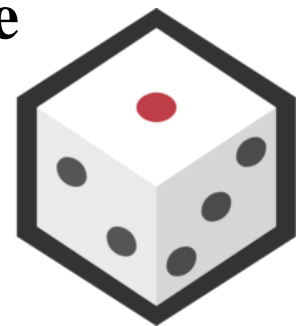# Statistics Learning for Graduate

Author: Ziyang Gong

Date: May 5, 2021

Version: 0.1.0

Facts are stubborn things, but statistics are pliable. —— Mark Twain

Contents

VI   Statistics Inference

VII   Convex Optimization

Part I

Calculus

Chapter 1 Limit Theory

Definition 1.1. Mapping

Let $X : \Omega_1 \to \Omega_2$ be a mapping.

1. *For every subset $B \in \Omega_2$, the inverse image of B is*

$$X^{-1}(B) = \{\omega : \omega \in \Omega_1, X(\omega) \in B\} := \{X \in B\}.$$

2. *For every class*

♣

Chapter 2 Differential Calculus

Chapter 3 Integral Calculus

Part II

Real Analysis

Chapter 4 Measure Theory

4.1 Semi-algebras, Algebras and Sigma-algebras

> **Definition 4.1. Semi-algebra**
>
> A nonempty class of $\mathcal{S}$ of subsets of Ω is an **semi-algebra** on Ω that satisfy
>
> 1. if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$.
> 2. if $A \in \mathcal{S}$, then A^C is a finite disjoint union of sets in $\mathcal{S}$, i.e.,
> $$A^C = \sum_{i=1}^{n} A_i, where A_i \in \mathcal{S}, A_i \cap A_j = \emptyset, i \neq j.$$

> **Definition 4.2. Algebra**
>
> A nonempty class $\mathcal{A}$ of subsets of Ω is an **algebra** on Ω that satisfy
>
> 1. if $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.
> 2. if $A_1, A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$.

> **Definition 4.3. σ-algebra**
>
> A nonempty class $\mathcal{F}$ of subsets of Ω is a σ-**algebra** on Ω that satisfy
>
> 1. if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
> 2. if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\cup_i A_i \in \mathcal{F}$.

Example 4.1 Special σ-algebra

1. **Trival σ-algebra** := $\{\emptyset, \Omega\}$. This is smallest σ-algebra.
2. **Power Set** := all subsets of σ, denoted by $\mathcal{P}(\Omega)$. This is the largest σ-algebra.
3. **The smallest σ-algebra containing** $A \in \Omega := \{\emptyset, A, A^C, \Omega\}$.

   It is easy to define (Lesbegue) measure on the semi-algebra $\mathcal{S}$, and then easily to extend it to the algebra $\overline{\mathcal{S}}$, finally, we can extend it further t o some σ-algebra (mostly consider the smallest one containing $\mathcal{S}$).

> **Lemma 4.1**
>
> If $\mathcal{S}$ is a semi-algebra, then
> $$\overline{\mathcal{S}} = \{finite\ disjoint\ unions\ of\ sets\ in\ \mathcal{S}\}$$

> *is an algebra, denoted by $\mathcal{A}(S)$, called **the algebra generated by** S.*

Proof   Let $A, B \in \overline{S}$, then $A = \sum_{i=1}^{n} A_i, B = \sum_{j=1}^{m} B_j$ with $A_i, B_i \in S$.

Intersection: For $A_i \cap B_j \in S$ by the definition of semi-algebra S, thus

$$A \cap B = \sum_{i=1}^{n} \sum_{j=1}^{m} A_i \cap B_j \in \overline{S}.$$

So $\overline{S}$ is closed under (finite) intersection.

Complement: For DeMorgan's Law, $A_i^C \in S$ by the definition of semi-algebra S and $\overline{S}$ closed under (finite) intersection that we just shown, thus

$$A^C = (\sum_{i=1}^{n} A_i)^C = \cap_{i=1}^{n} A_i^C \in \overline{S}.$$

So $\overline{S}$ is closed under complement.

Union: For DeMorgan's Law and $\overline{S}$ closed under (finite) intersection and complement that we just shown, thus

$$A \cup B = (A^C \cap B^C)^C \in \overline{S}.$$

So $\overline{S}$ is closed under (finite) union.

Hence, $\overline{S}$ is an algebra.

Theorem 4.1

*For any class $\mathcal{A}$, there exists a unique minimal σ-algebra containing $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$, called **the σ-algebra generated by** $\mathcal{A}$. In other words,*

  1. $\mathcal{A} \subset \sigma(\mathcal{A})$.

  2. For any σ-algebra $\mathcal{B}$ with $\mathcal{A} \subset \mathcal{B}$, $\sigma(\mathcal{A}) \subset \mathcal{B}$.

and $\sigma(\mathcal{A})$ is unique.

Proof   **Existence**:

Uniqueness:

Example 4.2 Borel σ-algebras generated from semi-algebras

  1.

4.2 Measure

> **Definition 4.4. Measure**
>
> **Measure** is a nonnegative countably additive set function, that is, a function $\mu : \mathcal{A} \to \mathbf{R}$ with
>
> 1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{A}$.
> 2. if $A_i \in \mathcal{A}$ is a countable sequence of disjoint sets, then
> $$\mu(\cup_i A_i) = \sum_i \mu(A_i).$$
>
> ♣

> **Definition 4.5. Measure Space**
>
> If μ is a measure on a σ-algebra $\mathcal{A}$ of subsets of Ω, the triplet $(\Omega, \mathcal{A}, \mu)$ is a **measure space**.
>
> ♣

Note *A measure space $(\Omega, \mathcal{A}, \mu)$ is a **probability space**, if $P(\Omega) = 1$.*

Property *Let μ be a measure on a σ-algebra $\mathcal{A}$*

1. **monotonieity** *if $A \subset B$, then $\mu(A) \leq \mu(B)$.*
2. **subadditivity** *if $A \subset \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} u(A_m)$.*
3. **continuity from below** *if $A_i \uparrow A$ (i.e. $A_1 \subset A_2 \subset \ldots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.*
4. **continuity from above** *if $A_i \downarrow A$ (i.e. $A_1 \supset A_2 \supset \ldots$ and $\cap_i A_i = A$), then $\mu(A_i) \downarrow \mu(A)$.*

Proof

Chapter 5 Lebesgue Integration

5.1 Properties of the Integral

> **Theorem 5.1. Jensen's Inequality**
>
> *Let (Ω, A, μ) be a probability space. If f is a real-valued function that is μ-integrable, and if φ is a convex function on the real line, then:*
>
> $$\varphi\left(\int_\Omega f\,\mathrm{d}\mu\right) \leq \int_\Omega \varphi(f)\,\mathrm{d}\mu. \tag{5.1}$$

Proof Let $x_0 = \int_\Omega f\,\mathrm{d}\mu$. Since the existence of subderivatives for convex functions, $\exists a, b \in R$, such that,

$$\forall x \in R, \varphi(x) \geq ax + b \text{ and } ax_0 + b = \varphi(x_0).$$

Then, we got

$$\int_\Omega \varphi(f)\,\mathrm{d}\mu \geq \int_\Omega af + b\,\mathrm{d}\mu = a\int_\Omega f\,\mathrm{d}\mu + b = ax_0 + b = \varphi\left(\int_\Omega f\,\mathrm{d}\mu\right).$$

> **Theorem 5.2. Hölder's Inequality**
>
> *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p, q \in [1, \infty]$ with $1/p + 1/q = 1$. Then, for all measurable functions f and g on Ω,*
>
> $$\int_\Omega |f \cdot g|\,\mathrm{d}\mu \leq \|f\|_p \|g\|_q. \tag{5.2}$$

Proof

> **Theorem 5.3. Minkowski's Inequality**
>
> *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and let $p \in [1, \infty]$. Then, for all measurable functions f and g on Ω,*
>
> $$\|f + g\|_p \leq \|f\|_p + \|g\|_p. \tag{5.3}$$

Proof Since $\varphi(x) = x^p$ is a convex function for $p \in [1, \infty)$. By it's definition,

$$|f + g|^p = \left|2 \cdot \frac{f}{2} + 2 \cdot \frac{g}{2}\right|^p \leq \frac{1}{2}|2f|^p + \frac{1}{2}|2g|^p = 2^{p-1}\left(|f|^p + |g|^p\right).$$

Therefore,

$$|f + g|^p < 2^{p-1} \left(|f|^p + |g|^p \right) < \infty.$$

By Hölder's Inequality (5.2),

$$\|f + g\|_p^p = \int |f + g|^p \mathrm{d}\mu$$

$$= \int |f + g| \cdot |f + g|^{p-1} \mathrm{d}\mu$$

$$\leq \int (|f| + |g|)|f + g|^{p-1} \mathrm{d}\mu$$

$$= \int |f||f + g|^{p-1} \mathrm{d}\mu + \int |g||f + g|^{p-1} \mathrm{d}\mu$$

$$\leq \left(\left(\int |f|^p \mathrm{d}\mu \right)^{\frac{1}{p}} + \left(\int |g|^p \mathrm{d}\mu \right)^{\frac{1}{p}} \right) \left(\int |f + g|^{(p-1)\left(\frac{p}{p-1}\right)} \mathrm{d}\mu \right)^{1-\frac{1}{p}}$$

$$= \left(\|f\|_p + \|g\|_p \right) \frac{\|f + g\|_p^p}{\|f + g\|_p}$$

which means, as $p \in [1, \infty)$,

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p.$$

When $p = \infty$,

$$a$$

> **Theorem 5.4. Bounded Convergence Theorem**
>
> ♡

> **Theorem 5.5. Fatou's Lemma**
>
> ♡

> **Theorem 5.6. Monotone Convergence Theorem**
>
> ♡

5.2 Product Measures

> **Theorem 5.7. Fubini's Theorem**
>
> ♡

Part III

Functional Analysis

Part IV

Probability Theory

Chapter 6 Random Variables

Introduction

❏ *Probability Space*   ❏ *Expected Value*

❏ *Random Variables*   ❏ *Independence*

❏ *Distributions*   ❏ *Characteristic Functions*

6.1 Probability Space

> **Definition 6.1. Probability Space**
>
> *A probability space is a triple $(\Omega, \mathcal{F}, P)$ consisting of:*
>
> 1. *the sample space Ω: an arbitrary non-empty set.*
> 2. *the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$: a set of subsets of Ω, called events.*
> 3. *the probability measure $P : \mathcal{F} \rightarrow [0,1]$: a function on $\mathcal{F}$ which is a measure function.*
>
> ♣

6.2 Random Variables

> **Definition 6.2. Random Variable**
>
> *A random variable is a measurable function $X : \Omega \rightarrow S$ from a set of possible outcomes $(\Omega, \mathcal{F})$ to a measurable space $(S, \mathcal{S})$, that is,*
>
> $$X^{-1}(B) \equiv \{\omega : X(\omega) \in B\} \in \mathcal{F} \quad \forall B \in \mathcal{S}. \tag{6.1}$$
>
> *Typically, $(S, \mathcal{S}) = (R^d, \mathcal{R}^d) \quad (d > 1)$.*
>
> ♣

How to prove that functions are measurable?

> **Theorem 6.1. I**
>
> *$\{\omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$ and $\mathcal{A}$ generates $\mathcal{S}$, then X is measurable.*
>
> ♡

1.

6.3 Distributions

6.3.1 Definition of Distributions

> **Definition 6.3. Distribution**
>
> A distribution of random variable X is a probability function $P : \mathcal{R} \rightarrow \mathbb{R}$ by setting
>
> $$\mu(A) = P(X \in A) = P\left(X^{-1}(A)\right), \quad for A \in \mathcal{R}. \tag{6.2}$$

> **Definition 6.4. Distribution Function**
>
> The distribution of a random variable X is usually described by giving its **distribution function**,
>
> $$F(x) = P(X \leq x). \tag{6.3}$$

> **Definition 6.5. Density Function**
>
> If the distribution function $F(x) = P(X \leq x)$ has the form
>
> $$F(x) = \int_{-\infty}^{x} f(y) \mathrm{d}y,$$
>
> that X has density function f.

6.3.2 Properties of Distributions

> **Theorem 6.2. Properties of Distribution Function**
>
> Any distribution function F has the following properties,
>
> 1. F is nondecreasing.
> 2. $\lim_{x \to \infty} F(x) = 1, \lim_{x \to -\infty} F(x) = 0$.
> 3. F is right continuous, i.e., $\lim_{y \downarrow x} F(y) = F(x)$.
> 4. If $F(x-) = \lim_{y \uparrow x} F(y)$, then $F(x-) = P(X < x)$.
> 5. $P(X = x) = F(x) - F(x-)$.

Proof

> **Theorem 6.3**
>
> If F satisfies (1), (2), and (3) in Theorem 6.2, then it is the distribution function of some random variable.

Proof

> **Theorem 6.4**
>
> *A distribution function has at most countably many discontinuities* ♡

Proof

6.3.3 Families of Distributions

6.4 Expected Value

> **Definition 6.6. Expectation** ♣

> **Theorem 6.5. Bounded Convergence Theorem** ♡

> **Theorem 6.6. Fatou's Lemma**
>
> *If $X_n \geq 0$, then*
>
> $$\liminf_{n\to\infty} E X_n \geq E\left(\liminf_{n\to\infty} X_n\right). \tag{6.4}$$
> ♡

> **Theorem 6.7. Monotone Convergence Theorem**
>
> *If $0 \leq X_n \uparrow X$, then*
>
> $$E X_n \uparrow E X. \tag{6.5}$$
> ♡

> **Theorem 6.8. Dominated Convergence Theorem**
>
> *If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n, and $EY < \infty$, then*
>
> $$E X_n \to E X. \tag{6.6}$$
> ♡

6.5 Independence

6.5.1 Definition of Independence

> **Definition 6.7. Independence**
>
> 1. *Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.*
> 2. *Two random variables X and Y are independent if for all $C, D \in \mathcal{R}$*
>
> $$P(X \in C, Y \in D) = P(X \in C)P(Y \in D). \qquad (6.7)$$
>
> 3. *Two σ-fields $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$ the events A and B are independent.* ♣

The second definition is a special case of the third.

> **Theorem 6.9**
>
> 1. *If X and Y are independent then $\sigma(X)$ and $\sigma(Y)$ are independent.*
> 2. *Conversely, if $\mathcal{F}$ and $\mathcal{G}$ are independent, $X \in \mathcal{F}$ and $Y \in \mathcal{G}$, then X and Y are independent.* ♡

The first definition is, in turn, a special case of the second.

> **Theorem 6.10**
>
> 1. *If A and B are independent, then so are A^c and B, A and B^c, and A^c and B^c.*
> 2. *Conversely, events A and B are independent if and only if their indicator random variables 1_A and 1_B are independent.* ♡

The definition of independence can be extended to the infinite collection.

> **Definition 6.8**
>
> *An infinite collection of objects (σ-fields, random variables, or sets) is said to be independent if every finite subcollection is,*
>
> 1. *σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ are independent if whenever $A_i \in \mathcal{F}_i$ for $i = 1, \ldots, n$, we have*
>
> $$P\left(\cap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i). \qquad (6.8)$$
>
> 2. *Random variables $X_1, \ldots, X_n$ are independent if whenever $B_i \in \mathcal{R}$ for $i = 1, \ldots, n$*

we have

$$P \left(\cap_{i=1}^{n} \{X_i \in B_i\} \right) = \prod_{i=1}^{n} P \left(X_i \in B_i \right). \tag{6.9}$$

3. Sets $A_1, \ldots, A_n$ are independent if whenever $I \subset \{1, \ldots, n\}$ we have

$$P \left(\cap_{i \in I} A_i \right) = \prod_{i \in I} P \left(A_i \right). \tag{6.10}$$

♣

6.5.2 Sufficient Conditions for Independence

6.5.3 Independence, Distribution, and Expectation

Theorem 6.11

Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has distribution μ_i, then $(X_1, \ldots, X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.

♡

Theorem 6.12

If $X_1, \ldots, X_n$ are independent and have

1. $X_i \geq 0$ for all i, or
2. $E |X_i| < \infty$ for all i.

then

$$E \left(\prod_{i=1}^{n} X_i \right) = \prod_{i=1}^{n} E X_i \tag{6.11}$$

♡

6.5.4 Sums of Independent Random Variables

Theorem 6.13. Convolution for Random Variables

1. If X and Y are independent, $F(x) = P(X \leq x)$, and $G(y) = P(Y \leq y)$, then

$$P(X + Y \leq z) = \int F(z - y) \mathrm{d}G(y). \tag{6.12}$$

2. If X and Y are independent, X with density f and Y with distribution function G, then $X + Y$ has density

$$h(x) = \int f(x - y) \mathrm{d}G(y). \tag{6.13}$$

Suppose Y has density g, the last formula can be written as

$$h(x) = \int f(x - y)g(y)\mathrm{d}y. \tag{6.14}$$

3. If X and Y are independent, integral-valued random variables, then

$$P(X + Y = n) = \sum_m P(X = m)P(Y = n - m). \tag{6.15}$$

6.6 Moments

Lemma 6.1

If $Y > 0$ and $p > 0$, then

$$E(Y^p) = \int_0^\infty py^{p-1}P(Y > y)\mathrm{d}y. \tag{6.16}$$

6.7 Characteristic Functions

6.7.1 Definition of Characteristic Functions

Definition 6.9. Characteristic Function

If X is a random variable, we define its characteristic function (ch.f) by

$$\varphi(t) = E\left(e^{itX}\right) = E(\cos tX) + iE(\sin tX). \tag{6.17}$$

Note *Eular Equation.*

6.7.2 Properties of Characteristic Functions

Theorem 6.14. Properties of Characteristic Function

Any characteristic function has the following properties:

1. $\varphi(0) = 1$,
2. $\varphi(-t) = \overline{\varphi(t)}$,
3. $|\varphi(t)| = |Ee^{itX}| \leq E|e^{itX}| = 1$,
4. $\varphi(t)$ is uniformly continuous on $(-\infty, \infty)$,

5. $Ee^{it(aX+b)} = e^{itb}\varphi(at)$,

6. If X_1 and X_2 are independent and have ch.f.'s φ_1 and φ_2, then $X_1 + X_2$ has ch.f. $\varphi_1(t)\varphi_2(t)$.

Proof

6.7.3 The Inversion Formula

The characteristic function uniquely determines the distribution. This and more is provided by:

> **Theorem 6.15. The Inversion Formula**
>
> Let $\varphi(t) = \int e^{itx}\mu(\mathrm{d}x)$ where μ is a probability measure. If $a < b$, then
> $$\lim_{T\to\infty}(2\pi)^{-1}\int_{-T}^{T}\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)dt = \mu(a,b) + \frac{1}{2}\mu(\{a,b\}) \qquad (6.18)$$

Proof

> **Theorem 6.16**
>
> If $\int |\varphi(t)|\mathrm{d}t < \infty$, then μ has bounded continuous density
> $$f(y) = \frac{1}{2\pi}\int e^{-ity}\varphi(t)\mathrm{d}t. \qquad (6.19)$$

Proof

6.7.4 Convergence in Distribution

> **Theorem 6.17. Lèvy's Continuity Theorem**
>
> Let $\mu_n, 1 \le n \le \infty$ be probability measures with ch.f. φ_n.
>
> 1. If $\mu_n \xrightarrow{d} \mu_\infty$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all t.
>
> 2. If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at 0, then the associated sequence of distributions μ_n is tight and converges weakly to the measure μ with characteristic function φ.

Proof

6.7.5 Moments and Derivatives

> **Theorem 6.18**
>
> *If $\int |x|^n \mu(dx) < \infty$, then its characteristic function φ has a continuous derivative of order n given by*
> $$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(dx). \qquad (6.20)$$

> **Theorem 6.19**
>
> *If $E|X|^2 < \infty$ then*
> $$\varphi(t) = 1 + itEX - t^2 E\left(X^2\right)/2 + o\left(t^2\right). \qquad (6.21)$$

> **Theorem 6.20**
>
> *If $\limsup_{h \downarrow 0} \{\varphi(h) - 2\varphi(0) + \varphi(-h)\}/h^2 > -\infty$, then*
> $$E|X|^2 < \infty. \qquad (6.22)$$

Chapter 7 Convergence of Random Variables

> **Introduction**
>
> ❏ *Convergence in Mean*   ❏ *Convergence in Distribution*
>
> ❏ *Convergence in Probability*   ❏ *Almost Sure Convergence*

7.1 Convergence in Mean

> **Definition 7.1. Convergence in Mean**
>
> *A sequence $\{X_n\}$ of real-valued random variables **converges in the r-th mean** $(r \geq 1)$ towards the random variable X, if*
>
> 1. *The r-th absolute moments $E(|X_n|^r)$ and $E(|X|^r)$ of $\{X_n\}$ and X exist,*
> 2. $\lim_{n \to \infty} E\left(|X_n - X|^r\right) = 0.$
>
> *Convergence in the r-th mean is denoted by*
>
> $$X_n \xrightarrow{L^r} X. \tag{7.1}$$ ♣

7.2 Convergence in Probability

7.2.1 Definition of Convergence in Probability

> **Definition 7.2. Convergence in Probability**
>
> *A sequence $\{X_n\}$ of real-valued random variables **converges in probability** towards the random variable X, if*
>
> $$\forall \varepsilon > 0, \quad \lim_{n \to \infty} P\left(|X_n - X| > \varepsilon\right) = 0. \tag{7.2}$$
>
> *Convergence in probability is denoted by*
>
> $$X_n \xrightarrow{p} X. \tag{7.3}$$ ♣

> **Definition 7.3. Convergence in Uninform**
>
> ♣

7.2.2 Properties of Convergence in Probability

7.3 Convergence in Distribution

7.3.1 Definition of Convergence in Distribution

> **Definition 7.4. Convergence in Distribution**
>
> *A sequence $\{X_n\}$ of real-valued random variables is said to **converge in distribution**, or **converge weakly**, or **converge in law** to a random variable X, if*
>
> $$\lim_{n \to \infty} F_n(x) = F(x), \tag{7.4}$$
>
> *for every number at $x \in \mathbb{R}$ which F is continuous. Here F_n and F are the cumulative distribution functions of random variables X_n and X, respectively. Convergence in distribution can be denoted as*
>
> $$X_n \xrightarrow{d} X, \text{ or } X_n \Rightarrow X. \tag{7.5}$$ ♣

Note

- *Convergence in Distribution is the weakest form of convergence typically discussed, since it is implied by all other types of convergence mentioned in this chapter.*
- *Convergence in Distribution does not imply that the sequence of corresponding probability density functions will also converge. However, according to Scheffé's theorem, convergence of the probability density functions implies convergence in distribution.*

7.3.2 Properties of Convergence in Distribution

> **Lemma 7.1**
>
> *If $F_n \xrightarrow{d} F_\infty$, then there are random variables $Y_n, 1 \leq n \leq \infty$, with distribution F_n so that*
>
> $$Y_n \xrightarrow{a.s.} Y_\infty. \tag{7.6}$$ ♡

> **Theorem 7.1. Portmanteau Lemma**
>
> *$\{X_n\}$ converges in distribution to X, if and only if any of the following statements are true,*
>
> - *$P(X_n \leq x) \to P(X \leq x)$, for all continuity points of the distribution of X.*

- $Ef(X_n) \to Ef(X)$, *for all bounded, continuous (Lipschitz) functions* f.
- $\liminf_{n\to\infty} P(X_n \in G) \geq P(X_\infty \in G)$, *for all open sets* G.
- $\limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K)$, *for all closed sets* K.
- $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$, *for all Borel sets* A *with* $P(X_\infty \in \partial A) = 0$.

Proof

Theorem 7.2. Continuous Mapping Theorem

Let g be a measurable function and $D_g = \{x : g$ is discontinuous at $x\}$ with $P(X \in D_g) = 0$, then,

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X),$$
$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X), \tag{7.7}$$
$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X).$$

If in addition g is bounded, then

$$Eg(X_n) \to Eg(X). \tag{7.8}$$

Proof

Theorem 7.3

If $X_n \xrightarrow{p} X$, then

$$X_n \xrightarrow{d} X, \tag{7.9}$$

and that, conversely, if $X_n \xrightarrow{d} c$, where c is a constant, then

$$X_n \xrightarrow{p} c. \tag{7.10}$$

Proof

1. $\forall \varepsilon > 0$, at fixed point x, since if $X_n \leq x$ and $|X_n - X| \leq \varepsilon$, then $X \leq x + \varepsilon$, then

$$\{X \leq x + \varepsilon\} \subset \{X_n \leq x\} \cup \{|X_n - X| > \varepsilon\},$$

similarily, if $X \leq x - \varepsilon$ and $|X_n - X| \leq \varepsilon$, then $X_n \leq x$, then

$$\{X_n \leq x\} \subset \{X \leq x - \varepsilon\} \cup \{|X_n - X| > \varepsilon\},$$

then, by the union bound,

$$P(X \leq x + \varepsilon) \leq P(X_n \leq x) + P(|X_n - X| > \varepsilon),$$
$$P(X_n \leq x) \leq P(X \leq x - \varepsilon) + P(|X_n - X| > \varepsilon).$$

So, we got

$$P\left(X \leq x + \varepsilon\right) - P\left(|X_n - X| > \varepsilon\right) \leq P\left(X_n \leq x\right)$$

$$\leq P\left(X \leq x - \varepsilon\right) + P\left(|X_n - X| > \varepsilon\right)$$

As $n \to \infty$, $P\left(|X_n - X| > \varepsilon\right) \to 0$, then

$$P\left(X \leq x - \varepsilon\right) \leq \lim_{n \to \infty} P\left(X_n \leq x\right) \leq P\left(X \leq x + \varepsilon\right)$$

$$\Rightarrow F(x - \varepsilon) \leq \lim_{n \to \infty} F_n(x) \leq F(x + \varepsilon)$$

.

By the property of distribution (Theorem 6.2), as $\varepsilon \to 0$, then

$$\lim_{n \to \infty} F_n(x) = F(x),$$

which means,

$$X_n \xrightarrow{d} X.$$

2. Since $X_n \xrightarrow{d} c$, where c is a constant, then $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P(X_n \leq c + \varepsilon) = 1 \Rightarrow \lim_{n \to \infty} P(X_n > c + \varepsilon) = 0$$

$$\lim_{n \to \infty} P(X_n \leq c - \varepsilon) = 0.$$

Therefore,

$$P\left(|X_n - c| < \varepsilon\right) = 0,$$

which means

$$X_n \xrightarrow{p} c.$$

> **Theorem 7.4. Slutsky's Theorem**
>
> *Let X_n, Y_n be sequences of random variables. If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then*
> 1. *$X_n + Y_n \xrightarrow{d} X + c$.*
> 2. *$X_n Y_n \xrightarrow{d} cX$.*
> 3. *$X_n / Y_n \xrightarrow{d} X/c$, provided that c is invertible.*

Proof

1. Since

2.

3.

Note *However that convergence in distribution of $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ does in general not imply convergence in distribution of $X_n + Y_n \xrightarrow{d} X + Y$ or of $X_n Y_n \xrightarrow{d} XY$.*

Theorem 7.5. Cramér-Wold Theorem

Theorem 7.6. Delta Method

Let $\{X_n\}$ be a sequence of random variables with

$$\sqrt{n}\,[X_n - \theta] \xrightarrow{d} \sigma\chi,$$

where θ and σ are finite, then for any function g with the property that $g'(\theta)$ exists and is non-zero valued,

$$\sqrt{n}\,[g\,(X_n) - g(\theta)] \xrightarrow{d} \sigma g'(\theta)\chi.$$

Proof   Under the assumption that $g'(\theta)$ is continuous.

Since, $g'(\theta)$ exists, with the first-order Taylor Approximation:

$$g(X_n) = g(\theta) + g'(\tilde{\theta})(X_n - \theta),$$

where $\tilde{\theta}$ lies between X_n and θ.

Since $X_n \xrightarrow{p} \theta$, and $|\tilde{\theta} - \theta| < |X_n - \theta|$, then

$$\tilde{\theta} \xrightarrow{p} \theta,$$

Since $g'(\theta)$ is continuous, by Continuous Mapping Theorem (7.2),

$$g'(\tilde{\theta}) \xrightarrow{p} g'(\theta).$$

and,

$$\sqrt{n}\,(g(X_n) - g(\theta)) = \sqrt{n}g'(\tilde{\theta})(X_n - \theta),$$

$$\sqrt{n}\,[X_n - \theta] \xrightarrow{d} \sigma\chi,$$

by Slutsky's Theorem (7.4),

$$\sqrt{n}\,[g\,(X_n) - g(\theta)] \xrightarrow{d} \sigma g'(\theta)\chi.$$

7.3.3 Limits of Sequences of Distributions $\{F_n\}$

Theorem 7.7. Helly's Selection Theorem

For every sequence F_n of distribution functions, there is a subsequence $F_{n}(k)$ and a right continuous nondecreasing function F so that $\lim_{k \to \infty} F_{n(k)}(y) = F(y)$ at all continuity points y of F.

Theorem 7.8

Every subsequential limit is the distribution function of a probability measure if and only if the sequence F_n is tight, i.e., for all $\epsilon > 0$ there is an M_ϵ so that

$$\limsup_{n \to \infty} 1 - F_n(M_\epsilon) + F_n(-M_\epsilon) \leq \epsilon. \tag{7.11}$$

7.4 Almost Sure Convergence

7.4.1 Definition of Almost Sure Convergence

Definition 7.5. Almost Sure Convergence

*A sequence $\{X_n\}$ of real-valued random variables converges **almost sure** or **almost everywhere** or **with probability 1** or **strongly** towards the random variable X, if*

$$P\left(\lim_{n \to \infty} X_n = X\right) = 1. \tag{7.12}$$

Almost sure convergence is denoted by

$$X_n \overset{a.s.}{\to} X. \tag{7.13}$$

Note

7.4.2 Properties of Almost Sure Convergence

Theorem 7.9

If $X_n \overset{a.s.}{\to} X$, then

$$X_n \overset{p}{\to} X. \tag{7.14}$$

Proof

Theorem 7.10

$X_n \overset{p}{\to} X$ if and only if for all subsequence $X_{n(m)}$ exists a further subsequence $X_{n(m_k)}$, such that

$$X_{n(m_k)} \overset{a.s.}{\to} X. \tag{7.15}$$

① $if\ r \geq s \geq 1, X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{L^s} X$

② $X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{p} X$

③ $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$

④ $if\ X = c, X_n \xrightarrow{d} X \Rightarrow X_n \xrightarrow{p} X$

⑤ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$

⑥ $X_n \xrightarrow{p} X \Rightarrow X_{n(m_k)} \xrightarrow{a.s.} X$

⑦ $X_n \xrightarrow{d} X \Rightarrow \exists Y_n \overset{d}{=} X_n, Y \overset{d}{=} X, s.t.\ Y_n \xrightarrow{a.s.} Y$

⑧ $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{d} X$

$X_n \xrightarrow{L^r} X \xrightarrow{①} X_n \xrightarrow{L^s} X$

$X_n \xrightarrow{p} X$

$X_n \xrightarrow{d} X \xrightarrow{⑦} X_n \xrightarrow{a.s.} X$
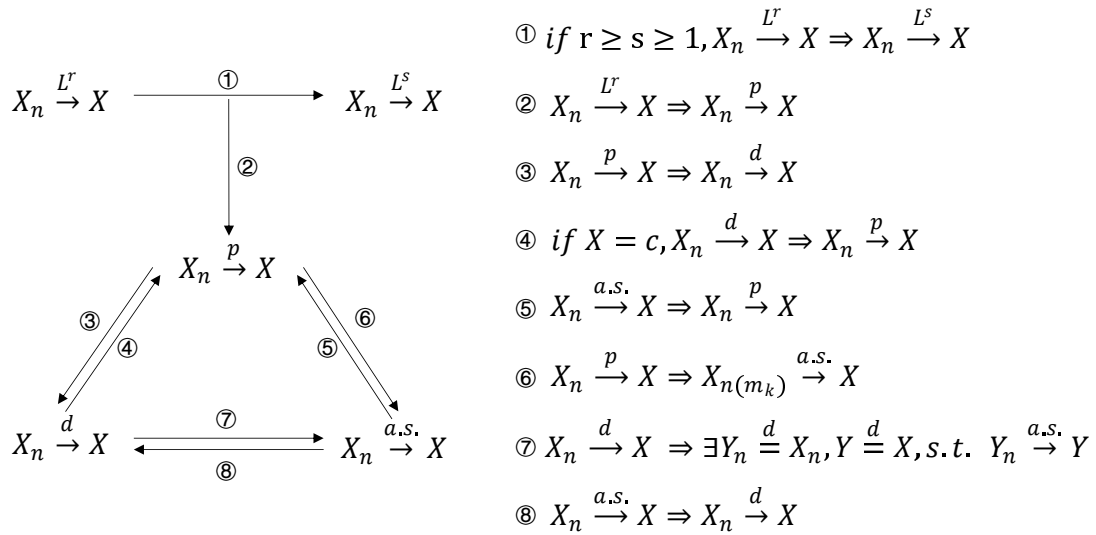
Figure 7.1: Relations of Convergence of Random Variables

7.5 Asymptotic Notation for Random Variables

Definition 7.6. Little o_p

A sequence $\{A_n\}$ of real-valued random variables is of smaller order in probability than a sequence $\{B_n\}$, if

$$\frac{A_n}{B_n} \xrightarrow{p} 0. \tag{7.16}$$

Smaller order in probability is denoted by

$$A_n = o_p(B_n). \tag{7.17}$$

Particularly,

$$A_n = o_p(1) \iff A_n \xrightarrow{p} 0. \tag{7.18}$$

♣

Definition 7.7. Big O_p

A sequence $\{A_n\}$ of real-valued random variables is of smaller order than or equal to a sequence $\{B_n\}$ in probability, if

$$\forall \varepsilon > 0\ \exists M_\varepsilon, \quad \lim_{n \to \infty} P\left(|A_n| \leq M_\varepsilon |B_n|\right) \geq 1 - \varepsilon. \tag{7.19}$$

Smaller order than or equal to in probability is denoted by

$$A_n = O_p(B_n). \tag{7.20}$$

♣

🍷 **Note**

Chapter 8 Law of Large Numbers

Introduction

❏ *Weak Law of Large Numbers*　　　　❏ *Uniform Law of Large Numbers*

❏ *Strong Law of Large Numbers*

8.1 Weak Law of Large Numbers

Lemma 8.1

If $p > 0$ and $E\,|Z_n|^p \to 0$, then

$$Z_n \xrightarrow{d} 0. \tag{8.1}$$

Proof

Theorem 8.1. Weak Law of Large Numbers with Finite Variances

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E X_i = \mu$ and $Var(X_i) \le C < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$, then

$$S_n/n \xrightarrow{L^2} \mu, \quad S_n/n \xrightarrow{p} \mu. \tag{8.2}$$

Proof

Theorem 8.2. Weak Law of Large Numbers without i.i.d.

Let $X_1, X_2, \ldots$ be random variables, Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu_n = E S_n$, $\sigma_n^2 = Var(S_n)$, if $\sigma_n^2/b_n^2 \to 0$, then

$$\frac{S_n - \mu_n}{b_n} \xrightarrow{p} 0. \tag{8.3}$$

Proof

Theorem 8.3. Weak Law of Large Numbers for Triangular Arrays

For each n, let $X_{n,m}, 1 \le m \le n$, be independent random variables. Suppose $b_n > 0$ with $b_n \to \infty$, $\bar{X}_{n,m} = X_{n,m} I_{\left(X_{n,m} \le b_n\right)}$, if

1. $\sum_{m=1}^{n} P\left(\left|X_{n,m}\right| > b_n\right) \to 0$, and
2. $b_n^{-2} \sum_{m=1}^{n} E \bar{X}_{n,m}^2 \to 0$.

Suppose $S_n = X_{n,1} + \cdots + X_{n,n}$ and $a_n = \sum_{m=1}^{n} E \bar{X}_{n,m}$, then

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0. \tag{8.4}$$

Proof

Theorem 8.4. Weak Law of Large Numbers by Feller

Let $X_1, X_2, \ldots$ be i.i.d. random variables with

$$\lim_{x \to 0} x P(|X_i| > x) = 0. \tag{8.5}$$

Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu_n = E\left(X_1 I_{(|X_1| < n)}\right)$, then

$$S_n/n - \mu_n \xrightarrow{p} 0. \tag{8.6}$$

Proof

Theorem 8.5. Weak Law of Large Numbers

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$,
$\mu = EX_i$, then

$$S_n/n \xrightarrow{p} \mu. \tag{8.7}$$

Proof

Note $E|X_i| = \infty$

8.2 Strong Law of Large Numbers

8.2.1 Borel-Cantelli Lemmas

Lemma 8.2. Borel-Cantelli Lemma

If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P(A_n \text{ i.o. }) = 0. \tag{8.8}$$

Lemma 8.3. The Second Borel-Cantelli Lemma

If $\{A_n\}$ are independent with $\sum_{n=1}^{\infty} P(A_n) = \infty$, then,

$$P(A_n \text{ i.o. }) = 1. \tag{8.9}$$

> **Corollary 8.1**
>
> Suppose $\{A_n\}$ are independent with $P(A_n) < 1, \forall n$. If $P\left(\cup_{n=1}^{\infty} A_n\right) = 1$ then
>
> $$\sum_{n=1}^{\infty} P(A_n) = \infty, \tag{8.10}$$
>
> and hence $P(A_n \text{ i.o. }) = 1$

Proof

8.2.2 Strong Law of Large Numbers

> **Theorem 8.6. Strong Law of Large Numbers**
>
> Let $X_1, X_2, \ldots$ be i.i.d. random variables with $E|X_i| < \infty$. Suppose $S_n = X_1 + X_2 + \ldots + X_n$, $\mu = EX_i$, then
>
> $$S_n/n \overset{a.s.}{\to} \mu. \tag{8.11}$$

8.3 Uniform Law of Large Numbers

> **Theorem 8.7. Uniform Law of Large Numbers**
>
> Suppose
>
> 1. Θ is compact.
> 2. $g(X_i, \theta)$ is continuous at each $\theta \in \Theta$ almost sure.
> 3. $g(X_i, \theta)$ is dominated by a function $G(X_i)$, i.e. $|g(X_i, \theta)| \leq G(X_i)$.
> 4. $EG(X_i) < \infty$.
>
> Then
>
> $$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta) \right| \overset{p}{\to} 0. \tag{8.12}$$

Proof Suppose

$$\Delta_\delta(X_i, \theta_0) = \sup_{\theta \in B(\theta_0, \delta)} g(X_i, \theta) - \inf_{\theta \in B(\theta_0, \delta)} g(X_i, \theta).$$

Since (i) $\Delta_\delta(X_i, \theta_0) \overset{a.s.}{\to} 0$ by condition (2), (ii) $\Delta_\delta(X_i, \theta_0) \leq 2 \sup_{\theta \in \Theta} |g(X_i, \theta)| \leq 2G(X_i)$ by condition (3) and (4). Then

$$E\Delta_\delta(X_i, \theta_0) \to 0, \text{ as } \delta \to 0.$$

So, for all $\theta \in \Theta$ and $\varepsilon > 0$, there exists $\delta_\varepsilon(\theta)$ such that

$$E\left[\Delta_{\delta_\varepsilon(\theta)}(X_i, \theta)\right] < \varepsilon.$$

Since Θ is compact, we can find a finite subcover, such that Θ is covered by

$$\cup_{k=1}^{K} B\left(\theta_k, \delta_\varepsilon(\theta_k)\right).$$

$$\sup_{\theta \in \Theta}\left[n^{-1}\sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta)\right]$$

$$= \max_k \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))}\left[n^{-1}\sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta)\right].$$

$$\leq \max_k\left[n^{-1}\sum_{i=1}^{n} \sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E\inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta)\right]$$

Since

$$E\left|\sup_{\theta \in B(\theta_k, \delta_c(\theta_k))} g(X_i, \theta)\right| \leq EG(X_i) < \infty,$$

by the Weak Law of Large Numbers (Theorem 8.5),

$$= o_p(1) + \max_k\left[E\sup_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta) - E\inf_{\theta \in B(\theta_k, \delta_\varepsilon(\theta_k))} g(X_i, \theta)\right]$$

$$= o_p(1) + \max_k E\Delta_{\delta_\varepsilon(\theta_k)}(X_i, \theta_k)$$

$$\leq o_p(1) + \varepsilon$$

By analogous argument,

$$\inf_{\theta \in \Theta}\left[n^{-1}\sum_{i=1}^{n} g(X_i, \theta) - Eg(X_i, \theta)\right] \geq o_p(1) - \varepsilon.$$

The desired result follows from the above equation by the fact that ε is chosen arbitrarily.

Chapter 9 Central Limit Theorems

9.1 The De Moivre-Laplace Theorem

Lemma 9.1. Stirling's Formula

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \text{ as } n \to \infty. \tag{9.1}$$

Proof

Lemma 9.2

If $c_j \to 0$, $a_j \to \infty$ and $a_j c_j \to \lambda$, then

$$\left(1 + c_j\right)^{a_j} \to e^{\lambda}. \tag{9.2}$$

Proof

Theorem 9.1. The De Moivre-Laplace Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and let $S_n = X_1 + \cdots + X_n$. If $a < b$, then as $m \to \infty$

$$P\left(a \le S_m/\sqrt{m} \le b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \mathrm{d}x. \tag{9.3}$$

Proof If n and k and integers

$$P(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n}$$

By lemma 9.1, we have

$$\binom{2n}{n+k} = \frac{(2n)!}{(n+k)!(n-k)!}$$

$$\sim \frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \cdot \frac{(2\pi(2n))^{1/2}}{(2\pi(n+k))^{1/2}(2\pi(n-k))^{1/2}}$$

Hence,

$$P(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n}$$

$$\sim \left(1 + \frac{k}{n}\right)^{-n-k} \cdot \left(1 - \frac{k}{n}\right)^{-n+k}$$

$$\cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2}$$

$$= \left(1 - \frac{k^2}{n^2}\right)^{-n} \cdot \left(1 + \frac{k}{n}\right)^{-k} \cdot \left(1 - \frac{k}{n}\right)^{k}$$

$$\cdot (\pi n)^{-1/2} \cdot \left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2}$$

Let $2k = x\sqrt{2n}$, i.e., $k = x\sqrt{\frac{n}{2}}$. By lemma 9.2, we have

$$\left(1 - \frac{k^2}{n^2}\right)^{-n} = \left(1 - x^2/2n\right)^{-n} \to e^{x^2/2}$$

$$\left(1 + \frac{k}{n}\right)^{-k} = (1 + x/\sqrt{2n})^{-x\sqrt{n/2}} \to e^{-x^2/2}$$

$$\left(1 - \frac{k}{n}\right)^{k} = (1 - x/\sqrt{2n})^{x\sqrt{n/2}} \to e^{-x^2/2}$$

For this choice of k, $k/n \to 0$, so

$$\left(1 + \frac{k}{n}\right)^{-1/2} \cdot \left(1 - \frac{k}{n}\right)^{-1/2} \to 1.$$

Putting things together, we have

$$P(S_{2n} = 2k) \sim (\pi n)^{-1/2} e^{-x^2/2}, \text{ as } \frac{2k}{\sqrt{2n}} \to x.$$

Therefore,

$$P\left(a\sqrt{2n} \le S_{2n} \le b\sqrt{2n}\right) = \sum_{m \in \left[a\sqrt{2n}, b\sqrt{2n}\right] \cap 2\mathbb{Z}} P(S_{2n} = m)$$

Let $m = x\sqrt{2n}$, we have that this is

$$\approx \sum_{x \in [a,b] \cap \left(2\mathbb{Z}/\sqrt{2n}\right)} (2\pi)^{-1/2} e^{-x^2/2} \cdot (2/n)^{1/2}$$

where $2\mathbb{Z}/\sqrt{2n} = \left\{2z/\sqrt{2n} : z \in \mathbb{Z}\right\}$. As $n \to \infty$, the sum just shown is

$$\approx \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \mathrm{d}x.$$

To remove the restriction to even integers, observe $S_{2n+1} = S_{2n} \pm 1$.

Let $m = 2n$, as $m \to \infty$,

$$P\left(a \leq S_m/\sqrt{m} \leq b\right) \to \int_a^b (2\pi)^{-1/2} e^{-x^2/2} \mathrm{d}x.$$

9.2 Central Limit Theorems

Theorem 9.2. Central Limit Theorem for i.i.d. Sequences

Let $X_1, X_2, \ldots$ be i.i.d. with $E X_i = \mu$, $Var(X_i) = \sigma^2 \in (0, \infty)$. Let $S_n = X_1 + X_2 + \ldots + X_n$,

then

$$\frac{S_n - n\mu}{\sigma n^{\frac{1}{2}}} \xrightarrow{d} \chi, \tag{9.4}$$

where χ has the standard normal distribution.

Proof

Theorem 9.3. The Linderberg-Feller Central Limit Theorem

For each n, let $X_{n,m}, 1 \leq m \leq n$, be independent random variables with $E X_{n,m} = 0$. If

1. $\sum_{m=1}^n E X_{n,m}^2 \to \sigma^2 > 0$.

2. $\forall \epsilon > 0, \lim_{n \to \infty} \sum_{m=1}^n E\left(\left|X_{n,m}\right|^2 ; \left|X_{n,m}\right| > \epsilon\right) = 0$

Then $S_n = X_{n,1} + \cdots + X_{n,n} \xrightarrow{d} \sigma\chi$ as $n \to \infty$.

Proof

Chapter 10 Exercises for Probability Theory and Examples

10.1 Measure Theory

✎ **Exercise 10.1**

1. Show that if $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ are σ-algebras, then $\cup_i \mathcal{F}_i$ is an algebra.
2. Give an example to show that $\cup_i \mathcal{F}_i$ need not be a σ-algebra.

Solution

1. ***Complement***: *Suppose $A \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A \in \mathcal{F}_i$. And each $\mathcal{F}_i$ is σ-algebra,*

$$A^c \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

   Finite Union: *Suppose $A_1, A_2 \in \cup_i \mathcal{F}_i$, since $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, assume $A_1 \in \mathcal{F}_i, A_2 \in \mathcal{F}_j$, such that,*

$$A_1, A_2 \in \mathcal{F}_{\max(i,j)}.$$

   Since each $\mathcal{F}_i$ is σ-algebra,

$$A_1 \cup A_2 \in \mathcal{F}_i \subset \cup_i \mathcal{F}_i.$$

2. *Let $\mathcal{F}_i$ be a Borel Set of $[1, 2 - \frac{1}{i}]$. Suppose $A_i = [1, 2 - \frac{1}{i}] \in \mathcal{F}_i$,*

$$\cup_i A_i = [1, 2) \notin \cup_i \mathcal{F}_i.$$

10.2 Laws of Large Numbers

10.3 Central Limit Theorems

✎ **Exercise 10.2** Let $g \geq 0$ be continuous. If $X_n \overset{d}{\to} X_\infty$, then

$$\liminf_{n \to \infty} E g\left(X_n\right) \geq E g\left(X_\infty\right).$$

Solution *Let $Y_n \overset{d}{=} X_n, 1 \leq n \leq \infty$ with $Y_n \overset{a.s.}{\to} Y_\infty$ (Lemma 7.1). Since $g \geq 0$ be continuous,*

$g(Y_n) \overset{a.s.}{\to} g(Y_\infty)$ *and* $g(Y_n) \geq 0$ *(Theorem 7.2), and the Fatou's Lemma (6.6) implies,*

$$\liminf_{n\to\infty} Eg(X_n) = \liminf_{n\to\infty} Eg(Y_n) \geq E\left(\liminf_{n\to\infty} g(Y_n)\right)$$

$$= Eg(Y_\infty) = Eg(X_\infty).$$

✎ **Exercise 10.3** Suppose g, h are continuous with $g(x) > 0$, and $|h(x)|/g(x) \to 0$ as $|x| \to \infty$. If $F_n \overset{d}{\to} F$ and $\int g(x)\mathrm{d}F_n(x) \leq C < \infty$, then

$$\int h(x)\mathrm{d}F_n(x) \to \int h(x)\mathrm{d}F(x).$$

Solution

$$\left|\int h(x)\mathrm{d}F_n(x) - \int h(x)\mathrm{d}F(x)\right| = \left|\int_{x\in[-M,M]} h(x)\mathrm{d}F_n(x) + \int_{x\notin[-M,M]} h(x)\mathrm{d}F_n(x)\right.$$

$$\left. - \int_{x\in[-M,M]} h(x)\mathrm{d}F(x) - \int_{x\notin[-M,M]} h(x)\mathrm{d}F(x)\right|$$

$$\leq \left|\int_{x\in[-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x\in[-M,M]} h(x)\mathrm{d}F(x)\right|$$

$$+ \left|\int_{x\notin[-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x\notin[-M,M]} h(x)\mathrm{d}F(x)\right|.$$

Let $X_n, 1 \leq n < \infty$, with distribution F_n, so that $X_n \overset{a.s.}{\to} X$ (Lemma 7.1).

$$\left|\int_{x\in[-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x\in[-M,M]} h(x)\mathrm{d}F(x)\right| = \left|E\left(h(X_n) - h(X)\right)I_{x\in[-M,M]}\right|.$$

By Continuity Mapping Theorem (7.2), $\lim_{n\to\infty}\left|E\left(h(X_n) - h(X)\right)I_{x\in[-M,M]}\right| = 0$.

Since

$$h(x)I_{x\notin[-M,M]} \leq g(x)\sup_{x\notin[-M,M]}\frac{h(x)}{g(x)},$$

and by Exercise 10.2

$$Eg(X) \leq \liminf_{n\to\infty} Eg(X_n) = \liminf_{n\to\infty}\int g(x)\mathrm{d}F_n(x) \leq C < \infty,$$

$$\left|\int_{x\notin[-M,M]} h(x)\mathrm{d}F_n(x) - \int_{x\notin[-M,M]} h(x)\mathrm{d}F(x)\right| = \left|E\left(h(X_n) - h(X)\right)I_{x\notin[-M,M]}\right|$$

$$\leq 2E\max(h(Xn), h(X))I_{x\notin[-M,M]} \leq 2C\sup_{x\notin[-M,M]}\frac{h(x)}{g(x)}.$$

Hence, let $M \to \infty$,

$$\lim_{n\to\infty}\left|\int h(x)\mathrm{d}F_n(x) - \int h(x)\mathrm{d}F(x)\right| \leq 2C\sup_{x\notin[-M,M]}\frac{h(x)}{g(x)} \to 0,$$

which means,

$$\int h(x)\mathrm{d}F_n(x) \to \int h(x)\mathrm{d}F(x).$$

✎ **Exercise 10.4** Let $X_1, X_2, \ldots$ be i.i.d. with $EX_i = 0$ and $EX_i^2 = \sigma^2 \in (0, \infty)$. Then

$$\sum_{m=1}^{n} X_m \Big/ \left(\sum_{m=1}^{n} X_m^2 \right)^{1/2} \xrightarrow{d} \chi.$$

✎ **Exercise 10.5** Show that if $|X_i| \leq M$ and $\sum_n \text{Var}(X_n) = \infty$, then

$$(S_n - ES_n) \Big/ \sqrt{\text{Var}(S_n)} \xrightarrow{d} \chi.$$

✎ **Exercise 10.6** Suppose $EX_i = 0, EX_i^2 = 1$ and $E|X_i|^{2+\delta} \leq C$ for some $0 < \delta, C < \infty$. Show that

$$S_n / \sqrt{n} \xrightarrow{d} \chi.$$

Part V

Stochastic Process

Chapter 11   Martingales

Chapter 12 Exercises for Probability Theory and Examples

Part VI

Statistics Inference

Chapter 13 Introduction

13.1 Populations and Samples

13.2 Statistics

13.2.1 Sufficient Statistics

> **Definition 13.1. Sufficient Statistics**
>
> *A statistic T is said to be sufficient for X, or for the family $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$ of possible distributions of X, or for θ, if the conditional distribution of X given $T = t$ is independent of θ for all t.* ♣

> **Theorem 13.1. Fisher–Neyman Factorization Theorem**
>
> *If the probability density function is $p_\theta(x)$, then T is sufficient for θ if and only if nonnegative functions g and h can be found such that*
>
> $$p_\theta(x) = h(x)g_\theta[T(x)].$$
> ♡

Proof

13.2.2 Complete Statistics

> **Definition 13.2. Complete Statistics**
>
> *A statistic T is said to be complete, if $Eg(T) = 0$ for all θ and some function g implies that $P(g(T) = 0 \mid \theta) = 1$ for all θ.* ♣

13.3 Estimators

13.3.1 Definition of Estimators

> **Definition 13.3. Estimator**
>
> An estimator is a real-valued function defined over the sample space, that is
>
> $$\delta : X \to \mathbb{R}. \tag{13.1}$$
>
> It is used to estimate an estimand, θ, a real-valued function of the parameter.

13.3.2 Properties of Estimators

Unbiasedness

> **Definition 13.4. Unbiasedness**
>
> An estimator $\hat{\theta}$ of θ is unbiased if
>
> $$E\hat{\theta} = \theta, \quad \forall\theta \in \Theta. \tag{13.2}$$

Note

- *Unbiased estimators of θ may not exist.*
-

Example 13.1 Nonexistence of Unbiased Estimator

Consistency

> **Definition 13.5. Consistency**
>
> An estimator $\hat{\theta}_n$ of θ is consistent if
>
> $$\lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = 0, \quad \forall\varepsilon > 0. \tag{13.3}$$

Example 13.2 Unbiased But Consistent

Example 13.3 Biased But Not Consistent

Asymptotic Normality

> **Definition 13.6. Asymptotic Normality**
>
> *An estimator $\hat{\theta}_n$ of θ is asymptotic normality if*
> $$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \sigma_\theta^2\right). \tag{13.4}$$
>
> ♣

Efficiency

> **Definition 13.7. Efficiency**
>
> ♣

Robustness

> **Definition 13.8. Robustness**
>
> ♣

Chapter 14 Maximum Likelihood Estimator

Suppose that $\mathbf{X}_n = (X_1, \ldots, X_n)$, where the X_i are i.i.d. with common density $p(x; \theta_0) \in \mathcal{P} = \{p(x; \theta) : \theta \in \Theta\}$.

We assume that

> θ_0 is identified in the sense that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $p(x; \theta) \neq p(x; \theta_0)$
> with respect to the dominating measure μ.

For fixed $\theta \in \Theta$, the joint density of $\mathbf{X}_n$ is equal to the product of the individual densities, i.e.,

$$p(\mathbf{X}_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta). \tag{14.1}$$

The maximum likelihood estimate for observed $\mathbf{X}_n$ is the value $\theta \in \Theta$ which maximizes $L(\theta; X_n) := p(\mathbf{X}_n; \theta)$, i.e.,

$$\hat{\theta}(\mathbf{X}_n) = \max_{\theta \in \Theta} L(\theta; X_n). \tag{14.2}$$

Equivalently, the MLE can be taken to be the maximum of the standardized log-likelihood,

$$\frac{l(\theta; \mathbf{X}_n)}{n} = \frac{\log L(\theta; \mathbf{X}_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i). \tag{14.3}$$

Define

$$Q(\theta; \mathbf{X}_n) := \frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i),$$
$$\hat{\theta}(\mathbf{X}_n) := \max_{\theta \in \Theta} Q(\theta; \mathbf{X}_n). \tag{14.4}$$

14.1 Consistency of MLE

By the Weak Law of Large Numbers (Theorem 8.5), we can get,

$$\frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i) \xrightarrow{p} E[l(\theta; X)]. \tag{14.5}$$

Suppose $Q_0(\theta) = E\left[l(\theta; X)\right]$, then we will show that $Q_0(\theta)$ is maximized at θ_0 (i.e., the truth).

> **Lemma 14.1**
>
> *If θ_0 is identified and $E_{\theta_0}\left[|\log p(X; \theta)|\right] < \infty, \forall \theta \in \Theta$, then $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$.*

Proof

> **Theorem 14.1. Consistency of MLE**
>
> *Suppose that $Q(\theta; X_n)$ is continuous in θ and there exists a function $Q_0(\theta)$ such that*
>
> 1. *$Q_0(\theta)$ is uniquely maximized at θ_0.*
> 2. *Θ is compact.*
> 3. *$Q_0(\theta)$ is continuous in θ.*
> 4. *$Q(\theta; X_n)$ converges uniformly in probability to $Q_0(\theta)$.*
>
> *then*
> $$\hat{\theta}(X_n) \xrightarrow{p} \theta_0. \tag{14.6}$$

Proof $\forall \varepsilon > 0$, let

$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_0\| < \epsilon\}.$$

Since $\Theta(\epsilon)$ is an open set, then $\Theta \cap \Theta(\epsilon)^C$ is a compact set (Assumption 2).

Since $Q_0(\theta)$ is a continuous function (Assumption 3), then

$$\theta^* := \sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} \{Q_0(\theta)\}$$

is a achieved for a θ in the compact set.

Since θ_0 is the unique maximized, let

$$Q_0(\theta_0) - Q_0(\theta^*) = \delta > 0.$$

1. For $\theta \in \Theta \cap \Theta(\epsilon)^C$. Let $A_n = \left\{\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} |Q(\theta; X_n) - Q_0(\theta)| < \frac{\delta}{2}\right\}$, then

$$
\begin{aligned}
A_n \Rightarrow Q(\theta; X_n) &< Q_0(\theta) + \frac{\delta}{2} \\
&\leq Q_0(\theta^*) + \frac{\delta}{2} \\
&= Q_0(\theta_0) - \frac{\delta}{2}
\end{aligned}
$$

2. For $\theta \in \Theta(\epsilon)$. Let $B_n = \left\{ \sup_{\theta \in \Theta(\epsilon)} |Q(\theta; X_n) - Q_0(\theta)| < \frac{\delta}{2} \right\}$, then

$$B_n \Rightarrow Q(\theta; X_n) > Q_0(\theta) - \frac{\delta}{2}, \forall \theta \in \Theta(\epsilon)$$

By Assumption 1,

$$Q(\theta_0; X_n) > Q_0(\theta_0) - \frac{\delta}{2}$$

If both A_n and B_n hold, then

$$\hat{\theta} \in \Theta(\epsilon).$$

By Assumption 4, we can concluded that $P(A_n \cap B_n) \to 1$, so

$$P(\hat{\theta} \in \Theta(\epsilon)) \to 1,$$

which means,

$$\hat{\theta}(X_n) \xrightarrow{p} \theta_0.$$

14.2 Asymptotic Normality of MLE

14.3 Efficiency of MLE

Chapter 15 Minimum-Variance Unbiased Estimator

> **Definition 15.1. UMVU Estimators**
>
> *An unbiased estimator $\delta(\boldsymbol{X})$ of $g(\theta)$ is the uniform minimum variance unbiased (UMVU) estimator of $g(\theta)$ if*
>
> $$Var_\theta\delta(\boldsymbol{X}) \le Var_\theta\delta'(\boldsymbol{X}), \quad \forall\theta \in \Theta, \tag{15.1}$$
>
> *where $\delta'(\boldsymbol{X})$ is any other unbiased estimator of $g(\theta)$.* ♣

Note *If there exists an unbiased estimator of g, the estimand g will be called U-estimable.*

1. If $T(\boldsymbol{X})$ is a complete sufficient statistic, estimator $\delta(\boldsymbol{X})$ that only depends on $T(\boldsymbol{X})$, then for any U-estimable function $g(\theta)$ with

$$E_\theta\delta(T(\boldsymbol{X})) = g(\theta), \quad \forall\theta \in \Theta, \tag{15.2}$$

hence, $\delta(T(\boldsymbol{X}))$ is the unique UMVU estimator of $g(\theta)$.

2. If $T(\boldsymbol{X})$ is a complete sufficient statistic and $\delta(\boldsymbol{X})$ is any unbiased estimator of $g(\theta)$, then the UMVU estimator of $g(\theta)$ can be obtained by

$$E\left[\delta(\boldsymbol{X}) \mid T(\boldsymbol{X})\right]. \tag{15.3}$$

Example 15.1 Estimating Polynomials of a Normal Variance Let $X_1, \ldots, X_n$ be distributed with joint density

$$\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \xi)^2\right]. \tag{15.4}$$

Discussing the UMVU estimators of ξ^r, σ^r, ξ/σ.

Solution

1. *σ is known:*

   Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the complete sufficient statistic of X_i, and

   $$E(\bar{X}) = \xi,$$

   then the UMVU estimator of ξ is $\bar{X}$.

   Therefore, the UMVU estimator of ξ^r is $\bar{X}^r$ and the UMVU estimator of ξ/σ is $\bar{X}/\sigma$.

2. *ξ is known:*

Since $s^r = \sum (x_i - \xi)^r$ is the complete sufficient statistic of X_i.

Assume

$$E\left[\frac{s^r}{\sigma^r}\right] = \frac{1}{K_{n,r}},$$

where $K_{n,r}$ is a constant depends on n, r.

Since $s^2/\sigma^2 \sim Ga(n/2, 1/2) = \chi^2(n)$, then

$$E\left[\frac{s^r}{\sigma^r}\right] = E\left[\left(\frac{s^2}{\sigma^2}\right)^{\frac{r}{2}}\right] = \int_0^\infty x^{\frac{r}{2}} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx = \frac{\Gamma\left(\frac{n+r}{2}\right)}{\Gamma(\frac{n}{2})} \cdot 2^{\frac{r}{2}}.$$

therefore,

$$K_{n,r} = \frac{\Gamma(\frac{n}{2})}{2^{\frac{r}{2}} \cdot \Gamma\left(\frac{n+r}{2}\right)}.$$

Hence,

$$E\left[s^r K_{n,r}\right] = \sigma^r \text{ and } E[\xi s^{-1} K_{n,-1}] = \xi/\sigma,$$

which means the UMVU estimator of σ^r is $s^r K_{n,r}$ and the UMVU estimator of ξ/σ is $\xi s^{-1} K_{n,-1}$.

3. **Both ξ and σ is unknown:**

   Since $(\bar{X}, s_x^r)$ are the complete sufficient statistic of X_i, where $s_x^2 = \sum (x_i - \bar{X})^r$.

   Since $s_x^2/\sigma^2 \sim \chi^2(n-1)$, then

   $$E\left[\frac{s_x^r}{\sigma^r}\right] = \frac{1}{K_{n-1,r}}.$$

   Hence,

   $$E\left[s_x^r K_{n-1,r}\right] = \sigma^r,$$

   which means the UMVU estimator of σ^r is $s_x^r K_{n-1,r}$, and

   $$E(\bar{X}^r) = \xi^r,$$

   which means the UMVU estimator of ξ^r is $\bar{X}^r$.

   Since $\bar{X}$ and s_x^r are independent, then

   $$E[\bar{X}s_x^{-1}K_{n-1,-1}] = \xi/\sigma$$

   which means the UMVU estimator of ξ/σ is $\bar{X}s_x^{-1}K_{n-1,-1}$.

Example 15.2 Let $X_1, \ldots, X_n$ be i.i.d sample from $U(\theta_1 - \theta_2, \theta_1 + \theta_2)$, where $\theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+$. Discussing the UMVU estimators of θ_1, θ_2.

Solution *Let $X_{(i)}$ be the i-th order statistic of X_i, then $(X_{(1)}, X_{(n)})$ is the complete and sufficient statistic for (θ_1, θ_2). Thus it suffices to find a function $(X_{(1)}, X_{(n)})$, which is unbiased of (θ_1, θ_2).*

Let

$$Y_i = \frac{X_i - (\theta_1 - \theta_2)}{2\theta_2} \sim U(0, 1),$$

and

$$Y_{(i)} = \frac{X_{(i)} - (\theta_1 - \theta_2)}{2\theta_2},$$

be the i-th order statistic of Y_i, then we got

$$E[X_{(1)}] = 2\theta_2 E[Y_{(1)}] + (\theta_1 - \theta_2)$$

$$= 2\theta_2 \int_0^1 ny(1-y)^{n-1} dy + (\theta_1 - \theta_2)$$

$$= \theta_1 - \frac{3n+1}{n+1}\theta_2$$

$$E[X_{(n)}] = 2\theta_2 E[Y_{(n)}] + (\theta_1 - \theta_2)$$

$$= 2\theta_2 \int_0^1 ny^n dy + (\theta_1 - \theta_2)$$

$$= \theta_1 + \frac{n-1}{n+1}\theta_2$$

.

Thus,

$$\theta_1 = E\left[\frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}\right],$$

$$\theta_2 = E\left[-\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}\right],$$

which means the UMVU estimator is

$$\hat{\theta}_1 = \frac{n-1}{4n}X_{(1)} + \frac{3n+1}{4n}X_{(n)}, \quad \hat{\theta}_2 = -\frac{n+1}{4n}X_{(1)} + \frac{n+1}{4n}X_{(n)}.$$

Chapter 16 Bayes Estimator

We shall look for some estimators that make the risk function $R(\theta, \delta)$ small in some overall sense. There are two way to solve it: minimize the average risk, minimize the maximum risk.

This chapter will discuss the first method, also known as, Bayes Estimator.

> **Definition 16.1. Bayes Estimator**
>
> *The Bayes Estimator δ with respect to Λ is minimizing the Bayes Risk of δ*
>
> $$r(\Lambda, \delta) = \int R(\theta, \delta)\, \mathrm{d}\Lambda(\theta) \tag{16.1}$$
>
> *where Λ is the probability distribution.* ♣

In Bayesian arguments, it is important to keep track of which variables are being conditioned on. Hence, the notations are as followed:

- The density of X will be denoted by $X \sim f(x \mid \theta)$.
- The prior distribution will be denoted by $\Pi \sim \pi(\theta \mid \lambda)$ or $\Lambda \sim \gamma(\lambda)$, where λ is another parameter (sometimes called a hyperparameter).
- The posterior distribution, which calculate the conditional distributions as that of θ given x and λ, or λ given x, which is denoted by $\Pi \sim \pi(\theta \mid x, \lambda)$ or $\Lambda \sim \gamma(\lambda \mid x)$, that is

$$\pi(\theta \mid x, \lambda) = \frac{f(x \mid \theta)\, \pi(\theta \mid \lambda)}{m(x \mid \lambda)}, \tag{16.2}$$

  where marginal distributions $m(x \mid \lambda) = \int f(x \mid \theta)\, \pi(\theta \mid \lambda)\, \mathrm{d}\theta$.

> **Theorem 16.1**
>
> *Let Θ have distribution Λ, and given $\Theta = \theta$, let X have distribution P_θ. Suppose, the following assumptions hold for the problem of estimating $g(\Theta)$ with non-negative loss function $L(\theta, d)$,*
>
> - *There exists an estimator δ_0 with finite risk.*
> - *For almost all x, there exists a value $\delta_\Lambda(x)$ minimizing*
>
> $$E\{L[\Theta, \delta(x)] \mid X = x\}. \tag{16.3}$$
>
> *Then, $\delta_\Lambda(x)$ is a Bayes Estimator.* ♡

⚗ **Note** *Improper prior*

> **Corollary 16.1**
>
> *Suppose the assumptions of Theorem 16.1 hold.*
>
> 1. *If $L(\theta, d) = [d - g(\theta)]^2$, then*
>
> $$\delta_\Lambda(x) = E[g(\Theta) \mid x]. \qquad (16.4)$$
>
> 2. *If $L(\theta, d) = w(\theta)[d - g(\theta)]^2$, then*
>
> $$\delta_\Lambda(x) = \frac{E[w(\theta) g(\Theta) \mid x]}{E[w(\theta) \mid x]}. \qquad (16.5)$$
>
> 3. *If $L(\theta, d) = |d - g(\theta)|$, then $\delta_\Lambda(x)$ is any median of the conditional distribution of Θ given x.*
>
> 4. *If*
>
> $$L(\theta, d) = \begin{cases} 0 \text{ when } |d - \theta| \le c \\ 1 \text{ when } |d - \theta| > c \end{cases},$$
>
> *then $\delta_\Lambda(x)$ is the midpoint of the interval I of length $2c$ which maxmizes $P(\Theta \in I \mid x)$.*
>
> ♡

Proof

> **Theorem 16.2**
> *Necessiary condition for Bayes Estimator*
>
> ♡

Methodologies have been developed to deal with the difficulty which sometimes incorporate frequentist measures to assess the choic of Λ.

- Empirical Bayes.
- Hierarchical Bayes.
- Robust Bayes.
- Objective Bayes.

16.1 Single-Prior Bayes

The Single-Prior Bayes model in a general form as

$$X \mid \theta \sim f(x \mid \theta),$$
$$\Theta \mid \gamma \sim \pi(\theta \mid \lambda), \qquad (16.6)$$

where we assume that the functional form of the prior and the value of λ is known (we will write it as $\gamma = \gamma_0$).

Given a loss function $L(\theta, d)$, we would then determine the estimator that minimizes

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, d\theta, \tag{16.7}$$

where $\pi(\theta \mid x)$ is posterior distribution given by

$$\pi(\theta \mid x) = \frac{f(x \mid \theta) \pi(\theta \mid \gamma_0)}{\int f(x \mid \theta) \pi(\theta \mid \gamma_0) \, d\theta}.$$

In general, this Bayes estimator under squared error loss is given by

$$E(\Theta \mid x) = \frac{\int \theta f(x \mid \theta) \pi(\theta \mid \gamma_0) \, d\theta}{\int f(x \mid \theta) \pi(\theta \mid \gamma_0) \, d\theta}. \tag{16.8}$$

Example 16.1 Consider

$$X_i \overset{\text{i.i.d}}{\sim} N(\mu, \Gamma^{-1}), \quad i = 1, 2, \ldots, n$$

$$\mu \sim N(0, 1),$$

$$\Gamma \sim \text{Gamma}(2, 1),$$

calculate the Single-Prior Bayes estimator under squared error loss.

Solution

$$p(X \mid \mu, \Gamma) = \Gamma^n (2\pi)^{-\frac{n}{2}} \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right],$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2}\right),$$

$$p(\Gamma) = \frac{1}{\Gamma(2)} \Gamma \exp(-\Gamma).$$

Therefore,

$$h(X, \mu, \Gamma) = C\Gamma^n \exp\left[-2\Gamma^2 \sum_{i=1}^{n} (x_i - \mu)^2\right] \exp\left(-\frac{\mu^2}{2}\right) \Gamma \exp(-\Gamma),$$

where $C = \frac{(2\pi)^{-\frac{n+1}{2}}}{\Gamma(2)}$.

For μ, we have

$$\pi(\mu \mid X, \Gamma) = \frac{h(X, \mu, \Gamma)}{p(\mu \mid X)}$$

For exponential families

Theorem 16.3

♡

16.2 Hierarchical Bayes

In a Hierarchical Bayes model, rather than specifying the prior distribution as a single function, we specify it in a **hierarchy**. Thus, the Hierarchical Bayes model in a general form as

$$X \mid \theta \sim f(x \mid \theta),$$
$$\Theta \mid \gamma \sim \pi(\theta \mid \lambda), \tag{16.9}$$
$$\Gamma \sim \psi(\gamma),$$

where we assume that $\psi(\cdot)$ is known and not dependent on any other unknown hyperparameters.

Note *We can continue this hierarchical modeling and add more stages to the model, but this is not ofthen done in practice.*

Given a loss function $L(\theta, d)$, we would then determine the estimator that minimizes

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, \mathrm{d}\theta, \tag{16.10}$$

where $\pi(\theta \mid x)$ is posterior distribution given by

$$\pi(\theta \mid x) = \frac{\int f(x \mid \theta) \pi(\theta \mid \gamma) \psi(\gamma) \, \mathrm{d}\gamma}{\int \int f(x \mid \theta) \pi(\theta \mid \gamma) \psi(\gamma) \, \mathrm{d}\theta \mathrm{d}\gamma}.$$

Note *The posterior distribution can also be writed as*

$$\pi(\theta \mid x) = \int \pi(\theta \mid x, \gamma) \pi(\gamma \mid x) \, \mathrm{d}\gamma,$$

where $\pi(\gamma \mid x)$ is the posterior distribution of Γ, unconditional on θ. The equation 16.10 can be writed as

$$\int L(\theta, d(x)) \pi(\theta \mid x) \, \mathrm{d}\theta = \int \left[\int L(\theta, d(x)) \pi(\theta \mid x, \gamma) \, \mathrm{d}\theta \right] \pi(\gamma \mid x) \, \mathrm{d}\gamma.$$

*which shows that **the Hierarchical Bayes estimator can be thought of as a mixture of Single-Prior estimators**.*

Example 16.2 Poisson Hierarchy Consider

$$X_i \mid \lambda \overset{\text{i.i.d}}{\sim} \text{Poisson}(\lambda), \quad i = 1, 2 \ldots, n$$
$$\lambda \mid b \sim \text{Gamma}(a, b), \ a \text{ known}, \tag{16.11}$$
$$\frac{1}{b} \sim \text{Gamma}(k, \tau),$$

calculate the Hierarchical Bayes estimator under squared error loss.

Theorem 16.4

For the Hierarchical Bayes model (16.9),

$$K\left[\pi\left(\lambda \mid x\right), \psi\left(\lambda\right)\right] < K\left[\pi\left(\theta \mid x\right), \pi\left(\theta\right)\right], \qquad (16.12)$$

where K is the Kullback-Leibler information for discrimination between two densities.

Proof

Note

16.3 Empirical Bayes

16.4 Bayes Prediction

Chapter 17 Hypothesis Testing

Part VII

Convex Optimization

Chapter 18 Convex Sets

18.1 Affine and Convex Sets

18.1.1 Affine Sets

Definition 18.1. Affine Set

*A nonempty set C is a **affine set** that satisfy*

$$\forall x_1, x_2 \in C, \theta \in \mathbf{R}, \theta x_1 + (1 - \theta)x_2 \in C.$$

♣

18.1.2 Convex Sets

Definition 18.2. Convex Set

*A nonempty set C is a **convex set** that satisfy*

$$\forall x_1, x_2 \in C, \theta \in [0, 1], \theta x_1 + (1 - \theta)x_2 \in C.$$

♣

Definition 18.3. Convex Hull

*The **convex hull** of a set C, denoted by conv C is a set of all convex combinations of points in C,*

$$conv\ C = \{\theta_1 x_1 + \ldots + \theta_k x_k | x_i \in C; \theta_i \geq 0, i = 1, \ldots, k; \theta_1 + \ldots + \theta_k = 1\}.$$

♣

Note *The convex hull conv C is always convex, which is the minimal convex set that contains C.*

18.1.3 Cones

Definition 18.4. Cone

*A nonempty set C is a **cone** that satisfy*

$$\forall x \in C, \theta \geq 0, \theta x \in C.$$

♣

Definition 18.5. Convex Cone

*A nonempty set C is a **convex cone** that satisfy*

$$\forall x_1, x_2 \in C, \theta_1, \theta_2 \geq 0, \theta_1 x_1 + \theta_2 x_2 \in C.$$

♣

18.2 Some Important Examples

Definition 18.6. Hyperplane

A hyperplane is a set of the form

$$\{x | a^T x = b\},$$

where $a \in \mathbf{R}^n, a \neq 0, b \in \mathbf{R}$.

♣

Definition 18.7. Halfspace

A hyperplane is a set of the form

$$\{x | a^T x \leq b\},$$

where $a \in \mathbf{R}^n, a \neq 0, b \in \mathbf{R}$.

♣

Definition 18.8. (Euclidean) Ball

A (Euclidean) ball in $\mathbf{R}^n$ with center x_c and radius r has the form

$$B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\} = \{x_c + ru | \|u\|_2 \leq 1\},$$

where $r > 0$.

♣

Definition 18.9. Ellipsoid

A Ellipsoid in $\mathbf{R}^n$ with center x_c has the form

$$\mathcal{E} = \{x | (x - x_c)^T P^{-1} (x - x_c) \leq 1\} = \{x_c + Au | \|u_2\| \leq 1\},$$

where $P \in \mathbf{S}_{++}^n$ (symmetric positive definite).

♣

18.3 Generalized Inequalities

18.3.1 Definition of Generalized Inequalities

Definition 18.10. Proper Cone

A cone $K \subseteq \mathbf{R}^n$ is proper cone, if

- *K is convex.*
- *K is closed.*
- *K is solid (nonempty interior).*
- *K is pointed (contains no line).*

♣

Definition 18.11. Generalized Inequalities

The partial ordering on $\mathbf{R}^n$ defined by proper cone K, if

$$y - x \in K, \tag{18.1}$$

which can be denoted by

$$x \preceq_K y \text{ or } y \succeq_K x. \tag{18.2}$$

The strict partial ordering on $\mathbf{R}^n$ defined by proper cone K, if

$$y - x \in \ \text{int } K, \tag{18.3}$$

which can be denoted by

$$x \prec_K y \text{ or } y \succ_K x. \tag{18.4}$$

♣

Note *When $K = \mathbf{R}_+$, the partial ordering $\preceq_K$ is the usual ordering $\leq$ on $\mathbf{R}$, and the strict partial ordering $\prec_K$ is the usual strict ordering $<$ on $\mathbf{R}$.*

18.3.2 Properties of Generalized Inequalities

Theorem 18.1. Properties of Generalized Inequalities

A generalized inequality $\preceq_K$ has the following properties:

- *Preserved under addition:*
- *Transitive:*
- *Preserved under nonnegative scaling:*
- *Reflexive:*

- *Antisymmetric:*

- *Preserved under limits:*

A strict generalized inequality $\prec_K$ has the following properties:

Chapter 19 Convex Optimization Problems

19.1 Generalized Inequality Constraints

> **Definition 19.1. With Generalized Inequality Constraints**
>
> *A convex optimization problem with generalized inequality constraints has the form*
>
> $$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \preceq_{K_i} 0, \quad i = 1, \ldots, m \\ & Ax = b \end{aligned} \tag{19.1}$$
>
> *where $f_0 : \mathbf{R}^n \to \mathbf{R}$, $K_i \in \mathbf{R}^{k_i}$ are proper conves, and $f_i : \mathbf{R}^n \to \mathbf{R}^{k_i}$ are K_i-convex.* ♣

19.1.1 Conic Form Problems

> **Definition 19.2. Conic Form Problem**
>
> *A conic form problem has the form*
>
> $$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Fx + g \preceq_K 0 \\ & Ax = b \end{aligned} \tag{19.2}$$
>
> ♣

19.1.2 Semidefinite Programming

19.2 Vector Optimization

Chapter 20 Unconstrained Minimization

20.1 Definition of Unconstrained Minimization

> **Definition 20.1. Unconstrained Minimization Problem**
>
> *The unconstrained minimization problem has the form*
>
> $$\min_x f(x) \tag{20.1}$$
>
> *where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable.* ♣

Note *We assume that the problem is solvable, i.e., there exists an optimal point x^*, such that,* $f(x^*) = \inf_x f(x)$.

Example 20.1Quadratic Minimization

Example 20.2Least Square Estimation

Example 20.3Unconstrained Geometric Programming

Example 20.4Analytic Center of Linear Inequalities

20.2 General Descent Method

20.3 Gradient Descent Method

20.4 Steepest Descent Method

20.5 Newton's Method

Chapter 21 Exercises for Convex Optimization

21.1 Convex Sets

✍ **Exercise 21.1** Solution set of a quadratic inequality Let $C \subseteq \mathbf{R}^n$ be the solution set of a quadratic inequality,

$$C = \{x \in \mathbf{R}^n | x^T A x + b^T x + c \leq 0\}$$

with $A \in \mathbf{S}^n$, $b \in \mathbf{R}^n$, and $c \in \mathbf{R}$.

1. Show that C is convex if $A \succeq 0$.

Solution

1. *We have to show that $\theta x + (1-\theta)y \in C$ for all $\theta \in [0,1]$ and $x, y \in C$.*

$$(\theta x + (1-\theta)y)^T A (\theta x + (1-\theta)y) + b^T (\theta x + (1-\theta)y) + c$$

$$= \theta^2 x^T A x + \theta(1-\theta)(y^T A x + x^T A y) + (1-\theta)^2 y^T A y + \theta b^T x + (1-\theta)b^T y + c$$

$$= \theta^2 (x^T A x + b^T x + c) + (1-\theta)^2 (y^T A y + b^T y + c) - \theta^2 (b^T x + c)$$

$$\quad - (1-\theta)^2 (b^T y + c) + \theta(1-\theta)(y^T A x + x^T A y) + \theta b^T x + (1-\theta)b^T y + c$$

$$\leq - \theta^2 (b^T x + c) - (1-\theta)^2 (b^T y + c) + \theta(1-\theta)(y^T A x + x^T A y)$$

$$\quad + \theta b^T x + (1-\theta)b^T y + c$$

$$= \theta(1-\theta)[(b^T x + c) + (b^T y + c) + x^T A x + y^T A y]$$

$$\leq \theta(1-\theta)(-x^T A x - y^T A y + x^T A x + y^T A y) \leq 0$$

Therefore, $\theta x + (1-\theta)y \in C$, which shows that C is convex if $A \succeq 0$.

Part VIII

Generalized Linear Model

Chapter 22 Introduction

Chapter 23 Binary Data

Chapter 24 Polytomous Data