(\theta^*, \rho)} \left| \frac{1}{n} \sum_{i=1}^{n} h_\theta(X_i) - \mathbb{E}[h_\theta(X)] \right|$$

for all large enough $n$ with probability 1. This goes to 0 as we have shown in the beginning. ■

Before resuming our discussion on asymptotic normality of $M$-estimators, let's recall our motivation.

**As previously seen.** Let $M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) \overset{\text{a.s.}}{\to} M(\theta) = \mathbb{E}[m_\theta(X)]$ for all $\theta \in \Theta$, and let $\hat{\theta}_n$ be the maximizer of $\theta \mapsto M_n(\theta)$. From Theorem 5.2.3, $\hat{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$ under assumptions:

- $\Theta$ is compact;
- $\theta \mapsto m_\theta(x)$ is upper semi-continuous for all $x$;
- $\theta^*$ is the unique maximizer of $\theta \mapsto M(\theta)$;
- $\mathbb{E}[\sup_{\theta \in B(\theta^*, \rho)} m_\theta(X)] < \infty$ for some $\rho > 0$,

While $\hat{\theta}_n$ is strongly consistent, we have seen that $\hat{\theta}_n$ doesn't need to be asymptotically normal, e.g., the uniform example. Furthermore, it's also unclear how to compute $\hat{\theta}_n$ in practice.

Toward this end, we will need additional assumptions in order to solve these two problems. Firstly, assume further that $\Theta \subseteq \mathbb{R}^m$ with $\theta^* \in \text{int}(\Theta)$,[4] and $\theta \mapsto m_\theta(x) \in C^2$ for any $x$.

**Notation.** The class of second-differentiable functions is denoted as $C^2$.

Furthermore, let $\psi_\theta(x) \coloneqq \nabla_\theta m_\theta(x)$ for all $x$ and assume

$$\mathbb{E}\left[ \sup_{\theta \in \overline{B(\theta^*, \rho)}} \|\nabla \psi_\theta(X)\|_{\max} \right] < \infty,$$

where the *max norm* is defined as $\|A\|_{\max} = \max_{i,j} |a_{ij}|$.

**Remark.** Since $\theta^* \in \text{int}(\Theta)$, $\theta^*$ is the maximizer of $M(\theta) = \mathbb{E}[m_\theta(X)]$.

[4] Before this, for consistency discussion, we can work with a general metric space.

From our assumption, $M$ is also differentiable, and $\theta^*$ is a root of

$$\nabla M(\theta) = \nabla \mathbb{E}[m_\theta(X)] = \mathbb{E}[\nabla m_\theta(X)] = \mathbb{E}[\psi_\theta(X)],$$

implying $\mathbb{E}[\psi_{\theta^*}(X)] = 0$.

**Note.** We can interchange $\nabla$ and $\mathbb{E}$ from our assumption on $\|\nabla \psi_\theta(X)\|_{\max}$.

Finally, assume that $\mathrm{Var}[\psi_{\theta^*}(X)] = J_*$ and $\mathbb{E}[\nabla \psi_\theta(X)] =: -J(\theta)$ exists for all $\theta \in B(\theta^*, \rho)$ for some $\rho > 0$ such that they are both invertible, and in particular, positive semi-definite.

**Note.** $J_*$ doesn't equal to $J(\theta^*)$ in general.

**Example.** Under the framework of MLE and with our assumptions, $J(\theta^*) = J_*$. This is something commonly used in the calculation of Cramér-Rao lower bound.

With these notations and new assumptions, we answer a slightly different version of the first problem.

**As previously seen.** In Theorem 5.2.3, we're searching for the (approximate) maximizer $\hat{\theta}_n$ over the entire space $\Theta$, hence it's possible that $\hat{\theta}_n$ isn't the actual root of $\nabla M_n(\theta)$.

Since in practice, when maximizing $M_n$, we will find the root of the derivative of $M_n$, i.e.,

$$\Psi_n(\theta) \coloneqq \nabla M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla m_\theta(X_i).$$

We want to know that if we restrict our attention to roots of $\Psi_n(\theta)$, can we still establish consistency?

**Problem.** Does there exist a sequence of roots $(\widetilde{\theta}_n)$ of $\Psi_n(\theta)$ that is consistent with $\theta^*$?

**Answer.** For every $x$, for some $\rho > 0$, consider for all $\theta \in \overline{B(\theta^*, \rho)}$ with the Taylor expansion

$$m_\theta(x) - m_{\theta^*}(x) = \psi_{\theta^*}(x)(\theta - \theta^*) + (\theta - \theta^*)^\top \left( \int_0^1 \int_0^1 \nabla \psi_{\theta^* + uv(\theta - \theta^*)}(xuv) u \,\mathrm{d}u \,\mathrm{d}v \right) (\theta - \theta^*).$$

Then, we apply Theorem 5.2.3 with $m_\theta(x) - m_{\theta^*}(x)$ by considering the metric space being $\overline{B(\theta^*, \rho)}$ instead of $\Theta$, i.e., we only need to check $\mathbb{E}[\sup_{\theta \in B(\theta^*, \rho)} m_\theta(X)] < \infty$, or equivalently,

$$\begin{aligned}
&\mathbb{E}\left[ \sup_{\theta \in B(\theta^*, \rho)} (m_\theta(X) - m_{\theta^*}(X)) \right] \\
&= \mathbb{E}\left[ \sup_{\theta \in B(\theta^*, \rho)} \psi_{\theta^*}(X)(\theta - \theta^*) + (\theta - \theta^*)^\top \left( \int_0^1 \int_0^1 \nabla \psi_{\theta^* + uv(\theta - \theta^*)}(Xuv) u \,\mathrm{d}u \,\mathrm{d}v \right) (\theta - \theta^*) \right] < \infty.
\end{aligned}$$

This is true since $\mathbb{E}[\psi_{\theta^*}(X)] = 0$ and $\mathbb{E}[\sup_{\theta \in \overline{B(\theta^*, \rho)}} \|\nabla \psi_\theta(X)\|_{\max}] < \infty$. Hence, there exists a maximizer $\widetilde{\theta}_n$ of $M_n(\theta) - M_n(\theta^*)$ such that $\widetilde{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$ by Theorem 5.2.3.

**Claim.** $\widetilde{\theta}_n$ is a root of $\Psi_n(\theta)$ for large enough $n$ with probability 1.

**Proof.** As $\widetilde{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$, for large $n$ and with probability 1, $\widetilde{\theta}_n$ will not be on the boundary of $\overline{B(\theta^*, \rho)}$, i.e., it's indeed a root of $\Psi_n(\theta)$ since if $\widetilde{\theta}_n$ is in the interior and is a maximizer, the first-order condition gives $\nabla M_n(\widetilde{\theta}_n) = \Psi_n(\widetilde{\theta}_n) = 0$. ⊛

Since $\widetilde{\theta}_n \overset{\text{a.s.}}{\to} \theta^*$ and $\widetilde{\theta}_n$'s are roots of $\Psi_n(\theta)$ for large enough $n$ with probability 1, we're done. ⊛

Therefore, we see that there indeed exists a sequence of roots $(\widetilde{\theta}_n)$ of $\Psi_n(\theta)$ that is consistent of $\theta^*$. Next, we answer the question about asymptotic normality, but for $\widetilde{\theta}_n$ this time.

**Problem.** Let $(\widetilde{\theta}_n)$ be a sequence of roots of $\Psi_n(\theta)$ such that $\widetilde{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$. Is $\widetilde{\theta}_n$ asymptotically normal?

**Answer.** By the first-order Taylor expansion,

$$\Psi_n(\widetilde{\theta}_n) = \frac{1}{n}\sum_{i=1}^n \psi_{\widetilde{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^n \psi_{\theta^*}(X_i) + \frac{1}{n}\sum_{i=1}^n \left(\int_0^1 \nabla\psi_{\theta^*+u(\widetilde{\theta}_n-\theta^*)}(X_i)\,\mathrm{d}u\right)(\widetilde{\theta}_n-\theta^*),$$

which is 0 for $n$ large enough since then $\Psi_n(\widetilde{\theta}_n)=0$. Multiplying both sides by $\sqrt{n}$, we have

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta^*}(X_i) = \sqrt{n}(\widetilde{\theta}_n-\theta^*)\cdot\frac{1}{n}\sum_{i=1}^n \left(\int_0^1 \nabla\psi_{\theta^*+u(\widetilde{\theta}_n-\theta^*)}(X_i)\,\mathrm{d}u\right).$$

By the multivariate central limit theorem, the left-hand side converges to $\mathcal{N}(0, J_*)$ where $J_* = \mathrm{Var}[\psi_{\theta^*}(X)]$. For the right-hand side, we see the following.

**Intuition.** If we can ignore $u(\widetilde{\theta}_n-\theta^*)$, then by the strong law of large number, we have

$$\frac{1}{n}\sum_{i=1}^n \int_0^1 \nabla\psi_{\theta^*}(X_i)\,\mathrm{d}u \xrightarrow{\text{a.s.}} \mathbb{E}[\nabla\psi_{\theta^*}(X)] = -J(\theta^*).$$

Indeed, by the uniform strong law of large number, with $\widetilde{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$, this can be rigorously justified. Combining everything together, we see that

$$\sqrt{n}(\widetilde{\theta}_n-\theta^*) \xrightarrow{D} -J(\theta^*)^{-1}\mathcal{N}(0, J_*),$$

hence we conclude that $\widetilde{\theta}_n$ is asymptotically normal. ⊛

The question now becomes how to obtain $\widetilde{\theta}_n$ in practice. While our numerical algorithm can find roots, we're not guaranteed to get $\widetilde{\theta}_n$ since there might be multiple roots. However, if we initialize our algorithm with some $\check{\theta}_n$ "close enough" to $\theta^*$, we're fine.

**Claim.** Let $(\check{\theta}_n)$ be a sequence such that $\sqrt{n}(\check{\theta}_n-\theta^*) = O_p(1)$. Then with one Newton-Raphson update, i.e., $\hat{\theta}_n \coloneqq \check{\theta}_n - \left(\nabla\Psi_n(\check{\theta}_n)\right)^{-1}\Psi_n(\check{\theta}_n)$, we have $\sqrt{n}(\hat{\theta}_n-\widetilde{\theta}_n) \xrightarrow{p} 0$.

**Proof.** Firstly, consider writing

$$\sqrt{n}(\hat{\theta}_n-\widetilde{\theta}_n) = \sqrt{n}(\hat{\theta}_n-\check{\theta}_n) + \sqrt{n}(\check{\theta}_n-\widetilde{\theta}_n).$$

By the definition, the first term is $-\sqrt{n}\left(\nabla\Psi_n(\check{\theta}_n)\right)^{-1}\Psi_n(\check{\theta}_n)$, which can be written as

$$-\sqrt{n}\left(\nabla\Psi_n(\check{\theta}_n)\right)^{-1}\left[\Psi_n(\widetilde{\theta}_n) + \left(\int_0^1 \nabla\Psi_n(\widetilde{\theta}_n+u(\check{\theta}_n-\widetilde{\theta})_n)\,\mathrm{d}u\right)(\check{\theta}_n-\widetilde{\theta}_n)\right],$$

and with $\Psi_n(\widetilde{\theta}_n)=0$, we then have

$$\sqrt{n}(\hat{\theta}_n-\check{\theta}_n) = -\sqrt{n}\underbrace{\left(\nabla\Psi_n(\check{\theta}_n)\right)^{-1}\left(\int_0^1 \nabla\Psi_n(\widetilde{\theta}_n+u(\check{\theta}_n-\widetilde{\theta})_n)\,\mathrm{d}u\right)}_{\coloneqq B_n}(\check{\theta}_n-\widetilde{\theta}_n) =: \sqrt{n}B_n(\check{\theta}_n-\widetilde{\theta}_n).$$

Hence, $\sqrt{n}(\hat{\theta}_n-\widetilde{\theta}_n) = (I-B_n)\sqrt{n}(\check{\theta}_n-\widetilde{\theta}_n)$. To show that this is $o_p(1)$, we first see that $\sqrt{n}(\check{\theta}_n-\widetilde{\theta}_n) = O_p(1)$ since $\widetilde{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$ and $\sqrt{n}(\check{\theta}_n-\theta^*)=O_p(1)$. On the other hand, for $I-B_n$, from the dominated convergence theorem,

$$\int_0^1 \nabla\Psi_n(\widetilde{\theta}_n+u(\check{\theta}_n-\widetilde{\theta}_n))\,\mathrm{d}u \to \mathbb{E}[\nabla\psi_{\theta^*}(X)],$$

and by the uniform strong law of large number,

$$\nabla\Psi_n(\check{\theta}_n) \to \mathbb{E}[\nabla\Psi_n(\theta^*)] = \mathbb{E}[\nabla\psi_{\theta^*}(X)],$$

which implies $(\nabla\Psi_n(\check{\theta}_n))^{-1} \to (\mathbb{E}[\nabla\psi_{\theta^*}(X)])^{-1}$ since inverse function is continuous. Combining both, we see that $B_n \to (\mathbb{E}[\nabla\psi_{\theta^*}(X)])^{-1}\mathbb{E}[\nabla\psi_{\theta^*}(X)] = I$, which gives

$$\sqrt{n}(\hat{\theta}_n - \widetilde{\theta}_n) = (I - B_n)\sqrt{n}(\check{\theta}_n - \widetilde{\theta}_n) = o(1)O_p(1) = o_p(1),$$

i.e., $\sqrt{n}(\hat{\theta}_n - \widetilde{\theta}_n) \xrightarrow{p} 0$. ⊛

# Appendix

# Bibliography


[Das08] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Science & Business Media, Feb. 6, 2008. 727 pp. ISBN: 978-0-387-75971-5. Google Books: sX4_AAAAQBAJ.

[Fer17] Thomas S. Ferguson. *A Course in Large Sample Theory*. Routledge, Sept. 6, 2017. 140 pp. ISBN: 978-1-351-47005-6. Google Books: clc0DwAAQBAJ.

[Leh04] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, Aug. 27, 2004. 640 pp. ISBN: 978-0-387-98595-4. Google Books: geIoxvgTXlEC.

[Ser09] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Sept. 25, 2009. 399 pp. ISBN: 978-0-470-31719-8. Google Books: enUouJ4EHzQC.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D (visited on 10/17/2023).