STAT575
Lrage Sample Theory

Pingbang Hu

January 30, 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

1 Introduction **2**
1.1 Parametrized Approach 2
1.2 Hypothesis Testing 3

2 Modes of Convergence **4**
2.1 Different Modes of Convergence 4
2.2 Weak Convergence 7

Chapter 1

Introduction

Lecture 1: Introduction to Large Sample Theory

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

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

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

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

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

1.1 Parametrized Approach

There are various ways of doing this task, one way is the so-called parametrized approach. By postulating a family of CDFs $\{F_\theta, \theta \in \Theta\}$ where 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 $\operatorname{Var}[X] < \infty$.

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

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

In this case, some other test is needed. A simple test would be looking at the sign of X_i, i.e., the sign test.

Example (Sign test)**.** We might reject H_0 if $\sum_{i=1}^{n} \mathbb{1}_{X_i > 0}$ is large. Note that under H_0, $\sum_{i=1}^{n} \mathbb{1}_{X_i > 0} \sim \operatorname{Bin}(n, 1/2)$, 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 convergence). (X_n) *point-wise converges* to X, denoted as $X_n \to X$, if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.

As previously seen. From analysis, $X_n(\omega) \to X(\omega)$ if and only if for every $\epsilon > 0$, there exists $n_0(\omega) \in \mathbb{N}$ such that for every $n \geq n_0$, $\|X_n(\omega) - X(\omega)\|_2 < \epsilon$.

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

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

In other words, almost-surely convergence means that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega \setminus N$ where $\mathbb{P}(N) = 0$. However, this might still be too strong.

Definition 2.1.3 (Convergence 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 (Converges 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 (Converges in L^p). Let $p > 0$, we say $X_n \overset{L^p}{\to} X$ if $\mathbb{E}\left[\|X_n - X\|^r\right] \to 0$ 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-cusrely convergence is useful.

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

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

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

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

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

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

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

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

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

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

2.1.2 Applications to Statistics

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

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 $n \to \infty$.

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

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

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

On the other hand, if X has a PMF f, then

$$\mathbb{E}\left[X\right] = \sum_x x f(x) = \sum_x x \Delta F(x),$$

where $f(x) = \Delta F(x) \equiv F(x) - F(x_-)$. And if X has a PDF f, then

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

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

where $F(\mathrm{d}x) := f(x)\mathrm{d}x$ in a measure-theoretical sense.

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

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

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

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

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

Another interesting problem is the supremum estimation.

Example (Supremum estimation)**.** Suppose there is a $\theta \in \mathbb{R}$ where distribution function F such that $F(\theta - \epsilon) < 1 = F(\theta)$ for all $\epsilon > 0$. This means $\theta = \sup_\omega X(\omega)$ since $\mathbb{P}(X \leq \theta - \epsilon) = F(\theta - \epsilon)$ and $F(\theta) = \mathbb{P}(X \leq \theta)$.[a] The natural estimator for θ would be $\hat{\theta}_n = \max_{1 \leq i \leq n} X_i$, and it's indeed strongly consistent.

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

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

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

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

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

Example. Consider any deterministic sequence (a_n) in $\mathbb{R}$ which converges to 0. Adding a_n to $\hat{\theta}_n$ will not change the convergence of $\hat{\theta}_n$. This shows that being consistent might not be enough in some cases.

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

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

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

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

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

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

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

Let's see some examples.

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

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

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

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

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

Proof. To show $\|(X_n, Y_n) - (X, Y)\| \to 0$, we have

$$\|(X_n, Y_n) - (X, Y)\| \leq \|X_n - X\| + \|Y_n - Y\|$$

for all $n \geq 1$.[a] The latter two terms goes to 0 (in whatever sense) by assumption. ⊛

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

Lecture 3: Weak Convergence Portmanteau theorem

2.2 Weak Convergence

25 Jan. 9:30

Another important mode of convergence is called weak convergence, which we'll be primarily working with. Again, consider working with a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. First we see the following.

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

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

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

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

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

Remark. If X_n has density f_n and X has density f, $f_n \to f$ implies $X_n \stackrel{\mathrm{TV}}{\to} X$.

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

Let's see some examples.

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

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

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

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

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

This suggests that we should look at something weaker.

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

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

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

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

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

2.2.1 Portmanteau Theorem

The following is our main tool of proving weak convergence.

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

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

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

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

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

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

Before we prove Portmanteau theorem, we need some result from analysis. Specifically, although we might think of we're working in Euclidean spaces, all our discussion can be extended to metric spaces. Recall some results from metric 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 $x, y \in S$,

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

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

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

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

Now, we let

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

We see that

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

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

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

for all $x, y \in S$. ■

Then we can prove the Portmanteau theorem.

Proof of Theorem 2.2.1. (a) $\Rightarrow$ (b) is clear. And note that (c) $\Leftrightarrow$ (d) since the following.

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

Proof. Since when A is open,

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

⊛

(c) $\Rightarrow$ (d) is exactly the same, hence (c) $\Leftrightarrow$ (d). Next, we prove (b) $\Rightarrow$ (d).

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

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

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

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

$$\limsup_{n \to \infty} \mathbb{P}(X_n \in A) \leq \mathbb{E}\left[g_k(X)\right] \to \mathbb{E}\left[\mathbb{1}_A(X)\right] = \mathbb{P}(X \in A)$$

as $k \to \infty$. ⊛

The proof will be continued...

Lecture 4: Continuous Mapping Theorem

Before continuing the proof of Portmanteau theorem, we need one additional tools. 26 Jan. 9:30

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

[a] Note that I can be uncountable.

Proof. Since we can write

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

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

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

which is a contradiction. ■

We now finish the proof of Portmanteau theorem.

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

Claim. (c) + (d) ⇒ (e).

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

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

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

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

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

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

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

Finally, we prove the following.

Claim. (e) ⇒ (a).

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

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

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

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

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

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

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

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

This finishes the proof. ■

2.2.2 Continuous Mapping Theorem

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

As previously seen. Given two metric spaces (S, 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

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

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

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

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

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

from Proposition 2.2.1, i.e.,

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

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

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

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

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

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

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

Finally, another characterization of weak convergence is the following.

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

We first see some applications.

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

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

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

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

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

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).