

STAT575  
Large Sample Theory

Pingbang Hu

April 16, 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: *Asymptotic 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

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Parametrized Approach . . . . .	2
1.2	Hypothesis Testing . . . . .	3
<b>2</b>	<b>Modes of Convergence</b>	<b>4</b>
2.1	Different Modes of Convergence . . . . .	4
2.2	Weak Convergence . . . . .	7
2.3	Convergence in Distribution . . . . .	13
2.4	Stochastic Boundedness . . . . .	16
2.5	Skorohod's Representation Theorem . . . . .	19
2.6	Characteristic Function . . . . .	21
<b>3</b>	<b>Fundamental Theorems of Probability</b>	<b>28</b>
3.1	Law of Large Number and Central Limit Theorem . . . . .	28
3.2	Inference for Population Mean and Variance . . . . .	30
3.3	Testing Normality . . . . .	33
3.4	A Quick Detour . . . . .	38
3.5	Inference for Population Quantiles . . . . .	40
3.6	Inference for Distribution Function . . . . .	45
<b>4</b>	<b>Lindeberg Central Limit Theorem</b>	<b>50</b>
4.1	Lindeberg Central Limit Theorem . . . . .	50
4.2	The Two-Sample Problem . . . . .	56
4.3	$U$ -Statistics . . . . .	61
4.4	Asymptotic Relative Efficiency of Tests . . . . .	65
4.5	Linear Rank Statistics . . . . .	70
<b>5</b>	<b>Maximum Likelihood Method and <math>M</math>-Estimation</b>	<b>77</b>
5.1	Maximum Likelihood Estimator . . . . .	77

# Chapter 1

## Introduction

### Lecture 1: Introduction to Large Sample Theory

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  $(\Omega, \mathcal{F}, \mathbb{P})$ . 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

16 Jan. 9:30

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

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  $(\Omega, \mathcal{F}, \mathbb{P})$  is the **Lebesgue-Stieltjes measure**  $\mu_F$  induced by the distribution function  $F$ . When doing integration, we will often denote

$$d\mu_F(x) = d\mathbb{P}(x) =: F(dx) =: dF(x) =: f(x)dx$$

**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  $n \rightarrow \infty$ .

### 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 “closest”, 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  $n \rightarrow \infty$ . 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  $n \rightarrow \infty$ . 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  $H_0: \mu = 0$ .

**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

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

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  $\text{Var}[X] < \infty$ .

**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  $\text{Var}[X]$  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  $H_0$ ,  $\sum_{i=1}^n \mathbb{1}_{X_i > 0} \sim \text{Bin}(n, 1/2)$ , 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

#### 2.1 Different Modes of Convergence

18 Jan. 9:30

Given a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , consider a sequence of  $d$ -dimensional random vectors  $(X_n)$  and a random vector  $X$ , i.e.,  $X_n, X: \Omega \rightarrow \mathbb{R}^d$ . 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  $X_n \rightarrow X$ , if  $X_n(\omega) \rightarrow X(\omega)$  for all  $\omega \in \Omega$ .<sup>a</sup>

<sup>a</sup>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  $X_n \xrightarrow{\text{a.s.}} X$ , if  $\mathbb{P}(X_n \rightarrow X) = 1$ .<sup>a</sup>

<sup>a</sup>I.e.,  $X_n(\omega) \rightarrow X(\omega)$  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  $X_n \xrightarrow{p} X$ , if for every  $\epsilon > 0$ ,  $\mathbb{P}(\|X_n - X\| > \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ .

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

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  $X_n \xrightarrow{\text{comp}} X$ , 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  $X_n \xrightarrow{L^p} X$ , if  $\mathbb{E}[\|X_n - X\|^p] \rightarrow 0$  as  $n \rightarrow \infty$ .

##### 2.1.1 Connection Between Modes of Convergence

We have the following connections between different modes of convergence.

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

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

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

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

From [Proposition 2.1.1](#), it's clear that  $\xrightarrow{\text{a.s.}}$  implies  $\xrightarrow{p}$  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,  $\xrightarrow{\text{comp}}$  implies  $\xrightarrow{\text{a.s.}}$  follows from the third equivalence. Lastly, the [convergence in  \$L^p\$](#)  implies the [convergence in probability](#) since

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

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

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

**Theorem 2.1.2 (Scheffé's theorem).** If  $X_n \xrightarrow{p} X$  and  $\limsup_{n \rightarrow \infty} \mathbb{E} [\|X_n\|^p] \leq \mathbb{E} [\|X\|^p] < \infty$ , then  $X_n \xrightarrow{L^p} X$ .

### 2.1.2 Consistent Estimator

Let  $(X_n) \stackrel{\text{i.i.d.}}{\sim} F$  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  $X_1, \dots, X_n$ , we estimate  $\theta$  by computing an estimator  $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$  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  $\hat{\theta}_n \xrightarrow{p} \theta$  as  $n \rightarrow \infty$ .

**Definition 2.1.7 (Strongly consistent).**  $\hat{\theta}_n$  is *strongly consistent* of  $\theta$  if  $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta$  as  $n \rightarrow \infty$ .

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

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

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

**Example (Mean estimation).** 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

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

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

$X$  has a pdf  $f$ , then

$$\mathbb{E}[X] = \int_0^\infty xf(x) dx = \int_0^\infty xF(dx).$$

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

On the other hand, if  $\text{Var}[X] < \infty$ , then  $\bar{X}_n \xrightarrow{L^2} \mathbb{E}[X]$ , which further implies  $\bar{X}_n \xrightarrow{p} \mathbb{E}[X]$ , hence  $\hat{\theta}_n$  is [consistent](#).<sup>a</sup>

<sup>a</sup>The latter is true even when  $\text{Var}[X] = \infty$  as we expect.

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

$$\frac{\text{Var}[X]}{n} = \text{Var}[\bar{X}_n] = \mathbb{E}[(\bar{X}_n - \mathbb{E}[X])^2] \rightarrow 0$$

as  $n \rightarrow \infty$ , which implies  $\bar{X}_n \xrightarrow{p} \mathbb{E}[X]$ . \*

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  $\mathbb{P}(X \leq \theta - \epsilon) = F(\theta - \epsilon)$  and  $F(\theta) = \mathbb{P}(X \leq \theta)$ .<sup>a</sup> Then  $\hat{\theta}_n = \max_{1 \leq i \leq n} X_i$  is indeed a [strongly consistent estimator](#) of  $\theta$ .

<sup>a</sup>Such a distribution exists, for example,  $\mathcal{U}(0, \theta)$ .

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

$$\begin{aligned} \mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) &= \mathbb{P}(\hat{\theta}_n > \theta + \epsilon) + \mathbb{P}(\hat{\theta}_n < \theta - \epsilon) \\ &= \mathbb{P}\left(\bigcup_{i=1}^n \{X_i > \theta + \epsilon\}\right) + \mathbb{P}\left(\bigcap_{i=1}^n \{X_i < \theta - \epsilon\}\right) \\ &\leq \sum_{i=1}^n \underbrace{\mathbb{P}(X_i > \theta + \epsilon)}_0 + \prod_{i=1}^n \mathbb{P}(X_i < \theta - \epsilon) = (\mathbb{P}(X_1 < \theta - \epsilon))^n \leq (F(\theta - \epsilon))^n \rightarrow 0 \end{aligned}$$

as  $n \rightarrow \infty$  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  $\hat{\theta}_n \rightarrow \theta$ .

**Example (Mean estimation for Gaussian).** Suppose  $X \sim \mathcal{N}(\theta, 1)$ . Then  $\hat{\theta}_n = \bar{X}_n \sim \mathcal{N}(\theta, 1/n)$ , 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,  $h: \mathbb{R}^d \rightarrow \mathbb{R}^m$ , and  $c \in \mathbb{R}^d$  constant.

(a) If  $X_n \rightarrow c$ , then  $h(X_n) \rightarrow h(c)$  if  $h$  is continuous at  $c$ .<sup>a</sup> This also holds for  $\xrightarrow{\text{a.s.}}$  and  $\xrightarrow{p}$ .

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

<sup>a</sup>This is an if and only if condition if this holds for any  $h$ .

Let's see some examples.



**Example.** If  $d = 1$ , and  $X_n \rightarrow \theta \neq 0$ . Then  $1/X_n \rightarrow 1/\theta$  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  $X_n \rightarrow X$  and  $Y_n \rightarrow Y$ , then  $(X_n, Y_n) \rightarrow (X, Y)$ .<sup>a</sup> The same holds for  $\xrightarrow{\text{a.s.}}$  and  $\xrightarrow{p}$ .

<sup>a</sup>The converse is also true since projections are continuous.

**Proof.**  $\|(X_n, Y_n) - (X, Y)\| \rightarrow 0$  since  $\|(X_n, Y_n) - (X, Y)\| \leq \|X_n - X\| + \|Y_n - Y\|$  for all  $n \geq 1$ .<sup>a</sup> The latter two terms go to 0 (in whatever sense) by assumption.  $\circledast$

<sup>a</sup>This can be seen from  $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ .

## Lecture 3: Weak Convergence Portmanteau Theorem

### 2.2 Weak Convergence

25 Jan. 9:30

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  $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$  (and  $(\Omega, \mathcal{F}, \mathbb{P})$ ) 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

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

The above makes sense even if  $X$  and  $Y$  are defined on different probability spaces, e.g., in our situation, consider a sequence of 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  $\text{TV}(X_n, X) \rightarrow 0$  as  $n \rightarrow \infty$ .

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

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

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

**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  $\theta_n \rightarrow \theta$ , then  $X_n \xrightarrow{\text{TV}} X \sim f_{\theta}$ . For example, if  $X_n \sim \text{Pois}(\theta_n)$  and  $\theta_n \rightarrow \theta$ , then  $X_n \xrightarrow{\text{TV}} X \sim \text{Pois}(\theta)$ .

#### 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  $X_n \xrightarrow{w} X$ , if for all bounded continuous  $g: \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\mathbb{E}_n[g(X_n)] \rightarrow \mathbb{E}[g(X)]$ .

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

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

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)  $X_n \xrightarrow{w} X$ .
- (b)  $\mathbb{E}_n[g(X_n)] \rightarrow \mathbb{E}[g(X)]$  for all bounded Lipschitz  $g: \mathbb{R}^d \rightarrow \mathbb{R}$ .
- (c)  $\mathbb{P}(X \in A) \leq \liminf_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A)$  for all  $A \subseteq \mathbb{R}^d$  open.
- (d)  $\mathbb{P}(X \in A) \geq \limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A)$  for all  $A \subseteq \mathbb{R}^d$  closed.
- (e)  $\mathbb{P}_n(X_n \in A) \rightarrow \mathbb{P}(X \in A)$  for all  $A \in \mathcal{F}$  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  $x, y \in S$ ,

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

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

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

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

$$\rho(x, A) \leq \rho(x, x_n) + \rho(x_n, A) \rightarrow 0,$$

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  $\rho(x, y_n) \rightarrow \rho(x, A) = 0$ , i.e.,  $x \in \bar{A}$ .  $\circledast$

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  $g_k: S \rightarrow \mathbb{R}$ , decreasing in  $k$  such that  $g_k(x) \searrow \mathbb{1}_A(x)$ .

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

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

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,

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

(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  $A$ ,

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

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

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

from our assumption (b). Finally, as  $k \rightarrow \infty$ , it goes to  $\mathbb{E}[\mathbb{1}_A(X)] = \mathbb{P}(X \in A)$  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.<sup>a</sup>

<sup>a</sup>Note that  $I$  can be uncountable.

**Proof.** It suffices to show  $|I_k| < \infty$  where  $I_k := \{i \in I : \mathbb{P}(A_i) \geq 1/k\}$  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  $(a) \Rightarrow (b) \Rightarrow (d) \Leftrightarrow (c)$ .

**Claim.**  $(c) + (d) \Rightarrow (e)$ .

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

$$\begin{aligned} \mathbb{P}(X \in A^\circ) &\leq \liminf_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A^\circ) \leq \liminf_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A) \\ &\leq \limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A) \leq \limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in \bar{A}) \leq \mathbb{P}(X \in \bar{A}) \end{aligned}$$

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

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

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

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

$$\text{and } \mathbb{P}(X \in A^\circ) \leq \mathbb{P}(X \in A) \leq \mathbb{P}(X \in \bar{A}), \mathbb{P}(X \in A) = \lim_{n \rightarrow \infty} \mathbb{P}_n(X_n \in A). \quad \circledast$$

Finally, we prove the following.

**Claim.**  $(e) \Rightarrow (a)$ .

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

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

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  $t \in [0, K]$ ,

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

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  $A := \{g > t\}$ , it suffices to show  $\mathbb{P}(X \in \partial\{g > t\}) = 0$  for almost all  $t$ . Firstly,

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

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. ◻

<sup>a</sup>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  $X_n \xrightarrow{w} X$ , we want to know when will  $h(X_n) \xrightarrow{w} h(X)$ . To develop the theorem of this, we need some more facts about metric spaces.

**As previously seen.** Given two metric spaces  $(S, \rho)$ ,  $(S', \rho')$ ,  $g: S \rightarrow S'$  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  $g: S \rightarrow S'$  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  $g: S \rightarrow S'$ , we denote the *continuous set* as  $C_g := \{x \in S: g \text{ is continuous at } x\}$ .

Then we have the following.

**Proposition 2.2.1.** Given  $g: S \rightarrow S'$  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  $X_n \xrightarrow{w} X$  and  $h: \mathbb{R}^d \rightarrow \mathbb{R}^m$ . If  $\mathbb{P}(X \in C_h) = 1$ , then  $h(X_n) \xrightarrow{w} h(X)$ .

**Proof.** Let  $A \subseteq \mathbb{R}^m$  be a closed set. Then from [Portmanteau theorem \(d\)](#), we need to show

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n(h(X_n) \in A) \leq \mathbb{P}(h(X) \in A).$$

Since  $\limsup_{n \rightarrow \infty} \mathbb{P}_n(h(X_n) \in A) = \limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in \{h \in A\})$ , implying

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n(h(X_n) \in A) \leq \limsup_{n \rightarrow \infty} \mathbb{P}_n(X_n \in \overline{\{h \in A\}}) \leq \mathbb{P}(X \in \overline{\{h \in A\}}),$$

where the last inequality follows again from [Portmanteau theorem \(d\)](#) since  $\overline{\{h \in A\}}$  is clearly closed and  $X_n \xrightarrow{w} X$ . Finally, as  $\mathbb{P}(X \in C_h) = 1$ ,

$$\mathbb{P}(X \in \overline{\{h \in A\}}) = \mathbb{P}(X \in \overline{\{h \in A\}} \cap C_h) \leq \mathbb{P}(X \in \{h \in \overline{A}\})$$

from [Proposition 2.2.1](#), i.e.,

$$\limsup_{n \rightarrow \infty} \mathbb{P}_n(h(X_n) \in A) \leq \mathbb{P}(X \in \{h \in \overline{A}\}) = \mathbb{P}(X \in \{h \in A\}) = \mathbb{P}(h(X) \in A)$$

since  $A$  is closed, hence we're done. ■

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

**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.<sup>a</sup>  $\circledast$

<sup>a</sup>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  $\epsilon > 0$ ,  $\mathbb{P}_n(\|X_n - Y_n\| > \epsilon) \rightarrow 0$  as  $n \rightarrow \infty$ . 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  $\epsilon > 0$ ,<sup>a</sup>

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

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

<sup>a</sup>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  $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n) = (\Omega, \mathcal{F}, \mathbb{P})$ .

**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](#).

**Proof of Theorem 2.2.3.** From [Portmanteau theorem \(b\)](#), we want to prove that  $\mathbb{E}_n[g(Y_n)] \rightarrow \mathbb{E}[g(X)]$  for all bounded and Lipschitz  $g: \mathbb{R}^d \rightarrow \mathbb{R}$ . 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

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

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

Another characterization regards the difference between marginal and joint [weak convergence](#).

1 Feb. 9:30

**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  $\xrightarrow{a.s.}$ .

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

However, this does not imply  $(X_n, Y_n) \xrightarrow{w} (X, X)$  since otherwise, by **continuous mapping theorem**,  $X_n + Y_n \xrightarrow{w} X + X = 2X$ , 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  $X_n \xrightarrow{w} X$  in  $\mathbb{R}^d$  and  $Y_n \xrightarrow{P} c$  in  $\mathbb{R}^m$ ,<sup>a</sup> then  $(X_n, Y_n) \xrightarrow{w} (X, c)$ .

<sup>a</sup>Recall that from **Corollary 2.2.1**, for a constant  $c$ , **weak convergence** is equivalent to **convergence in probability**.

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

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

**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  $(t_1, \dots, t_d) \in C_{F_X}$ ,

$$F_{X_n}(t_1, \dots, t_d) \rightarrow F_X(t_1, \dots, t_d).$$

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

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

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

**Proof.** Since  $X_n \xrightarrow{TV} X$  means  $\mathbb{P}_n(X_n \in A) \rightarrow \mathbb{P}(X \in A)$  uniformly in  $A$ , but  $X_n \xrightarrow{D} X$  only requires the above holds for  $A$  in the form of  $(-\infty, t_1] \times \dots \times (-\infty, t_d]$ , 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  $n \rightarrow \infty$ ,

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

**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

$$f_n \rightarrow f \Leftrightarrow X_n \xrightarrow{\text{TV}} X \Leftrightarrow X_n \xrightarrow{D} X.$$

**Proof.** The forward implications are clear, so we just need to show  $X_n \xrightarrow{D} X$  implies  $f_n \rightarrow f$ . 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  $t \pm \epsilon \in C_X$ ,  $X_n \xrightarrow{D} X$  implies  $\mathbb{P}_n(X_n \leq t + \epsilon) \rightarrow \mathbb{P}(X \leq t + \epsilon)$ . The same holds for  $t - \epsilon$ , hence

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

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  $X_n \xrightarrow{D} c$ , we don't have to show  $\mathbb{P}_n(X_n \leq c) \rightarrow \mathbb{P}(X \leq c) = 1$ . 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  $X_n \xrightarrow{D} X$  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,  $X_n \xrightarrow{D} X$  is equivalent as

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

### 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  $\mathbb{R}^d$ ,  $X_n \xrightarrow{w} X$  if and only if  $X_n \xrightarrow{D} X$ .

**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  $t \in C_{F_X}$ ,  $\mathbb{P}_n(X_n \leq t) \rightarrow \mathbb{P}(X \leq t)$ . 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  $A = (-\infty, t]$ ,  $X_n \xrightarrow{w} X$  is equivalently to  $\mathbb{P}_n(X_n \leq t) \rightarrow \mathbb{P}(X \leq t)$  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  $x \in \mathbb{R}^d$ ,

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

**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  $n \geq 1$  and  $k \geq 1$ ,

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

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

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

By taking  $k \rightarrow \infty$ ,  $F_X(x_k) \rightarrow F_X(x^-)$ , while  $F_X(y_k) \rightarrow F_X(x)$ ,<sup>a</sup> and we're done.

<sup>a</sup>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  $\liminf$ .

2 Feb. 17:30

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

$$\liminf_{n \rightarrow \infty} (x_n + y_n) \geq \liminf_{n \rightarrow \infty} x_n + \liminf_{n \rightarrow \infty} y_n,$$

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. From [Portmanteau theorem \(c\)](#), it suffices to show that for every open  $A \subseteq \mathbb{R}$ , we have

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

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

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

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

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

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

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

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  $t_n \rightarrow t \in C_{F_X}$ , then  $\mathbb{P}_n(X_n \leq t_n) \rightarrow \mathbb{P}(X \leq t)$ .

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

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

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  $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ , 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  $(\Omega, \mathcal{F}, \mathbb{P})$  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  $M \rightarrow \infty$ ,  $\mathbb{P}(\|X\| = \infty) \geq \epsilon$ , 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.  $\blacksquare$

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)**,

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

In other words,  $\liminf_{n \rightarrow \infty} \mathbb{P}_n(\|X_n\| < M) > 1 - \epsilon$ , 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.  $\blacksquare$

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  $a_n \rightarrow 0$  and  $b_n$  is bounded,  $a_n b_n \rightarrow 0$ .

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  $\mathbb{P}_n(|X_n Y_n| > \epsilon) \rightarrow 0$ . 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,  $\mathbb{P}_n(|X_n| > \epsilon/M) \rightarrow 0$  as  $n \rightarrow \infty$ .

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

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  $b_n = \sqrt{n/p(1-p)} \rightarrow \infty$ ,  $T_n = X_n/n$ , and  $\theta = p$ , we have

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

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  $b_n \rightarrow \infty$  be a positive deterministic sequence. If  $b_n(T_n - \theta) \xrightarrow{D} Y$ , then  $T_n \xrightarrow{P} \theta$ . Moreover, if  $g: \mathbb{R}^d \rightarrow \mathbb{R}^m$  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  $b_n \rightarrow \infty$ ,  $T_n \xrightarrow{P} \theta$  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  $\theta$ ,

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

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

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

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

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

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](#).  $\otimes$

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.<sup>a</sup> Combining the above, the result follows.  $\blacksquare$

<sup>a</sup>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  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  on which we can define random vectors  $(Y_n)$  and  $Y$  such that  $Y_n \stackrel{D}{=} X_n$  for all  $n$  and  $Y \stackrel{D}{=} X$ , and  $\tilde{\mathbb{P}}(Y_n \rightarrow Y) = 1$ .

**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  $(\Omega, \mathcal{F}, \mathbb{P})$ . 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  $p^{\text{th}}$  quantile of  $X$  is defined as any  $t \in \mathbb{R}$  such that

$$\mathbb{P}(X \leq t) \geq p \geq \mathbb{P}(X < t).$$

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

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

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

**Remark.**  $t \geq F^{-1}(p)$  if and only if  $F(t) \geq p$ ; 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  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ . Then,  $F^{-1}(U) =: Y$  is a random variable with cdf  $F$ .

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

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

⊛

### 2.5.2 Proof of Skorohod's representation theorem

Now we can prove [Skorohod's representation theorem](#).

**Proof of Theorem 2.5.1.** Consider  $\tilde{\Omega} = (0, 1)$ , and  $\tilde{\mathbb{P}}((a, b)) = b - a$  for all  $a < b$ . Then, we can define  $U(p) = p$  for all  $p \in \tilde{\Omega}$ , 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  $Y_n \stackrel{D}{=} X_n$  and  $Y \stackrel{D}{=} X$ , so we just need to show  $\tilde{\mathbb{P}}(Y_n \rightarrow Y) = 1$ .

**Claim.** It's equivalent to  $\tilde{\mathbb{P}}(F_{X_n}(Z) < p) \rightarrow \tilde{\mathbb{P}}(F_X(Z) < p)$  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,  $Y_n(p) \rightarrow Y(p)$  if and only if  $\Phi(Y_n(p)) \rightarrow \Phi(Y(p))$  as  $\Phi$  is strictly increasing and continuous, or equivalently,

$$\Phi(Y_n(p)) = \tilde{\mathbb{P}}(Z \leq Y_n(p)) \rightarrow \tilde{\mathbb{P}}(Z \leq Y(p)) = \Phi(Y(p)).$$

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

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

which holds if and only if  $\tilde{\mathbb{P}}(F_{X_n}(Z) < p) \rightarrow \tilde{\mathbb{P}}(F_X(Z) < p)$ .<sup>a</sup> ⊗

<sup>a</sup>Follows from [the remark](#). Explicitly, firstly, it's equivalent to  $\tilde{\mathbb{P}}(Z \geq F_{X_n}^{-1}(p)) \rightarrow \tilde{\mathbb{P}}(Z \geq F_X^{-1}(p))$ , and with  $\tilde{\mathbb{P}}(Z \geq F_{X_n}^{-1}(p)) = \tilde{\mathbb{P}}(F_{X_n}(Z) \geq p)$  and  $\tilde{\mathbb{P}}(Z \geq F_X^{-1}(p)) = \tilde{\mathbb{P}}(F_X(Z) \geq p)$ , the result follows.

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

$$\tilde{\mathbb{P}}(F_{X_n}(Z) \rightarrow F_X(Z)) = 1,$$

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

$$\tilde{\mathbb{P}}(F_{X_n}(Z) \leq p) \rightarrow \tilde{\mathbb{P}}(F_X(Z) \leq p),$$

and also  $\tilde{\mathbb{P}}(F_{X_n}(Z) < p) \rightarrow \tilde{\mathbb{P}}(F_X(Z) < p)$  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  $X_n \xrightarrow{D} X$ , from [Skorohod's representation theorem](#), we can obtain  $Y_n \xrightarrow{a.s.} Y$  on  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  such that  $X_n \stackrel{D}{=} Y_n$  and  $X \stackrel{D}{=} Y$ . Then for any bounded and continuous  $g$ ,

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

by the [bounded convergence theorem](#), which proves  $X_n \xrightarrow{w} X$ . ⊗

Another application is to generalize [Fatou's lemma](#).

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

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

<sup>a</sup>Can be on different probability spaces.

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

$$\mathbb{E}[g(X)] = \tilde{\mathbb{E}}[g(Y)] \leq \liminf_{n \rightarrow \infty} \tilde{\mathbb{E}}[g(Y_n)] = \liminf_{n \rightarrow \infty} \mathbb{E}_n[g(X_n)].$$

The following is well-known from real analysis [dominated convergence theorem](#).

**Theorem 2.5.2.** If  $X_n \xrightarrow{\text{a.s.}} X$ ,  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  is continuous and  $(g(X_n))$  is uniformly integrable<sup>a</sup> if and only if  $\mathbb{E}_n[g(X_n)] \rightarrow \mathbb{E}[g(X)]$ .

<sup>a</sup>I.e.,  $\lim_{t \rightarrow \infty} \sup_{n \geq 1} \mathbb{E}[|g(X_n)| \mathbb{1}_{|g(X_n)| \geq t}] = 0$ .

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

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

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

**Remark.** [Proposition 2.5.2](#) can be proved with [Skorohod's representation theorem](#) also.

## 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  $X_n \xrightarrow{D} X$ , 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.**  $(x_n) \rightarrow x$  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  $\rho(x_{n_k}, x) \geq \epsilon$  for every  $k \geq 1$ . 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.**  $X_n \xrightarrow{w} X$  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](#).

**Theorem 2.6.3.** If  $X_n \xrightarrow{w} X$  and  $X_n \xrightarrow{w} Y$ , then  $X \stackrel{D}{=} Y$ . More generally, if  $X_n \xrightarrow{w} X$  and  $Y_n \xrightarrow{w} Y$ , with  $X_n \stackrel{D}{=} Y_n$  for all  $n \geq 1$ ,  $X \stackrel{D}{=} Y$ .

in  $L_1$ ?

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

the in  $L_1$  version?

**Proof.** For every  $n \geq 1$ ,  $\mathbb{E}_n[g(X_n)] = \mathbb{E}_n[g(Y_n)]$  for all  $g: \mathbb{R}^d \rightarrow \mathbb{R}$ . If  $g$  is bounded and continuous,  $\mathbb{E}_n[g(X_n)] \rightarrow \mathbb{E}[g(X)]$  and  $\mathbb{E}_n[g(Y_n)] \rightarrow \mathbb{E}[g(Y)]$ . To show that  $X \stackrel{D}{=} Y$ , we want to show  $F_X = F_Y$ , or  $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$  for all  $B \in \mathcal{F} = \mathcal{B}(\mathbb{R}^d)$ . 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.,

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

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

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  $X \stackrel{D}{=} Y$ ?

**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  $\phi_X: \mathbb{R}^d \rightarrow \mathbb{C}$  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

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

**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,

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

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  $X \stackrel{D}{=} Y$ . 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  $X \stackrel{D}{=} Y$ .

**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  $X + \sigma Z \stackrel{D}{=} Y + \sigma Z'$ . Now, for  $\sigma = 1/k$ ,  $k \in \mathbb{N}$ ,

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

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  $X \stackrel{D}{=} Y$  from [Theorem 2.6.3](#).  $\otimes$

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

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

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

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

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

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

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

**Note.** Now showing  $X \stackrel{D}{=} Y$  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,  $F_{X_n}(t) \rightarrow F(t)$  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  $M_\epsilon \in C_F$ ,  $F_{X_n}(M_\epsilon) \rightarrow F(M_\epsilon)$ . 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  $\epsilon \rightarrow 0$ ,  $F(+\infty) = 1$ . Similarly,  $F(-\infty) = 0$ . ■

We now state the theorem.

**Theorem 2.6.6** (Lévy-Cramer continuity theorem). If  $\phi_{X_n}(t) \rightarrow \phi(t)$  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  $Y_n \xrightarrow{w} Y$  and  $Z_n \xrightarrow{w} Z$  are two subsequences of  $X_n$  such that  $Y \neq Z$ . But since  $\phi_{Y_n}(t) \rightarrow \phi_Y(t)$  and  $\phi_{Z_n}(t) \rightarrow \phi_Z(t)$ , with the fact that  $(\phi_{Y_n}(t))$  and  $(\phi_{Z_n}(t))$  are subsequences of  $(\phi_{X_n}(t))$  for every  $t$ , as  $\phi_{X_n}(t) \rightarrow \phi(t)$ , 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,  $Y \stackrel{D}{=} Z$ .

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

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

13 Feb. 9:30

**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

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

We claim that we can find an  $n_0 \in \mathbb{N}$  such that for every  $n \geq n_0$ ,  $\mathbb{P}_n(|X_n| \geq 2/\delta) < \epsilon$ .<sup>a</sup> 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

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

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,

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

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  $\phi_{X_n}(t) \rightarrow \phi(t)$  for all  $t$ , we have  $|1 - \phi_{X_n}(t)| \rightarrow |1 - \phi(t)|$  point-wise, hence by the **bounded convergence theorem**,

$$\frac{1}{\delta} \int_{-\delta}^{\delta} |1 - \phi_{X_n}(t)| dt \rightarrow \frac{1}{\delta} \int_{-\delta}^{\delta} |1 - \phi(t)| dt < \frac{\epsilon}{2}$$

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

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

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

<sup>a</sup>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

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

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

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

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

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

**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  $g: \mathbb{R} \rightarrow \mathbb{R}$  has bounded support, the same calculation

from the proof of [uniqueness theorem](#) gives

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

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

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \int_a^b \int \lim_{\sigma \searrow 0} e^{-ixu - \sigma^2 u^2/2} \phi_X(u) du dx = \frac{1}{2\pi} \int_a^b \int e^{-ixu} \phi_X(u) du dx.$$

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

$$\int_a^b \int \sup_{\sigma > 0} |e^{-ixu - \sigma^2 u^2/2} \phi_X(u)| du dx < \infty$$

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

$$\int_a^b \int_{\mathbb{R}} |\phi_X(u)| \sup_{\sigma > 0} |e^{-\sigma^2 u^2/2}| du dx \leq \int_a^b \int_{\mathbb{R}} |\phi_X(u)| du dx$$

which is finite since  $\int |\phi_X(u)| du < \infty$ . ■

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

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

$$|f_{X_n}(x) - f_X(x)| \leq \frac{1}{2\pi} \int_{\mathbb{R}} |e^{-ixu}| \cdot |\phi_{X_n}(u) - \phi_X(u)| du, \leq \frac{1}{2\pi} \int_{\mathbb{R}} |\phi_{X_n}(u) - \phi_X(u)| du$$

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.,

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

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

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

which goes to 0 as  $h \rightarrow 0$  since  $|e^{ihX} - 1| \leq 2$  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

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

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

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

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

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

Since  $G$  is differentiable,  $G(h) - G(0) = \int_0^h G'(y) dy$ , hence

$$G(h) - G(0) - G'(0)h = \int_0^h G'(y) - G'(0) dy = h \int_0^1 G'(uh) - G'(0) du = h \int_0^1 iX e^{iuhX} - iX du$$

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

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

Finally, taking the limit as  $h \rightarrow 0$ , with the fact that  $\mathbb{E}[|X|] < \infty$  and  $\int_0^1 |e^{ihuX} - 1| du \leq 2$ , we see that  $|X| \int_0^1 |e^{ihuX} - 1| du \leq 2|X|$ , 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,

$$\lim_{h \rightarrow 0} \mathbb{E} \left[ |X| \int_0^1 |e^{ihuX} - 1| du \right] = \mathbb{E} \left[ |X| \lim_{h \rightarrow 0} \int_0^1 |e^{ihuX} - 1| du \right] = 0$$

since  $\lim_{h \rightarrow 0} \int_0^1 |e^{ihuX} - 1| du = 0$ , 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.<sup>a</sup>

<sup>a</sup>This is a generalization of [Proposition 2.6.1](#).

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

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

which goes to 0 as  $h \rightarrow 0$  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** (Khinchin'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  $\bar{X}_n \xrightarrow{P} \mathbb{E}[X]$ .

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

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

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

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

Since  $(1 + a_n)^n \rightarrow e^c$  if  $na_n \rightarrow c$ , it remains to show  $\int_0^1 \phi'(ut/n) du \rightarrow ic$ . First,  $\phi'(t)$  is continuous at 0 from [Corollary 2.6.2](#),<sup>a</sup> as  $n \rightarrow \infty$

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

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

$$\int_0^1 \phi'(ut/n) du \rightarrow \int_0^1 ic du = ic$$

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

<sup>a</sup>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  $\bar{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$ .

### 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,

$$g(x) = g(0) + g'(0)x + x \int_0^1 g'(ux) - g'(0) du.$$

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

$$\begin{aligned} g(x) &= g(0) + g'(0)x + x \int_0^1 \int_0^{ux} g''(y) dy du \\ &= g(0) + g'(0)x + x^2 \int_0^1 \int_0^1 g''(xuv) u dv du. \end{aligned} \quad y = uxv, dy = uxdv$$

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  $\mathbb{E}[X_i] =: \mu$ ,  $\text{Var}[X_i] =: \sigma^2 < \infty$  for all  $1 \leq i \leq n$ . Then

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

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

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

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

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

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  $(\phi(t/\sqrt{n}))^n \rightarrow e^{-t^2/2}$ . We see that as  $n \rightarrow \infty$ , the integrand

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

Hence, from the [bounded convergence theorem](#),

$$\int_0^1 \int_0^1 u\phi''(ut/\sqrt{n}) du dv \rightarrow \int_0^1 \int_0^1 -u du dv = -\frac{1}{2},$$

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  $\mathbb{E}[X] = \mu < \infty$ ,  $\text{Var}[X] = \sigma^2$ . If, also,  $X_i$ 's are Gaussian,  $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ , i.e.,

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

In this case, a natural estimator of  $\mu$  is  $\bar{X}_n$ , and we have the distribution of  $\bar{X}_n - \mu$ , 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

$$T_n := \frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1} \xrightarrow{\text{TY}} \mathcal{N}(0, 1)$$

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

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

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

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

**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  $Y_i := X_i - \mu$  for all  $i$  (and also  $Y = X - \mu$ ), as  $n \rightarrow \infty$ ,

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

since  $\frac{1}{n} \sum_{i=1}^n Y_i^2 \xrightarrow{P} \mathbb{E}[Y^2] = \text{Var}[X] = \sigma^2 < \infty$  as  $X \in L^2$ , and  $(\bar{Y}_n)^2 \xrightarrow{P} (\mathbb{E}[Y])^2 = 0$ , 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  $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$  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  $\bar{X}_n \pm Z_{\alpha/2} s_n / \sqrt{n}$  and  $\bar{X}_n \pm Z_{\alpha/2} \hat{\sigma}_n / \sqrt{n}$  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  $X, X_1, \dots, X_n$  be i.i.d. samples. If they are Gaussian,

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

where  $(Z_{n-1}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ . 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,

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

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

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

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  $Y = X - \mu$ ),

$$\begin{aligned} \hat{\sigma}_n^2 &= \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}_n^2 \Rightarrow \hat{\sigma}_n^2 - \sigma^2 = \frac{1}{n} \sum_{i=1}^n (Y_i^2 - \sigma^2) - \bar{Y}_n^2 \\ &\Rightarrow \sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i^2 - \sigma^2) - \frac{(\sqrt{n}\bar{Y}_n)^2}{\sqrt{n}}. \end{aligned}$$

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

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

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

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

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,  $\mathbb{E}[Z^4] \geq (\mathbb{E}[Z^2])^2 \geq 1$ , hence indeed, the assumption might not be true in general.

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

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

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

Following the intuition, let's find such *consistent* estimators. Let  $Y := X - \mu = X - \mathbb{E}[X]$  (and also  $Y_i = X_i - \mu$  as usual),  $\mu_k := \mathbb{E}[Y^k] = \mathbb{E}[(X - \mu)^k]$  for all  $k \geq 2$ , and finally  $\tilde{\mu}_k = \mu_k/\sigma^k = \mathbb{E}[(X - \mu)^k/\sigma^k]$ . 20 Feb. 9:30

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

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

The task is then the following.

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

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

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

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

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  $\bar{X}_n$ .

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  $\bar{X}_n =: \bar{X}$  and  $\bar{Y}_n =: \bar{Y}$ . Then

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

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

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

By the [weak law of large number](#),  $\bar{Y}^k \xrightarrow{P} \mathbb{E}[Y^k] = \mu_k$  and  $(-\bar{Y})^{k-\ell} \xrightarrow{P} 0$  for  $\ell < k$  from  $-\bar{Y} \xrightarrow{P} 0$  (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. ■

[Proposition 3.2.2](#) and [Proposition 3.2.3](#) imply the following.

**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

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

### 3.3 Testing Normality

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

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

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,

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

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  $\mathcal{N}(0, v_k) = \mathcal{N}(0, \text{Var}[Y^k - \mu_k - k \mu_k Y])$  by [central limit theorem](#) and [Slutsky's theorem](#), where

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

since  $\text{Cov}[Y, Y^k] = \mathbb{E}[Y \cdot Y^k] - \mathbb{E}[Y] \mathbb{E}[Y^k] = \mathbb{E}[Y^{k+1}] = \mu_{k+1}$  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  $H_0: X \sim \mathcal{N}(\mu, \sigma^2)$  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  $\tilde{\mu}_k = 0$ ) 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

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

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  $\tilde{v}_k := v_k/\sigma^{2k} > 0$ , then  $\sqrt{n}M_k/\hat{\sigma}_n^k = \sqrt{n}\tilde{M}_k \xrightarrow{D} \mathcal{N}(0, \tilde{v}_k)$ .

**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  $H_0: X \sim \mathcal{N}(\mu, \sigma^2)$ ,

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

**Proof.** From the symmetry of normal distribution, [Corollary 3.3.1](#) gives

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

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

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

which provides the result. \*

<sup>a</sup>More generally,  $\text{Var}[Y^k] = \mathbb{E}[Y^{2k}] - (\mathbb{E}[Y^k])^2 = \mathbb{E}[Y^{2k}] = \mu_{2k}$  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  $g(M_k, \hat{\sigma}_n^2) = M_k/\hat{\sigma}_n^k = \tilde{M}_k$  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

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

which implies  $X_n \xrightarrow{D} X$  by the [Lévy-Cramér continuity theorem](#). ■

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

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

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

## Lecture 12: Asymptotic Joint Distribution by Multivariate CLT

22 Feb. 9:30

**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  $Z \sim \mathcal{N}(0, \Sigma)$ . We see that from the [univariate central limit theorem](#), the left-hand side is

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n t \cdot X_i \xrightarrow{D} \mathcal{N}(0, \text{Var}[t \cdot X_i]),$$

and since  $\text{Var}[t \cdot X] = t^\top \text{Var}[X] t = t^\top \Sigma t = \text{Var}[t \cdot Z]$ , hence 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  $\tilde{M}_k = M_k / \hat{\sigma}_n^k$  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  $Y = X - \mu$ ),  $\sigma^2 = \text{Var}[X]$ ,  $\mu_k = \mathbb{E}[Y^k]$ , and  $\tilde{\mu}_k = \mu_k / \sigma^k$ . Let's start with  $k = 1$ , i.e., compute the asymptotic law of  $\bar{X}_n / \hat{\sigma}_n$ . In this case, we have proved the following.

**As previously seen.** From [Proposition 3.2.1](#) and [Proposition 3.2.2](#),

- $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$  from  $\sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ , 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  $\tilde{\mu}_4 > 1$ .<sup>a</sup>

<sup>a</sup>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  $X \in L^2$ ,

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

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

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

**Remark** (Asymptotically independent). We know that when  $X$  is Gaussian,  $\bar{X}_n$  and  $s_n^2$  are independent. Related back to [Corollary 3.3.1](#), when their [skewness](#) is 0,  $\bar{X}_n$  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  $\bar{X}_n$  and  $\hat{\sigma}_n^2$ , but not  $\hat{\sigma}_n$ . This is fine since we can further apply the [delta method](#) with  $g(\bar{X}_n, \hat{\sigma}_n^2) := \bar{X}_n / \hat{\sigma}_n$  to get the distribution of  $\bar{X}_n / \hat{\sigma}_n$ . However,

let's leave the application of the [delta method](#) to the general  $k$ . We note the following.

**Note.** The actual characterization of  $\bar{X}_n$  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  $\sqrt{n}(M_k - \mu_k) \rightarrow \mathcal{N}(0, \text{Var}[Y^k - k\mu_{k-1}Y])$  if  $X \in L^{2k}$  and the variance is strictly positive.

This implies that for  $X \in L^k$  for any  $k > 2$ ,<sup>1</sup>

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

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

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

**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  $\widetilde{v}_k := \text{Var}[-\frac{k}{2}\widetilde{\mu}_k Z^2 + Z^k - k\widetilde{\mu}_{k-1}Z] > 0$ , 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  $g(x, y) := y/x^{k/2}$  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  $\widetilde{\mu}_k = \mu_k/\sigma^k$ ,  $\partial g/\partial x = -k y x^{-k/2-1}/2$ , 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  $Z_i := (X_i - \mu)/\sigma = Y_i/\sigma$ , proving the first claim. Then by the [multivariate central limit](#)

<sup>1</sup>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](#).

theorem and [Slutsky's theorem](#), the above further [converges in distribution](#) to  $\mathcal{N}(0, \tilde{v}_k)$  when

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

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.

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

$$\left( \sqrt{\frac{n}{\tilde{v}_3}} \tilde{M}_3 \right)^2 + \left( \sqrt{\frac{n}{\tilde{v}_4}} (\tilde{M}_4 - \tilde{\mu}_4) \right)^2 \xrightarrow{D} \chi_2^2.$$

**Proof.** One can write down  $\sqrt{n}(\tilde{M}_\ell - \tilde{\mu}_\ell)$  for even  $\ell$ , and also  $\sqrt{n}(\tilde{M}_k - \tilde{\mu}_k) = \sqrt{n}\tilde{M}_k$  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  $\tilde{v}_k, \tilde{v}_\ell < \infty$ ,

$$\frac{n}{\tilde{v}_k} \tilde{M}_k^2 + \frac{n}{\tilde{v}_\ell} (\tilde{M}_\ell - \tilde{\mu}_\ell)^2 \xrightarrow{D} \chi_2^2.$$

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

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

As  $\sigma^2 = \theta$  and  $\mu_4 = 3\theta^2 + \theta$ , we see that  $\bar{X}_n$  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  $T_n^1$  and  $T_n^2$ ,

$$\sqrt{n}(T_n^i - \theta) \xrightarrow{D} \sigma_i^2(\theta) Z \sim \mathcal{N}(0, \sigma_i^2(\theta))$$

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

$$\mathbb{P}(\theta \in I_n^i) := \mathbb{P}\left(\theta \in T_n^i \pm Z_{\alpha/2} \frac{\sigma_i(\theta)}{\sqrt{n}}\right) \cong 1 - \alpha$$

where  $I_n^i := T_n^i \pm Z_{\alpha/2} \sigma_i(\theta) / \sqrt{n}$ . Let  $n_i(\gamma)$  be the value of  $n$  such that  $|I_n^i| = \gamma$ , 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.,  $n_1(\gamma)/n_2(\gamma) \cong \sigma_1^2(\theta)/\sigma_2^2(\theta)$ . 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  $T_n^1$  and  $T_n^2$  for  $\theta$  such that  $\sqrt{n}(T_n^i - \theta) \xrightarrow{D} \mathcal{N}(0, \sigma_i^2(\theta))$  is defined as

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

### 3.4.2 Variance Stabilizing Transformation

Continuing on the previous [example](#), say we use  $\bar{X}_n$  as the estimator of  $\theta$ . We have

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

**As previously seen.** As the asymptotic distribution depends on  $\theta$ , we don't directly get a confidence interval. Indeed, usually we will first write  $\sqrt{n}/\sqrt{\theta}(\bar{X}_n - \theta) \xrightarrow{D} Z \sim \mathcal{N}(0, 1)$ , replace  $\sqrt{\theta}$  by  $\sqrt{\bar{X}_n}$ , 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  $g: \mathbb{R} \rightarrow \mathbb{R}$  differentiable at  $\theta$  and  $g'(\theta) \neq 0$ ,

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

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

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

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

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

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

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

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

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

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

**Remark.** The above can be easily generalized.

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

- $\sqrt{\theta}$  by  $h(\theta)$ , a positive function;<sup>a</sup>
- $\sqrt{n}$  by  $b_n$ , a positive divergent strictly increasing sequence;
- $\bar{X}_n$  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. \*

<sup>a</sup>We don't need continuity since we don't need  $h(T_n)$  when doing the variance stabilizing transformation.

## Lecture 13: Bahadur's Representation for Quantiles

### 3.5 Inference for Population Quantiles

27 Feb. 9:30

Let  $X, X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$  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  $F^{-1}(p) = \inf\{t \in \mathbb{R} : F(t) \geq p\}$ .

**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  $p^{\text{th}}$ -sample quantile

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

**Remark.** If  $F$  is continuous, then apart from a null set we have

$$\hat{\theta}_p = \inf\{X_{(i)} : \hat{F}_n(X_{(i)}) = i/n \geq p\} = \inf\left\{t \in \mathbb{R} : \sum_{i=1}^n \mathbb{1}_{X_i \leq t} \geq \lceil np \rceil\right\} = X_{(\lceil np \rceil)}.$$

**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  $\sum_{i=1}^n \mathbb{1}_{X_i \leq t} \in \mathbb{N}$ . \*

#### 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  $\epsilon > 0$ ,

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

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  $p_n \rightarrow p$ , then  $\hat{\theta}_{p_n} \xrightarrow{P} \theta_p$ .

**Proof.** We want to show that for any  $\epsilon > 0$ ,  $\mathbb{P}(|\hat{\theta}_{p_n} - \theta_p| > \epsilon) \rightarrow 0$ . 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,  $\hat{\theta}_{p_n} = \hat{F}_n^{-1}(p_n) > \theta_p + \epsilon$ , 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  $n \rightarrow \infty$  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  $\sqrt{n}(p_n - p) \rightarrow c \in [0, \infty)$ , 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  $\sqrt{n}(p_n - p) \rightarrow 0$ , 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  $\sqrt{n}(p_n - p) \rightarrow c$ , 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  $\sqrt{n}(\ell_n - p) \rightarrow -c$  and  $\sqrt{n}(u_n - p) \rightarrow c$ . If  $F'(\theta_p) =: f(\theta_p) > 0$ , then  $\mathbb{P}(\hat{\theta}_{\ell_n} \leq \theta_p \leq \hat{\theta}_{u_n}) \rightarrow 1 - \alpha$ .

**Proof.** First, consider  $\ell_n$ . Since  $\sqrt{n}(\ell_n - p) \rightarrow -c$ , 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  $\text{Var}[X] = \sigma^2 < \infty$ . Then both  $\hat{\theta}_{1/2}$  and  $\bar{X}_n$  are two possible estimators of  $\mu$ , and in this case, we might want to look at the [asymptotic relative efficiency](#). Specifically,

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

Let's summarize the above in the following.

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

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  $\bar{X}_n$  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.,

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

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

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

**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.,

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

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

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

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

where

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

## 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

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

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  $p = F(\theta_p)$ ,

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

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  $t \in \mathbb{R}$ ,

$$\mathbb{P}(W_n \leq t, U_n \geq t + \epsilon) + \mathbb{P}(U_n \leq t, W_n \geq t + \epsilon) \rightarrow 0,$$

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  $\epsilon > 0$ ,  $\mathbb{P}(|W_n - U_n| > \epsilon) \rightarrow 0$ . Firstly, observe that for every fixed  $\epsilon > 0$ , if  $b - a < \epsilon/2$ ,

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

since the left-hand side is equal to

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

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

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

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  $\bigcup_{i \in I} [a_i, b_i] \supseteq [-M, M]$ ,

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

Since  $I$  is finite, together with the first observation, implies  $\limsup_{n \rightarrow \infty} \mathbb{P}(|U_n - W_n|) \leq \delta$ . As  $\delta$  is arbitrary, letting  $\delta \rightarrow 0$  completes the proof.  $\blacksquare$

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.,  $Z_n(t) := \sqrt{n}(F(t) - \hat{F}_n(t))$ . We have seen that  $\mathbb{E}[Z_n(t)] = 0$  and  $\text{Var}[Z_n(t)] = F(t)(1 - F(t))$ , and  $Z_n(t) \xrightarrow{D} \mathcal{N}(0, F(t)(1 - F(t)))$  by [central limit theorem](#).

**Claim.** For any  $t, s \in \mathbb{R}$ ,  $\text{Var}[Z_n(t) - Z_n(s)] = \mathbb{E}[(Z_n(t) - Z_n(s))^2] \leq |F(t) - F(s)|$ . Hence, if  $s_n \rightarrow s$  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  $\text{Var}[Z_n(t) - Z_n(s)] = \text{Var}[\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}] \leq \mathbb{E}[|\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}|^2]$  where

$$|\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}| = \begin{cases} 1, & \text{if } s < X \leq t \text{ or } t < X \leq s; \\ 0, & \text{otherwise.} \end{cases}$$

Hence, as  $|\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}| = |\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}|^2$ ,

$$\begin{aligned} \mathbb{E}[|\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}|^2] & = \mathbb{P}(s < X \leq t) + \mathbb{P}(t < X \leq s) \\ & = (F(t) - F(s))^+ + (F(s) - F(t))^+ = |F(t) - F(s)|, \end{aligned}$$

i.e.,  $|\mathbb{1}_{X \leq t} - \mathbb{1}_{X \leq s}| \sim \text{Ber}(|F(t) - F(s)|)$ .  $\circledast$

From [Lemma 3.5.1](#), it suffices to show  $\mathbb{P}(W_n \leq t, U_n \geq t + \epsilon) \rightarrow 0$  and  $\mathbb{P}(U_n \leq t, W_n \geq t + \epsilon) \rightarrow 0$  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

$$\begin{aligned} W_n \leq t & \Leftrightarrow \sqrt{n}(\hat{\theta}_{p_n} - \theta_p) - \sqrt{n} \frac{p_n - p}{f(\theta_p)} \leq t \\ & \Leftrightarrow \hat{\theta}_{p_n} = \hat{F}_n^{-1}(p_n) \leq \theta_p + \frac{t}{\sqrt{n}} + \frac{p_n - p}{f(\theta_p)} =: \theta_p + \delta_n \quad \delta_n := \frac{t}{\sqrt{n}} + \frac{p_n - p}{f(\theta_p)} \end{aligned}$$

From the property of  $\hat{F}_n^{-1}$ ,  $p_n \leq \hat{F}_n(\theta_p + \delta_n)$ ,

$$\Leftrightarrow \sqrt{n}(p_n - F(\theta_p + \delta_n)) \leq \sqrt{n}(\hat{F}_n(\theta_p + \delta_n) - F(\theta_p + \delta_n)) = -Z_n(\theta_p + \delta_n),$$

which can be written as

$$Z_n(\theta_p + \delta_n) \leq \sqrt{n}(F(\theta_p + \delta_n) - p_n) \Leftrightarrow \frac{Z_n(\theta_p + \delta_n)}{f(\theta_p)} \leq \frac{\sqrt{n}(F(\theta_p + \delta_n) - p_n)}{f(\theta_p)} =: t_n.$$

Putting everything together, with  $U_n = Z_n(\theta_p)/f(\theta_p)$ , we have

$$\begin{aligned} \mathbb{P}(W_n \leq t, U_n \geq t + \epsilon) &= \mathbb{P}(Z_n(\theta_p + \delta_n) \leq t_n f(\theta_p), Z_n(\theta_p) \geq f(\theta_p)(t + \epsilon)) \\ &\leq \mathbb{P}(Z_n(\theta_p + \delta_n) - Z_n(\theta_p) \leq (t_n - t - \epsilon)f(\theta_p)) \\ &= \mathbb{P}\left(\frac{Z_n(\theta_p + \delta_n) - Z_n(\theta_p)}{f(\theta_p)} - (t_n - t) \leq -\epsilon\right), \end{aligned}$$

which goes to 0 as  $n \rightarrow \infty$  if  $t_n \rightarrow t$  since from the previous [claim](#):

- let  $s_n := \theta_p + \delta_n$ ,  $s := \theta_p$ , with  $F$  being continuous at  $s$  and  $\delta_n \rightarrow 0$ ,  $Z_n(\theta_p + \delta_n) - Z_n(\theta_p) \xrightarrow{P} 0$ ;
- if further,  $t_n \rightarrow t$ , the left-hand side goes to 0, and the inequality tends to be vacuous.

**Claim.** Indeed,  $t_n \rightarrow t$ .

**Proof.** We want to show that

$$t_n = \frac{F(\theta_p + \delta_n) - p_n}{f(\theta_p)/\sqrt{n}} \rightarrow t.$$

By assumption, as  $\delta_n \rightarrow 0$  and  $F'(\theta_p) = f(\theta_p)$ ,

$$\frac{F(\theta_p + \delta_n) - F(\theta_p)}{\delta_n} \rightarrow f(\theta_p) \Leftrightarrow \frac{F(\theta_p + \delta_n) - F(\theta_p) - \delta_n f(\theta_p)}{\delta_n} \rightarrow 0,$$

i.e.,  $F(\theta_p + \delta_n) = F(\theta_p) + \delta_n f(\theta_p) + o(\delta_n)$ . 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  $o(\delta_n) = \delta_n o(1)$ , 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  $t_n = t + o(1) \rightarrow t$ . ⊗

The second claim  $\mathbb{P}(U_n \leq t, W_n \geq t + \epsilon) \rightarrow 0$  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

$$\begin{pmatrix} \sqrt{n}(\hat{F}_n(t_1) - F(t_1)) \\ \vdots \\ \sqrt{n}(\hat{F}_n(t_m) - F(t_m)) \end{pmatrix} = \begin{pmatrix} Z_n(t_1) \\ \vdots \\ Z_n(t_m) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \mathbb{1}_{X_i \leq t_1} - F(t_1) \\ \vdots \\ \mathbb{1}_{X_i \leq t_m} - F(t_m) \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \Sigma)$$

from [multivariate central limit theorem](#) where

$$\Sigma_{ij} = \text{Cov}[\mathbb{1}_{X \leq t_i}, \mathbb{1}_{X \leq t_j}] = \mathbb{P}(X \leq t_i \wedge X \leq t_j) - \mathbb{P}(X \leq t_i)\mathbb{P}(X \leq t_j).$$

## 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  $F$ ,  $\|\hat{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$  as  $n \rightarrow \infty$ .

**Proof.** Fix some  $\epsilon > 0$ , and let  $\epsilon > 2/k$  for some  $k \in \mathbb{N}$ . Then, for finitely many  $t_1, \dots, t_{k-1}$ ,  $\hat{F}_n(t_i) \xrightarrow{\text{a.s.}} F(t_i)$  and  $\hat{F}_n(t_i^-) \xrightarrow{\text{a.s.}} F(t_i^-)$  for every  $1 \leq i \leq k-1$  from the [strong law of large number](#). This means that there exists  $n_0$  such that for  $n \geq n_0$ , 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  $1 \leq i \leq k-1$ , 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  $n \geq n_0$ ,

$$\hat{F}_n(t) - F(t) \leq \hat{F}_n(t_i^-) - F(t_{i-1}) = \hat{F}_n(t_i^-) - F(t_i^-) + F(t_i^-) - F(t_{i-1}) \leq \frac{1}{k} + \frac{i}{k} - \frac{i-1}{k} = \frac{2}{k} < \epsilon$$

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*

$$Z_n(t) := \sqrt{n}(F(t) - \hat{F}_n(t))$$

for  $t \in \mathbb{R}$  introduced in the [proof](#) of [Bahadur representation theorem](#). Recall the following.

**As previously seen.** We have seen that

$$Z_n(t) := \sqrt{n}(F(t) - \hat{F}_n(t)) \xrightarrow{D} B_F(t) := \mathcal{N}(0, F(t)(1 - F(t))).$$

Furthermore, for any  $t_1, \dots, t_m \in \mathbb{R}$ ,

$$(Z_n(t_1), \dots, Z_n(t_m)) \xrightarrow{D} (B_F(t_1), \dots, B_F(t_m)) \sim \mathcal{N}(0, \Sigma_F(t_1, \dots, t_m))$$

where for  $1 \leq i \leq j \leq m$ ,

$$\text{Cov}[B_F(t_i), B_F(t_j)] = \text{Cov}[\mathbb{1}_{X \leq t_i}, \mathbb{1}_{X \leq t_j}] = F(t_i \wedge t_j) - F(t_i)F(t_j).$$

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  $t, s \in \mathbb{R}$ ,

$$\mathbb{E}[B_F(t)] = 0 \text{ and } \text{Cov}[B_F(t), B_F(s)] = F(t \wedge s) - F(t)F(s).$$

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  $B := B_F$  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

$$(Z_n) := (t \mapsto Z_n(t))_{n \geq 1} \xrightarrow{w} B_F := \{t \mapsto B(t)\}_{t \in \mathbb{R}}.$$

One immediate consequence is the following.

**Proposition 3.6.1.** If  $T: \mathcal{D} \rightarrow \mathbb{R}$  is continuous,<sup>a</sup> then  $T(Z_n) \xrightarrow{D} T(B_F)$ .

<sup>a</sup>I.e.,  $T(G_n) \rightarrow T(F)$  if  $(G_n), F \in \mathcal{D}$  such that  $\|G_n - F\|_\infty \rightarrow 0$ .

**Proof.** Since  $Z_n \xrightarrow{w} B_F$ , [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,  $\mathbb{P}(\sqrt{n}\|F - \hat{F}_n\|_\infty \geq d_\alpha) \rightarrow \alpha$ , i.e.,

$$\mathbb{P}\left(\forall t \in \mathbb{R}: \hat{F}_n(t) - \frac{d_\alpha}{\sqrt{n}} \leq F(t) \leq \hat{F}_n(t) + \frac{d_\alpha}{\sqrt{n}}\right) \rightarrow 1 - \alpha.$$

Another application of [Proposition 3.6.1](#) is the following.

**Corollary 3.6.2.** If  $F$  is continuous, then

$$n \int_{\mathbb{R}} \left(\hat{F}_n(t) - F(t)\right)^2 F(dt) \xrightarrow{D} \int_{\mathbb{R}} B_F^2(t) F(dt).$$

**Proof.** From [Proposition 3.6.1](#), it suffices to show that  $G \mapsto \int_{\mathbb{R}} G^2 dF$  is continuous for  $G \in \mathcal{D}$ . Firstly, let  $(G_n), G \in \mathcal{D}$  such that  $\|G_n - G\|_\infty \rightarrow 0$ . Then

$$\begin{aligned} |T(G_n) - T(G)| &= \left| \int_{\mathbb{R}} G_n^2 - G^2 dF \right| \\ &\leq \int_{\mathbb{R}} |G_n^2 - G^2| dF \leq \|G_n - G\|_\infty \int_{\mathbb{R}} |G_n(t) + G(t)| F(dt) \leq 2\|G_n - G\|_\infty \end{aligned}$$

since  $\|G_n - G\|_\infty = \sup_t |G_n(t) - G(t)|$ . As  $\|G_n - G\|_\infty \rightarrow 0$ , 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  $H_0: F = F_0$  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*

$$C_n := \int_{\mathbb{R}} \left( \hat{F}_n(t) - F_0(t) \right)^2 F_0(dt).$$

For any  $\alpha \in (0, 1)$ , we reject  $H_0$  when  $nC_n > c_\alpha^2$ , where  $c_\alpha$  is defined as  $\alpha =: \mathbb{P}(\int_0^1 B_F^2(t) dt \geq c_\alpha^2)$ .

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  $H_0: F = F_0$  using the [Kolmogorov-Smirnov statistic](#), then for any  $F \neq F_0$ ,  $\mathbb{P}_F(\text{reject}) = \mathbb{P}(\sqrt{n}K_n > d_\alpha) \rightarrow 1$  as  $n \rightarrow \infty$ .

**Proof.** For any metric  $d$ , we have

$$\mathbb{P}_F(\sqrt{nd}(\hat{F}_n, F_0) \geq d_\alpha) \geq \mathbb{P}_F(\sqrt{n}(d(\hat{F}_n, F) - d(F, F_0)) \geq d_\alpha) = \mathbb{P}_F(\sqrt{nd}(\hat{F}_n, F) \geq d_\alpha + \sqrt{nd}(F, F_0)).$$

As  $F \neq F_0$ ,  $d(F, F_0) > 0$  is a fixed number, so  $\sqrt{nd}(F, F_0) \rightarrow \infty$ . 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  $H_0: F = F_0$  using the [Cramér-von Mises statistic](#), then for any  $F \neq F_0$ ,  $\mathbb{P}_F(\text{reject}) = \mathbb{P}(nC_n > c_\alpha^2) \rightarrow 1$  as  $n \rightarrow \infty$ .

We omit the proof of [Proposition 3.6.3](#), instead, we focus on a related result regarding the representation of  $C_n$ . The proof of [Proposition 3.6.4](#) can be done similarly. First, recall the following.

**As previously seen.** From [Corollary 3.6.2](#), under  $H_0: F = F_0$ ,

$$nC_n = n \int_{\mathbb{R}} (\hat{F}_n - F_0)^2 dF_0 \xrightarrow{D} \int_{\mathbb{R}} B_F^2 dF_0.$$

However, if  $F \neq F_0$ , what's the distribution now? Consider

$$h(F) := \int_{\mathbb{R}} (F - F_0)^2 dF_0.$$

Since  $h$  is continuous,  $C_n = h(\hat{F}_n) \rightarrow h(F)$ . As 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

$$\sqrt{n} \left( C_n - \int_{\mathbb{R}} (F - F_0)^2 dF_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) + o_p(1).$$

If we further have  $F \neq F_0$ , then the above converges to  $\mathcal{N}(0, \text{Var}[g(X)])$ .

**Proof.** We first note that the left-hand side is just  $\sqrt{n}(h(\hat{F}_n) - h(F))$ . Now, since

$$\begin{aligned} h(\hat{F}_n) &= C_n = \int_{\mathbb{R}} (\hat{F}_n - F_0)^2 dF_0 \\ &= \int_{\mathbb{R}} (\hat{F}_n - F + F - F_0)^2 dF_0 \\ &= \int_{\mathbb{R}} (\hat{F}_n - F)^2 dF_0 + \underbrace{\int_{\mathbb{R}} (F - F_0)^2 dF_0}_{h(F)} + 2 \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) dF_0, \end{aligned}$$

we have

$$\sqrt{n} \left( h(\hat{F}_n) - h(F) \right) = \sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)^2 dF_0 + 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) dF_0.$$

As  $n \int_{\mathbb{R}} (\hat{F}_n - F)^2 dF_0 \xrightarrow{w} \int_{\mathbb{R}} B_F^2 dF_0$ , [Proposition 2.4.2](#) implies

$$\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)^2 dF_0 = \frac{n}{\sqrt{n}} \int_{\mathbb{R}} (\hat{F}_n - F)^2 dF_0 = \frac{O_p(1)}{\sqrt{n}} = o_p(1),$$

which gives

$$\sqrt{n} \left( h(\hat{F}_n) - h(F) \right) = 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) dF_0 + o_p(1) =: \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) + o_p(1)$$

where we define the function  $g: \mathbb{R} \rightarrow \mathbb{R}$  as

$$g(x) := 2 \int_{\mathbb{R}} (\mathbb{1}_{x \leq t} - F(t))(F(t) - F_0(t)) F_0(dt)$$

since

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) &= \frac{2}{\sqrt{n}} \sum_{i=1}^n \int_{\mathbb{R}} (\mathbb{1}_{X_i \leq t} - F(t))(F(t) - F_0(t)) F_0(dt) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} \sum_{i=1}^n (\mathbb{1}_{X_i \leq t} - F(t))(F(t) - F_0(t)) F_0(dt) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} (F(t) - F_0(t)) - nF(t)(F(t) - F_0(t)) F_0(dt) \\ &= \frac{2}{\sqrt{n}} \int_{\mathbb{R}} n \left( \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} (F(t) - F_0(t)) - F(t)(F(t) - F_0(t)) \right) F_0(dt) \\ &= 2\sqrt{n} \int_{\mathbb{R}} (\hat{F}_n - F)(F - F_0) dF_0. \end{aligned}$$

To show  $\mathbb{E}[g(X)] = 0$ , as  $F(t), F_0(t), \mathbb{1}_{x \leq t}$  are all bounded by 1, [Fubini's theorem](#) gives

$$\mathbb{E}[g(X)] = 2 \int_{\mathbb{R}} (\mathbb{P}(X \leq t) - F(t))(F(t) - F_0(t)) F_0(dt) = 0$$

since  $\mathbb{P}(X \leq t) = F(t)$ . Finally, when  $F \neq F_0$ ,  $0 < \mathbb{E}[g^2(X)] < \infty$  follows from the same calculation, hence [central limit theorem](#) gives the distributional result.  $\blacksquare$

# Chapter 4

## Lindeberg Central Limit Theorem

### Lecture 16: Lindeberg Central Limit Theorem

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  $S_n = X_1 + \dots + X_n$ 's are a sum of independent, but not necessarily identically distributed, random variables whose distribution may vary with  $n$ .

19 Mar. 9:30

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

$$S_n \sim \text{Pois}\left(\sum_{i=1}^n \frac{1}{i^2}\right) \xrightarrow{\text{TV}} \text{Pois}\left(\sum_{i=1}^{\infty} \frac{1}{i^2}\right) = \text{Pois}\left(\frac{\pi^2}{6}\right),$$

which does not go to normal as we expected since  $X_i$  are not identically distributed.

On the other hand, something trickier can happen when  $X_i$  are "seemingly" identically distributed.

**Example.** Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(1/n)$  for every  $n \geq 1$ . But since  $S_n \sim \text{Pois}(1)$  for all  $n \geq 1$ ,

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}[S_n]}} \not\xrightarrow{D} \mathcal{N}(0, 1).$$

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  $(X_{nk_n}) = (X_{n1}, \dots, X_{nk_n})$  with  $\text{Var}[X_{nj}] < \infty$  for all  $1 \leq j \leq k_n$  and  $n$ . Again, we define  $S_n = X_{n1} + \dots + X_{nk_n}$ . In picture, we have something like a *triangular array* of random variables:

$$\begin{aligned} n = 1: & \quad X_{11}, \dots, X_{1k_1}; \\ n = 2: & \quad X_{21}, X_{22}, \dots, X_{2k_2}; \\ & \quad \vdots \\ n: & \quad X_{n1}, X_{n2}, X_{n3}, \dots, X_{nk_n}. \end{aligned}$$

**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  $Y_{nj} := (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\text{Var}[S_n]}$  for every  $1 \leq j \leq k_n$ . If the [Lindeberg condition](#) holds, then

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}[S_n]}} = \sum_{j=1}^{k_n} \frac{X_{nj} - \mathbb{E}[X_{nj}]}{\sqrt{\text{Var}[S_n]}} =: \sum_{j=1}^{k_n} Y_{nj} \xrightarrow{D} \mathcal{N}(0, 1).$$

**Note.** In the above notation, for all  $n \geq 1$ ,  $\mathbb{E}[Y_{nj}] = 0$  for all  $1 \leq j \leq k_n$  and  $\sum_{j=1}^{k_n} \text{Var}[Y_{nj}] = 1$ .

#### 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  $n \rightarrow \infty$ ,

$$\frac{\max_{1 \leq j \leq k_n} \text{Var}[X_{nj}]}{\text{Var}[S_n]} \rightarrow 0.$$

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,  $\max_{1 \leq j \leq k_n} \text{Var}[X_{nj}] = 1/n \rightarrow 0$ , 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  $Y_{nj} := (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\text{Var}[S_n]}$  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  $n \rightarrow \infty$ ,

$$\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] \rightarrow 0.$$

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  $\max_{1 \leq j \leq k_n} \text{Var}[Y_{nj}] \rightarrow 0$  as  $n \rightarrow \infty$ . Firstly, for any  $n$  and every  $1 \leq j \leq k_n$ , by splitting up the expectation, for every  $\epsilon > 0$ , we have  $\text{Var}[Y_{nj}] \leq \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] + \epsilon^2$ . Then with the [Lindeberg condition](#), we have

$$\max_{1 \leq j \leq k_n} \text{Var}[Y_{nj}] \leq \sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] + \epsilon^2 \rightarrow \epsilon^2,$$

i.e.,  $\limsup_{n \rightarrow \infty} \max_{1 \leq j \leq k_n} \text{Var}[Y_{nj}] \leq \epsilon^2$ . By letting  $\epsilon \rightarrow 0$ , we complete the proof.  $\blacksquare$

While the UAN is insufficient, it's also not necessary.

**Remark.** Let  $(X_n)$  be a sequence of independent Gaussian variables with  $\text{Var}[X_i] = 1/i^2$  for all  $i \geq 1$ . Then indeed,  $S_n \sim \mathcal{N}$ . However, the [uniform asymptotic negligibility](#) does not hold since

$$\frac{\max_{1 \leq i \leq n} \text{Var}[X_i]}{\text{Var}[S_n]} = \frac{1}{\sum_{i=1}^n 1/i^2} \rightarrow \frac{6}{\pi^2} > 0.$$

### 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  $|w_i|, |z_i| \leq 1$  for all  $1 \leq i \leq n$ , we have

$$\left| \prod_{i=1}^n z_i - \prod_{i=1}^n w_i \right| \leq \sum_{i=1}^n |w_i - z_i|.$$

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  $x \in \mathbb{R}$ ,  $|e^{ix} - (1 + ix)| \leq x^2/2$  and  $|e^{ix} - 1 - ix - (ix)^2/2| \leq x^2$ .

**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  $|z| \leq \epsilon < 1$ ,  $|e^z - 1 - z - z^2/2| \leq |z|^3/(1 - \epsilon)$ .

**Proof.** Since

$$\left| e^z - 1 - z - \frac{z^2}{2} \right| \leq \left| \sum_{n=3}^{\infty} \frac{z^n}{n!} \right| \leq |z|^3 \sum_{n=0}^{\infty} |z|^n = |z|^3 \cdot \frac{1}{1 - |z|} \leq \frac{|z|^3}{1 - \epsilon},$$

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  $\delta \in (0, 1)$ ,  $|e^{iz} - (1 + iz)| \leq \delta|z|$ .

**Proof.** Since

$$|e^{iz} - 1 - iz| = \left| \sum_{n=2}^{\infty} \frac{(iz)^n}{n!} \right| \leq \sum_{n=2}^{\infty} |z|^n = |z|^2 \sum_{n=0}^{\infty} |z|^n = \frac{|z|^2}{1 - |z|} = \frac{|z|}{1 - |z|} |z| < \frac{\delta/2}{2 - \delta} |z| \leq \delta|z|,$$

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  $\phi_{nj}(t) := \mathbb{E}[e^{itX_{nj}}]$  for  $t \in \mathbb{R}$ . We want to show that

$$\sum_{j=1}^{k_n} Y_{nj} \xrightarrow{D} \mathcal{N}(0, 1) \Leftrightarrow \prod_{j=1}^{k_n} \phi_{nj}(t) \rightarrow e^{-t^2/2}$$

for every  $t \in \mathbb{R}$  from the [uniqueness theorem](#). Fix  $t \in \mathbb{R}$ , from triangle inequality, it suffices to show

$$\left| \prod_{j=1}^{k_n} \phi_{nj}(t) - \prod_{j=1}^{k_n} e^{\phi_{nj}(t)-1} \right| + \left| \prod_{j=1}^{k_n} e^{\phi_{nj}(t)-1} - e^{-t^2/2} \right| \rightarrow 0.$$

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

$$\left| \prod_{j=1}^{k_n} \phi_{nj}(t) - \prod_{j=1}^{k_n} e^{\phi_{nj}(t)-1} \right| \leq \sum_{j=1}^{k_n} \left| \phi_{nj}(t) - e^{\phi_{nj}(t)-1} \right| = \sum_{j=1}^{k_n} \left| e^{\phi_{nj}(t)-1} - (\phi_{nj}(t) - 1) - 1 \right|.$$

Let  $z_j := \phi_{nj}(t) - 1$ , 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  $\max_{1 \leq j \leq k_n} |z_j| < \delta/2$  for large enough  $n$ .

**Claim.** For any  $\delta \in (0, 1)$ ,  $\max_{1 \leq j \leq k_n} |\phi_{nj}(t) - 1| \leq \delta/2$  for  $n$  large enough.

**Proof.** As  $\mathbb{E}[Y_{nj}] = 0$  for all  $1 \leq j \leq k_n$ , by using [Lemma 4.1.2](#), we have

$$\begin{aligned} \max_{1 \leq j \leq k_n} |\phi_{nj}(t) - 1| &= \max_{1 \leq j \leq k_n} |\mathbb{E}[e^{itY_{nj}} - 1 - itY_{nj}]| \\ &\leq \max_{1 \leq j \leq k_n} \mathbb{E}[|e^{itY_{nj}} - (1 + itY_{nj})|] \leq \frac{t^2}{2} \max_{1 \leq j \leq k_n} \mathbb{E}[Y_{nj}^2] \end{aligned}$$

From the [Lindeberg condition](#),  $\max_{1 \leq j \leq k_n} \mathbb{E}[Y_{nj}^2] \rightarrow 0$ , hence we're done.  $\circledast$

Hence, for any  $\delta \in (0, 1)$ , when  $n$  is large enough, [Lemma 4.1.4](#) and the above calculation gives,

$$\left| \prod_{j=1}^{k_n} \phi_{nj}(t) - \prod_{j=1}^{k_n} e^{\phi_{nj}(t)-1} \right| \leq \delta \sum_{j=1}^{k_n} |\phi_{nj}(t) - 1| \leq \delta \cdot \frac{t^2}{2} \sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = \frac{\delta t^2}{2}$$

since  $\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = \sum_{j=1}^{k_n} \text{Var}[Y_{nj}] = 1$ . By letting  $n \rightarrow \infty$ , and  $\delta \rightarrow 0$ , we see that the first term indeed goes to 0 as  $n \rightarrow \infty$ . As for the second term, it suffices to show that for every  $t \in \mathbb{R}$ ,

$$\sum_{j=1}^{k_n} (\phi_{nj}(t) - 1) \rightarrow -\frac{t^2}{2} \Leftrightarrow \sum_{j=1}^{k_n} (\phi_{nj}(t) - 1) + \frac{t^2}{2} \rightarrow 0 \Leftrightarrow \sum_{j=1}^{k_n} \left[ (\phi_{nj}(t) - 1) + \frac{t^2}{2} \text{Var}[Y_{nj}] \right] \rightarrow 0$$

as  $\sum_{j=1}^{k_n} \text{Var}[Y_{nj}] = 1$ . Since  $Y_{nj}$  is centered, we have  $\mathbb{E}[Y_{nj}] = 0$  and  $\text{Var}[Y_{nj}] = \mathbb{E}[Y_{nj}^2]$ , hence

$$\sum_{j=1}^{k_n} \left[ (\phi_{nj}(t) - 1) + \frac{t^2}{2} \text{Var}[Y_{nj}] \right] = \sum_{j=1}^{k_n} \mathbb{E} \left[ e^{itY_{nj}} - 1 - itY_{nj} - \frac{(itY_{nj})^2}{2} \right].$$

To bound this, we decompose it via the event  $|Y_{nj}| > \epsilon$  (for some  $\epsilon > 0$  to be determined later) as

$$\sum_{j=1}^{k_n} \mathbb{E} \left[ \left( e^{itY_{nj}} - 1 - itY_{nj} - \frac{(itY_{nj})^2}{2} \right) \mathbb{1}_{|Y_{nj}| > \epsilon} \right] + \sum_{j=1}^{k_n} \mathbb{E} \left[ \left( e^{itY_{nj}} - 1 - itY_{nj} - \frac{(itY_{nj})^2}{2} \right) \mathbb{1}_{|Y_{nj}| \leq \epsilon} \right].$$

We then see that

- from [Lemma 4.1.2](#), with  $x := tY_{nj}$ , we can bound the first term as

$$\sum_{j=1}^{k_n} \mathbb{E} \left[ \left( e^{itY_{nj}} - 1 - itY_{nj} - \frac{(itY_{nj})^2}{2} \right) \mathbb{1}_{|Y_{nj}| > \epsilon} \right] \leq t^2 \sum_{j=1}^{k_n} \mathbb{E} [Y_{nj}^2 \mathbb{1}_{|Y_{nj}| > \epsilon}] \rightarrow 0$$

as  $n \rightarrow \infty$  by the [Lindeberg condition](#);

- from [Lemma 4.1.3](#), with  $z := itY_{nj}$  such that  $|z| = |tY_{nj}| \leq |t|\epsilon$  under the event. Let  $\epsilon$  be defined such that  $|t|\epsilon < 1$ , then we can bound the second term by

$$\frac{1}{1-\epsilon} \sum_{j=1}^{k_n} \mathbb{E}[|tY_{nj}|^3 \cdot \mathbb{1}_{|Y_{nj}| \leq \epsilon}] \leq \frac{|t|^3}{1-\epsilon} \sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^2 \cdot \mathbb{1}_{|Y_{nj}| \leq \epsilon}] \leq \frac{|t|^3}{1-\epsilon} \sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^2] = \frac{|t|^3}{1-\epsilon} \epsilon$$

since again  $\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = \sum_{j=1}^{k_n} \text{Var}[Y_{nj}] = 1$ . By letting  $\epsilon \rightarrow 0$ ,  $|t|^3\epsilon/(1-\epsilon) \rightarrow 0$  as desired.

Hence, we see that both terms go to 0 when  $n \rightarrow \infty$ , so the second term indeed go to 0  $\blacksquare$

### 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** (Hajek-Sidak central limit theorem). If  $X_{ni} = c_{ni}Z_i^a$  for all  $1 \leq i \leq k_n$  where  $(Z_{k_n})$  are i.i.d. with  $\mathbb{E}[Z_i] = \mu$  and  $\text{Var}[Z_i] = \sigma^2$ . If  $\max_{1 \leq i \leq k_n} c_{ni}^2 / \sum_{i=1}^{k_n} c_{ni}^2 \rightarrow 0$  as  $n \rightarrow \infty$ , then the [Lindeberg condition](#) holds.

<sup>a</sup>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  $\text{Var}[S_n] = \sum_{j=1}^{k_n} c_{nj}^2 \sigma^2$ , hence with our usual notation, we define

$$Y_{nj} := \frac{c_{nj}(Z_j - \mu)}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2} \sigma} =: \frac{c_{nj}}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}} W_j,$$

where  $W_j := (Z_j - \mu)/\sigma$ . 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,

$$\leq \sum_{j=1}^{k_n} \frac{c_{nj}^2}{\sum_{i=1}^{k_n} c_{ni}^2} \mathbb{E} \left[ W_j^2 \mathbb{1}_{\max_{1 \leq k \leq k_n} \frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}} |W_k| > \epsilon} \right]$$

which makes the term in the expectation i.i.d., so we can replace both  $W_j$  and  $W_k$  by  $W := W_1$ ,

$$\begin{aligned} &= \sum_{j=1}^{k_n} \frac{c_{nj}^2}{\sum_{i=1}^{k_n} c_{ni}^2} \mathbb{E} \left[ W^2 \mathbb{1}_{\max_{1 \leq k \leq k_n} \frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}} |W| > \epsilon} \right] \\ &= \mathbb{E} \left[ W^2 \mathbb{1}_{\max_{1 \leq k \leq k_n} \frac{|c_{nk}|}{\sqrt{\sum_{i=1}^{k_n} c_{ni}^2}} |W| > \epsilon} \right]. \end{aligned}$$

Hence, it reduces to show  $\mathbb{E}[W^2 \mathbb{1}_{|W| > x}] \rightarrow 0$  as  $n \rightarrow \infty$  for  $x := \epsilon \sqrt{\sum_{i=1}^{k_n} c_{ni}^2} / \max_{1 \leq k \leq k_n} |c_{nk}|$ . From our assumption, for any  $\epsilon > 0$ ,  $x \rightarrow \infty$  as  $n \rightarrow \infty$ , hence the expectation indeed goes to 0 as long as  $W$  has finite second moment, which is indeed the case by our assumption.  $\blacksquare$

We see that [Hajek-Sidak central limit theorem](#) is very common in practice.

**Intuition.** Often time for every  $n$ ,  $c_{ni} = c_n$ , the same for every  $1 \leq i \leq k_n$ . In this case, we may write  $X_{ni} = c_n \cdot X_{ni}/c_n =: c_n Z_i$  such that  $Z_i := X_{ni}/c_n$  is i.i.d. distributed, ready for checking the [Hajek-Sidak condition](#).



Let's see three examples.

**Example (Uniform distribution).** For every  $n \geq 1$ , let  $(X_{nk_n}) \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(-c_n, c_n)$  for some  $c_n > 0$ . Then we see that  $Z_i := X_{ni}/c_n \sim \mathcal{U}(-1, 1)$  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  $Z_i := X_{ni}/c_n$  is now i.i.d. distributed.

**Example (Exponential distribution).** For every  $n \geq 1$ , let  $(X_{nk_n}) \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(1/c_n)$  for some  $c_n > 0$ , hence  $Z_i := X_{ni}/c_n \sim \text{Exp}(1)$  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  $\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^{2+\delta}] \rightarrow 0$  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

$$\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^2 \mathbb{1}_{|Y_{nj}| > \epsilon}] \leq \sum_{j=1}^{k_n} \mathbb{E} \left[ \left( \frac{|Y_{nj}|}{\epsilon} \right)^\delta |Y_{nj}|^2 \mathbb{1}_{|Y_{nj}| > \epsilon} \right] \leq \sum_{j=1}^{k_n} \mathbb{E} \left[ \left( \frac{|Y_{nj}|}{\epsilon} \right)^\delta |Y_{nj}|^2 \right] \rightarrow 0$$

by taking  $\epsilon^{-\delta}$  out and then the result follows from the assumption.  $\blacksquare$

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  $C_n/\sqrt{\text{Var}[S_n]} \rightarrow 0$ , the [Lindeberg condition](#) holds. In particular, when  $C_n =: C$ , it suffices to check  $\text{Var}[S_n] \rightarrow \infty$ .

**Proof.** We see that for every  $1 \leq j \leq k_n$ ,

$$|Y_{nj}| = \frac{|X_{nj} - \mathbb{E}[X_{nj}]|}{\sqrt{\text{Var}[S_n]}} \leq \frac{2C_n}{\sqrt{\text{Var}[S_n]}}.$$

In this case, for any  $\delta > 0$ , recall that  $\sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = 1$ , hence

$$\sum_{j=1}^{k_n} \mathbb{E}[|Y_{nj}|^{2+\delta}] \leq \left( \frac{2C_n}{\sqrt{\text{Var}[S_n]}} \right)^\delta \sum_{j=1}^{k_n} \mathbb{E}[Y_{nj}^2] = \left( \frac{2C_n}{\sqrt{\text{Var}[S_n]}} \right)^\delta \rightarrow 0,$$

hence the [Lyapunov's condition](#) holds.  $\blacksquare$

**Example (Bernoulli distribution).** For every  $n \geq 1$ , let  $X_{ni} \sim \text{Ber}(p_{ni})$  for all  $1 \leq i \leq k_n$ . Since  $X_{ni} \leq 1$ , from [Corollary 4.1.3](#), it suffices to check

$$\text{Var}[S_n] = \sum_{i=1}^{k_n} p_{ni}(1 - p_{ni}) \rightarrow \infty.$$

- If  $p_{ni} = 1/i$ : then  $\text{Var}[S_n] = \sum_{i=1}^{k_n} 1/i - \sum_{i=1}^{k_n} 1/i^2 \rightarrow \infty$ .
- If  $p_{ni} = p_n$ : then  $\text{Var}[S_n] = k_n p_n(1 - p_n)$ . In particular, for  $p_n = 1/n$  and  $k_n = n$ ,

$$\text{Var}[S_n] = n \cdot \frac{1}{n} \cdot \frac{n-1}{n} \rightarrow 1 \neq \infty.$$

In general, if  $np_n \rightarrow \lambda > 0$ , then the sum  $S_n$  [converges](#) to  $\text{Pois}(\lambda)$ .

## 4.2 The Two-Sample Problem

With this new tool in hand, i.e., the [Lindeberg central limit theorem](#), we will now apply it to study the [two-sample problem](#). First, let's see one illustrative problem, 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<sup>a</sup>

$$R_i = \sum_{j=1}^i \mathbb{1}_{X_j < X_i}.$$

In particular, if  $R_i = i$  ( $R_i = 1$ ), then  $X_i$  is the largest (smallest) of  $X_1, \dots, X_i$ . ⊛

<sup>a</sup>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 4.2.1.** Let  $(X_n) \stackrel{\text{i.i.d.}}{\sim} F$  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  $F$  is continuous (hence no ties),

$$\mathbb{P}(X_{\sigma(1)} < X_{\sigma(2)} < \dots < X_{\sigma(i)}) = \frac{1}{i!}$$

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  $\mathbb{E}[\sum_{i=1}^n R_i] = \sum_{i=1}^n \frac{i}{2} = n(n+1)/4$  and

$$\text{Var} \left[ \sum_{i=1}^n R_i \right] = \sum_{i=1}^n \frac{i^2 - 1}{12} = \frac{n(2n^2 + 3n - 5)}{72} \sim \frac{n^3}{36},$$

hence if the [Lindeberg central limit theorem](#) holds, then we're done. We check [Corollary 4.1.3](#): by noting that  $|R_i| \leq n$ , we indeed have  $n/\sqrt{\text{Var}[\sum_{i=1}^n R_i]} \rightarrow 0$ . ■

**Remark (Record).** We may instead consider the *record*  $Z_i = \mathbb{1}_{R_i=i}$ .<sup>a</sup> 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  $\sum_{i=1}^n Z_i$ .

<sup>a</sup>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  $H_0$ .

We now state the famous [two-sample problem](#) formally.

**Problem 4.2.1 (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 the task of testing the effect of a treatment. Ideally, we have the treatment group and the control group (the two populations), and we observe  $Y_i^T$ 's from the treatment group, and  $Y_i^C$ 's from the control group. It's common to pair the data together and get  $(Y_1^T, Y_1^C), (Y_2^T, Y_2^C), \dots, (Y_n^T, Y_n^C)$ .

Let's use the [treatment effect testing problem](#) as our running example. Let  $X_i := Y_i^T - Y_i^C$  for all  $1 \leq i \leq n$ . 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  $H_0: \mu = 0$ .

#### 4.2.1 Student's $t$ -test

Since we assume  $X_i$ 's are i.i.d., we may use the  $t$ -test, i.e., we reject  $H_0$  if the  [\$t\$ -statistic](#)

$$T_n = \frac{\sqrt{n} \bar{X}_n}{\hat{\sigma}_n} = \frac{\sum_{i=1}^n X_i}{\sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2}} = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}} > Z_\alpha$$

for some  $\alpha > 0$ . Since  $\text{Var}[X_i] < \infty$ ,  $T_n \xrightarrow{D} \mathcal{N}(0, 1)$  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. ⊛

The above answer is pretty vague, which is intended. The key idea is the following:

**Intuition.** 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

By conditioning on  $|X_j|$ 's, we overcome the non i.i.d. problem: consider the following  $H_0$  such that

26 Mar. 9:30

$$\forall 1 \leq i \leq n: \mathbb{P}(X_i = \pm |X_i| \mid |X_j|, 1 \leq j \leq n) = \frac{1}{2}.$$

In other words,  $\mathbb{P}(\text{sgn}(X_i) = \pm 1 \mid |X_j|, 1 \leq j \leq n) = 1/2$ . We see that  $\text{sgn}(X_i)$ 's are now i.i.d.

**Intuition.** 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.

Then, by writing  $\sum_{i=1}^n X_i = \sum_{i=1}^n |X_i| \text{sgn}(X_i)$  and treating  $|X_i|$ 's as constants, under  $H_0$ ,

$$\frac{\sum_{i=1}^n |X_i| \text{sgn}(X_i)}{\sqrt{\sum_{i=1}^n X_i^2}} \xrightarrow{D} \mathcal{N}(0, 1) \quad (4.1)$$

by the [Lindeberg central limit theorem](#). We can then use the left-hand side for our test as usual.

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  $T_n \xrightarrow{D} \mathcal{N}(0, 1)$  under  $H_0$ ?

Clearly, the answer to this problem 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  $\text{Var}[X_i]$ , or even  $\mathbb{E}[X_i]$ , doesn't exist?

It turns out that this is something we will solve along the way when we deal with the i.i.d. problem.

### 4.2.2 Sign Test

Our first goal is to avoid conditioning on  $|X_j|$ 's. Inspired by the above, consider the following.

**Intuition.** We may reject  $H_0$  if  $\sum_{i=1}^n \text{sgn}(X_i)$  is large, i.e., we simply discard all  $X_i$ 's in Equation 4.1.

This motivates us to define the **sign statistic**.

**Definition 4.2.1 (Sign statistic).** Given a sample  $X_1, \dots, X_n$ , the *sign statistic* is defined as

$$\text{sign}_n := \sum_{i=1}^n \text{sgn}(X_i) = 2 \sum_{i=1}^n \mathbb{1}_{X_i > 0} - n.$$

Observe that under  $H_0$ , the distribution of  $\text{sign}_n$  is independent of  $F$  since  $\sum_{i=1}^n \mathbb{1}_{X_i > 0} \stackrel{H_0}{\sim} \text{Bin}(n, 1/2)$ , i.e., we can conduct the hypothesis test non-asymptotically. On the other hand, under  $H_0$ , we also have asymptotic normality, i.e.,

$$\frac{1}{\sqrt{n}} \text{sign}_n = \frac{\sum_{i=1}^n \text{sgn}(X_i)}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$$

from the usual **central limit theorem**, so we can also reject  $H_0$  when  $\sum_{i=1}^n \text{sgn}(x_i)/\sqrt{n} \geq Z_\alpha$ .

### 4.2.3 Wilcoxon Signed-Rank Test

The sign test with the **sign statistic**  $\text{sign}_n$  is simple, but since we're not using the information of  $|X_j|$  at all, it's questionable that how powerful it is. To address this, let's again allow conditioning on  $|X_j|$ 's but design another test to get some sense. In this case, we can utilize  $|X_j|$ 's as follows.

**Intuition.** Replace  $|X_i|$ 's in Equation 4.1 by their "rank" information.

In particular, consider the *rank*  $R_i$  for  $1 \leq i \leq n$

$$R_i := \sum_{j=1}^n \mathbb{1}_{|X_j| \leq |X_i|}.$$

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 test).** Given a sample  $X_1, \dots, X_n$ , the *Wilcoxon signed-rank statistic* is defined as

$$W_n = \sum_{i=1}^n R_i \text{sgn}(X_i).$$

Then, the corresponding Wilcoxon signed-rank test is to reject  $H_0$  if  $W_n$  is "large." To figure out the critical value, since we condition on  $\{|X_j|\}_{j=1}^n$ , without loss of generality,  $R_i = i$  for every  $1 \leq i \leq n$ .<sup>1</sup> In this case,  $W_n = \sum_{i=1}^n i \text{sgn}(X_i)$ , hence under  $H_0$ ,

$$\frac{\sum_{i=1}^n i \text{sgn}(X_i)}{\sqrt{\sum_{i=1}^n i^2}} \xrightarrow{D} \mathcal{N}(0, 1)$$

by the **Lindeberg central limit theorem**. We can then use this to test  $H_0$ .

**Note.** Under  $H_0$ , this test statistic is again independent of  $F$ .

Without conditioning on  $|X_j|$ 's, the above doesn't hold. However, some sign tests are still possible.

<sup>1</sup>This is doable since  $|X_j|$ 's are treated as constants.

**Example (Yet another sign test).** To test whether the data is i.i.d., consider  $H_0: \text{sgn}(X_i) = \epsilon_i$  where  $\epsilon_i$  is a Rademacher random variable, i.e.,  $\mathbb{P}(X > 0) = \mathbb{P}(X < 0) = 1/2$  if  $X$  is symmetric.

Indeed, while the above doesn't need  $\mathbb{E}[X_i] < \infty$ , but if we assume  $\text{Var}[X_i] < \infty$ , then the Wilcoxon signed-rank test still work, just that now we have its original form, i.e.,  $W_n = \sum_{i=1}^n R_i \text{sgn}(X_i)$ . Let's first see a useful characterization of  $W_n$  to get started.

**Proposition 4.2.1.** Let  $h(x_1, x_2) := \mathbb{1}_{x_1+x_2 \geq 0}$ . Then we have

$$W_n - \sum_{i=1}^n \text{sgn}(X_i) = \sum_{i \neq j} \left( h(X_i, X_j) - \frac{1}{2} \right).$$

**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  $1 \leq i, j \leq n: i \neq j$  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  $S := S_1 = S_2$ , 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  $\mathbb{1}_{X_i > 0} - \mathbb{1}_{X_i < 0} = \text{sgn}(X_i)$ , we have

$$W_n - \sum_{i=1}^n \text{sgn}(X_i) = \frac{1}{2} \sum_{i \neq j} (\mathbb{1}_{X_i + X_j \geq 0} - \mathbb{1}_{X_i + X_j < 0}) = \frac{1}{2} \sum_{i \neq j} (2\mathbb{1}_{X_i + X_j \geq 0} - 1).$$

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

$$H_0: X, X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F \text{ where } F \text{ is continuous such that } X \stackrel{D}{=} -X,$$

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 exist.

**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),

$$W_n - \sum_{i=1}^n \text{sgn}(X_i) = \sum_{i \neq j} \left( h(X_i, X_j) - \frac{1}{2} \right)$$

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 close to the familiar form.

Indeed, under  $H_0$ ,  $\mathbb{E}[h(X_i, X_j)] = 1/2$  for  $i \neq j$  as

$$\begin{aligned} \mathbb{E}_{H_0}[h(X_i, X_j)] &= \mathbb{E}_{H_0}[h(X_1, X_2)] = \mathbb{P}(X_1 + X_2 \geq 0) \\ &= \mathbb{P}(X_1 \geq -X_2) \\ &= \int_{\mathbb{R}} \mathbb{P}(X_1 \geq -x) F(dx) = \int_{\mathbb{R}} \mathbb{P}(X \leq x) F(dx) = \mathbb{E}[F(X)] = \frac{1}{2} \end{aligned}$$

since  $F$  is continuous, and  $F(X) \sim \mathcal{U}(0, 1)$  as we have shown in the homework. Hence, under  $H_0$ ,

$$\sum_{i \neq j} \left( h(X_i, X_j) - \frac{1}{2} \right) = 2 \sum_{i < j} \left( h(X_i, X_j) - \frac{1}{2} \right) = 2 \sum_{i < j} (h(X_i, X_j) - \mathbb{E}[h(X_i, X_j)]) .$$

By rescaling by the number of terms in the summation, we're looking at

$$\frac{1}{\binom{n}{2}} \left( W_n - \sum_{i=1}^n \text{sgn}(X_i) \right) = \frac{2}{\binom{n}{2}} \sum_{i < j} (h(X_i, X_j) - \mathbb{E}[h(X_i, X_j)]) =: 2 \left( U_n - \frac{1}{2} \right)$$

where we define  $U_n$  for some permutation symmetric function  $h^2$  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.,

$$\frac{\sqrt{n}}{\binom{n}{2}} \left( W_n - \sum_{i=1}^n \text{sgn}(X_i) \right) = 2\sqrt{n} \left( U_n - \frac{1}{2} \right) \xrightarrow{D} \mathcal{N}(0, 1). \quad (4.2)$$

**Note.** This asymptotic result will imply an asymptotic distribution for  $W_n$  exactly alone since

$$2 \frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^n \text{sgn}(X_i) = \frac{2}{n-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \text{sgn}(X_i) \right) \xrightarrow{P} 0.$$

<sup>2</sup>Indeed, since  $\mathbb{E}[h(X_1, X_2)] = \mathbb{P}(X_1 + X_2 \geq 0)$ , so  $h(X_1, X_2) = h(X_2, X_1)$ .

### 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. Recall the following proposition we have shown in the homework.

**Proposition 4.3.1.** If  $\text{Var}[S_n]/\text{Var}[\tilde{S}_n] \rightarrow 1$  and  $\text{Corr}(S_n, \tilde{S}_n) \rightarrow 1$ , then

$$\left| \frac{S_n - \mathbb{E}[S_n]}{\sqrt{\text{Var}[S_n]}} - \frac{\tilde{S}_n - \mathbb{E}[\tilde{S}_n]}{\sqrt{\text{Var}[\tilde{S}_n]}} \right| \xrightarrow{p} 0.$$

Hence, to show  $(S_n - \mathbb{E}[S_n])/\sqrt{\text{Var}[S_n]} \xrightarrow{D} Y$ , we may find  $(\tilde{S}_n)$  such that  $(\tilde{S}_n - \mathbb{E}[\tilde{S}_n])/\sqrt{\text{Var}[\tilde{S}_n]} \xrightarrow{D} Y$  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.

## Lecture 19: Projection and $U$ -Statistic

### 4.3.1 Projection

28 Mar. 9:30

A natural approach to find such  $(\tilde{S}_n)$  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 the *projection* of  $Y \in L^2$  onto  $K$  if  $\|Y^* - Y\| \leq \|Z - Y\|$  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  $K$ .

**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  $Z \in K$ ,  $\mathbb{E}[(Y^* - Y)Z] = 0$ , with  $Y^* - Z \in K$ ,

$$= \|Y - Y^*\|^2 + \|Y^* - Z\|^2,$$

which is greater than  $\|Y - Y^*\|^2$ . ■

**Corollary 4.3.1.** Under the setup of [Proposition 4.3.2](#), if in addition,  $1 \in K$ , then  $\text{Cov}[Y, Y^*] = \text{Var}[Y^*]$ . In particular,  $\text{Corr}(Y, Y^*) = \sqrt{\text{Var}[Y^*]/\text{Var}[Y]}$ .

**Proof.** By taking  $Z = 1$  in [Proposition 4.3.2](#),  $\mathbb{E}[Y^*] = \mathbb{E}[Y]$ . On the other hand, for  $Z = Y^*$ , we have  $\mathbb{E}[YY^*] = \mathbb{E}[(Y^*)^2]$ . Hence,

$$\text{Cov}[Y, Y^*] = \mathbb{E}[YY^*] - \mathbb{E}[Y]\mathbb{E}[Y^*] = \mathbb{E}[(Y^*)^2] - \mathbb{E}[Y^*]^2 = \text{Var}[Y^*].$$

The correlation is then  $\text{Corr}(Y, Y^*) = \text{Var}[Y^*]/\sqrt{\text{Var}[Y]\text{Var}[Y^*]} = \sqrt{\text{Var}[Y^*]/\text{Var}[Y]}$ . ■

Following the intuition, we may select  $\tilde{S}_n$  as a [projection](#) of  $S_n$  onto some linear subspace  $K_n \subseteq L^2$  where  $1 \in K_n$ . If such  $K_n$ 's are found, as long as  $\text{Var}[S_n]/\text{Var}[\tilde{S}_n] \rightarrow 1$ , from [Corollary 4.3.1](#),  $\text{Corr}(Y, Y^*) \rightarrow 1$  is automatically satisfied, hence [Proposition 4.3.1](#) can be applied, i.e.:

**Intuition.** Find  $K_n$ 's such that  $\text{Var}[S_n]/\text{Var}[\tilde{S}_n] \rightarrow 1$ , and  $\tilde{S}_n - \mathbb{E}[\tilde{S}_n]$  [converges](#) somewhere.

Let's first see some examples of how to find [projections](#) in practice.

**Example.** Let  $X$  be a random vector, and let  $K = \{g(X) : g(X) \in L^2\}$ . Then  $Y^* = \mathbb{E}[Y | X]$ . More generally, let  $X_1, \dots, X_n$  be a sequence of random vectors, and let

$$K_n = \{g(X_1, \dots, X_n) : g(X_1, \dots, X_n) \in L^2\}.$$

Then  $Y^* = \mathbb{E}[Y | X_1, \dots, X_n]$ .

**Proof.** From Proposition 4.3.2, it suffices to show that for all  $g \in K$ ,

$$\mathbb{E}[\mathbb{E}[Y | X] \cdot g(X)] = \mathbb{E}[\mathbb{E}[g(X) \cdot Y | X]] = \mathbb{E}[g(X) \cdot Y]$$

from the basic properties for conditional expectation.  $\circledast$

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

$$K_n = \left\{ \sum_{i=1}^n g_i(X_i) : g_i(X_i) \in L^2 \text{ for all } 1 \leq i \leq n \right\}.$$

In this case,  $Y^* - \mathbb{E}[Y] = \sum_{i=1}^n \mathbb{E}[Y - \mathbb{E}[Y] | X_i]$ .

**Proof.** We want to show that for all  $g \in K_n$ ,  $\mathbb{E}[Y^* \sum_{i=1}^n g_i(X_i)] = \mathbb{E}[Y \sum_{i=1}^n g_i(X_i)]$ . Equivalently,

$$\mathbb{E} \left[ (Y^* - \mathbb{E}[Y^*]) \sum_{i=1}^n g_i(X_i) \right] = \mathbb{E} \left[ (Y - \mathbb{E}[Y]) \sum_{i=1}^n g_i(X_i) \right]$$

since  $\mathbb{E}[Y^*] = \mathbb{E}[Y]$ . Expanding the left-hand side, we have

$$\begin{aligned} \mathbb{E} \left[ (Y^* - \mathbb{E}[Y^*]) \sum_{i=1}^n g_i(X_i) \right] &= \mathbb{E} \left[ \left( \sum_{i=1}^n \mathbb{E}[Y - \mathbb{E}[Y] | X_i] \right) \sum_{j=1}^n g_j(X_j) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E} [\mathbb{E}[(Y - \mathbb{E}[Y]) | X_i] g_j(X_j)] \end{aligned}$$

observe that for  $i \neq j$ ,  $\mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y]) | X_i]] = \mathbb{E}[Y - \mathbb{E}[Y]]$  is independent of  $g_j(X_j)$ , hence

$$\begin{aligned} &= \sum_{i \neq j} \mathbb{E}[(Y - \mathbb{E}[Y])] \mathbb{E}[g_j(X_j)] + \sum_{i=1}^n \mathbb{E} [\mathbb{E}[Y - \mathbb{E}[Y] | X_i] \cdot g_i(X_i)] \\ &= \sum_{i=1}^n \mathbb{E} [\mathbb{E}[(Y - \mathbb{E}[Y]) \cdot g_i(X_i) | X_i]] \\ &= \mathbb{E} \left[ \sum_{i=1}^n (Y - \mathbb{E}[Y]) g_i(X_i) \right], \end{aligned}$$

where we again use the property of conditional expectation.  $\circledast$

### 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  $h : \mathbb{R}^m \rightarrow \mathbb{R}$ , the  $U$ -statistic is defined as

$$U_n = \frac{1}{\binom{n}{m}} \sum_{\substack{i_1, \dots, i_m \\ \{i_1, \dots, i_m\} \subseteq [n]}} h(X_{i_1}, \dots, X_{i_m}).$$



**Remark.** The  $U$ -statistic is an unbiased estimator of  $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$ .

**Notation.** Let  $[n] := \{1, 2, \dots, n\}$  for  $n \in \mathbb{N}$ .

**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  $(W_n - \sum_{i=1}^n \text{sgn}(X_i)) / \binom{n}{2} = 2U_n - 1$  where  $U_n$  is an  $U$ -statistic defined as

$$U_n = \frac{1}{\binom{n}{2}} \sum_{i < j} h(X_i, X_j).$$

**Note.** In the above, we see that  $i < j$  is the same as  $\{i, j\} \subseteq [n]$ . Indeed, one can assume  $i_1 < \dots < i_m$  without loss of generality.

Our goal is clear now: given the projection  $U_n^*$  of  $U_n$  onto  $K_n = \{\sum_{i=1}^n g_i(X_i)\}$ , we want to show

$$\left| \frac{U_n - \mathbb{E}[U_n]}{\sqrt{\text{Var}[U_n]}} - \frac{U_n^* - \mathbb{E}[U_n^*]}{\sqrt{\text{Var}[U_n]}} \right| = \left| \frac{U_n - \theta}{\sqrt{\text{Var}[U_n]}} - \frac{U_n^* - \theta}{\sqrt{\text{Var}[U_n]}} \right| \xrightarrow{P} 0$$

since  $\mathbb{E}[U_n] = \mathbb{E}[U_n^*] = \theta$  (recall Corollary 4.3.1 with  $Z = 1$ ). Let's first find  $U_n^*$ . From the last example,

$$\begin{aligned} U_n^* - \theta &= \sum_{i=1}^n \mathbb{E}[U_n - \theta \mid X_i] \\ &= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_1, \dots, i_m\} \subseteq [n]} \mathbb{E}[h(X_{i_1}, \dots, X_{i_m}) - \theta \mid X_k] \\ &= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_1, \dots, i_m\} \subseteq [n]} (\mathbb{E}[h(X_{i_1}, \dots, X_{i_m}) \mid X_k] - \theta). \end{aligned}$$

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,

$$= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_2, \dots, i_m\} \subseteq [n] \setminus \{k\}} (\mathbb{E}[h(X_k, X_{i_2}, \dots, X_{i_m}) \mid X_k] - \theta)$$

as  $X_i$ 's are i.i.d. and  $h$  is permutation symmetry, by letting  $\tilde{h}(x) := \mathbb{E}[h(x, X_2, \dots, X_m)]$ , we have

$$\begin{aligned} &= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \sum_{\{i_2, \dots, i_m\} \subseteq [n] \setminus \{k\}} (\tilde{h}(X_k) - \theta) \\ &= \sum_{k=1}^n \frac{1}{\binom{n}{m}} \binom{n-1}{m-1} (\tilde{h}(X_k) - \theta) \\ &= \frac{m}{n} \sum_{k=1}^n (\tilde{h}(X_k) - \theta). \end{aligned}$$

**Note.**  $\tilde{h}(X_k)$ 's are i.i.d., so we can apply the central limit theorem.

In particular, if  $\text{Var}[\tilde{h}(X)] \in (0, \infty)$ , then

$$\sqrt{n}(U_n^* - \theta) \xrightarrow{D} \mathcal{N}(0, m^2 \text{Var}[\tilde{h}(X)]),$$

hence, our second goal is achieved. We just need to show that  $\text{Var}[U_n] / \text{Var}[U_n^*] \rightarrow 1$ . Firstly, since

$$\text{Var}[U_n^*] = \frac{m^2}{n} \text{Var}[\tilde{h}(X)],$$

it reduces to show  $\text{Var}[U_n] \sim m^2 \text{Var}[\tilde{h}(X_i)]/n$ . Recall the usual rule for variance.

**As previously seen.** For  $S_n = \sum_{i=1}^n X_i$ ,  $\text{Var}[S_n] = \sum_i \text{Var}[X_i] + \sum_{i \neq j} \text{Cov}[X_i, X_j]$ .

In our case, we see that

$$\begin{aligned} \binom{n}{m}^2 \text{Var}[U_n] &= \sum_{\{i_1, \dots, i_m\} \subseteq [n]} \text{Var}[h(X_{i_1}, \dots, X_{i_m})] \\ &\quad + \sum_{\substack{\{i_1, \dots, i_m\}, \{i'_1, \dots, i'_m\} \subseteq [n] \\ \{i_1, \dots, i_m\} \neq \{i'_1, \dots, i'_m\}}} \text{Cov}[h(X_{i_1}, \dots, X_{i_m}), h(X_{i'_1}, \dots, X_{i'_m})]. \end{aligned}$$

The variance terms are clear since for every subset, they're all the same, hence

$$\sum_{\{i_1, \dots, i_m\} \subseteq [n]} \text{Var}[h(X_{i_1}, \dots, X_{i_m})] = \binom{n}{m} \text{Var}[h(X_1, \dots, X_m)] =: J_{0,m}.$$

For the covariance terms, we observe the following.

**Intuition.** If  $\{i_1, \dots, i_m\} \cap \{i'_1, \dots, i'_m\} = \emptyset$ , the covariance vanishes from independence.

Hence, we might consider iterate through subsets of  $i$ 's and  $i'$ 's with  $r \geq 1$  common indices.

**Example.** 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

$$\text{Cov}[h(X_1, \dots, X_r, X_{i_{r+1}}, \dots, X_{i_m}), h(X_1, \dots, X_r, X_{i'_{r+1}}, \dots, X_{i'_m})]$$

where  $\{i_{r+1}, \dots, i_m\} \cap \{i'_{r+1}, \dots, i'_m\} = \emptyset$ . 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.

Let denote the covariance for a particular  $r$  by  $J_{r,m}$ . Then, the second summation becomes

$$\sum_{\substack{\{i_1, \dots, i_m\}, \{i'_1, \dots, i'_m\} \subseteq [n] \\ \{i_1, \dots, i_m\} \neq \{i'_1, \dots, i'_m\}}} \text{Cov}[h(X_{i_1}, \dots, X_{i_m}), h(X_{i'_1}, \dots, X_{i'_m})] = \sum_{r=1}^{m-1} \binom{n}{m} \binom{m}{r} \binom{n-m}{m-r} J_{r,m},$$

where the summation is from  $r = 1$  to  $m - 1$  since it's impossible to have  $r = m$  as we require  $\{i_1, \dots, i_m\} \neq \{i'_1, \dots, i'_m\}$ . 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.

## Lecture 20: Comparing Different Tests for Symmetry

Continuing on the calculation, we now need to calculate  $J_{r,m}$ , which is

$$J_{r,m} = \text{Cov}[h(X_1, \dots, X_r, X_{r+1}, \dots, X_m), h(X_1, \dots, X_r, X_{m+1}, \dots, X_{2m-r})],$$

which implies

$$\text{Var}[U_n] = \frac{1}{\binom{n}{m}} \left( J_{0,m} + \sum_{r=1}^{m-1} \binom{m}{r} \binom{n-m}{m-r} J_{r,m} \right) \sim \frac{\binom{m}{1} \binom{n-m}{m-1}}{\binom{n}{m}} J_{1,m} \sim \frac{m^2}{n} J_{1,m}$$

with some algebra. It remains to show that  $\text{Var}[\tilde{h}(X)] = J_{1,m}$ . Indeed,

$$\begin{aligned}
 J_{1,m} &= \text{Cov}[(X_1, X_2, \dots, X_m), h(X_1, X_{m+1}, \dots, X_{2m-1})] \\
 &= \mathbb{E}[(h(X_1, X_2, \dots, X_m) - \theta) \cdot (h(X_1, X_{m+1}, \dots, X_{2m-1}) - \theta)] \\
 &= \int \mathbb{E}[h(x, X_2, \dots, X_m) - \theta] \cdot \mathbb{E}[h(x, X_{m+1}, \dots, X_{2m-1})] F(dx) \\
 &= \int (\tilde{h}(x) - \theta)(\tilde{h}(x) - \theta) F(dx) \\
 &= \mathbb{E}[(\tilde{h}(X) - \theta)^2] \\
 &= \text{Var}[\tilde{h}(X)].
 \end{aligned}$$

**Example.** Let  $\theta = \mathbb{E}[h(X_1, X_2)] = \mathbb{P}(X_1 + X_2 > 0)$ . Then,  $h(x) = \mathbb{P}(x + X > 0) = 1 - F(-x)$  with

$$\text{Var}[\tilde{h}(X)] = \text{Var}[F(-X)] > 0$$

when  $X$  is not trivial. E.g., if  $H_0: X \stackrel{D}{=} -X$  and  $X$  is continuous, under  $H_0$ ,  $\text{Var}[\tilde{h}(X)] = \text{Var}[F(X)] = 1/12$  as  $F(X) \sim \mathcal{U}(0, 1)$  since  $F$  is assumed to be continuous. Moreover,

$$\theta = \mathbb{P}(X_1 > -X_2) = \mathbb{E}[\tilde{h}(X)] = \mathbb{E}[1 - F(-X)] = 1 - \frac{1}{2} = \frac{1}{2}.$$

Therefore,  $\sqrt{n}(U_n - 1/2) \stackrel{D}{\rightarrow} \mathcal{N}(0, 2^2/12) = \mathcal{N}(0, 1/3)$ .

**Remark.** Recall that  $W_n = \sum_{k=1}^n \text{sgn}(X_k) R_k$  and Equation 4.2. Now, we conclude that

$$\frac{\sqrt{n}}{n(n-1)} \left( W_n - \sum_{i=1}^n \text{sgn}(X_i) \right) = \sqrt{n} \left( U_n - \frac{1}{2} \right) \stackrel{D}{\rightarrow} \mathcal{N} \left( 0, \frac{1}{3} \right).$$

This implies that under  $H_0$ ,  $W_n/n^{3/2} \stackrel{D}{\rightarrow} \mathcal{N}(0, 1/3)$ .

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,<sup>a</sup> the order statistic is complete and sufficient.

<sup>a</sup>Since we didn't assume anything about  $F$  excepts for the continuity.

## 4.4 Asymptotic Relative Efficiency of Tests

For the [two-sample problem](#), we have three potential tests, i.e., [t-statistic](#)  $T_n = \sqrt{n}\bar{X}_n/\hat{\sigma}_n$ , the sign test statistic  $\text{sign}_n = \sum_{i=1}^n \text{sgn}(X_i)/\sqrt{n}$ , and the [Wilcoxon signed-rank statistic](#)  $W_n = \sum_{i=1}^n R_i \text{sgn}(X_i)$ . A natural question is the following.

**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  $H_0: \theta = \theta_0$  and  $H_1: \theta > \theta_0$ .

**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  $\theta$ ,

$$\frac{T_n - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} \stackrel{D}{\rightarrow} \mathcal{N}(0, 1).$$

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  $\mathbb{P}_{\theta_0}(\text{reject}) \rightarrow \alpha$  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  $H_1$ ,

$$\mathbb{P}_\theta(\text{reject}) = \mathbb{P}_\theta \left( \frac{T_n - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} \geq \frac{\mu(\theta_0) - \mu(\theta)}{\sigma(\theta)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta)} \right) \xrightarrow{\theta \rightarrow \theta_0} 1$$

as  $n \rightarrow \infty$  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

$$1 - \beta = \mathbb{P}_{\theta^*} \left( \frac{T_n - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} \geq \frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)} \right) \rightarrow 1 - \Phi \left( \frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)} \right),$$

where the convergent is not rigorous since the right-hand side still depend on  $n$ . Anyway, this leads to<sup>3</sup>

$$Z_{1-\beta} = -Z_\beta = \frac{\mu(\theta_0) - \mu(\theta^*)}{\sigma(\theta^*)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta^*)} \Rightarrow \sqrt{n} \frac{\mu(\theta^*) - \mu(\theta_0)}{\sigma(\theta^*)} = Z_\beta + \frac{\sigma(\theta_0)}{\sigma(\theta^*)} Z_\alpha.$$

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  $\theta_n \rightarrow \theta_0$  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  $\sigma(\theta_n) \rightarrow \sigma(\theta_0)$ , the right-hand side becomes  $Z_\alpha + Z_\beta$ . If we further assume that  $\mu$  is differentiable at  $\theta_0$ , with  $\sqrt{n}(\mu(\theta_n) - \mu(\theta_0)) = \mu'(\theta_0)\sqrt{n}(\theta_n - \theta_0) + \sqrt{n} \cdot o(\theta_n - \theta_0)$ ,

$$\sqrt{n}(\theta_n - \theta_0) \rightarrow \frac{Z_\alpha + Z_\beta}{\mu'(\theta_0)/\sigma(\theta_0)}.$$

Let  $n^*$  be the  $n$  such that  $\theta_{n^*} = \theta_n = \theta^*$ . Then, we have

$$\sqrt{n^*} \rightarrow \frac{Z_\alpha + Z_\beta}{\frac{\mu'(\theta_0)}{\sigma(\theta_0)}(\theta_{n^*} - \theta_0)}.$$

<sup>3</sup>Recall that  $Z_\alpha$  is defined for the right-tail.

**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.  $\circledast$

**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  $\tilde{T}_n$  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  $\sqrt{n}(\theta_n - \theta_0) \rightarrow \xi$ , and suppose there exists  $\mu(\theta)$  and  $\sigma(\theta)$  such that

$$\sqrt{n} \frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)} \xrightarrow{D} \mathcal{N}(0, 1) \Leftrightarrow \mathbb{P}_{\theta_n} \left( \frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} \leq x \right) \rightarrow \Phi(x)$$

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

$$\mathbb{P}_{\theta_n}(\text{reject}) = \mathbb{P}_{\theta_n} \left( T_n > \mu(\theta_0) + \frac{\sigma(\theta_0)}{\sqrt{n}} Z_\alpha \right) = \mathbb{P}_{\theta_n} \left( \frac{T_n - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} > \frac{\mu(\theta_0) - \mu(\theta_n)}{\sigma(\theta_n)/\sqrt{n}} + Z_\alpha \frac{\sigma(\theta_0)}{\sigma(\theta_n)} \right).$$

If  $\mu$  is differentiable at  $\theta_0$  and  $\sigma$  is continuous at  $\theta_0$ , then as  $\sqrt{n}(\theta_n - \theta_0) \rightarrow \xi$ , 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

$$\sqrt{n^*} = \frac{Z_\alpha + Z_\beta}{\frac{\mu'(\theta_0)}{\sigma(\theta_0)}(\theta^* - \theta_0)}.$$

**Remark.** This confirms our heuristic argument. Moreover,  $n^*$  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  $\tilde{T}$  which has  $\tilde{\mu}$ ,  $\tilde{\sigma}$ , and  $\tilde{n}^*$ , such that it also satisfies all the assumptions, i.e., asymptotic normality, differentiability for  $\mu$ , and continuity for  $\sigma$ , 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  $\tilde{T}_n$  is defined as

$$\text{ARE}(T, \tilde{T}) = \frac{n^*}{\tilde{n}^*} = \left( \frac{\tilde{\mu}'(\theta_0)/\tilde{\sigma}(\theta_0)}{\mu'(\theta_0)/\sigma(\theta_0)} \right)^2.$$

Let compare the  $t$ -test, sign test, and Wilcoxon signed-rank test on the problem of testing symmetry.

**Problem.** Let  $\epsilon, \epsilon_1, \dots, \epsilon_n \stackrel{\text{i.i.d.}}{\sim} F$  where  $F$  is continuous and symmetric around 0, i.e.,  $\epsilon \stackrel{D}{=} -\epsilon$ . Furthermore, assuming  $X_{n1}, \dots, X_{nn} \stackrel{\text{i.i.d.}}{\sim} \theta_n + \epsilon_i$  such that  $\sqrt{n}(\theta_n - \theta_0) = \xi$  for some fixed  $\xi \geq 0$ . We're interested in testing whether  $H_0: X_{n1} \stackrel{D}{=} -X_{n1}$ . In other words,  $H_0: \theta_0 = 0$ .

**Example** (Sign test). The **slope** of the averaged **sign statistic**  $\overline{\text{sign}}_n := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_{ni} > 0}$  is  $2f(0)$ .

**Proof.** We first see that

$$\overline{\text{sign}}_n := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_{ni} > 0} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\epsilon_i > -\theta_n}.$$

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

$$\text{Var} \left[ \sum_{i=1}^n \mathbb{1}_{\epsilon_i > -\theta_n} \right] = nF(\theta_n)(1 - F(\theta_n)) \rightarrow nF(\theta_0)(1 - F(\theta_0)) \rightarrow \infty.$$

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  $\overline{\text{sign}}_n$  is  $\mu'(0)/\sigma(0) = 2f(0)$ . \*

Note that for the sign test, we don't need any moment assumption. However, for  $t$ -test, we need the second moment to exist.

**Example** ( $t$ -test). Suppose  $\text{Var}[\epsilon] = \sigma^2 < \infty$ , then the **slope** of the “normalized”  **$t$ -statistic**  $\tilde{T}_n := T_n/\sqrt{n}$  is  $1/\sigma$ .

**Proof.** Firstly, since  $\tilde{T}_n := T_n/\sqrt{n} = \bar{X}_n/\hat{\sigma}_n$ , with

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

we have

$$\tilde{T}_n = \frac{\bar{X}_n}{\hat{\sigma}_n} = \frac{\theta_n + \bar{\epsilon}_n}{\hat{\sigma}_n} = \left( \frac{\theta_n}{\hat{\sigma}_n} - \frac{\theta_n}{\sigma} \right) + \left( \frac{\bar{\epsilon}_n}{\hat{\sigma}_n} + \frac{\theta_n}{\sigma} \right) \Rightarrow \sqrt{n} \left( \tilde{T}_n - \frac{\theta_n}{\sigma} \right) = \sqrt{n} \theta_n \left( \frac{1}{\hat{\sigma}_n} - \frac{1}{\sigma} \right) + \sqrt{n} \frac{\bar{\epsilon}_n}{\hat{\sigma}_n}.$$

Since  $\theta_0 = 0$  and  $\sqrt{n}(\theta_n - \theta_0) = \xi$ , we have  $\sqrt{n}\theta_n \rightarrow \xi \geq 0$ ,  $1/\hat{\sigma}_n - 1/\sigma \xrightarrow{P} 0$ , so the first term goes

to 0. On the other hand, by the usual [central limit theorem](#),  $\sqrt{n}\bar{\epsilon}_n \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ , hence

$$\sqrt{n} \left( \tilde{T}_n - \frac{\theta_n}{\sigma} \right) = \sqrt{n} \frac{\tilde{T}_n - \frac{\theta_n}{\sigma}}{1} \xrightarrow{D} \mathcal{N}(0, 1).$$

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  $\tilde{T}_n$  is  $\mu'(0)/\sigma(0) = 1/\sigma$ . \*

This gives us a way to compare  $\overline{\text{sign}_n}$  and  $\tilde{T}_n$ , i.e.,

$$\text{ARE}(\tilde{T}_n, \overline{\text{sign}_n}) = (2f(0)\sigma)^2$$

**As previously seen.** From [Proposition 3.5.2](#),  $\text{ARE}(\bar{X}_n, \hat{\theta}_{1/2}) = (2f(0)\sigma)^2$ , exactly the same!

Let's see some actual example, where we can borrow from the previous calculation.<sup>4</sup>

**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

$$\text{ARE}(\tilde{T}_n, \overline{\text{sign}_n}) = \left( \frac{2\sigma}{\sqrt{2\pi}\sigma} \right)^2 = \frac{2}{\pi} < 1$$

**Example (Laplace).** If  $\epsilon \sim \text{Laplace}(\mu, b)$  with  $\sigma^2 = 2b^2$ , since  $f(0) = 1/\sqrt{2}\sigma$ , hence

$$\text{ARE}(\tilde{T}_n, \overline{\text{sign}_n}) = 4\sigma^2 \frac{1}{2\sigma^2} = 2 > 1,$$

**Example (Uniform).** If  $\epsilon \sim \mathcal{U}(-c, c)$  for some  $c$  such that  $\text{Var}[\epsilon] = \sigma^2$ , we have

$$\text{ARE}(\tilde{T}_n, \overline{\text{sign}_n}) = \frac{1}{3}.$$

**Proof.** We see that  $c$  should be

$$\sigma^2 = \frac{(2c)^2}{12} = \frac{c^2}{3} \Rightarrow c = \sqrt{3}\sigma,$$

so  $f(0) = 1/2c = 1/2\sqrt{3}\sigma$ . 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.** When testing  $X \stackrel{D}{=} -X$ , from [Equation 4.2](#), we have

$$\frac{\sqrt{n}}{\binom{n}{2}} \left( W_n - \sum_{i=1}^n \text{sgn}(X_i) \right) = 2\sqrt{n} \left( U_n - \frac{1}{2} \right) \xrightarrow{D} \mathcal{N}(0, 1)$$

where  $U_n := \frac{1}{\binom{n}{2}} \sum_{i < j} h(X_i, X_j)$  where  $h(x_1, x_2) = \mathbb{1}_{x_1 + x_2 \geq 0}$ .

In our case, since  $X_i = \theta_n + \epsilon_i$ , we have

$$U_n = \frac{1}{\binom{n}{2}} \sum_{\{i,j\} \subseteq [n]} \mathbb{1}_{X_i + X_j > 0} = \frac{1}{\binom{n}{2}} \sum_{\{i,j\} \subseteq [n]} \mathbb{1}_{\epsilon_i + \epsilon_j > -2\theta_n}.$$

<sup>4</sup>The previous example include [normal](#) and [Laplace](#).

**Example (Wilcoxon signed-rank test).** The **slope** of the corresponding  **$U$ -statistic**  $U_n$  of the **Wilcoxon signed-rank statistic**  $W_n$  is  $2\sqrt{3} \cdot \int f^2(x) dx$ .

**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 the theory of  **$U$ -statistic**,

$$U_n - \mathbb{E}[h_n(\epsilon_1, \epsilon_2)] = \frac{2}{n} \sum_{k=1}^n \left( \tilde{h}_n(\epsilon_k) - \mathbb{E}[\tilde{h}_n(\epsilon_k)] \right) + o_p(1)$$

where the 2 comes from  $m$ , i.e., the number of arguments in  $h$ . Here, we have

$$\tilde{h}_n(x) = \mathbb{E}[h(x, \epsilon)] = \mathbb{P}(x + \epsilon > -2\theta_n) = \mathbb{P}(\epsilon > -x - 2\theta_n) = \mathbb{P}(\epsilon \leq x + 2\theta_n) = F(x + 2\theta_n),$$

with  $\epsilon_i$ 's being i.i.d., by multiplying and dividing both sides by  $\sqrt{n}$  and  $\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}$ ,

$$\sqrt{n} \frac{U_n - \mathbb{E}[h(\epsilon_1, \epsilon_2)]}{\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}} = \frac{2}{\sqrt{n}} \sum_{k=1}^n \frac{\tilde{h}_n(\epsilon_k) - \mathbb{E}[F(\epsilon + 2\theta_n)]}{\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}} + o_p(1).$$

We finally see that by the usual **central limit theorem**, as long as  $\text{Var}[F(\epsilon + 2\theta_n)] > 0$ , we have

$$\sqrt{n} \frac{U_n - \mathbb{E}[F(\epsilon + 2\theta_n)]}{2\sqrt{\text{Var}[F(\epsilon + 2\theta_n)]}} \xrightarrow{D} \mathcal{N}(0, 1).$$

We see that

- $\mu(\theta) = \mathbb{E}[F(\epsilon + 2\theta)]$ , if we assume  $f$  exists, it's just

$$\mu(\theta) = \int_{\mathbb{R}} F(x + 2\theta) F(dx) = \int_{\mathbb{R}} F(x + 2\theta) f(x) dx.$$

If we further assume that we can interchange the derivative and the integral, we then have  $\mu'(\theta) = \int_{\mathbb{R}} 2f(x + 2\theta) f(x) dx$ , giving  $\mu'(0) = 2 \int_{\mathbb{R}} f^2(x) dx$ .

- $\sigma(\theta) = 2\sqrt{\text{Var}[F(\epsilon + 2\theta)]}$ , hence  $\sigma(0) = 2\sqrt{\text{Var}[F(\epsilon)]} = 2 \cdot \sqrt{1/12} = 1/\sqrt{3}$ .

Hence, the **slope** of  $U_n$  is  $2\sqrt{3} \cdot \int_{\mathbb{R}} f^2(x) dx$ . ⊛

This gives us a way to compare  $U_n$  and  $\tilde{T}_n$ , i.e.,

$$\text{ARE}(\tilde{T}_n, U_n) = \frac{2\sqrt{3} \int_{\mathbb{R}} f^2(x) dx}{1/\sigma} = 2\sqrt{3}\sigma \int_{\mathbb{R}} f^2(x) dx.$$

## Lecture 22: The Theory of Linear Rank Statistics

### 4.5 Linear Rank Statistics

9 Apr. 9:30

Consider  $X_1, \dots, X_N, \dots \stackrel{\text{i.i.d.}}{\sim} F$  such that  $F$  is continuous, and for  $1 \leq i \leq N$ , let the *rank* be

$$R_{Ni} := \sum_{j=1}^N \mathbb{1}_{X_j \leq X_i},$$

i.e.,  $R_{Ni}$  is the rank of the  $i^{\text{th}}$  observation among the first  $N$ .

**Remark.** Clearly, as  $N$  varies,  $R_{Ni}$  is distributed differently. In particular,  $R_{Ni} \sim \mathcal{U}([N])$  for all  $1 \leq i \leq N$ , but they are not independent. On the other hand, since  $F$  is continuous,  $U_i := F(X_i) \stackrel{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1)$  for all  $i \geq 1$ , and we have  $R_{Ni} = \sum_{j=1}^N \mathbb{1}_{U_j \leq U_i}$  almost surely.

Now, let's revisit the **two-sample problem**.



**Example (Two-sample problem).** Consider two samples  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$  and  $Y_1, \dots, Y_m \stackrel{\text{i.i.d.}}{\sim} G$ , and we want to test whether  $H_0: F = G$ . Set  $X_{n+i} = Y_i$  for  $1 \leq i \leq m$ , which gives

$$X_1, \dots, X_n, X_{n+1} = Y_1, \dots, X_{n+m} = Y_m$$

with  $N := n + m$ . Under  $H_0$ ,  $X_1, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} F$ . Naively, we will reject  $H_0$  when  $\bar{Y}_m - \bar{X}_n$  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

$$c_{Ni} = \begin{cases} 1, & \text{if } n < i \leq N; \\ -1, & \text{if } 1 \leq i \leq n. \end{cases}$$

From our experience, one might replace  $X_i$  by  $R_{Ni}$ , i.e., consider

$$\sum_{i=1}^N c_{Ni} R_{Ni} = \sum_{i=1}^N (2\tilde{c}_{Ni} - 1) R_{Ni} = 2 \sum_{i=1}^N \tilde{c}_{Ni} R_{Ni} - \sum_{i=1}^N R_{Ni} = 2 \sum_{i=n+1}^N R_{Ni} - \frac{N(N+1)}{2}$$

since  $\sum_{i=1}^N R_{Ni} = \sum_{i=1}^N i$ , and we define

$$\tilde{c}_{Ni} = \frac{c_{Ni} + 1}{2} = \begin{cases} 1, & \text{if } n < i \leq N; \\ 0, & \text{if } 1 \leq i \leq n. \end{cases}$$

Observe that  $\sum_{i=n+1}^N R_{Ni}$  is just a Wilcoxon two-sample rank statistic.

This suggests we look into the following.

**Definition 4.5.1 (Linear rank statistic).** Consider  $X_1, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} F$  where  $F$  is continuous. The *linear rank statistic* is defined as

$$\sum_{i=1}^N c_{Ni} \alpha_N(R_{Ni}),$$

where  $\alpha_N(i) =: \alpha_N$  for  $1 \leq i \leq N$  and  $c_{N1}, \dots, c_{NN}$  are all constants.

**Remark.** The distribution of any *linear rank statistics* is independent of  $F$ !

**Example (Median statistic).** We can also consider  $\sum_{i=n+1}^N \mathbb{1}_{R_{Ni} \geq (N+1)/2}$ .

**Example (Simple random sampling).** Given a finite population  $\{x_1, \dots, x_N\}$ , say we want to estimate the population average  $(x_1 + \dots + x_N)/N$ . To do this, 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, we observe that  $(R_{N1}, \dots, R_{NN})$  is equally likely to take any permutation of  $[N]$ . Hence, the above indicators can be defined as  $\mathbb{1}_{R_{Ni} \leq n}$ , which suggests  $\alpha_N(R_{Ni}) := \mathbb{1}_{R_{Ni} \leq n}$  in the above notation.

**Notation.** We write  $R_{\sim N} := (R_{N1}, \dots, R_{NN})$ .

All these examples motivates us to develop a general theory for the *linear rank statistic* in the form of  $T_N := \sum_{i=1}^N c_{Ni} \alpha_N(R_{Ni})$ . It helps to compute the expectation and the variance of  $T_N$  to get some sense what it really is. Consider the following notations.

**Notation.** We write  $\bar{\alpha}_N$  and  $\bar{c}_N$  to be the population mean of  $\alpha_{Ni}$  and  $c_{Ni}$ , i.e.,

$$\bar{\alpha}_N := \frac{1}{N} \sum_{i=1}^N \alpha_N(i), \text{ and } \bar{c}_N := \frac{1}{N} \sum_{i=1}^N c_{Ni}.$$

Moreover, let  $\sigma_{N\alpha}^2$  and  $\sigma_{Nc}^2$  to be the population variance of  $\alpha_{Ni}$  and  $c_{Ni}$  respectively, i.e.,<sup>a</sup>

$$\sigma_{N\alpha}^2 = \frac{1}{N} \sum_{i=1}^N (\alpha_N(i) - \bar{\alpha}_N)^2, \text{ and } \sigma_{Nc}^2 = \frac{1}{N} \sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2.$$

<sup>a</sup>Recall that for any individual  $R_{Ni}$ , marginally they're identically distributed.

#### 4.5.1 Asymptotically Normality of Linear Rank Statistics

Let's first compute the expectation.

**Claim.**  $\mathbb{E}[T_N] = N\bar{\alpha}_N\bar{c}_N$ .

**Proof.** Since marginally,  $R_{Ni}$ 's are just uniform over  $[N]$ , hence the expectation of  $T_N$  is

$$\mathbb{E}[T_N] = \sum_{i=1}^N c_{Ni} \mathbb{E}[\alpha_N(R_{Ni})] = \sum_{i=1}^N c_{Ni} \mathbb{E}[\alpha_N(R_{N1})] = \sum_{i=1}^N c_{Ni} \sum_{j=1}^N \frac{\alpha_N(j)}{N} =: N\bar{\alpha}_N\bar{c}_N$$

as  $\bar{\alpha}_N = \mathbb{E}[\alpha_N(R_{N1})] = \frac{1}{N} \sum_{i=1}^N \alpha_{Ni}$  and  $\bar{c}_N = \frac{1}{N} \sum_{i=1}^N c_{Ni}$ . \*

Computing the variance is a bit more challenging, but still doable.

**Claim.**  $\text{Var}[T_N] = \frac{N^2}{N-1} \sigma_{Nc}^2 \sigma_{N\alpha}^2$ .

**Proof.** Let's first center  $T_N$ , which gives

$$\begin{aligned} T_N - \mathbb{E}[T_N] &= \sum_{i=1}^N c_{Ni} \alpha_N(R_{Ni}) - \sum_{i=1}^N c_{Ni} \bar{\alpha}_N \\ &= \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni}) + \bar{c}_N \sum_{i=1}^N \alpha_N(R_{Ni}) - \sum_{i=1}^N c_{Ni} \bar{\alpha}_N \end{aligned}$$

and with  $\sum_{i=1}^N \alpha_N(R_{Ni}) = \sum_{i=1}^N \alpha_N(i)$ , the last two terms cancel out, and we have

$$= \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni}).$$

Then, by the definition of variance,

$$\begin{aligned} \text{Var}[T_N] &= \text{Var} \left[ \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni}) \right] \\ &= \sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 \text{Var}[\alpha_N(R_{Ni})] + \sum_{i \neq j} (c_{Ni} - \bar{c}_N)(c_{Nj} - \bar{c}_N) \text{Cov}[\alpha_N(R_{Ni}), \alpha_N(R_{Nj})]. \end{aligned}$$

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  $\{(i, j) : 1 \leq i \neq j \leq N\}$ .

Hence, we can replace  $\text{Cov}[\alpha_N(R_{Ni}), \alpha_N(R_{Nj})]$  by  $\text{Cov}[\alpha_N(R_{N1}), \alpha_N(R_{N2})]$ , and focus only on  $\sum_{i \neq j} (c_{Ni} - \bar{c}_N)(c_{Nj} - \bar{c}_N)$ . In particular, we have the following.

**Note.** For any sequence  $(x_N)$ , we have  $\sum_{i \neq j} (x_i - \bar{x}_N)(x_j - \bar{x}_N) = -\sum_{i=1}^N (x_i - \bar{x}_N)^2$ .

**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

$$0 = \left( \sum_{i=1}^N (x_i - \bar{x}_N) \right)^2 = \sum_{i=1}^N (x_i - \bar{x}_N)^2 + \sum_{i \neq j} (x_i - \bar{x}_N)(x_j - \bar{x}_N).$$

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,

$$\begin{aligned} & \sum_{i \neq j} (c_{Ni} - \bar{c}_N)(c_{Nj} - \bar{c}_N) \text{Cov}[\alpha_N(R_{Ni}), \alpha_N(R_{Nj})] \\ &= \text{Cov}[\alpha_N(R_{N1}), \alpha_N(R_{N2})] \cdot \sum_{i \neq j} (c_{Ni} - \bar{c}_N)(c_{Nj} - \bar{c}_N) \\ &= \left[ \frac{1}{N(N-1)} \sum_{i \neq j} (\alpha_{Ni} - \bar{\alpha}_N)(\alpha_{Nj} - \bar{\alpha}_N) \right] (-N\sigma_{Nc}^2) \\ &= \left[ -\frac{1}{N(N-1)} \sum_{i=1}^N (\alpha_{Ni} - \bar{\alpha}_N)^2 \right] (-N\sigma_{Nc}^2) = \frac{N}{N-1} \sigma_{N\alpha}^2 \sigma_{Nc}^2. \end{aligned}$$

Putting everything together, we have

$$\text{Var}[T_N] = N\sigma_{Nc}^2 \sigma_{N\alpha}^2 + \frac{N}{N-1} \sigma_{N\alpha}^2 \sigma_{Nc}^2 = \frac{N^2}{N-1} \sigma_{Nc}^2 \sigma_{N\alpha}^2,$$

which gives the desired result. ⊗

With the above calculation, we now want to establish the asymptotic normality for the **linear rank statistic**. Specifically, we consider a special form of  $\alpha_N(i)$  given some  $\phi: [0, 1] \rightarrow \mathbb{R}$ , i.e., for  $1 \leq i \leq N$ ,

$$\alpha_N(i) = \phi\left(\frac{i}{N+1}\right).$$

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  $U_{N(1)} \leq \dots \leq U_{N(N)}$  for the uniform  $U_i$ 's. Then, for  $1 \leq i \leq N$ ,  $\mathbb{E}[U_{N(i)}] = i/(N+1)$ , implying  $\alpha_N(i) = \phi(\mathbb{E}[U_{N(i)}])$ .

**Example.** The **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  $\alpha_N(R_{Ni}) = \mathbb{E}[\phi(U_i) \mid R_{\sim N}]$ .

The above shows that  $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$  is more convenient since we will have

$$T_N = \sum_{i=1}^N c_{Ni} \alpha_N(R_{Ni}) = \sum_{i=1}^N c_{Ni} \mathbb{E}[\phi(U_i) \mid R_{\sim N}],$$

which is a conditional expectation. We can then apply the theory of **projection**.

## Lecture 23: Asymptotically Normality of Linear Rank Statistics

We will now work with  $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$ . It'll be convenient to consider

11 Apr. 9:30

$$T_N = \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni}) + \bar{c}_N \sum_{i=1}^N \alpha_N(R_{Ni}) = \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni}) + N \bar{c}_N \bar{\alpha}_N,$$

with the fact that  $\mathbb{E}[T_N] = N \bar{c}_N \bar{\alpha}_N$ , we see that

$$T_N - \mathbb{E}[T_N] = \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \alpha_N(R_{Ni})$$

and since  $\alpha_N(R_{Ni}) = \mathbb{E}[\phi(U_i) \mid R_{\sim N}]$ , we further have

$$= \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \mathbb{E}[\phi(U_i) \mid R_{\sim N}] = \mathbb{E} \left[ \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \phi(U_i) \mid R_{\sim N} \right] =: \mathbb{E}[\tilde{T}_N \mid R_{\sim N}]$$

where we let  $\tilde{T}_N := \sum_{i=1}^N (c_{Ni} - \bar{c}_N) \phi(U_i)$ .

**Remark.** We can easily have asymptotically normality for  $\tilde{T}_N$ .

**Proof.** We see that by the [Hajek-Sidak central limit theorem](#), if  $\phi(U_i)$ 's are i.i.d. such that

$$0 < \mathbb{E}[\phi^2(U)] = \int_0^1 \phi^2(u) du < \infty, \text{ and } \max_{1 \leq i \leq N} \frac{(c_{Ni} - \bar{c}_N)^2}{\sum_{j=1}^N (c_{Nj} - \bar{c}_N)^2} \rightarrow 0,$$

then  $\tilde{T}_N / \sqrt{\text{Var}[\tilde{T}_N]} \xrightarrow{D} \mathcal{N}(0, 1)$  as  $\mathbb{E}[\tilde{T}_N] = 0$ . ⊗

We can then apply the [projection](#) theory we have developed. In particular, to show  $T_N$  is asymptotically normal, from [Corollary 4.3.1](#), it suffices to show that  $\text{Var}[T_N] / \text{Var}[\tilde{T}_N] \rightarrow 1$  as  $n \rightarrow \infty$ . Now,

- $\text{Var}[T_N] = \frac{N^2}{N-1} \sigma_{N\alpha}^2 \alpha_{Nc}^2$ ;
- $\text{Var}[\tilde{T}_N] = N \sigma_{Nc}^2 \text{Var}[\phi(U_1)]$  since

$$\text{Var}[\tilde{T}_N] = \text{Var}[\phi(U_1)] \cdot \sum_{i=1}^N (c_{Ni} - \bar{c}_N)^2 = N \sigma_{Nc}^2 \text{Var}[\phi(U)],$$

so it suffices to show  $\sigma_{N\alpha}^2 = \text{Var}[\alpha_N(R_{N1})] \rightarrow \text{Var}[\phi(U_1)]$ .

**Note.** To show  $\text{Var}[X] \rightarrow \text{Var}[Y]$ , it suffices to show  $X \xrightarrow{L^2} Y$  as this will imply convergence for both the first and second moments, hence the variance.

In particular, it reduces to show  $\alpha_N(R_{N1}) = \mathbb{E}[\phi(U_1) \mid R_{\sim N}] \xrightarrow{L^2} \phi(U_1)$ . Firstly, we write

$$\mathbb{E}[\phi(U_1) \mid R_{\sim N}] = \mathbb{E}[\phi(U_1) \mid R_{\sim 1}, \dots, R_{\sim N}]$$

as the condition on the right-hand side is equivalent to  $R_{\sim N}, U_2, \dots, U_N$ , and  $U_1$  is independent of  $U_2, \dots, U_N$ . From the theory of martingale, we have

$$\mathbb{E}[\phi(U_1) \mid R_{\sim 1}, \dots, R_{\sim N}] \xrightarrow{\text{a.s.}} \mathbb{E}[\phi(U_1) \mid R_{\sim N}, N \geq 1],$$

which is 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  $\{R_{\sim N}\}_{N \geq 1}$ . This is proved the following.

**Claim.** Knowing the first components of  $R_{\sim N}$  for every  $N \geq 1$ , i.e.,  $R_{N1}$ , determines  $U_1$ .

**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](#), we will have  $U_1 \xrightarrow{L^2} R_{N1}/(N+1)$  if and only if the ratio between the variances converges to 1. Indeed,

$$\frac{\text{Var}[U_1]}{\text{Var}\left[\frac{R_{N1}}{N+1}\right]} = \frac{1/12}{\frac{1}{(N+1)^2} \text{Var}[R_{N1}]} = \frac{1/12}{\frac{1}{(N+1)^2} \frac{N^2-1}{12}} = \frac{N+1}{N-1} \rightarrow 1$$

since  $R_{N1} \sim \mathcal{U}([N])$ , hence we're done.  $\circledast$

This concludes that for  $\alpha_N(i) = \mathbb{E}[\phi(U_{N(i)})]$ ,  $T_N$  is asymptotically normal. To study the connection to  $\alpha'_{Ni} = \phi(i/(N+1))$ , consider writing

$$T_N = \sum_{i=1}^N c_{Ni} \alpha_N(R_{Ni}), \text{ and } T'_N = \sum_{i=1}^N c_{Ni} \alpha'_N(R_{Ni}),$$

and we know  $T_N$  is asymptotically normal. To show  $T'_N$  is also asymptotically normal, i.e.,

$$\frac{T'_N - \mathbb{E}[T'_N]}{\sqrt{\text{Var}[T'_N]}} \xrightarrow{D} \mathcal{N}(0, 1).$$

It suffices to show that

$$\frac{T_N - \mathbb{E}[T_N]}{\sqrt{\text{Var}[T_N]}} - \frac{T'_N - \mathbb{E}[T'_N]}{\sqrt{\text{Var}[T'_N]}} = \frac{\text{Var}[T_N - T'_N]}{\text{Var}[T_N]} \xrightarrow{L^2} 0.$$

**As previously seen.** We have  $\text{Var}[T_N] = \frac{N^2}{N-1} \sigma_{Nc}^2 \sigma_{N\alpha}^2$ .

This directly gives  $\text{Var}[T_N - T'_N] = \frac{N^2}{N-1} \sigma_{Nc}^2 \sigma_{N(\alpha-\alpha')}^2$  since we can write

$$T_N - T'_N = \sum_{i=1}^N c_{Ni} (\alpha_N - \alpha'_N)(R_{Ni}),$$

and the same calculation applies. Hence, it suffices to show  $\sigma_{N(\alpha-\alpha')}^2 / \sigma_{N\alpha}^2 \rightarrow 0$ .

**As previously seen.** We have proved that  $\sigma_{N\alpha}^2 \rightarrow \text{Var}[\phi(U_1)] > 0$ .

Hence, we just need to show that  $\sigma_{N(\alpha-\alpha')}^2 = \text{Var}[(\alpha_N - \alpha'_N)(R_{N1})] \rightarrow 0$ , which again, suffice to show that  $(\alpha_N - \alpha'_N)(R_{N1}) \xrightarrow{L^2} 0$ , i.e.,

$$\mathbb{E}[\phi(U_1) \mid R_{\sim N}] - \phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{L^2} 0.$$

With  $\mathbb{E}[\phi(U_1) \mid R_{\sim N}] \xrightarrow{L^2} \phi(U_1)$ , it reduces to show  $\phi(R_{N1}/(N+1)) \xrightarrow{L^2} \phi(U_1)$ .

**Example.** If  $\phi = \text{id}$ , then we have showed that  $R_{N1}/(N+1) \xrightarrow{L^2} U_1$  in the [previous claim](#).

If we assume that the above holds almost surely, then for an almost surely continuous  $\phi$ , we will have

$$\phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{\text{a.s.}} \phi(U_1) \xrightarrow{?} \phi\left(\frac{R_{N1}}{N+1}\right) \xrightarrow{L^2} \phi(U_1).$$

The implication can be provided by [Scheffé's theorem](#), i.e., we check<sup>5</sup>

$$\limsup_{N \rightarrow \infty} \mathbb{E}\left[\phi^2\left(\frac{R_{N1}}{N+1}\right)\right] = \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \phi^2\left(\frac{i}{N+1}\right) \leq \mathbb{E}[\phi^2(U_1)] = \int_0^1 \phi^2(u) du$$

since  $R_{N1} \sim \mathcal{U}([N])$  as we mentioned.

<sup>5</sup>[Scheffé's theorem](#) actually requires  $\mathbb{E}[\phi^2(R_{N1}/(N+1))] \rightarrow \mathbb{E}[\phi^2(U_1)] < \infty$ . However, one direction is trivial by [Fatou's lemma](#), so we only have another.

**Intuition.** We can understand the left-hand side as a Riemann integral.

**Example.** Both conditions, i.e.,  $\phi(R_{N1}/(N+1)) \xrightarrow{\text{a.s.}} \phi(U_1)$  and the condition for [Scheffé's theorem](#) are satisfied when  $\phi$  is increasing.

**Proof.** Since we can write

$$\begin{aligned} \int_0^1 \phi^2(u) \, du &= \lim_{N \rightarrow \infty} \sum_{i=1}^{N+1} \int_{\frac{i-1}{N+1}}^{\frac{i}{N+1}} \phi^2(u) \, du \\ &\geq \lim_{N \rightarrow \infty} \sum_{i=1}^{N+1} \phi^2\left(\frac{i-1}{N+1}\right) \cdot \frac{1}{N+1} \geq \lim_{N \rightarrow \infty} \frac{N}{N+1} \frac{1}{N} \sum_{i=1}^N \phi^2\left(\frac{i}{N+1}\right), \end{aligned}$$

which is what we want. \*

## Chapter 5

# Maximum Likelihood Method and $M$ -Estimation

### Lecture 24: The Maximum Likelihood Estimator

Consider  $X, X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$ , 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  $T(\hat{F}_n)$ . 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  $T(F) = \mu_k = \int (x - \mu)^k F(dx)$  where  $\mu = \int x F(dx)$ , which is estimated by the sample  $k^{\text{th}}$ -central moment  $T(\hat{F}_n) = M_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k$ .

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 estimator).** Given  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$  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  $\frac{1}{n} \sum_{i=1}^n \log(g_\theta(X_i))$ .<sup>1</sup> As  $n \rightarrow \infty$ , from the **strong law of large number**, we know that for all  $\theta \in \Theta$ ,

$$\frac{1}{n} \sum_{i=1}^n \log(g_\theta(X_i)) \xrightarrow{\text{a.s.}} \mathbb{E}[\log(g_\theta(X))] =: M(\theta) \in [-\infty, \infty].$$

<sup>1</sup>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

$$M(\theta) = \int_{\mathbb{R}} \log g_{\theta}(x) f(x) dx = \int_{\mathbb{R}} \log \frac{g_{\theta}(x)}{f(x)} f(x) dx + \int_{\mathbb{R}} \log f(x) f(x) dx.$$

The second term is independent of  $\theta$ , and the first term is just the **Kullback-Leibler divergence** since

$$\int_{\mathbb{R}} \log \frac{g_{\theta}(x)}{f(x)} f(x) dx = - \int_{\mathbb{R}} \log \frac{f(x)}{g_{\theta}(x)} f(x) dx = -\text{KL}(f \| g_{\theta}).$$

**Remark.** The maximizer of  $M(\theta)$  is minimizing the **KL divergence**  $\text{KL}(f \| g_{\theta})$ .

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  $\mathbb{P}(0 \leq X \leq u) = 1$  for some  $u > 0$  such that for all  $\epsilon > 0$ ,  $F(u - \epsilon) < 1 = F(u)$ . Consider the family of uniform distribution  $\{G_{\theta} = \mathcal{U}(0, \theta) : \theta > 0\}$ . The likelihood function is given by

$$\Theta \ni \theta \mapsto \prod_{i=1}^n \frac{1}{\theta^n} \mathbb{1}_{\max_{1 \leq i \leq n} X_i \leq \theta},$$

hence the maximizer of the likelihood function is given by  $\hat{\theta}_n = \max_{1 \leq i \leq n} X_i$ .

On the other hand,  $M(\theta) = \mathbb{E}[\log g_{\theta}(X)]$  for all  $\theta > 0$ , and want to find  $\sup_{\theta > 0} M(\theta)$ . But if  $\theta < u$ ,  $\mathbb{P}(g_{\theta}(x) = 0) > 0$ , then  $M(\theta) = -\infty$ , hence

$$\sup_{\theta > 0} M(\theta) = \sup_{\theta \geq u} M(\theta) = \sup_{\theta \geq u} \log \left( \frac{1}{\theta} \right),$$

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

- $\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^n -(X_i - \theta)^2 = \bar{X}_n$ ;
- $\theta^* = \arg \max_{\theta \in \Theta} \mathbb{E}[-(X - \theta)^2] = \mathbb{E}[X]$ .

From **strong law of large number**,  $\bar{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}[X]$  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](#),  $\bar{X}_n$  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  $G \sim \text{Cauchy}$  with location 0 and scale 1, i.e.,  $g(x) = \frac{1}{\pi(1+x^2)}$ , then

- $\hat{\theta}_n = \arg \max_{\theta \in \Theta} \sum_{i=1}^n -\log(1 + (X_i - \theta)^2)$ ;
- $\theta^* = \arg \max_{\theta \in \Theta} \mathbb{E}[-\log(1 + (X - \theta)^2)]$ .

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  $m(x) = x^2 \mathbb{1}_{|x| \leq k} + (2k|x| - k^2) \mathbb{1}_{|x| > k}$  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).**

fill

**Intuition.** This allows us to not specifying the family  $\{G_\theta: \theta \in \Theta\}$ .

More generally, consider  $X_1, \dots, X_n$ , not necessarily i.i.d. For all  $\theta \in \Theta$  where  $\Theta$  is a metric space, consider a function  $M_n(\theta)$  of  $X_1, \dots, X_n$  such that  $M_n(\theta) \rightarrow M(\theta) \in [-\infty, \infty)$  as  $n \rightarrow \infty$ , where the convergence can be either [almost surely](#) or just [in probability](#).

**Note.** Again, we need to at least assume that  $M(\theta) > -\infty$  for some  $\theta \in \Theta$ .

Let  $\hat{\theta}_n$  be an “approximate maximizer” of  $\theta \mapsto M_n(\theta)$ , i.e.,

$$M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - J_n$$

where  $J_n \xrightarrow{\text{a.s.}} 0$  (or  $\xrightarrow{P}$ ) as  $n \rightarrow \infty$ .

**Problem.** Where does  $\hat{\theta}_n$  converge, if anywhere?

We will need some assumptions to develop a theory of  $M$ -estimator to answer the above question.

(A1)  $M$  has a unique maximizer  $\theta^*$ , and for all  $\epsilon > 0$ ,  $M(\theta^*) > \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta)$ ,<sup>2</sup> i.e., it's well-separated.

(A2)  $\mathbb{P}(\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) - \sup_{\theta \in B(\theta^*, \epsilon)} M_n(\theta) \geq \delta) \rightarrow 0$ .

We can also consider a stronger version of (A2):

(A3)  $\mathbb{P}(\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \geq \sup_{\theta \in B(\theta^*, \epsilon)} M_n(\theta) + \delta \text{ i.o.}) = 0$  where i.o. means happens infinitely often (in terms of  $n$ ).

<sup>2</sup>Specifically,  $B(\theta^*, \epsilon) = \{\theta \in \Theta: d(\theta, \theta^*) < \epsilon\}$ , where  $d$  is the metric on  $\Theta$ .

Let's understand them one-by-one. Firstly, we see that if  $\tilde{\theta} \neq \theta^*$ , then  $\tilde{\theta} \notin B(\theta^*, \epsilon)$  for some  $\epsilon > 0$ . Hence, (A1) implies  $M(\theta^*) > \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \geq M(\tilde{\theta})$ . To satisfy (A1), consider the following.

**Definition 5.1.3 (Upper semi-continuous).** A function  $M: \Theta \rightarrow \mathbb{R}$  is *upper semi-continuous* if for  $\theta_n \rightarrow \theta$ , we have  $M(\theta) \geq \limsup_{n \rightarrow \infty} M(\theta_n)$ .

Upper semi-continuity is useful since we have the following result from analysis.

**Theorem 5.1.1.** 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 5.1.1 allows us to prove (A1).

**Corollary 5.1.1.** (A1) is satisfied if  $\Theta$  is compact,  $M$  is upper semi-continuous, and the maximizer  $\theta^*$  is unique.

**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 5.1.1, we see that  $\sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta)$  is obtained by a maximizer  $\theta^*$ , which is unique as assumed. ■

**Remark.** 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, we might start by considering  $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \rightarrow 0$  first, where the convergence can be either  $\xrightarrow{\text{a.s.}}$  or  $\xrightarrow{P}$ .

**Claim.**  $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$  is stronger than (A2).

**Proof.** By the triangle inequality,  $M_n(\theta) \leq |M_n(\theta) - M(\theta)| + M(\theta)$ , hence

$$\sup_{\theta \in \Theta} M_n(\theta) \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| + \sup_{\theta \in \Theta} M(\theta),$$

implying  $\sup_{\theta \in \Theta} M_n(\theta) - \sup_{\theta \in \Theta} M(\theta) \leq \sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)|$ . Hence,  $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$  implies

$$\mathbb{P} \left( \sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) - \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \geq \delta \right) \rightarrow 0$$

for any  $\delta > 0$ , which is exactly (A2). \*

**Intuition.** (A2) uniformly controls the convergence.

In this regard, (A3) is motivated by (A2) by modifying (A2) as

$$\begin{aligned} & \mathbb{P} \left( \sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \geq \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) + \delta \text{ i.o.} \right) = 0 \\ & \Leftrightarrow \mathbb{P} \left( \limsup_{n \rightarrow \infty} \sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \geq \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) + \delta \right) = 0 \\ & \Leftrightarrow \mathbb{P} \left( \limsup_{n \rightarrow \infty} \sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \leq \sup_{\theta \notin B(\theta^*, \epsilon)} M(\theta) \right) = 1. \end{aligned}$$

**Theorem 5.1.2.** 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  $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$ .

**Proof.** We will need to show that for all  $\epsilon > 0$ ,  $\mathbb{P}(d(\hat{\theta}_n, \theta^*) \geq \epsilon) \rightarrow 0$  or  $\mathbb{P}(d(\hat{\theta}_n, \theta^*) > \epsilon \text{ i.o.}) = 0$ . Firstly, since  $d(\hat{\theta}_n, \theta^*) \geq \epsilon$  implies  $\hat{\theta}_n \notin B(\theta^*, \epsilon)$ , hence

$$\sup_{\theta \notin B(\theta^*, \epsilon)} M_n(\theta) \geq M_n(\hat{\theta}_n) \geq \sup_{\theta \in \Theta} M_n(\theta) - J_n \geq M_n(\theta^*) - J_n.$$

■

# 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: [clcODwAAQBAJ](#).
- [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/CB09780511802256](#). URL: <https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D> (visited on 10/17/2023).