# STAT575
# Lrage Sample Theory

Pingbang Hu

February 28, 2024

**Abstract**

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

We list some references of this course, although we will not follow any particular book page by page: *Asymnptotic Statistics* [Vaa], *Asymptotic Theory of Statistics and Probability* [Das], *A course in Large Sample Theory* [Fer], *Approximation Theorems of Mathematical Statistics* [Ser], and *Elements of Large-Sample Theory* [Leh].

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

# Contents

# Chapter 1

# Introduction

## Lecture 1: Introduction to Large Sample Theory

Say we first collect $n$ data points $x_1, \ldots, x_n \in \mathbb{R}^d$, large sample theory concerns with the limiting theory as $n \to \infty$. We may treat $x_i$ as a realization of a random vector $X_i$ on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this course, we will primarily consider the case that $X_i$'s are i.i.d., i.e., independent and identically distributed from a distribution function, or the *cumulative density function* (cdf) $F$ such that 16 Jan. 9:30

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

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

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

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

Hence, the goal is to compute this by collecting data $x_i$'s, which is a statistical inference problem.

## 1.1 Parametrized Approach

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

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

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

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

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

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

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

## 1.2 Hypothesis Testing

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

**Example.** Consider a controlled group $Z$ and a treatment group $Y$, and we observe $Z_1, \ldots, Z_n$, and $Y_1, \ldots, Y_n$, 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 method is the so-called $t$-test. Let $s_n$ to be the sample standard derivation, then we can compute

$$T_n = \frac{\overline{X}_n}{s_n/\sqrt{n}} \sim t_{n-1}$$

as long as $X$ is Gaussian, i.e., the $t$-statistics for $H_0$. What if $X$ is not an Gaussian? We will show that even if $X$ is not Gaussian, this result is "approximately valid" when $n$ is "large enough" as long as $\mathrm{Var}[X] < \infty$.

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

What if now $\mathrm{Var}[X]$ doesn't exit? When the population has a heavy tail distribution, then second moment may not exit.

**Example** (Cauchy distribution)**.** The 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 $X_i$, i.e., the sign test.

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

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

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

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

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

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

# Chapter 2

# Modes of Convergence

## Lecture 2: Modes of Convergence

## 2.1 Different Modes of Convergence

18 Jan. 9:30

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

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

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

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

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

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

However, this might still be too strong.

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

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

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

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

Finally, we have the following.

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

### 2.1.1 Connection Between Modes of Convergence

We have the following connections between different modes of convergence.

$$\text{completely} \Longrightarrow \text{almost-surely} \Longrightarrow \text{in probability} \Longleftarrow \text{in } L^p$$

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

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

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

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

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

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

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

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

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

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

### 2.1.2 Consistent Estimator

Let $(X_n) \overset{\text{i.i.d.}}{\sim} F$ where $F$ is a distribution function. Say we're interested in some aspect of $F$, for example, some parameter $\theta = T(F) \in \mathbb{R}^m$. By collecting data $X_1, \ldots, X_n$, we estimate $\theta$ by computing an estimator $\hat{\theta}_n$ of $\theta$.[1] There are some properties we might want for $\hat{\theta}_n$.

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

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

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

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

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

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

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

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

[1] $\hat{\theta}_n$ is a function of $X_i$'s.

$X$ has a pdf $f$, then

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

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

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

[a] The latter is true even without $\mathrm{Var}[X] = \infty$ as we expect.

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

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

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

Another interesting problem is the supremum estimation.

**Example** (Supremum estimation)**.** Suppose there is a $\theta \in \mathbb{R}$ where 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 $F(\theta) = \mathbb{P}(X \leq \theta)$.[a] Then $\hat{\theta}_n = \max_{1 \leq i \leq n} X_i$ is indeed a strongly consistent estimator of $\theta$.

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

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

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

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

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

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

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

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

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

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

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

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

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

Let's see some examples.

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

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

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

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

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

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

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

# Lecture 3: Weak Convergence Portmanteau Theorem

## 2.2 Weak Convergence

25 Jan. 9:30

All convergences we have discussed are in some senses "point-wise" but not "distribution-wise", and the latter is more powerful. Consider working with a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and the following.

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

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

Returning to our situation, consider a sequence or random variables $(X_n)$ and a random variable $X$.

**Remark.** If $X_n$ has density $f_n$ and $X$ has density $f$, then $\mathrm{TV}(X_n, X) = \frac{1}{2} \int |f_n - f|$.

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

**Remark.** If $X_n$ and $X$ have densities $f_n$ and $f$, $f_n \to f$ implies $X_n \overset{\text{TV}}{\to} X$ from Scheffé's theorem.

**Note.** The above could make sense even if $X_n$ is defined on different $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ for every $n$.

Let's see some examples.

**Example.** Consider $X_n \sim \mathrm{Bin}(n, p_n)$ such that $np_n \to \lambda \in \mathbb{R}$. As this happens,

$$X_n \sim \mathrm{Bin}(n, p_n) \overset{\text{TV}}{\to} X \sim \mathrm{Pois}(\lambda).$$

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

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

**Example.** Let $X_n \sim \mathcal{U}\{0, 1/n, \ldots, (n-1)/n\}$, 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 $X_n \overset{\text{w}}{\to} X$, if for all bounded continuous $g\colon \mathbb{R}^d \to \mathbb{R}$, $\mathbb{E}\left[g(X_n)\right] \to \mathbb{E}\left[g(X)\right]$.

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

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

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

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

### 2.2.1 Portmanteau Theorem

The following is our main tool of proving weak convergence.

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

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

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

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

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

(e) $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ for all $A$ 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 $x, y \in S$,

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

**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. ⊛

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

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

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

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

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

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

**Proof.** Since $A$ is closed, $A = \overline{A}$ and

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

Now, we let

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

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,

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

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

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

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

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

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

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

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

from our assumption (b). Finally, as $k \to \infty$, 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.[a]

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

**Proof.** Since we can write

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

hence it suffices to show $|I_k| < \infty$ for any $k \geq 1$. Indeed, for any $k$, $|I_k| \leq k$. Suppose not. Then there exists a countable $J_k \subseteq I_k$ such that $|J_k| > k$, implying

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

which is a contradiction. ■

We now finish the proof of Portmanteau theorem.

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

**Claim.** (c) + (d) ⇒ (e).

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

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

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

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

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

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

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

Finally, we prove the following.

**Claim.** (e) ⇒ (a).

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

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

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

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

Observe that $\mathbb{P}(g(X_n) > t) = \mathbb{P}(X_n \in \{g > t\})$ 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,

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

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

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

This finishes the proof. ■

### 2.2.2 Continuous Mapping Theorem

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

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

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

It's clear that the following holds.

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

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

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

Then we have the following.

**Proposition 2.2.1.** Given $g\colon S \to S'$ between metric spaces and $A \subseteq S'$,

$$C_g \cap \overline{\{g \in A\}} \subseteq \{g \in \overline{A}\}.$$

**Proof.** Let $x \in C_g \cap \overline{\{g \in A\}}$. Since $x \in \overline{\{g \in A\}}$, there exists $(x_n) \in \{g \in A\}$ such that $x_n \overset{\rho}{\to} x$. Moreover, $x \in C_g$ implies $g$ is continuous at $x$, hence $g(x_n) \overset{\rho'}{\to} g(x)$, 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 $X_n \overset{w}{\to} X$ and $h\colon \mathbb{R}^d \to \mathbb{R}^m$. If $\mathbb{P}(X \in C_h) = 1$, then $h(X_n) \overset{w}{\to} h(X)$.

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

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

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

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

where the last inequality follows again from Portmanteau theorem (d) since $\overline{\{h \in A\}}$ is clearly closed and $X_n \overset{w}{\to} X$. 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.,

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

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

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

**Proof.** For the case of $X^2 \xrightarrow{w} X^2$, continuous mapping theorem clearly applies with $h(x) = x^2$. For the first case, consider

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

This means $C_h = \mathbb{R} \setminus \{0\}$. Then, we just need to show $\mathbb{P}(X \in C_h) = 1$ and apply continuous mapping theorem. Observe that this is the same as asking $\mathbb{P}(X = 0) = 0$, which is true when $X$ is continuous.[a] ⊛

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

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 $\epsilon > 0$, $\mathbb{P}(\|X_n - Y_n\| > \epsilon) \to 0$ as $n \to \infty$. Then, $Y_n \xrightarrow{w} X$.

We first see some applications.

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

**Proof.** By considering $X_n = X$ for all $n$, Theorem 2.2.3 implies that if $Y_n \xrightarrow{p} X$, $Y_n \xrightarrow{w} X$. Conversely, if $Y_n \xrightarrow{w} c$, from Portmanteau theorem (c), for any fixed $\epsilon > 0$,[a]

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

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

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

# Lecture 5: Convergence in Distribution and Weak Convergence

Now we prove Theorem 2.2.3. 1 Feb. 9:30

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

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

Since $X_n \xrightarrow{w} X$, the second term goes to 0. As for the first term, since $Y_n$ and $X_n$ are in the same probability space, we see that

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

As $n \to \infty$, we finally have

$$\limsup_{n \to \infty} |\mathbb{E}[g(Y_n)] - \mathbb{E}[g(X)]| \leq K\epsilon$$

for all $\epsilon > 0$, by letting $\epsilon \to 0$, we're done. ■

We can now apply Theorem 2.2.3 to prove something similar as we have seen before in the case of convergence in probability.

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

Now, in the case of weak convergence, from continuous mapping theorem, we see that 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.** Consider a random variable $X$ on $(\Omega, \mathscr{F}, \mathbb{P})$, and 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 $\mathbb{R}^m$,[a] then $(X_n, Y_n) \xrightarrow{w} (X, c)$.

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

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

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 as we wish. Combining both with Theorem 2.2.3 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. ■

### 2.2.3 Convergence in Distribution

So far, the notions of convergence we have talked about applies to general probability space, which needs not to be in $\mathbb{R}^d$ in general. However, traditionally, the case in $\mathbb{R}^d$ is considered first.

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

This allows us to define the following.

**Definition 2.2.4** (Converge in distribution)**.** Let $(X_n)$ and $X$ be random variables in $\mathbb{R}^d$. Then $(X_n)$ *converges in distribution* to $X$, denoted as $X_n \xrightarrow{D} X$, if for all $(t_1, \dots, t_d) \in C_{F_X}$,

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

**Note.** $X_n$ and $X$ (in $\mathbb{R}^d$) do not have to be on the same probability space.

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

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

same for $F_X$. So this reduces to the form we're familiar with, i.e., $\mathbb{P}(X_n \in A)$ for some $A$.

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

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

There are more classical results that are worth mentioning.

**Remark** (De Moivre central limit theorem). Let $X_n \sim \text{Bin}(n, p)$, then for every $t \in \mathbb{R}$, as $n \to \infty$,

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

**Proposition 2.2.2.** Let $X_n$ and $X$ be in $\mathbb{Z}$ such that $f_n$ and $f$ are their corresponding pmf's, then

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

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

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

for some $\epsilon > 0$ small enough. Now, as $t \pm \epsilon$ are in $C_X$ clearly, $X_n \overset{D}{\to} X$ implies

$$\mathbb{P}(X_n \leq t + \epsilon) \to \mathbb{P}(X \leq t + \epsilon),$$

and the same holds for $t - \epsilon$, hence

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

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

Now, the problem one might have is the following.

**Problem.** Why not defined for all $t \in \mathbb{R}^d$, rather than $t \in C_{F_X}$?

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

If $X_n \overset{D}{\to} X$ and $X$ is continuous, then $F_{X_n}$ converges to $F_X$ not only point-wise, but uniformly.

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

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

Now we have seen various remarks and clarifications about convergence in distribution, the upshot is that, it is actually just a renaming of weak convergence in $\mathbb{R}^d$!

**Theorem 2.2.5.** Given $X_n$ and $X$ in $\mathbb{R}^d$, then $X_n \overset{w}{\to} X$ if and only if $X_n \overset{D}{\to} X$.

**Proof.** We prove for the case of $d = 1$, then it's easy to see the same holds for $d \geq 1$. For the forward direction, we want to show that for all $t \in C_{F_X}$, $\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$. Note that $\mathbb{P}(X \leq t) = \mathbb{P}(X \in (-\infty, t])$ and $\mathbb{P}(X_n \leq t) = \mathbb{P}(X_n \in (-\infty, t])$, hence, from Portmanteau theorem (e) with $A = (-\infty, t]$, $X_n \overset{w}{\to} X$ is equivalently as saying $\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$ if

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

is 0. This is indeed the case since $t \in C_{F_X}$, hence we're done.

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

**Lemma 2.2.3.** $X_n \xrightarrow{D} X$ if and only if for all $x \in \mathbb{R}^d$,

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

**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 $k \geq 1$,

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

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

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

*The proof will be continued...*

# Lecture 6: Stochastic Boundedness and Delta Theorem

Before we finish the proof of Theorem 2.2.5, we recall one important characterization of lim inf. 2 Feb. 17:30

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

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

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

We can then finish the proof of Theorem 2.2.5.

**Proof of Theorem 2.2.5 (cont.)** Now we can prove the backward direction. Form Portmanteau theorem (c), it suffices to show that for every open $A \subseteq \mathbb{R}$, we have

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

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

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

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

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

**Proof.** Observe that $\mathbb{P}(X \in (a, b)) = F_X(b^-) - F_X(a)$, with Lemma 2.2.3, we further have

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

which proves the claim. ⊛

This proves the case of $d = 1$. ■

Theorem 2.2.5 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, which uses weak convergence's result but now prove something about the distribution.

**Proposition 2.2.3.** If $X_n \xrightarrow{D} X$ and $t_n \to t \in C_{F_X}$, then $\mathbb{P}(X_n \leq t_n) \to \mathbb{P}(X \leq t)$.

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

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

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

### 2.2.4 Stochastic Boundedness

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

**Definition 2.2.5** (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 $i \in I$,

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

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

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

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

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

**Proposition 2.2.4.** If $(X_i)_{i\in I}$ is bounded in $L^p$ for some $p > 0$, i.e., $\sup_{i\in I} \mathbb{E}[\|X_i\|^p] < \infty$, then $(X_i)_{i\in I}$ is bounded in probability.

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

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

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

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

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

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

**Proposition 2.2.5.** If $X_n \xrightarrow{D} X$, then $(X_n)$ is bounded in probability.

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

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

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

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

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

The generalization is the following.

**Proposition 2.2.6.** 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 $\mathbb{P}(|X_n Y_n| > \epsilon) \to 0$. This is because

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

for any $M$. Now, we see that

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

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

We often write the following.

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

**Remark.** Proposition 2.2.6 means $o_p(1) \times O_p(1) = o_p(1)$.

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

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

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

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

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

**Theorem 2.2.6** (Delta method)**.** Let $\theta \in \mathbb{R}^d$, $(T_n)$ and $Y$ be random vectors in $\mathbb{R}^d$, and $b_n \to \infty$ be a positive deterministic sequence. If $b_n(T_n - \theta) \overset{D}{\to} Y$, then $T_n \overset{p}{\to} \theta$. Moreover, if $g \colon \mathbb{R}^d \to \mathbb{R}^m$ is differentiable at $\theta$, $b_n(g(T_n) - g(\theta)) \overset{D}{\to} \nabla g(\theta) Y$.

**Proof.** We first observe that $\|b_n(T_n - \theta)\| \in O_p(1)$ since $b_n(T_n - \theta) \overset{D}{\to} Y$, with continuous mapping theorem and the fact that $\|\cdot\|$ is continuous, $\|b_n(T_n - \theta)\| \overset{p}{\to} \|Y\|$, so $\|b_n(T_n - \theta)\| \in O_p(1)$ by Proposition 2.2.5. With this, as $b_n \to \infty$, $T_n \overset{p}{\to} \theta$ since

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

as $o(1)O_p(1) = o_p(1)$ from Proposition 2.2.6. For the second claim, since $g$ is differentiable at $\theta$,

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

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

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

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

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

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

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

We observe the following.

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

**Proof.** Since it implies that $b_n(g(T_n) - g(\theta))$ has the same weak limit as $\nabla g(\theta) b_n(T_n - \theta)$ from Theorem 2.2.3, i.e., $\nabla g(\theta) Y$ from our assumption with continuous mapping theorem. ⊛

It's enough to show $\|h(T_n)\| = o_p(1)$ since we know that $\|b_n(T_n - \theta)\| = O_p(1)$ and $o_p(1)O_p(1) = o_p(1)$ from Proposition 2.2.6. Indeed, as $T_n \overset{p}{\to} \theta$, $h(T_n) \overset{p}{\to} h(\theta) = 0$ again by continuous mapping theorem with $h$ being continuous at $\theta$. This further implies $\|h(T_n)\| \overset{p}{\to} 0$ as we desired.[a] Combining the above, the result follows. ∎

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

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

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

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

# Lecture 7: Skorohod's Representation Theorem

## 2.2.5 Skorohod's Representation Theorem

6 Feb. 9:30

So far, we have seen the following.

$$\begin{tikzcd}
\text{completely} \arrow[r, Rightarrow] & \text{almost-surely} \arrow[r, Rightarrow] & \text{in probability} \arrow[r, Rightarrow] & \text{weakly} \arrow[d, Leftrightarrow] \\
& & \text{in } L^p \arrow[u, Rightarrow] & \text{in distribution}
\end{tikzcd}$$

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

**Theorem 2.2.7** (Skorohod's representation theorem)**.** If $X_n \overset{D}{\to} X$, there exists $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$ on which we can define random vectors $(Y_n)$ and $Y$ such that $Y_n \overset{D}{=} X_n$ for all $n$ and $Y \overset{D}{=} X$, and $\widetilde{\mathbb{P}}(Y_n \to Y) = 1$.

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

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

**Definition 2.2.6** (Quantile)**.** A $p^{th}$ *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 2.2.7** (Quantile function)**.** The *quantile function* of $X \sim F$ is defined as

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

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 $(\widetilde{\Omega}, \widetilde{\mathscr{F}}, \widetilde{\mathbb{P}})$. Then, $F^{-1}(U) =: Y$ is a random variable with cdf $F$.

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

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

⊛

Now we can prove Skorohod's representation theorem.

**Proof of Theorem 2.2.7.** Consider $\widetilde{\Omega} = (0,1)$, and $\widetilde{\mathbb{P}}((a,b)) = b - a$ for all $a < b$. Then, we can define $U(p) = p$ for all $p \in \widetilde{\Omega}$, i.e., $U \sim \mathcal{U}(0,1)$. Define $Y_n = F_{X_n}^{-1}(U)$ and $Y = F_X^{-1}(U)$ from the quantile functions. Denote $\Phi$ be the cdf of $\mathcal{N}(0,1)$, and let $Z = \Phi^{-1}(U)$.

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

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

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

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

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

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

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

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

Now we show $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$ for almost all $p$'s. Since $X_n \overset{D}{\to} X$ means $F_{X_n}(t) \to F_X(t)$, from Lemma 2.2.3, it further implies $F_{X_n}(t^-) \to F_X(t^-)$ for all $t \in C_{F_X}$. Note that $\widetilde{\mathbb{P}}(Z \in C_{F_X}) = 1$ since there can be only countably many discontinuities of $F_X$. Hence,

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

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

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

and also $\widetilde{\mathbb{P}}(F_{X_n}(Z) < p) \to \widetilde{\mathbb{P}}(F_X(Z) < p)$ from Lemma 2.2.3. 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 2.2.5.

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

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

Another application is to generalize Fatou's lemma.

**Proposition 2.2.7** (Fatou's lemma)**.** Let $X_n \overset{D}{\to} X$[a] and $g\colon \mathbb{R}^d \to [0,\infty)$ continuous. Then

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

[a] Can be on different probability spaces.

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

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

■

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

**Theorem 2.2.8.** If $X_n \xrightarrow{\text{a.s.}} X$, $g\colon \mathbb{R}^d \to \mathbb{R}$ is continuous and $(g(X_n))$ is uniformly integrable[a] if and only if $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$.

in $L_1$?

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

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

**Proposition 2.2.8.** If $X_n \xrightarrow{w} X$ and $\mathbb{P}(X \in C_g) = 1$ where $g\colon \mathbb{R}^d \to \mathbb{R}$ such that $(g(X_n))$ is uniformly integrable, then $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$.

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

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

**Remark.** Proposition 2.2.8 can be proved with Skorohod's representation theorem also.

the in $L_1$ version?

### 2.2.6 Proving Distributional Convergence

We often want to prove $X_n \xrightarrow{D} X$, which is not efficient 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 2.2.9.** $(x_n) \to x$ if and only if every subsequence of $(x_n)$ has a subsequence that converges to the same limit $x$.

**Proof.** The forward direction is clear. For the backward direction, if not, there exists $(x_{n_k})$ and $\epsilon > 0$ such that $\rho(x_{n_k}, x) \geq \epsilon$ 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 2.2.10.** $X_n \xrightarrow{w} X$ 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 2.2.9. ■

## Lecture 8: Characteristic Functions

We see other similar theorems apart from Theorem 2.2.10. 8 Feb. 9:30

**Theorem 2.2.11.** If $X_n \xrightarrow{w} X$ and $X_n \xrightarrow{w} Y$, then $X \overset{D}{=} Y$. More generally, if $X_n \xrightarrow{w} X$ and $Y_n \xrightarrow{w} Y$, with $X_n \overset{D}{=} Y_n$ for all $n \geq 1$, $X \overset{D}{=} Y$.

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

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

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

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

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

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

## 2.3 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 2.3.1** (Characteristic function). The *characteristic function* of a $d$-dimensional random vector $X$ is defined as $\phi_X \colon \mathbb{R}^d \to \mathbb{C}$ where $t \in \mathbb{R}^d$ such that

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

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

If we write $\phi_X$ explicitly, we have

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

**Remark.** Characteristic functions are bounded.

**Proof.** Since

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

⊛

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

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

We make one more remark for future reference.

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

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

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

### 2.3.1 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 2.3.1** (Uniqueness). If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}^d$, then $X \overset{D}{=} Y$. The converse is trivial.

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

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

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

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

With Corollary 2.2.2, since $Z/k \overset{p}{\to} 0$ (and also $Z'/k \overset{p}{\to} 0$), we have $X + Z/k \overset{D}{\to} X$ and $Y + Z'/k \overset{D}{\to} Y$, which implies $X \overset{D}{=} Y$ from Theorem 2.2.11. ⊛

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

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

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

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

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

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

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

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

### 2.3.2 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 2.3.2** (Prokhorov's theorem)**.** If $(X_n) = O_p(1)$, then there exists a weakly convergent subsequence of $(X_n)$.

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

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

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

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

We now state the theorem.

**Theorem 2.3.3 (Lévy-Cramer continuity theorem).** If $\phi_{X_n}(t) \to \phi(t)$ for all $t \in \mathbb{R}^d$, then all weakly convergent subsequences of $(X_n)$ have the same weak limit. Furthermore, if also $\phi$ is continuous at 0, then there exists $X$ such that $\phi = \phi_X$ and $X_n \overset{D}{\to} X$.

**Proof.** Let's start with the first claim. Suppose $Y_n \overset{w}{\to} Y$ and $Z_n \overset{w}{\to} Z$ are two subsequences of $X_n$ such that $Y \neq Z$. But since $\phi_{Y_n}(t) \to \phi_Y(t)$ and $\phi_{Z_n}(t) \to \phi_Z(t)$, with the fact that $(\phi_{Y_n}(t))$ and $(\phi_{Z_n}(t))$ are subsequences of $(\phi_{X_n}(t))$ for every $t$, as $\phi_{X_n}(t) \to \phi(t)$, both subsequences need to converge to the same limit, i.e.,

$$\phi_Y(t) = \phi(t) = \phi_Z(t)$$

for all $t \in \mathbb{R}^d$. From the uniqueness theorem, $Y \overset{D}{=} Z$.

For the second part, 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 we can show $(X_n) = O_p(1)$ from our assumption, 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 2.3.3 (cont.)** Fix $\epsilon > 0$. Then there exists $\delta > 0$ such that for all $|t| < \delta$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### 2.3.3 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 2.3.4** (Feller's inversion formula)**.** Let $X$ be a $d$-dimensional random vector with the characteristic function $\phi_X$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which will imply the result since a cdf is characterized by its values in $C_{F_X}$, i.e., if the above equality is true, it will be valid for all $a, b \in \mathbb{R}$. To do so, dominated convergence theorem states that

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

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

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

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

■

**Corollary 2.3.1.** Given $X_n$ and $X$ such that $\phi_X$ and $\phi_{X_n}$ are integrable. If $\phi_{X_n} \xrightarrow{L^1} \phi_X$,[a] then $X_n \xrightarrow{\mathrm{TV}} X$.

[a] I.e., $\int_{\mathbb{R}} |\phi_{X_n}(t) - \phi_X(t)| \,\mathrm{d}t \to 0$.

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

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

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

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

**Proposition 2.3.1.** For any random vector $X$, $\phi_X$ is uniformly continuous, i.e.,

$$\lim_{h \to 0} \sup_t |\phi_X(t+h) - \phi_X(t)| = 0.$$

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

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

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

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

**Theorem 2.3.5.** 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, $\phi_X^{(p)}(0) = i^p \mathbb{E}[X^p]$.

**Proof.** We prove the case $p = 1$ since for $p > 1$, it can be shown similarly with induction. It's enough to prove

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

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

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

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

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

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

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

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

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

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

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

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

**Remark.** Theorem 2.3.5 implies $\sup_t|\phi_X^{(p)}(t)|\leq \mathbb{E}[|X|^p]$, which is finite by the assumption $X\in L^p$.

A more useful consequence is the following, generalizing Proposition 2.3.1.

**Corollary 2.3.2.** If $X\in L^p$ for some $p\in\mathbb{N}$, then $\phi_X^{(p)}$ is uniformly continuous.

**Proof.** To show that $\phi_X^{(p)}$ is uniformly continuous, we show that $\sup_{t\in\mathbb{R}}|\phi^{(p)}(t+h)-\phi_X^{(p)}(t)|\to 0$ as $h\to 0$. But this is clear since for any $h\in\mathbb{R}$, with Theorem 2.3.5,

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

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

# Lecture 10: Law of Large Number and Central Limit Theorem

15 Feb. 9:30

## 2.4 Fundamental Theorems of Probability

With the tools we developed, we now prove two of the fundamental theorems of probability, i.e., the weak law of large number, and also the central limit theorem. We start from the first one.

**Theorem 2.4.1** (Khintchin's weak law of large number)**.** Let $(X_n)$ be i.i.d. random vectors, and $X$ be a random vector with $\mathbb{E}[|X|] < \infty$. Then $\overline{X}_n \overset{p}{\to} \mathbb{E}[X]$.

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

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

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

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

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

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

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

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

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

**Remark.** We don't need to assume finite first moment since assuming $\phi$ is differentiable at 0 such that $\phi'(0) = ic$ is enough.

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 2.3.5.

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

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

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

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

We now state the theorem.

**Theorem 2.4.2** (Lindeberg-Lévy central limit theorem)**.** Let $(X_n)$ be i.i.d. random variables (i.e., $d = 1$) with $\mathbb{E}[X_i] =: \mu$, $\mathrm{Var}[X_i] =: \sigma^2 < \infty$ for all $1 \leq i \leq n$. Then

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

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

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

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

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

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

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

Hence, from the bounded convergence theorem,

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

which shows the result. ■

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

## 2.5 Application of Inferences

### 2.5.1 Inference for Population Mean

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

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

**Intuition.** We make the distribution independent of parameters to get a confidence interval.

Since the left-hand side is not an estimator now since it depends on $\sigma$. If we replace it by the sample standard deviation $s_n$, as $n \to \infty$,

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

where $T_n$ follows $t$-distribution with $n-1$ degrees of freedom.

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

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

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

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

**Proposition 2.5.1.** If $X \in L^2$, then $\hat{\sigma}_n^2$ and $s_n^2$ are both consistent estimators of $\sigma^2$. Furthermore, $T_n \xrightarrow{D} \mathcal{N}(0,1)$, and the same holds if $s_n$ is replaced by $\hat{\sigma}_n$ in the definition of $T_n$.

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

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

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

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

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

[a]We need to check whether the first moment exists.

Proposition 2.5.1 says that for mean estimation, even if the data is not Gaussian, we're fine.

**Corollary 2.5.1.** If $X \in L^2$, then $\overline{X}_n \pm Z_{\alpha/2} s_n/\sqrt{n}$ and $\overline{X}_n \pm Z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$ are both asymptotically valid $100(1-\alpha)\%$ confidence intervals for $\mu$.

### 2.5.2 Inference for Population Variance

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

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

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

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

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

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

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

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

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

**Proposition 2.5.2.** 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 $Y_i \coloneqq X_i - \mu$ (and also $Y = X - \mu$),

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

As $n \to \infty$, $\hat{\sigma}_n^2/\sqrt{n} \xrightarrow{p} 0$, with the fact that $\sqrt{n}\overline{Y}_n$ converges in distribution from the central limit theorem, it's also bounded in probability from Proposition 2.2.5, hence

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

This proves the first claim. Now, from the central limit theorem, assuming $\operatorname{Var}[Y_i^2] < \infty$, we have

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

which implies $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \xrightarrow{D} \mathcal{N}(0, \operatorname{Var}[Y^2])$ from what we have shown, where

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

which proves the result with our assumption. 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 holds for replacing $\hat{\sigma}_n^2$ by $s_n^2$ as well. ∎

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

**Definition 2.5.1** (Kurtosis). The *Kurtosis* of a random variable $X$ is defined as $\mathbb{E}[((X-\mu)/\sigma)^4]$.

**Definition 2.5.2** (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 2.5.2 requires $\mathbb{E}[Z^4] > 1$. However, from Jensen's inequality, $\mathbb{E}[Z^4] \geq \left(\mathbb{E}[Z^2]\right)^2 \geq 1$, hence indeed, the assumption might not be true in general.

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

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

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

The takeaway is when $X$ is not a normal, and if the Kurtosis of $X$ is different from the normal, 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 $\widetilde{\mu}_k = \mu_k/\sigma^k = \mathbb{E}\left[(X-\mu)^k/\sigma^k\right]$. 20 Feb. 9:30

**As previously seen.** In this notation, Proposition 2.5.2 gives $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \to \mathcal{N}(0, (\widetilde{\mu}_4 - 1)\sigma^4)$, i.e.,

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

The task is then the following.

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

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

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

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

The above essentially comes from the following fact.

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

We now show that this still yields a consistent estimator.

**Proposition 2.5.3.** If $X \in L^k$ for $k > 2$, then $M_k \xrightarrow{p} \mathbb{E}[Y^k] = \mu_k$. Same for $\widetilde{M}_k$ and $\widetilde{\mu}_k$.

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

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

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

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

By the weak law of large number, $\overline{Y^k} \xrightarrow{p} \mathbb{E}[Y^k] = \mu_k$ and $(-\overline{Y})^{k-\ell} \xrightarrow{p} 0$ since $-\overline{Y} \xrightarrow{p} 0$ with 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 2.5.3 implies the following.

**Corollary 2.5.2.** 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

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

### 2.5.3 Asymptotic Distribution of Sample Central Moments

Furthermore, as we will see, 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 2.5.1.** 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 $v_k := \operatorname{Var}[Y^k - k\mu_{k-1} Y] > 0$, then $\sqrt{n}(M_k - \mu_k) \xrightarrow{D} \mathcal{N}(0, v_k)$.

**Proof.** Firstly,

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

We see that

- $\overline{Y}\sqrt{n}$ is asymptotically normal, hence is bounded in probability from Proposition 2.2.5;
- $\overline{Y^\ell} \xrightarrow{p} \mathbb{E}[Y^\ell] = O(1)$;
- $1/\sqrt{n}^{k-\ell-1} = o(1)$ for $\ell < k-1$.

Combining, every term in the summation is $O(1)O_p(1)o(1) = o_p(1)$ 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,

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

Observe that $Y_i^k - \mu_k - k\mu_{k-1} Y_i$'s are i.i.d., the whole thing converges to $\mathcal{N}(0, \operatorname{Var}\left[Y^k - \mu_k - k\mu_k Y\right])$ in distribution by central limit theorem, where the variance can be further simplified as

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

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

**Note.** Theorem 2.5.1 doesn't give the same result for $\widetilde{M}_k = M_k / \hat{\sigma}_n^k$, which requires we to obtain 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 2.5.1 does give an asymptotic distribution for $\widetilde{M}_k$.

### 2.5.4 Testing Normality with Odd Moments

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

**Intuition.** With Proposition 2.5.3, the idea is that to reject $H_0$ if $|\widetilde{M}_k| = |M_k/\hat{\sigma}_n^k|$ is "large".

In this regard, Theorem 2.5.1 for $M_k$ is not enough, we really need $\widetilde{M}_k$.

**As previously seen.** Theorem 2.5.1 gives $\sqrt{n}(M_k - \mu_k) \overset{D}{\to} \mathcal{N}(0, \operatorname{Var}\left[Y^k - k\mu_{k-1}Y\right]) = \mathcal{N}(0, v_k)$.

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

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

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

Then, by Slutsky's theorem, $\sqrt{n}M_k/\hat{\sigma}_n^k$ also converges to this normal. Since all we use is the fact that $\mu_k = 0$ for odd $k$ and Theorem 2.5.1, let's write this general result as a corollary.

**Corollary 2.5.3.** If $X \in L^{2k}$ for some odd $k > 2$ such that $\mu_k = 0$ and $v_k/\sigma^k > 0$, then $\sqrt{n}M_k/\hat{\sigma}_n^k \overset{D}{\to} \mathcal{N}(0, v_k/\sigma^{2k})$.

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

Before we see one concrete example, note the following property when $\mu_k = 0$ for odd $k > 2$.

**Note.** For odd $k$, $\operatorname{Var}[Y^k] = \mathbb{E}[Y^{2k}] - (\mathbb{E}[Y^k])^2 = \mathbb{E}[Y^{2k}] = \mu_{2k}$ since $(\mathbb{E}[Y^k])^2 = \mu_k^2 = 0$.

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

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

**Proof.** From Corollary 2.5.3 and the symmetry of normal distribution,

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

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

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

which provides the result. ⊛

For even $k$ or odd $k$ but $\mu_k \neq 0$, we really need to work out the joint. Since we know the asymptotic distribution of both $M_k$ and $\hat{\sigma}_n^2$, the joint can be obtained by the delta method with $g(M_k, \hat{\sigma}_n^2) = |M_k/\hat{\sigma}^k|$ and the "multivariate" version of Theorem 2.4.2 since we now have two quantities.

### 2.5.5 Multivariate Central Limit Theorem

Our next goal is to prove the multivariate central limit theorem, i.e., the high dimensional generalization of Theorem 2.4.2. We first need the following tool.

**Theorem 2.5.2** (Cramér-Wold device). Let $(X_n)$ be a sequence of random vectors and $X$ be a random vector in $\mathbb{R}^d$. Then $X_n \overset{D}{\to} X$ if and only if $t \cdot X_n \overset{D}{\to} t \cdot X$ 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 $t \cdot X_n \overset{D}{\to} t \cdot X$. Then

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

which implies $X_n \overset{D}{\to} X$ by the Lévy-Cramer continuity theorem. ∎

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

**Theorem 2.5.3** (Multivariate central limit theorem). Let $(X_n)$ be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = \mu \in \mathbb{R}^d$, $\operatorname{Var}[X_i] = \Sigma \in \mathbb{R}^{d\times d}$ for all $1 \leq i \leq n$. Then

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

## Lecture 12: Asymptotic Joint Distribution by Multivariate CLT

22 Feb. 9:30

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

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

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

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

and since $\operatorname{Var}[t \cdot X] = t^\top \operatorname{Var}[X_i] t = t^\top \Sigma t = \operatorname{Var}[t \cdot Z]$, hence we're done. ∎

### 2.5.6 Testing Normality with General Moments

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

**As previously seen.** From Proposition 2.5.1 and Proposition 2.5.2,

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

[a]The former asymptotic distribution result needs the assumption of $X \in L^4$ and $\widetilde{\mu}_4 > 1$.

We see that we have results for $\overline{X}_n$ and $\hat{\sigma}_n^2$, but not $\hat{\sigma}_n$. This is fine since we can

1. compute the joint distribution of $\overline{X}_n$ and $\hat{\sigma}_n^2$;
2. apply the delta method with $g(\overline{X}_n, \hat{\sigma}_n^2) \coloneqq \overline{X}_n / \hat{\sigma}_n$ to get the distribution of $\overline{X}_n / \hat{\sigma}_n$.

Specifically, we have the following.

**Proposition 2.5.4.** If $X \in L^2$,

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

Moreover, if $X \in L^4$ and $\widetilde{\mu}_4 = \mu_4 - \sigma^4 > 1$, then

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n \\ \hat{\sigma}_n^2\end{pmatrix} - \begin{pmatrix}\mu \\ \sigma^2\end{pmatrix}\right) \overset{D}{\to} \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \operatorname{Var}\left[\begin{pmatrix}Y \\ Y^2\end{pmatrix}\right]\right)$$

by the multivariate central limit theorem with

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

**Remark** (Asymptotically independent)**.** Related back to Corollary 2.5.3, when their skewness is 0, $\overline{X}_n$ and $\hat{\sigma}_n^2$ (or $s_n^2$) are asymptotically independent.

Let's leave the application of the delta method to the general $k$. We note the following.

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

Next, we compute the asymptotic law of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$ for general $k > 2$. Following a similar calculation, for $\hat{\sigma}_n^k$, we can again use the result for $\hat{\sigma}_n^2$, i.e., Proposition 2.5.2.

**As previously seen.** From Theorem 2.5.1, if $X \in L^k$,

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

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

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

$$\begin{aligned} Y \coloneqq \sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_n^2 \\ M_k\end{pmatrix} - \begin{pmatrix}\sigma^2 \\ \mu_k\end{pmatrix}\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}Y_i^2 - \sigma^2 \\ Y_i^k - \mu_k - k\mu_{k-1}Y_i\end{pmatrix} + o_p(1) \\ &\overset{D}{\to} \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\operatorname{Var}[Y^2] & \operatorname{Cov}[Y^2, Y^k - k\mu_{k-1}Y] \\ \operatorname{Cov}[Y^2, Y^k - k\mu_{k-1}Y] & \operatorname{Var}[Y^k - k\mu_{k-1}Y]\end{pmatrix}\right) \end{aligned}$$

where the multivariate central limit theorem can be applied when $X \in L^{2k}$ and $\operatorname{Var}[Y^k - k\mu_{k-1}Y] > 0$. This "$Y$" will be used in the delta method later.[2]

**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 2.5.3.

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

**Theorem 2.5.4.** Let $X \in L^k$ for some $k > 2$. Then for $Z = (X - \mu)/\sigma = Y/\sigma$,

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

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

[2] This is not exact since $Y$ should be the random vector corresponding the asymptotic distribution on the right-hand side. But this is fine in the end as we will soon see.

**Proof.** Since Proposition 2.5.2 is for $\hat{\sigma}_n^2$ but not $\hat{\sigma}_n^k$, we need to use delta method by considering

$$\widetilde{M}_k = \frac{M_k}{\hat{\sigma}_n^k} =: g(\hat{\sigma}_n^2, M_k)$$

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 $\widetilde{\mu}_k = \mu_k/\sigma^k$, $\partial g/\partial x = -kyx^{-k/2-1}/2$, and $\partial g/\partial y = x^{-k/2}$. From delta method and the above calculation, with $\widetilde{\mu}_k = g(\sigma^2, \mu_k)$, we get $\sqrt{n}(g(\hat{\sigma}_n^2, M_k) - g(\sigma^2, \mu_k)) \overset{D}{\to} \nabla g Y$, 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 $Z_i := (X_i - \mu)/\sigma = Y_i/\sigma$. Then by the multivariate central limit theorem and Slutsky's theorem, the above further converges to $\mathcal{N}(0, \widetilde{v}_k)$ when

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

as we assumed. ■

Compared to the last time when we do this for odd $k$ and $\mu_k = 0$ (Corollary 2.5.3), we only get an asymptotic distribution, not an explicit decomposition.

**Note.** It's more convenient to use delta method in this way, i.e., not use the actual $Y$ corresponding to the distribution, but use the term in the limiting sequence with $o_p(1)$.

With this explicit formula, we now see one example.

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

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

**Proof.** One can write down $\sqrt{n}(\widetilde{M}_\ell - \widetilde{\mu}_\ell)$ for even $\ell$, and also $\sqrt{n}(\widetilde{M}_k - \widetilde{\mu}_k) = \sqrt{n}\widetilde{M}_k$ for odd $k$, and see that while they both converge to $\mathcal{N}(0, 1)$, their covariance is 0, i.e., asymptotically independent, so the square of them add up to $\chi_2^2$. ⊛

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

$$\frac{n}{\widetilde{v}_k} \widetilde{M}_k^2 + \frac{n}{\widetilde{v}_\ell} (\widetilde{M}_\ell - \widetilde{\mu}_\ell)^2 \overset{D}{\to} \chi_2^2$$

from exactly the same argument.

### 2.5.7 Asymptotic Relative Efficiency

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

- $\sqrt{n}(\overline{X}_n - \theta) \overset{D}{\to} \mathcal{N}(0, \sigma^2)$;

- $\sqrt{n}(\hat{\sigma}_n^2 - \theta) \overset{D}{\to} \mathcal{N}(0, \mu_4 - \sigma^4)$.

As $\sigma^2 = \theta$ and $\mu_4 = 3\theta^2 + \theta$, we see that $\overline{X}_n$ is better since its variance is smaller. To further quantify how much better is it, we can ask how many data we need such that we get a similar precision. Consider

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

for two statistics $T_i$, $i = 1, 2$. This implies

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

Let $I_n^{(i)} := T_n^i \pm Z_{\alpha/2}\sigma_i(\theta)/\sqrt{n}$, and let $n_i$ be the value of $n$ such that $|I_n^{(i)}| = \gamma$,

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

i.e., $n_1/n_2 = \sigma_1(\theta)^2/\sigma_2(\theta)^2$. We called this the asymptotic relative efficiency $\mathrm{ARE}(T^1, T^2)$.

**Definition 2.5.3 (Asymptotic relative efficiency).** The *asymptotic relative efficiency* between two statistics $T_n^1$ and $T_n^2$ for $\theta$ such that $\sqrt{n}(T_n^i - \theta) \overset{D}{\to} \mathcal{N}(0, \sigma_i^2(\theta))$ is defined as

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

### 2.5.8 Variance Stabilizing Transformation

Now, say we use $\overline{X}_n$ as the estimator of $\theta$. We have $\sqrt{n}(\overline{X}_n - \theta) \overset{D}{\to} \sqrt{\theta}\mathcal{N}(0, 1) = \mathcal{N}(0, \theta)$.

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

To get around this, we first write

$$\sqrt{n}(\overline{X}_n - \theta) \overset{D}{\to} \sqrt{\theta}Z \sim \mathcal{N}(0, \theta)$$

for $Z \sim \mathcal{N}(0, 1)$. Then, instead of writing this as

$$\frac{\sqrt{n}}{\sqrt{\theta}}(\overline{X}_n - \theta) \overset{D}{\to} Z$$

and replace $\sqrt{\theta}$ on by some estimator, observe that from delta method with some $g$,

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

Suppose $g'(\theta)\sqrt{\theta} = c$ is some constant for every $\theta > 0$, our goal is also achieved, i.e.,

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

**Claim.** There exists $g$ such that $c = 1/2$.

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

Then, the asymptotic confidence interval for $g(\theta)$ with confidence level $1 - \alpha$ is just

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

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

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

In our case, $g^{-1}(u) = u^2$. This is the so-called *variance stabilizing transformation*, and can be easily generalized.

# Lecture 13: Inference for Population Quantiles

## 2.5.9 Inference for Population Quantiles

27 Feb. 9:30

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

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

Consider the empirical CDF

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

for all $t \in \mathbb{R}$, and by weak law of large number, $\hat{F}_n(t) \overset{p}{\to} \mathbb{P}(X \leq t) = F(t)$, i.e., for all $\epsilon > 0$, $\mathbb{P}(|\hat{F}_n(t) - F(t)| > \epsilon) \to 0$. We note the following.

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

**Note.** Hoeffding's inequality implies that this sum of Bernoulli's converges exponentially fast.

Now, let's estimate $\theta_p = F^{-1}(p)$ by

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

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

**Notation** (Order statistic)**.** The *order statistics* of $X_i$'s is defined as $X_{(1)} \leq \cdots \leq X_{(n)}$.

**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 2.5.5.** If $F(\theta_p + \epsilon) > F(\theta_p) \geq p$ for any $\epsilon > 0$, then $\hat{\theta}_p \overset{p}{\to} \theta_p$. More generally, if $p_n \to p$, then $\hat{\theta}_{p_n} \overset{p}{\to} \theta_p$.

**Proof.** We want to show that for any $\epsilon > 0$, $\mathbb{P}(|\hat{\theta}_{p_n} - \theta_p| > \epsilon) \to 0$. We see that

$$\mathbb{P}(|\hat{\theta}_{p_n} - \theta_p| > \epsilon) = \mathbb{P}(\hat{\theta}_{p_n} > \theta_p + \epsilon) + \mathbb{P}(\hat{\theta}_{p_n} < \theta_p - \epsilon).$$

For the first term, $\hat{\theta}_{p_n} = \hat{F}_n^{-1}(p_n) > \theta + \epsilon$, hence $p_n > \hat{F}_n(\theta_p + \epsilon)$, i.e.,

$$p_n - p + p - F(\theta_p + \epsilon) > \hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon).$$

Since $p < F(\theta_p + \epsilon)$, let $-\delta \coloneqq p - F(\theta_p + \epsilon)$ 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 $|p_n - p| < \delta/2$. Hence, $|\hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon)| > \delta/2$, 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 we have seen. The second term can be proved similarly. ∎

If $F$ is differentiable, we can actually say more in this situation.

**Theorem 2.5.6** (Bahadur's representation). If $F'(\theta_p) \coloneqq f(\theta_p) > 0$ 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. In the case of $p_n \neq p$, Bahadur's representation shows

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

We see that if $\sqrt{n}(p_n - p) \to c$, then

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

On the other hand, if $\sqrt{n}(p_n - p) \to 0$, then

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

which gives an asymptotically valid $100(1-\alpha)\%$ confidence interval for $\theta_p$ as

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

**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).

However, to implement this confidence interval, we need to estimate $f(\theta_p)$ consistently. On the other hand, 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

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

Consider $\hat{\theta}_{\ell_n}$ first. It's equivalent to say

$$\sqrt{n}(\hat{\theta}_{\ell_n} - \hat{\theta}_p) \to -Z_{\alpha/2} \frac{\sqrt{p(1-p)}}{f(\theta_p)},$$

and if $\sqrt{n}(\ell_n - p) \to c$, e.g., $\ell_n = p + c/\sqrt{n}$, then

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

This suggests $\ell_n = p - Z_{\alpha/2}\sqrt{p(1-p)}/\sqrt{n}$, and similarly, $u_n = p + Z_{\alpha/2}\sqrt{p(1-p)}/\sqrt{n}$. Another implication is the following. Firstly, consider $p = 1/2$, and in this case,

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

**Definition 2.5.4** (Median). When $p = 1/2$, $\theta_{1/2}$ is called the *median*.

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

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

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

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

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

⊛

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

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

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

⊛

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

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

where

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

**Remark.** Bahadur's representation is needed since if only have their asymptotic distribution, their joint is not necessary asymptotically independent, hence we really need their representations.

# Lecture 14

29 Feb. 9:30

# Appendix

# Bibliography


[Das] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Science & Business Media. 727 pp. ISBN: 978-0-387-75971-5. Google Books: sX4_AAAAQBAJ.

[Fer] Thomas S. Ferguson. *A Course in Large Sample Theory*. Routledge. 140 pp. ISBN: 978-1-351-47005-6. Google Books: clc0DwAAQBAJ.

[Leh] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media. 640 pp. ISBN: 978-0-387-98595-4. Google Books: geIoxvgTXlEC.

[Ser] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons. 399 pp. ISBN: 978-0-470-31719-8. Google Books: enUouJ4EHzQC.

[Vaa] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D.