STAT575
Lrage Sample Theory

Pingbang Hu

March 19, 2024

Abstract

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

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

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

Contents

Chapter 1

Introduction

Lecture 1: Introduction to Large Sample Theory

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

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

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

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

1.1 Parametrized Approach

There are various ways of doing this task, one way is the so-called parametrized approach. By postulating a family of cdfs $\{F_\theta, \theta \in \Theta\}$ where Θ 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 θ^* 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 θ^* such that F_{θ^*} is, in some sense, the "closest" to F among all possible F_θ (minimizing divergence, to be precise).

1.2 Hypothesis Testing

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

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

To do this, a well-known 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)**.** Cauchy distribution doesn't have finite moment of order greater than 1.

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

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

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

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

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

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

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

Chapter 2

Modes of Convergence

Lecture 2: Modes of Convergence

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.

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

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

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

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

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

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

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

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

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 $\lVert X_n - X \rVert \leq Z$ for all $n \geq 1$ where $\mathbb{E}\left[\lVert Z \rVert^p\right] < \infty$, then $X_n \overset{L^p}{\to} X$.

Theorem 2.1.2 (Scheffé's theorem)**.** If $X_n \overset{p}{\to} X$ and $\limsup_{n\to\infty} \mathbb{E}\left[\lVert X_n \rVert^p\right] \leq \mathbb{E}\left[\lVert X \rVert^p\right] < \infty$, 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 $X_1, \dots, X_n$, we estimate θ by computing an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ of θ. There are some properties we might want for $\hat{\theta}_n$.

Definition 2.1.6 (Consistent)**.** $\hat{\theta}_n$ is *consistent* of θ 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 θ 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 θ *in mean squared error* if $\hat{\theta}_n \overset{L^2}{\to} \theta$.

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

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

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

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

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

X has a pdf f, then

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

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 θ.

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 when $\mathrm{Var}[X] = \infty$ as we expect.

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

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

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

Another interesting problem is the supremum estimation.

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

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

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

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

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., $\sqrt{n}(\hat{\theta}_n - \theta) \sim \mathcal{N}(0, 1)$, i.e., we can write down a confidence interval such as $\hat{\theta}_n \pm 1.96/\sqrt{n}$ with 95% confidence level for θ.

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

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

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

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

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

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

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

Lecture 3: Weak Convergence Portmanteau Theorem

2.2 Weak Convergence

25 Jan. 9:30

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

2.2.1 Convergence in Total Variation

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

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

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

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

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

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

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

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

2.2.2 Weak Convergence

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

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

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

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

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

2.2.3 Portmanteau Theorem

The following is our main tool of proving weak convergence.

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

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

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

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

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

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

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

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

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

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

Then, consider

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

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

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

for all $x, y \in S$. ■

Then we can prove the Portmanteau theorem.

Proof of Theorem 2.2.1. (a) ⇒ (b) is clear, and we start by proving (c) ⇔ (d).

Claim. (c) ⇔ (d).

Proof. We first prove that (d) ⇒ (c). Since when A is open,

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

(c) ⇒ (d) is exactly the same, hence (c) ⇔ (d). ⊛

Next, we prove (b) ⇒ (d), which gives us (a) ⇒ (b) ⇒ (d) ⇔ (c).

Claim. (b) ⇒ (d).

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

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

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

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

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

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

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

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

which is a contradiction. ■

We now finish the proof of Portmanteau theorem.

Proof of Theorem 2.2.1 (cont.) We already proved (a) $\Rightarrow$ (b) $\Rightarrow$ (d) $\Leftrightarrow$ (c).

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

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

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

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

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

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

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

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

Finally, we prove the following.

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

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

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

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

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

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

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

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

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

This finishes the proof. ■

2.2.4 Continuous Mapping Theorem

A common scenario is that given a nice function h (in terms of continuity), if $X_n \overset{w}{\to} X$, 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, ρ), (S', ρ'), $g\colon S \to S'$ 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

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

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

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

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

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

since A is closed, hence we're done. ∎

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

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

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

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

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

2.2.5 Slutsky's Theorem

Another useful theorem for proving weak convergence is the following.

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

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

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

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

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

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

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

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

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

Lecture 5: Convergence in Distribution and Weak Convergence

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

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

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

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

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

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

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

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

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

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

However, this does not imply $(X_n, Y_n) \overset{w}{\to} (X, X)$ since otherwise, by continuous mapping theorem, $X_n + Y_n \overset{w}{\to} X + X = 2X$, 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 $X_n \overset{w}{\to} X$ in $\mathbb{R}^d$ and $Y_n \overset{p}{\to} c$ in $\mathbb{R}^m$,[a] then $(X_n, Y_n) \overset{w}{\to} (X, c)$.

[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 $(X_n, c) \overset{w}{\to} (X, c)$. Indeed, since for every continuous and bounded $g \colon \mathbb{R}^{d+m} \to \mathbb{R}$, from $X_n \overset{w}{\to} X$ with $g(\cdot, c)$ being continuous and bounded, $\mathbb{E}_n\left[g(X_n, c)\right] \to \mathbb{E}\left[g(X, c)\right]$.

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

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

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

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

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

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

2.3 Convergence in Distribution

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

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

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

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

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

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

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

Remark. $X_n \overset{\text{TV}}{\to} X$ implies $X_n \overset{D}{\to} X$.

Proof. Since $X_n \overset{\text{TV}}{\to} X$ means $\mathbb{P}_n(X_n \in A) \to \mathbb{P}(X \in A)$ 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's central limit theorem)**.** Let $X_n \sim \mathrm{Bin}(n,p)$, then for every $t \in \mathbb{R}$, as $n \to \infty$,

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

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

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

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

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

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

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

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

One important remark is the following.

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

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

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

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

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

2.3.1 Equivalency of Convergence in Distribution and Weak Convergence

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

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

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

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

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

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

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

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

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

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

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 2.3.1, we recall one important characterization of $\liminf$. 2 Feb. 17:30

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

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

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 2.3.1.

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

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

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

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

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

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

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

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

which proves the claim. ⊛

This proves the case of $d = 1$. ■

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

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

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

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

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

2.4 Stochastic Boundedness

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

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

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

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

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

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

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

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

2.4.1 Sufficient Conditions for Stochastic Boundedness

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The generalization is the following.

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

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

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

for any M. Now, we see that

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

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

We often write the following.

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

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

2.4.2 Delta Method

Let's see one important application which combines the above. Consider an estimator T_n of θ, and a deterministic sequence b_n which goes to ∞. 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 2.4.1 (Delta method)**.** Let $\theta \in \mathbb{R}^d$, (T_n) and Y be random vectors in $\mathbb{R}^d$, and $b_n \to \infty$ be a positive deterministic sequence. If $b_n(T_n - \theta) \overset{D}{\to} Y$, then $T_n \overset{p}{\to} \theta$. Moreover, if $g\colon \mathbb{R}^d \to \mathbb{R}^m$ is differentiable at θ, $b_n(g(T_n) - g(\theta)) \overset{D}{\to} \nabla g(\theta) Y$, where $\nabla g \in \mathbb{R}^{d\times m}$ is the Jacobian of g.

Proof. We first observe that $\|b_n(T_n - \theta)\| \in O_p(1)$ since $b_n(T_n - \theta) \overset{D}{\to} Y$, 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 2.4.2. 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 2.4.3. For the second claim, since g is differentiable at θ,

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

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 θ. Rewriting everything, we have

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

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

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

We observe the following.

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

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

It's enough to show $\|h(T_n)\| = o_p(1)$ since we know that $\|b_n(T_n - \theta)\| = O_p(1)$ and $o_p(1)O_p(1) = o_p(1)$ from Proposition 2.4.3. Indeed, as $T_n \overset{p}{\to} \theta$, $h(T_n) \overset{p}{\to} h(\theta) = 0$ again by continuous mapping theorem with h being continuous at θ. 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 θ.

Lecture 7: Skorohod's Representation Theorem

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 one might not expect.

Theorem 2.5.1 (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.

2.5.1 Quantile Function

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

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

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

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

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

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

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 $t \in \mathbb{R}$,

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

⊛

2.5.2 Proof of Skorohod's representation theorem

Now we can prove Skorohod's representation theorem.

Proof of Theorem 2.5.1. Consider $\widetilde{\Omega} = (0, 1)$, and $\widetilde{\mathbb{P}}((a, b)) = b - a$ for all $a < b$. Then, we can define $U(p) = p$ for all $p \in \widetilde{\Omega}$, i.e., $U \sim \mathcal{U}(0, 1)$. Define $Y_n = F_{X_n}^{-1}(U)$ and $Y = F_X^{-1}(U)$ from the quantile functions. Denote Φ 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 Φ 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 $X_n \xrightarrow{D} X$ means $F_{X_n}(t) \to F_X(t)$, from Lemma 2.3.1, it further implies $F_{X_n}(t^-) \to F_X(t^-)$ for all $t \in C_{F_X}$. Note that $\widetilde{\mathbb{P}}(Z \in C_{F_X}) = 1$ since there can be only countably many discontinuities of F_X. Hence,

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

i.e., converges almost surely, which implies $F_{X_n}(Z) \xrightarrow{D} F_X(Z)$, 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 2.3.1. Again, as F_X can have only countably many discontinuities, this holds for almost all p's, which is what we want to show. ■

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

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

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

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

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

Another application is to generalize Fatou's lemma.

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

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

[a]Can be on different probability spaces.

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

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

■

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

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

in L_1?

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

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

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

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

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

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

the in L_1 version?

2.6 Characteristic Function

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

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

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

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

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

Proof. Mimicking the proof as in Theorem 2.6.1. ■

Lecture 8: Characteristic Functions

We see other similar theorems apart from Theorem 2.6.2.

8 Feb. 9:30

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

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

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

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

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

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

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

2.6.1 Characteristic Function

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

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

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

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

If we write ϕ_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 ϕ_X is meaningful for any random vector X, unlike the moment generating function.

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

We make one more remark for future reference.

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

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

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

2.6.2 Uniqueness Theorem

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

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

Proof. Consider $d = 1$. Observe that if we can write F_X in terms of only ϕ_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 ϕ_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 ϕ_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 2.6.3. ⊛

Hence, our goal now is to write $F_{X+\sigma Z}$ in terms of ϕ_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 $t = (x-y)/\sigma$,

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

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

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

2.6.3 Continuity Theorem

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

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

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

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

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

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

We now state the theorem.

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

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

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

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

The proof will be continued...

Lecture 9: Proof of Lévy-Cramer Continuity Theorem

We now finish the proof of Lévy-Cramer continuity theorem. 13 Feb. 9:30

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

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

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

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

We claim that we can find an $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $\mathbb{P}_n(|X_n| \geq 2/\delta) < \epsilon$.[a] To bound $|X_n|$ with ϕ_{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 δx, for any $x \neq 0$, we have

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

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

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

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 ϵ for all $n \geq n_0$. As $\phi_{X_n}(t) \to \phi(t)$ for all t, we have $|1 - \phi_{X_n}(t)| \to |1 - \phi(t)|$ point-wise, hence by the bounded convergence theorem,

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

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

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

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

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

2.6.4 Inversion Theorem

On the other hand, another way to prove Lévy-Cramer continuity theorem is to directly calculate the pdf of X, given ϕ_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 ϕ_X and calculate $\mathbb{E}[g(X)]$ in terms of ϕ_X will tell us which condition should we add.

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

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

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

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

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

(c) If further, ϕ_X is absolute integrable, then X has a pdf

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

Proof. The proof is based on uniqueness theorem.

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

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

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

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

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

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

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

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

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

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

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

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

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

■

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

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

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

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

2.6.5 Properties of Characteristic Function

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

Proposition 2.6.1. For any random vector X, ϕ_X is uniformly continuous, i.e.,

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

Proof. We see that for any h,

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

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 ϕ_X and its connection to moments.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[a] This is a generalization of Proposition 2.6.1.

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

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

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

Chapter 3

Fundamental Theorems of Probability

Lecture 10: WLLN and CLT, and Applications to Inferences

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

3.1 Law of Large Number and Central Limit Theorem

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

3.1.1 Weak Law of Large Number

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

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

Proof. Since $c := \mathbb{E}[X]$ 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, ϕ is differentiable such that

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

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

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

With the fact that $\sup_t |\phi'(t)| \leq \mathbb{E}[|X|]$, 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. ■

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

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

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

3.1.2 Central Limit Theorem

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

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

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

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

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

We now state the theorem.

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

3.2 Inference for Population Mean and Variance

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

3.2.1 Population Mean

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

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

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

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

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

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

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

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

We see that when X is Gaussian, an asymptotically valid $100(1-\alpha)\%$ confidence interval for μ 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 3.2.1.** If $X \in L^2$, then both $\hat{\sigma}_n^2$ and s_n^2 are consistent estimators of σ^2. Furthermore, $T_n \overset{D}{\to} \mathcal{N}(0,1)$, and the same holds if s_n is replaced by $\hat{\sigma}_n$ in the definition of T_n.

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

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

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

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

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

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

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

3.2.2 Population Variance

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

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

where $(Z_{n-1}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Firstly, since $\mathbb{E}[Z_i^2] = \operatorname{Var}[Z_i] + (\mathbb{E}[Z_i])^2 = 1$, 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,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

implying $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \overset{D}{\to} \mathcal{N}(0, \operatorname{Var}[Y^2])$ from the first claim and Slutsky's theorem, 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 second claim. Finally, we note that

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

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

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

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

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

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

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

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

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

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

The takeaway is when X is not a normal (or when the Kurtosis of X is different from 3), then the distribution of $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2)$ is different. Specifically, if the Kurtosis exists and is not equal to 1, then an asymptotically valid $100(1-\alpha)\%$ confidence interval for σ^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 $Y \coloneqq X - \mu = X - \mathbb{E}[X]$ (and also $Y_i = X_i - \mu$ as usual), $\mu_k \coloneqq \mathbb{E}[Y^k] = \mathbb{E}[(X-\mu)^k]$ for all $k \geq 2$, and finally $\widetilde{\mu}_k = \mu_k/\sigma^k = \mathbb{E}\left[(X-\mu)^k/\sigma^k\right]$. 20 Feb. 9:30

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

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

The task is then the following.

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

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

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

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

The above essentially is motivated from the following observation.

Intuition. If we know μ, 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 μ, we need to use $\overline{X}_n$.

We now show that this still yields a consistent estimator.

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

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

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

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

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

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

Proposition 3.2.2 and Proposition 3.2.3 imply the following.

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

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

3.3 Testing Normality

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

3.3.1 Asymptotic Distribution of Sample Central Moments

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3.2 Testing Normality with Odd Moments

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

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

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

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

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

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

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 3.3.1, let's write this general result as a corollary.

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

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

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

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

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

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

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 $\operatorname{Var}[Y^3] = \mu_{2\times 3} = \mu_6$,[a] the variance of the normal is further equal to

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

which provides the result. ⊛

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

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

3.3.3 Multivariate Central Limit Theorem

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

Theorem 3.3.2 (Cramér-Wold device). Let (X_n) be a sequence of random vectors and X be a random vector in $\mathbb{R}^d$. Then $X_n \overset{D}{\to} X$ 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 3.3.3 (Multivariate central limit theorem). Let (X_n) be i.i.d. random vectors in $\mathbb{R}^d$ with $\mathbb{E}[X_i] = \mu \in \mathbb{R}^d$, $\operatorname{Var}[X_i] = \Sigma \in \mathbb{R}^{d\times d}$ for all $1 \le i \le n$. Then

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

Lecture 12: Asymptotic Joint Distribution by Multivariate CLT

22 Feb. 9:30

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

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

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. ■

3.3.4 Testing Normality with General Moments

With multivariate central limit theorem, we can now generalize Corollary 3.3.1, i.e., finding the asymptotic distribution of $\widetilde{M}_k = M_k/\hat{\sigma}_n^k$ for general k. Recall the setup, where we let (X_n) and X be i.i.d. random variable, $Y_i = X_i - \mu$ (and $Y = X - \mu$), $\sigma^2 = \operatorname{Var}[X]$, $\mu_k = \mathbb{E}[Y^k]$, 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 3.2.1 and Proposition 3.2.2,

- $\sqrt{n}(\overline{X}_n - \mu) \overset{D}{\to} \mathcal{N}(0,\sigma^2)$ 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 $X \in L^4$ and $\widetilde{\mu}_4 > 1$.[a]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

by letting $Z_i \coloneqq (X_i - \mu)/\sigma = Y_i/\sigma$, proving the first claim. Then by the multivariate central limit

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

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

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

as we assumed. ■

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

Example. Consider using both $\widetilde{M}_3$ 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 ℓ, 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 χ^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 $\widetilde{v}_k, \widetilde{v}_\ell < \infty$,

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

3.4 A Quick Detour

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

3.4.1 Asymptotic Relative Efficiency

First, consider the following illustrative example.

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

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

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 ask how many data we need to we get a similar precision: consider the problem of estimating a scalar parameter θ such that for two estimators T^1_n and T^2_n,

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

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

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

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

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

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

Definition 3.4.1 (Asymptotic relative efficiency). The *asymptotic relative efficiency* between two estimators T_n^1 and T_n^2 for θ such that $\sqrt{n}(T_n^i - \theta) \overset{D}{\to} \mathcal{N}(0, \sigma_i^2(\theta))$ is defined as

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

3.4.2 Variance Stabilizing Transformation

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

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

As previously seen. As the asymptotic distribution depends on θ, we don't directly get a confidence interval. Indeed, usually we will first write $\sqrt{n}/\sqrt{\theta}(\overline{X}_n - \theta) \overset{D}{\to} Z \sim \mathcal{N}(0,1)$, replace $\sqrt{\theta}$ by $\sqrt{\overline{X}_n}$, and apply continuous mapping theorem and Slutsky's theorem to get a confidence interval.

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

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

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

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

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

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

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

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

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

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

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

Remark. The above can be easily generalized.

Proof. Consider estimating a scalar parameter θ in some open interval Θ, where we replace:

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

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

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

Lecture 13: Bahadur's Representation for Quantiles

3.5 Inference for Population Quantiles

27 Feb. 9:30

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

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

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

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

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

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

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

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

Proof. Since F is continuous, with probability 1 there are no ties among X_i's, hence $\hat{F}_n(t)$ has jumps of size $1/n$ at every order statistics $X_{(i)}$. Finally, the ceiling can be taken since $\sum_{i=1}^{n} \mathbb{1}_{X_i \geq t} \in \mathbb{N}$. ⊛

3.5.1 Consistency

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

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

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

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

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

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

Theorem 3.5.1. If $F(\theta_p + \epsilon) > F(\theta_p) \geq p$ for any $\epsilon > 0$, then $\hat{\theta}_p \overset{p}{\to} \theta_p$. 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 $p_n > \hat{F}_n(\theta_p + \epsilon)$, which gives

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

Since $p < F(\theta_p + \epsilon)$, let $-\delta \coloneqq p - F(\theta_p + \epsilon)$ 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 $|p_n - p| < \delta/2$, which implies $|\hat{F}_n(\theta_p + \epsilon) - F(\theta_p + \epsilon)| > \delta/2$, i.e.,

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

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

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

3.5.2 Bahadur's Representation Theorem

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

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

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

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

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

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

and

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

Proof. From Bahadur's representation shows

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

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

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

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

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

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

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

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

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

3.5.3 Confidence Intervals

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

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

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

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

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

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

The above construction works due to the following.

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

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

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

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

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

3.5.4 Estimating the Center of a Distribution

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

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

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

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

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

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

Let's summarize the above in the following.

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

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

Example. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\overline{X}$ is a better estimator of μ 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.,

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

⊛

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

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

⊛

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 and 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

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

Lecture 14: Proof of Bahadur's Representation Theorem

3.5.5 Proof of Bahadur's Representation Theorem

29 Feb. 9:30

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

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

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

We now start the proof.

Proof of Theorem 3.5.2. Firstly, we write

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

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

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

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

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

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

then $U_n - W_n \overset{p}{\to} 0$.

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

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

since the left-hand side is equal to

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

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

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

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

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

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

Clearly, $U_n = O_p(1)$ since it converges in distribution, so we can try to apply Lemma 3.5.1. First, we study the numerator of U_n, i.e., $Z_n(t) := \sqrt{n}(F(t) - \hat{F}_n(t))$. We have seen that $\mathbb{E}[Z_n(t)] = 0$ and $\operatorname{Var}[Z_n(t)] = F(t)(1 - F(t))$, and $Z_n(t) \xrightarrow{D} \mathcal{N}(0, F(t)(1 - F(t)))$ by central limit theorem.

Claim. For any $t, s \in \mathbb{R}$, $\operatorname{Var}[Z_n(t) - Z_n(s)] = \mathbb{E}[(Z_n(t) - Z_n(s))^2] \le |F(t) - F(s)|$. Hence, if $s_n \to s$ and F is continuous at s, $Z_n(s_n) - Z_n(s) \xrightarrow{L^2} 0$, hence $Z_n(s_n) - Z_n(s) \xrightarrow{p} 0$.

Proof. Observe that $\operatorname{Var}[Z_n(t) - Z_n(s)] = \operatorname{Var}[\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}] \le \mathbb{E}[|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}|^2]$ where

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

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

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

i.e., $|\mathbb{1}_{X \le t} - \mathbb{1}_{X \le s}| \sim \operatorname{Ber}(|F(t) - F(s)|)$. ⊛

From Lemma 3.5.1, it suffices to show $\mathbb{P}(W_n \le t, U_n \ge t + \epsilon) \to 0$ and $\mathbb{P}(U_n \le t, W_n \ge t + \epsilon) \to 0$ for every $t \in \mathbb{R}$ and $\epsilon > 0$. Let's show the first one only. Fix $t \in \mathbb{R}$ and $\epsilon > 0$, then

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

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

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

which can be written as

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

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

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

which goes to 0 as $n\to\infty$ if $t_n\to t$ since from the previous claim:

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

Claim. Indeed, $t_n\to t$.

Proof. We want to show that

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

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

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

i.e., $F(\theta_p+\delta_n)=F(\theta_p)+\delta_n f(\theta_p)+o(\delta_n)$. Since $F(\theta_p)=p$ and $\delta_n = t/\sqrt{n}+(p_n-p)/f(\theta_p)$,

$$F(\theta_p+\delta_n) = p+\left(\frac{t}{\sqrt{n}}+\frac{p_n-p}{f(\theta_p)}\right)f(\theta_p)+o(\delta_n) = p+\frac{t}{\sqrt{n}}f(\theta_p)+(p_n-p)+o(\delta_n).$$

Rearranging, with $o(\delta_n)\cdot\sqrt{n}/f(\theta_p)=\sqrt{n}o(\delta_n)$ from $f(\theta_p)>0$, we have

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

Finally, since $o(\delta_n)=\delta_n o(1)$, with

$$\sqrt{n}\delta_n = \sqrt{n}\left(\frac{t}{\sqrt{n}}+\frac{p_n-p}{f(\theta_p)}\right) = t+\frac{\sqrt{n}(p_n-p)}{f(\theta_p)} = O(1)$$

from our assumption, we have $\sqrt{n}o(\delta_n)=O(1)o(1)=o(1)$, hence $t_n = t+o(1)\to t$. ⊛

The second claim $\mathbb{P}(U_n\leq t, W_n\geq t+\epsilon)\to 0$ can be proved similarly, hence we're done. ■

3.6 Inference for Distribution Function

Since we estimate F by $\hat{F}_n$ when estimating θ_p by $\hat{\theta}_p$, one might just as well focus on the former task.

3.6.1 Consistency

Since $\sqrt{n}(\hat{F}_n(t)-F(t))\xrightarrow{D}\mathcal{N}(0,F(t)(1-F(t)))$ for any fixed t, so given $t_1,\dots,t_m\in\mathbb{R}$, we have

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

from multivariate central limit theorem where

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

Lecture 15: Inference for Cumulative Density Function

Surprisingly, this consistency property holds uniformly over t. 5 Mar. 9:30

Theorem 3.6.1 (Glivenko-Cantelli). Given a cdf F, $\|\hat{F}_n - F\|_\infty \overset{\text{a.s.}}{\to} 0$ as $n \to \infty$.

Proof. Fix some $\epsilon > 0$, and let $\epsilon > 2/k$ for some $k \in \mathbb{N}$. Then, for finitely many $t_1, \ldots, t_{k-1}$, $\hat{F}_n(t_i) \overset{\text{a.s.}}{\to} F(t_i)$ and $\hat{F}_n(t_i^-) \overset{\text{a.s.}}{\to} F(t_i^-)$ for every $1 \le i \le k-1$ from the strong law of large number. This means that there exists n_0 such that for $n \ge n_0$, for every $\omega \notin N$ such that $\mathbb{P}(N) = 0$,

$$|\hat{F}_n(t_i, \omega) - F(t_i, \omega)| < \frac{1}{k}, \text{ and } |\hat{F}_n(t_i^-, \omega) - F(t_i^-, \omega)| < \frac{1}{k},$$

so when $t \in \{t_i\}_{i=1}^{k-1}$ for some finite k, the desired bound is established. In particular, consider $t_i = \inf\{t \in \mathbb{R} \colon F(t) > i/k\}$ for $1 \le i \le k-1$, and $t_0 \coloneqq -\infty$, $t_k = \infty$. Then for any $t \in \mathbb{R} \setminus \{t_i\}_{i=1}^{k-1}$, there exists a unique i such that $t \in (t_{i-1}, t_i)$, and furthermore, for all $n \ge n_0$,

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

Similarly, we can show $\hat{F}_n(t) - F(t) > -\epsilon$ for all $t \in \mathbb{R} \setminus \{t_i\}_{i=1}^{k-1}$, which completes the proof. ■

3.6.2 Donsker's Theorem

On the other hand, for distributional result, first recall the *empirical process*

$$Z_n(t) \coloneqq \sqrt{n}(F(t) - \hat{F}(t))$$

for $t \in \mathbb{R}$ introduced in the proof of Bahadur representation theorem. Recall the following.

As previously seen. We have seen that

$$Z_n(t) \coloneqq \sqrt{n}(F(t) - \hat{F}_n(t)) \overset{D}{\to} B_F(t) \coloneqq \mathcal{N}(0, F(t)(1 - F(t))).$$

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

$$(Z_n(t_1), \ldots, Z_n(t_m)) \overset{D}{\to} (B_F(t_1), \ldots, B_F(t_m)) \sim \mathcal{N}(0, \Sigma_F(t_1, \ldots, t_m))$$

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

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

We now ask the same question, i.e., does the convergence hold uniformly over t?

Intuition. As the theory of weak convergence applies to sequences of random elements that take values in general metric spaces, it's reasonable to conjecture that (Z_n) converges weakly to some random process B_F with index set $\mathbb{R}$, i.e., $B_F = \{B_F(t)\}_{t \in \mathbb{R}}$, such that for every $t, s \in \mathbb{R}$,

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

The conjecture is indeed correct, and it's an extension of Donsker's theorem.

Note. The formal setup is to view each Z_n as a random element that takes values on the space $\mathcal{D}$ of right continuous functions with left limits with the norm $\|\cdot\|_\infty$.

Example (Brownian bridge). A *Brownian bridge* is $B \coloneqq B_F$ when $F(t) = t$, i.e., $F \sim \mathcal{U}([0,1])$.

Note. For any cdf F, $B_F(t)$ is just B index at $F(t)$ for any $t \in \mathbb{R}$, i.e., $B_F(t) = B(F(t))$.

Taking this convergence as granted, i.e., $(Z_n) \coloneqq (t \mapsto Z_n(t))_{n \geq 1} \overset{w}{\to} B_F \coloneqq \{t \mapsto B(t)\}_{t \in \mathbb{R}}$. One immediate consequence is the following.

Proposition 3.6.1. If $T \colon \mathcal{D} \to \mathbb{R}$ is continuous,[a] then $T(Z_n) \overset{D}{\to} T(B_F)$.

[a] I.e., $T(G_n) \to T(F)$ if $(G_n), F \in \mathcal{D}$ such that $\|G_n - F\|_\infty \to 0$.

Proof. Since $Z_n \overset{w}{\to} B_F$, continuous mapping theorem proves the result. ■

3.6.3 Confidence Bands and Goodness of Fit Tests

One immediate application is the following.

Corollary 3.6.1. We have $\sqrt{n}\|\hat{F}_n - F\|_\infty \overset{D}{\to} \|B_F\|_\infty$.

Proof. From Proposition 3.6.1, we just note that $\|\cdot\|_\infty$ is continuous since it's a norm. ■

In particular, Corollary 3.6.1 allows us to do inference.

Example (Confidence bands). Consider $\alpha = \mathbb{P}(\|B_F\|_\infty \geq d_\alpha)$ for some d_α, then if F is continuous, $\mathbb{P}(\sqrt{n}\|F - \hat{F}_n\|_\infty \geq d_\alpha) \to \alpha$, i.e.,

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

Another application of Proposition 3.6.1 is the following.

Corollary 3.6.2. We have

$$n \int_{\mathbb{R}} \left(\hat{F}_n(t) - F(t)\right)^2 F(\mathrm{d}t) \overset{D}{\to} \int_{\mathbb{R}} B_F^2(t) F(\mathrm{d}t).$$

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

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

since $\|G_n - G\|_\infty = \sup_t |G_n(t) - G(t)|$. As $\|G_n - G\|_\infty \to 0$, we're done. ■

Suppose F is continuous, then Corollary 3.6.1 and Corollary 3.6.2 suggest we can test the null hypothesis $H_0 \colon F = F_0$ using the *Kolmogorov-Smirnov statistic*

$$K_n \coloneqq \|\hat{F}_n - F_0\|_\infty,$$

and the *Cramér-von Mises statistic*

$$C_n \coloneqq \int_{\mathbb{R}} \left(\hat{F}_n(t) - F_0(t)\right)^2 F_0(\mathrm{d}t).$$

Specifically, we

- reject H_0 when $\sqrt{n}\|\hat{F}_n - F_0\|_\infty = \sqrt{n} K_n \geq d_\alpha$, where d_α is defined as $\alpha = \mathbb{P}(\|B_F\|_\infty > d_\alpha)$;

- reject H_0 when $n\int_{\mathbb{R}}(\hat{F}_n - F_0)^2 \, \mathrm{d}F_0 = nC_n \geq c_\alpha^2$, where c_α is defined as $\alpha = \mathbb{P}(\int_0^1 B_F^2(t)\,\mathrm{d}t \geq c_\alpha^2)$.

It's clear that under H_0, the above two tests will reject with probability approaching α. On the other hand, we have the following.

Proposition 3.6.2. Suppose F is continuous. Consider the test $H_0\colon F = F_0$ using Kolmogorov-Smirnov statistic, then for any $F \neq F_0$, $\mathbb{P}_F(\text{reject}) = \mathbb{P}(\lVert B_F \rVert_\infty \geq d_\alpha) \to 1$ as $n \to \infty$.

Proof. For any metric d, we have

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

As $F \neq F_0$, $d(F, F_0) > 0$ is a fixed number, so $\sqrt{n}d(F, F_0) \to \infty$; on the other hand, since $\sqrt{n}\lVert \hat{F}_n, F\rVert_\infty = O_p(1)$ Corollary 3.6.1 with $d = \lVert\cdot\rVert_\infty$, the right-hand side goes to 1. ■

The same result can be obtained for the case of using the Cramér-von Mises statistic as follows.

As previously seen. From Corollary 3.6.2, under $H_0\colon F = F_0$,

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

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

$$h(F) \coloneqq \int_{\mathbb{R}} (F - F_0)^2 \, \mathrm{d}F_0.$$

Since h is continuous, $C_n = h(\hat{F}_n) \to h(F)$. As for a distributional result, we have the following.

Proposition 3.6.3. There is a function g so that $\mathbb{E}[g(X)] = 0$ and

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

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

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

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

we have

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

As $n\int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 \xrightarrow{w} \int_{\mathbb{R}} B_F^2 \, \mathrm{d}F_0$, Proposition 2.4.2 implies

$$\sqrt{n}\int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 = \frac{n}{\sqrt{n}}\int_{\mathbb{R}} (\hat{F}_n - F)^2 \, \mathrm{d}F_0 = \frac{O_p(1)}{\sqrt{n}} = o_p(1),$$

which gives

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

where we define the function $g\colon \mathbb{R}\to\mathbb{R}$ as

$$g(x) \coloneqq 2\int_{\mathbb{R}} (\mathbb{1}_{x\leq t} - F(t))(F(t)-F_0(t))F_0(\mathrm{d}t)$$

since

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

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

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

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

Lecture 16: Lindeberg Central Limit Theorem

3.7 Lindeberg Central Limit Theorem

19 Mar. 9:30

Now, we consider the case that when (X_n) are only independent but not identically distributed with $\operatorname{Var}[X_i]<\infty$ for all $i\geq 1$. I.e., we want to investigate $S_n = X_1+\cdots+X_n$'s asymptotic distribution.

3.7.1 Illustrative Examples Regarding Poisson Distribution

Let's first see some examples regarding Poisson distribution.

Example. If $X_i\sim \mathrm{Pois}(1/i^2)$ for all $i\geq 1$, then

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

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

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

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

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

This does not contradict to central limit theorem since $\mathrm{Pois}(1/n)$ depends on n.

In general, for any $n\geq 1$, let $K_n\nearrow\infty$ be the number of independent random variables in the sequence $X_{n1},\dots,X_{nK_n}$ with $\operatorname{Var}[X_{nj}]<\infty$ for all $1\leq j\leq K_n$ and n. Again, we define $S_n = X_{n1}+\cdots+X_{nK_n}$.

In picture, we have something like

$$\begin{aligned}
&n = 1: && X_{11}, \dots, X_{1K_1};\\
&n = 2: && X_{21}, X_{22}, \dots, X_{1K_2};\\
&\vdots\\
&n: && X_{n1}, X_{n2}, X_{n3}, \dots, X_{nK_n}.
\end{aligned}$$

Note the following.

Remark. For different n, (X_{nj}) can be defined on different probability space.

Note. As a special case, we previously have $X_{nj} = X_j$ for all $1 \leq j \leq n$, i.e., $K_n = n$.

3.7.2 Lindeberg Condition and Lindeberg Central Limit Theorem

The goal of this section is to establish the following.

Theorem 3.7.1 (Lindeberg central limit theorem)**.** For every $n \geq 1$, let (X_{nK_n}) be a sequence of independent variables with $K_n \nearrow \infty$ and let $Y_{nj} \coloneqq (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\operatorname{Var}[S_n]}$ for every $1 \leq j \leq K_n$. If the Lindeberg condition holds, then

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\operatorname{Var}[S_n]}} = \sum_{j=1}^{K_n} \frac{X_{nj} - \mathbb{E}[X_{nj}]}{\sqrt{\operatorname{Var}[S_n]}} \eqqcolon \sum_{j=1}^{K_n} Y_{nj} \overset{D}{\to} \mathcal{N}(0, 1)$$

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

We first explain the sufficient condition stated in the Lindeberg central limit theorem. Firstly, a weaker but more intuitive notion one might consider is the following.

Definition 3.7.1 (Uniform asymptotic negligibility)**.** Given a (family of) sequence (X_{nK_n}), we say it satisfies the *uniform asymptotic negligibility* (*UAN*), if as $n \to \infty$,

$$\frac{\max_{1 \leq j \leq K_n} \operatorname{Var}[X_{nj}]}{\operatorname{Var}[S_n]} \to 0.$$

However, as we have seen in the second examples, UAN doesn't suffice for the Lindeberg central limit theorem to hold since in this case, $\max_{i \leq j \leq n} \operatorname{Var}[X_{nj}] = 1/n \to 0$, but we know that Lindeberg central limit theorem fail. Hence, we consider the following stronger notion.

Definition 3.7.2 (Lindeberg condition)**.** Given a (family of) sequence (X_{nK_n}), let $Y_{nj} \coloneqq (X_{nj} - \mathbb{E}[X_{nj}])/\sqrt{\operatorname{Var}[S_n]}$ for every $1 \leq j \leq K_n$ and every $n \geq 1$. Then we say (X_{nK_n}) satisfies the *Lindeberg condition* if for every $\epsilon > 0$, as $n \to \infty$,

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

Indeed, Lindeberg condition is stronger than uniform asymptotic negligibility.

Proposition 3.7.1. The Lindeberg condition implies uniform asymptotic negligibility.

Proof. We want to prove that $\max_{1 \leq j \leq K_n} \operatorname{Var}[Y_n] \to 0$ as $n \to \infty$. Firstly, for any n and every $1 \leq j \leq K_n$, by splitting up the expectation, for every $\epsilon > 0$, we have $\operatorname{Var}[Y_{nj}] \leq \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}| > \epsilon}] + \epsilon^2$.

Then with the Lindeberg condition, we have

$$\max_{i\leq j\leq K_n} \operatorname{Var}[Y_{nj}] \leq \sum_{j=1}^{K_n} \mathbb{E}[Y_{nj}^2 \cdot \mathbb{1}_{|Y_{nj}|>\epsilon}] + \epsilon^2 \to \epsilon^2,$$

i.e., $\limsup_{n\to\infty} \max_{1\leq j\leq K_n} \operatorname{Var}[Y_{nj}] \leq \epsilon^2$. By letting $\epsilon \to 0$, we complete the proof. ■

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

Remark. Let (X_n) be a sequence of independent Gaussian variables with $\operatorname{Var}[X_i] = 1/i^2$ for all $i \geq 1$. Then indeed, $S_n \sim \mathcal{N}$. However,

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

hence uniform asymptotic negligibility does not hold.

Remark. If the usual central limit theorem holds with uniform asymptotic negligibility, it will imply the Lindeberg central limit theorem.

Now, to prove the Lindeberg central limit theorem, we will need the following lemmas.

Lemma 3.7.1. For any $w_1, \dots, w_n, z_1, \dots, z_n \in \mathbb{C}$ such that $|w_i|, |z_i| \leq 1$ for all $1 \leq i \leq n$, we have

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

It turns out that the following uniform bounds are useful.

Lemma 3.7.2. For any $x \in \mathbb{R}$, $|e^{ix} - (1 + ix)| \leq x^2/2$ and $|e^{ix} - 1 - ix - (ix)^2/2| \leq x^2$.

Proof. Recall the specific form of Taylor expansion we used before, which gives

$$e^{ix} = 1 + ix + (ix)^2 \int_0^1 \int_0^1 e^{iuvx} u \,\mathrm{d}u \,\mathrm{d}v = 1 + ix + \frac{(ix)^2}{2} + (ix)^2 \int_0^1 \int_0^1 (e^{iuvx} - 1) u \,\mathrm{d}u \,\mathrm{d}v,$$

which gives both inequalities by bounding the two integrals differently. ■

On the other hand, when $|z|$ is small enough, we have the following tighter bounds.

Lemma 3.7.3. For any $z \in \mathbb{C}$ such that $|z| \leq \epsilon < 1$, $|e^z - 1 - z - z^2/2| \leq |z|^3/(1 - \epsilon)$.

Proof. Since

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

where the series converges from the fact that $|z| < 1$. ■

Lemma 3.7.4. For any $z \in \mathbb{C}$ such that $|z| < \delta/2$ where $\delta \in (0, 1)$, $|e^{iz} - (1 + iz)| \leq \delta|z|$.

Proof. Since

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

where the series converges from the fact that $|z| < \delta/2 < 1$. ■

Finally, recall the following.

As previously seen. From Equation 2.1, for $Z \sim \mathcal{N}(0,1)$, $\phi_Z(t) = e^{-t^2/2}$.

We can now prove the Lindeberg central limit theorem.

Proof of Theorem 3.7.1. Let $\phi_{nj}(t) := \mathbb{E}[e^{itX_{nj}}]$ for $t \in \mathbb{R}$. We want to show that

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

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

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

Firstly, consider the first term, and recall what we have shown in the homework.

As previously seen. If ϕ is a characteristic function, so is $e^{\lambda(\phi-1)}$ for any $\lambda > 0$.

Hence, $e^{\phi_{nj}(t)-1}$ is a characteristic function, so both $\phi_{nj}(t)$ and $e^{\phi_{nj}(t)-1}$ are bounded by 1. This means we can apply Lemma 3.7.1 and get

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

Let $z_j := \phi_{nj}(t) - 1$, then the above is just $\sum_{j=1}^{K_n} |e^{z_j} - (z_j + 1)|$, suggesting Lemma 3.7.4. Fixing some $\delta \in (0,1)$, we show that $\max_{1 \leq j \leq K_n} |z_j| < \delta/2$ for large enough n.

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

Proof. As $\mathbb{E}[Y_{nj}] = 0$ for all $1 \leq j \leq K_n$, by using Lemma 3.7.2, we have

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

From the Lindeberg condition, $\max_{1 \leq j \leq K_n} \mathbb{E}[Y_{nj}^2] \to 0$, hence we're done. ⊛

Hence, for any $\delta \in (0,1)$, when n is large enough, Lemma 3.7.4 and the above calculation gives,

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

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

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

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

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

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

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

We then see that

- from Lemma 3.7.2, with $x \coloneqq tY_{nj}$, we can bound the first term as

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

  as $n \to \infty$ by the Lindeberg condition;
- from Lemma 3.7.3, with $z \coloneqq itY_{nj}$ such that $|z| = |tY_{nj}| \leq |t|\epsilon$ under the event. Let ϵ be defined such that $|t|\epsilon < 1$, then we can bound the second term by

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

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

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

Corollary 3.7.1 (Hajek-Sidak central limit theorem). If $X_{ni} = c_{ni}X_i$ for all $1 \leq i \leq n =: K_n$ where $X, (X_n)$ are i.i.d. with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$. If $\max_{1\leq i\leq n} c_n^2 \ll \sum_{i=1}^n c_{ni}^2$, then the Lindeberg condition holds.

Appendix

Bibliography

[Das08] Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer Science & Business Media, Feb. 6, 2008. 727 pp. ISBN: 978-0-387-75971-5. Google Books: sX4_AAAAQBAJ.

[Fer17] Thomas S. Ferguson. *A Course in Large Sample Theory*. Routledge, Sept. 6, 2017. 140 pp. ISBN: 978-1-351-47005-6. Google Books: clc0DwAAQBAJ.

[Leh04] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer Science & Business Media, Aug. 27, 2004. 640 pp. ISBN: 978-0-387-98595-4. Google Books: geIoxvgTXlEC.

[Ser09] Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Sept. 25, 2009. 399 pp. ISBN: 978-0-470-31719-8. Google Books: enUouJ4EHzQC.

[Vaa98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 1998. ISBN: 978-0-521-78450-4. DOI: 10.1017/CBO9780511802256. URL: https://www.cambridge.org/core/books/asymptotic-statistics/A3C7DAD3F7E66A1FA60E9C8FE132EE1D (visited on 10/17/2023).