An Introduction to Advanced Probability and Statistics

Junhui Qian

Preface

This booklet introduces advanced probability and statistics to first-year Ph.D. students in economics.

In preparation of this text, I borrow heavily from the lecture notes of Yoosoon Chang and Joon Y. Park, who taught me econometrics at Rice University. All errors are mine.

Shanghai, China,
October 2011

Junhui Qian
junhuiq@gmail.com

Contents

Chapter 1

Introduction to Probability

In this chapter we lay down the measure-theoretic foundation of probability.

1.1 Probability Triple

We first introduce the well known *probability triple*, $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is a sigma-field of a collection of subsets of Ω, and $\mathbb{P}$ is a probability measure. We define and characterize each of the probability triple in the following.

The sample space Ω is a set of *outcomes* from a random experiment. For instance, in a coin tossing experiment, the sample space is obviously $\{H, T\}$, where H denotes head and T denotes tail. For another example, the sample space may be an interval, say $\Omega = [0, 1]$, on the real line, and any outcome $\omega \in \Omega$ is a real number randomly selected from the interval.

To introduce sigma-field, we first define

Definition 1.1.1 (Field (or Algebra)) *A collection of subsets $\mathcal{F}$ is called a field or an algebra, if the following holds.*

(a) $\Omega \in \mathcal{F}$

(b) $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$

(c) $E_1, ..., E_m \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{m} E_n \in \mathcal{F}$

Note that (c) says that a field is closed under finite union. In contrast, a sigma-field, which is defined as follows, is closed under countable union.

Definition 1.1.2 (sigma-field (or sigma-algebra)) *A collection of subsets $\mathcal{F}$ is called a σ-field or a σ-algebra, if the following holds.*

(a) $\Omega \in \mathcal{F}$

(b) $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$

(c) $E_1, E_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$

Remarks:

- In both definitions, (a) and (b) imply that the empty set $\emptyset \in \mathcal{F}$
- (b) and (c) implies that if $E_1, E_2, \ldots \in \mathcal{F} \Rightarrow \bigcap_{n=1}^{\infty} E_n \in \mathcal{F}$, since $\cap_n E_n = (\cup_n E_n^c)^c$.
- A σ-field is a field; a field is a σ-field only when Ω is finite.
- An arbitrary intersection of σ-fields is still a σ-field. (Exercise 1)

In the following, we may interchangeably write sigma-field as σ-field. An element E of the σ-field $\mathcal{F}$ in the probability triple is called an event. For an example, if we toss a coin twice, then the sample space would be $\Omega = \{HH, HT, TH, TT\}$. A σ-field (or field) would be

$$\begin{aligned}\mathcal{F} \;=\; & \{\emptyset, \Omega, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \\ & \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \\ & \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}\}.\end{aligned}$$

The event $\{HH\}$ would be described as “two heads in a row”. The event $\{HT, TT\}$ would be described as “the second throw obtains tail”.

For an example of infinite sample space, we may consider a thought experiment of tossing a coin for infinitely many times. The sample space would be $\Omega = \{(r_1, r_2, \ldots,)|r_i = 1 \text{ or } 0\}$, where 1 stands for head and 0 stands for tail. One example of an event would be $\{r_1 = 1, r_2 = 1\}$, which says that the first two throws give heads in a row.

A sigma-field can be generated from a collection of subsets of Ω, a field for example. We define

Definition 1.1.3 (Generated σ-field) *Let $\mathcal{S}$ be a collection of subsets of Ω. The σ-field generated by $\mathcal{S}$, $\sigma(\mathcal{S})$, is defined to be the intersection of all the σ-fields containing $\mathcal{S}$.*

In other words, $\sigma(\mathcal{S})$ is the smallest σ-field containing $\mathcal{S}$.

Now we introduce the axiomatic definition of probability measure.

Definition 1.1.4 (Probability Measure) *A set function $\mathbb{P}$ on a σ-field $\mathcal{F}$ is a probability measure if it satisfies:*

(1) $\mathbb{P}(E) \geq 0 \ \ \forall E \in \mathcal{F}$

(2) $\mathbb{P}(\Omega) = 1$

(3) If $E_1, E_2, \ldots$ are disjoint, then $\mathbb{P}\left(\bigcup_n E_n\right) = \sum_n \mathbb{P}(E_n)$.

Properties of Probability Measure

(a) $\mathbb{P}(\emptyset) = 0$

(b) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

(c) $A \subset B \ \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$

(d) $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

(e) $A_n \subset A_{n+1}$ for $n = 1, 2, \ldots$, $\Rightarrow \mathbb{P}(A_n) \uparrow P(\cup_{n=1}^{\infty} A_n)$

(f) $A_n \supset A_{n+1}$ for $n = 1, 2, \ldots$, $\Rightarrow \mathbb{P}(A_n) \downarrow P(\cap_{n=1}^{\infty} A_n)$

(g) $\mathbb{P}(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$

Proof: (a)-(c) are trivial.

(d) Write $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, a union of disjoint sets. By adding and subtracting $\mathbb{P}(A \cap B)$, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, using the fact that $A = (A \cap B) \cup (A \cap B^c)$, also a disjoint union.

(e) Define $B_1 = A_1$ and $B_n = A_{n+1} - A_n$ for $n \geq 2$. We have $A_n = \bigcup_{j=1}^{n} B_j$ and $\bigcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} B_j$. Then it follows from

$$\mathbb{P}(A_n) = \sum_{j=1}^{n} \mathbb{P}(B_j) = \sum_{j=1}^{\infty} \mathbb{P}(B_j) - \sum_{j=n+1}^{\infty} \mathbb{P}(B_j) = \mathbb{P}(\bigcup_{n=1}^{\infty} A_n) - \sum_{j=n+1}^{\infty} \mathbb{P}(B_j).$$

(f) Note that $A_n^c \subset A_{n+1}^c$, use (e).

(g) Extend (d).

Note that we may write $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$, if A_n is monotone increasing, and $\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n$, if A_n is monotone decreasing.

1.2 Conditional Probability and Independence

Definition 1.2.1 (Conditional Probability) *For an event $F \in \mathcal{F}$ that satisfies $\mathbb{P}(F) > 0$, we define the conditional probability of another event E given F by*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

- For a fixed event F, the function $\mathbb{Q}(\cdot) = \mathbb{P}(\cdot|F)$ is a probability. All properties of probability measure hold for $\mathbb{Q}$.

- The probability of intersection can be defined via conditional probability:

  $$\mathbb{P}(E \cap F) = \mathbb{P}(E|F)\,\mathbb{P}(F),$$

  and

  $$\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E|F \cap G)\,\mathbb{P}(F|G)\,\mathbb{P}(G).$$

- If $\{F_n\}$ is a partition of Ω, ie, $F_n's$ are disjoint and $\bigcup_n F_n = \Omega$. Then the following *theorem of total probability* holds,

  $$\mathbb{P}(E) = \sum_n \mathbb{P}(E|F_n)\,\mathbb{P}(F_n), \text{ for all event } E.$$

- The *Bayes Formula* follows from $\mathbb{P}(E \cap F) = \mathbb{P}(E|F)\,\mathbb{P}(F) = \mathbb{P}(F|E)\,\mathbb{P}(E)$,

  $$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F)\,\mathbb{P}(F)}{\mathbb{P}(E)},$$

  and

  $$\mathbb{P}(F_k|E) = \frac{\mathbb{P}(E|F_k)\,\mathbb{P}(F_k)}{\sum_n \mathbb{P}(E|F_n)\,\mathbb{P}(F_n)}.$$

Definition 1.2.2 (Independence of Events) *Events E and F are called independent if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F)$.*

- We may equivalently define independence as

  $$\mathbb{P}(E|F) = \mathbb{P}(F), \text{ when } \mathbb{P}(F) > 0$$

- $E_1, E_2, \ldots$ are said to be independent if, for any $(i_1, \ldots, i_k)$,

$$\mathbb{P}\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right) = \bigcap_{j=1}^{k} \mathbb{P}\left(E_{i_j}\right)$$

- Let $E, E_1, E_2, \ldots$ be independent events. Then E and $\sigma(E_1, E_2, \ldots)$ are independent, ie, for any $S \in \sigma(E_1, E_2, \ldots)$, $\mathbb{P}(E \cap S) = \mathbb{P}(E)\,\mathbb{P}(S)$.

- Let $E_1, E_2, \ldots, F_1, F_2, \ldots$ be independent events. If $E \in \sigma(E_1, E_2, \ldots)$, then $E, F_1, F_2, \ldots$ are independent; furthermore, $\sigma(E_1, E_2, \ldots)$ and $\sigma(F_1, F_2, \ldots)$ are independent.

1.3 Limits of Events

limsup and liminf First recall that for a series of real numbers $\{x_n\}$, we define

$$\begin{aligned}\limsup_{n\to\infty} x_n &= \inf_k \left\{\sup_{n\geq k} x_n\right\}\\ \liminf_{n\to\infty} x_n &= \sup_k \left\{\inf_{n\geq k} x_n\right\}.\end{aligned}$$

And we say that $x_n \to x \in [-\infty, \infty]$ if $\limsup x_n = \liminf x_n = x$.

Definition 1.3.1 (limsup of Events) *For a sequence of events (E_n), we define*

$$\begin{aligned}\limsup_{n\to\infty} E_n &= \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} E_n\\ &= \{\omega |\ \forall k,\ \exists n(\omega) \geq k\ s.t.\ \omega \in E_n\}\\ &= \{\omega |\ \omega \in E_n \text{ for infinitely many } n.\}\\ &= \{\omega |\ E_n\ i.o.\},\end{aligned}$$

where i.o. denotes "infinitely often".

We may intuitively interpret $\limsup_{n\to\infty} E_n$ as the event that E_n occurs infinitely often.

Definition 1.3.2 (liminf of Events) *We define*

$$\begin{aligned}
\liminf_{n\to\infty} E_n &= \bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty} E_n \\
&= \{\omega|\ \exists\ k(\omega),\ \omega \in E_n\ \forall n \geq k\} \\
&= \{\omega|\ \omega \in E_n \textit{ for all large } n.\} \\
&= \{\omega|\ E_n \textit{ e.v.}\},
\end{aligned}$$

where e.v. denotes "eventually".

It is obvious that It is obvious that $(\liminf E_n)^c = \limsup E_n^c$ and $(\limsup E_n)^c = \liminf E_n^c$. When $\limsup E_n = \liminf E_n$, we say (E_n) has a limit $\lim E_n$.

Lemma 1.3.3 (Fatou's Lemma) *We have*

$$\mathbb{P}(\liminf E_n) \leq \liminf \mathbb{P}(E_n) \leq \limsup \mathbb{P}(E_n) \leq \mathbb{P}(\limsup E_n).$$

Proof: Note that $\bigcap_{n=k}^{\infty} E_n$ is monotone increasing and $\bigcap_{n=k}^{\infty} E_n \uparrow \bigcup_{k=1}^{\infty}\bigcap_{n=k}^{\infty} E_n$. Hence $\mathbb{P}(E_k) \geq \mathbb{P}(\bigcap_{n=k}^{\infty} E_n) \uparrow \mathbb{P}(\liminf E_n)$. The third inequality can be similarly proved. And the second inequality is obvious.

Lemma 1.3.4 (Borel-Cantelli Lemma) *Let $E_1, E_2, \ldots \in \mathcal{F}$, then*

(i) $\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \Rightarrow \mathbb{P}(\limsup E_n) = 0$;

(ii) if $\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty$, *and if* $\{E_n\}$ *are independent, then* $\mathbb{P}(\limsup E_n) = 1$.

Proof: (i) $\mathbb{P}(\limsup E_n) \leq \mathbb{P}(\bigcup_{n\geq k} E_n) \leq \sum_{n=k}^{\infty} \mathbb{P}(E_n) \to 0$.

(ii) For $m, n \in \mathbb{N}$, using $1 - x \leq \exp(-x),\ \forall x \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{n=k}^{\infty} E_n^c\right) &\leq \mathbb{P}\left(\bigcap_{n=k}^{k+m} E_n^c\right) \\
&= \prod_{n=k}^{k+m} \mathbb{P}\left(E_n^c\right) = \prod_{n=k}^{k+m} \left(1 - \mathbb{P}\left(E_n\right)\right) \\
&\leq \exp\left(-\sum_{n=k}^{k+m} \mathbb{P}\left(E_n\right)\right) \to 0,
\end{aligned}$$

as $m \to \infty$. Since $\mathbb{P}\left(\bigcup_k \bigcap_{n\geq k} E_n^c\right) \leq \sum_k \mathbb{P}\left(\bigcap_{n\geq k} E_n^c\right) = 0$, $\mathbb{P}\left(\limsup E_n\right) = 1 - \mathbb{P}\left(\bigcup_{k\geq 1}\bigcap_{n\geq k} E_n^c\right) = 1$.

Remarks:

- (ii) does not hold if $\{E_n\}$ are not independent. To give a counter example, consider infinite coin tossing. Let $E_1 = E_2 = \cdots = \{r_1 = 1\}$, the events that the first coin is head, then $\{E_n\}$ is not independent and $\mathbb{P}(\limsup E_n) = \mathbb{P}(r_1 = 1) = 1/2$.
- Let H_n be the event that the n-th tossing comes up head. We have $\mathbb{P}(H_n) = 1/2$ and $\sum_n \mathbb{P}(H_n) = \infty$. Hence $\mathbb{P}(H_n \ i.o.) = 1$, and $\mathbb{P}(H_n^c \ e.v.) = 1 - \mathbb{P}(H_n \ i.o.) = 0$.
- Let $B_n = H_{2^n+1} \cap H_{2^n+2} \cap \cdots \cap H_{2^n+\log_2 n}$. B_n is independent, and since $\mathbb{P}(B_n) = (1/2)^{\log_2 n} = 1/n$, $\sum_n \mathbb{P}(B_n) = \infty$. Hence $\mathbb{P}(B_n \ i.o.) = 1$.
- But if $B_n = H_{2^n+1} \cap H_{2^n+2} \cap \cdots \cap H_{2^n+2\log_2 n}$, $\mathbb{P}(B_n \ i.o.) = 0$.
- Let $B_n = H_n \cap H_{n+1}$, we also have $\mathbb{P}(B_n \ i.o.) = 1$. To show this, consider B_{2k}, which is independent.

Why σ-field? You may already see that events such as $\limsup E_n$ and $\liminf E_n$ are very interesting events. To make meaningful probabilistic statements about these events, we need to make sure that they are contained in $\mathcal{F}$, on which $\mathbb{P}$ is defined. This is why we require $\mathcal{F}$ to be a σ-field, which is closed to infinite unions and intersections.

Definition 1.3.5 (Tail Fields) *For a sequence of events $E_1, E_2, \ldots$, the* tail field *is given by*

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(E_n, E_{n+1}, \ldots).$$

- For any n, an event $E \in \mathcal{T}$ depends on events $E_n, E_{n+1}, \ldots$. Any finite number of events are irrelevant.
- In the infinite coin tossing experiment,
  - $\limsup H_n$, obtain infinitely many heads
  - $\liminf H_n$, obtain only finitely many heads
  - $\limsup H_{2^n}$ infinitely many heads on tosses 2, 4, 8, …
  - $\{\lim_{n\to\infty} 1/n \sum_{i=1}^{\infty} r_i \le 1/3\}$
  - $\{r_n = r_{n+1} = \cdots = r_{n+m}\}$, m fixed.

Theorem 1.3.6 (Kolmogrov Zero-One Law) *Let a sequence of events $E_1, E_2, \ldots$ be independent with a tail field $\mathcal{T}$. If an event $E \in \mathcal{T}$, then $\mathbb{P}(E) = 0$ or 1.*

Proof: Since $E \in \mathcal{T} \subset \sigma(E_n, E_{n+1}, \ldots)$, $E, E_1, E_2, \ldots, E_{n-1}$ are independent. This is true for all n, so $E, E_1, E_2, \ldots$ are independent. Hence E and $\sigma(E_1, E_2, \ldots)$ are independent, ie, for all $S \in \sigma(E_1, E_2, \ldots)$, S and E are independent. On the other hand, $E \in \mathcal{T} \subset \sigma(E_1, E_2, \ldots)$. It follows that E is independent of itself! So $\mathbb{P}(E \cap E) = \mathbb{P}^2(E) = \mathbb{P}(E)$, which implies $\mathbb{P}(E) = 0$ or 1.

1.4 Construction of Probability Measure

σ-fields are extremely complicated, hence the difficulty of directly assigning probability to their elements, events. Instead, we work on simpler classes.

Definition 1.4.1 (π-system) *A class of subsets of Ω, $\mathcal{P}$, is a π-system if the following holds:*

$$E, F \in \mathcal{P} \;\Rightarrow\; E \cap F \in \mathcal{P}.$$

For example, the collection $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system.

Definition 1.4.2 (λ-system) *A class of subsets of Ω, $\mathcal{L}$, is a λ-system if*

(a) $\Omega \in \mathcal{L}$

(b) If $E, F \in \mathcal{L}$, and $E \subset F$, then $F - E \in \mathcal{L}$

(c) If $E_1, E_2, \ldots \in \mathcal{L}$ and $E_n \uparrow E$, then $E \in \mathcal{L}$.

- If $E \in \mathcal{L}$, then $E^c \in \mathcal{L}$. It follows from (a) and (b).
- $\mathcal{L}$ is closed under countable union only for monotone increasing events.

Theorem 1.4.3 *A class $\mathcal{F}$ of subsets of Ω is a σ-field if and only if $\mathcal{F}$ is both a π-system and a λ-system.*

Proof: "only if" is trivial. To show "if", it suffices to show that for any $E_1, E_2, \ldots \in \mathcal{F}$, $\bigcup_n E_n \in \mathcal{F}$. We indeed have:

$$\left(\bigcap_{k=1}^{n} E_k^c\right)^c = \bigcup_{k=1}^{n} E_k \uparrow \bigcup_n E_n.$$

Notation: Let $\mathcal{S}$ be a class of subsets of Ω. $\sigma(\mathcal{S})$ is the σ-field generated by $\mathcal{S}$. $\pi(\mathcal{S})$ is the π-system generated by $\mathcal{S}$, meaning that $\pi(\mathcal{S})$ is the intersection of all π-system that contain $\mathcal{S}$. $\lambda(\mathcal{S})$ is similarly defined as the λ-system generated by $\mathcal{S}$.

We have

$$\pi(\mathcal{S}) \subset \sigma(\mathcal{S}) \quad \text{and} \quad \lambda(\mathcal{S}) \subset \sigma(\mathcal{S}).$$

Lemma 1.4.4 (Dynkin's Lemma) *Let $\mathcal{P}$ be a π-system, then $\lambda(\mathcal{P}) = \sigma(\mathcal{P})$.*

Proof: It suffices to show that $\lambda(\mathcal{P})$ is a π-system.

- For an arbitrary $C \in \mathcal{P}$, define

$$\mathcal{D}_C = \{B \in \lambda(\mathcal{P}) | B \cap C \in \lambda(\mathcal{P})\ \}.$$

- We have $\mathcal{P} \subset \mathcal{D}_C$, since for any $E \in \mathcal{P} \subset \lambda(\mathcal{P})$, $E \cap C \in \mathcal{P} \subset \lambda(\mathcal{P})$, hence $E \in \mathcal{D}_C$.

- For any $C \in \mathcal{P}$, $\mathcal{D}_C$ is a λ-system.
  - $\Omega \in \mathcal{D}_C$
  - If $B_1, B_2 \in \mathcal{D}_C$ and $B_1 \subset B_2$, then $(B_2 - B_1) \cap C = B_2 \cap C - B_1 \cap C$. Since $B_1 \cap C, B_2 \cap C \in \lambda(\mathcal{P})$ and $(B_1 \cap C) \subset (B_2 \cap C)$, $(B_2 - B_1) \cap C\ \in \lambda(\mathcal{P})$. Hence $(B_2 - B_1)\ \in \mathcal{D}_C$.
  - If $B_1, B_2, \ldots \in \mathcal{D}_C$, and $B_n \uparrow B$, then $(B_n \cap C)\ \uparrow (B \cap C)\ \in \lambda(\mathcal{P})$. Hence $B \in \mathcal{D}_C$.

- Thus, for any $C \in \mathcal{P}$, $\mathcal{D}_C$ is a λ-system containing $\mathcal{P}$. And it is obvious that $\lambda(\mathcal{P}) \subset \mathcal{D}_C$.

- Now for any $A \in \lambda(\mathcal{P})$, we define

$$\mathcal{D}_A = \{B \in \lambda(\mathcal{P}) | B \cap A \in \lambda(\mathcal{P})\}.$$

  By definition, $\mathcal{D}_A \subset \lambda(\mathcal{P})$.

- We have $\mathcal{P} \subset \mathcal{D}_A$, since if $E \in \mathcal{P}$, then $E \cap A \in \lambda(\mathcal{P})$, since $A \in \lambda(\mathcal{P}) \subset \mathcal{D}_C$ for all $C \in \mathcal{P}$.

- We can check that $\mathcal{D}_A$ is a λ-system that contains $\mathcal{P}$, hence $\lambda(\mathcal{P}) \subset \mathcal{D}_A$. We thus have $\mathcal{D}_A = \lambda(\mathcal{P})$, which means that for any $A, B \in \lambda(\mathcal{P})$, $A \cap B \in \lambda(\mathcal{P})$. Thus $\lambda(\mathcal{P})$ is a π-system. Q.E.D.

Remark: If $\mathcal{P}$ is a π-system, and $\mathcal{L}$ is a λ-system that contains $\mathcal{P}$, then $\sigma(\mathcal{P}) \subset \mathcal{L}$. To see why, note that $\lambda(\mathcal{P}) = \sigma(\mathcal{P})$ is the smallest λ-system that contains $\mathcal{P}$.

Theorem 1.4.5 (Uniqueness of Extension) *Let $\mathcal{P}$ be a π-system on Ω, and $\mathbb{P}_1$ and $\mathbb{P}_2$ be probability measures on $\sigma(\mathcal{P})$. If $\mathbb{P}_1$ and $\mathbb{P}_2$ agree on $\mathcal{P}$, then they agree on $\sigma(\mathcal{P})$.*

Proof: Let $\mathcal{D} = \{E \in \sigma(\mathcal{P}) | \mathbb{P}_1(E) = \mathbb{P}_2(E)\}$. $\mathcal{D}$ is a λ-system, since

- $\Omega \in \mathcal{D}$,
- $E, F \in \mathcal{D}$ and $E \subset F$ imply $F - E \in \mathcal{D}$, since
$$\mathbb{P}_1(F - E) = \mathbb{P}_1(F) - \mathbb{P}_1(E) = \mathbb{P}_2(F) - \mathbb{P}_2(E) = \mathbb{P}_2(F - E).$$
- If $E_1, E_2, \ldots \in \mathcal{D}$ and $E_n \uparrow E$, then $E \in \mathcal{D}$, since
$$\mathbb{P}_1(E) = \lim \mathbb{P}_1(E_n) = \lim \mathbb{P}_2(E_n) = \mathbb{P}_2(E).$$

The fact that $\mathbb{P}_1$ and $\mathbb{P}_2$ agree on $\mathcal{P}$ implies that $\mathcal{P} \subset \mathcal{D}$. The remark following Dynkin's lemma shows that $\sigma(\mathcal{P}) \subset \mathcal{D}$. On the other hand, by definition, $\mathcal{D} \subset \sigma(\mathcal{P})$. Hence $\mathcal{D} = \sigma(\mathcal{P})$. Q.E.D.

Borel σ-field The Borel σ-field is the σ-field generated by the family of open subsets (on a topological space). To probability theory, the most important Borel σ-field is the σ-field generated by the open subsets of $\mathbb{R}$ of real numbers, which we denote $\mathcal{B}(\mathbb{R})$.

Almost every subset of $\mathbb{R}$ that we can think of is in $\mathcal{B}(\mathbb{R})$, the elements of which may be quite complicated. As it is difficult for economic agents to assign probabilities to complicated sets, we often have to consider "simpler" systems of sets, π-system, for example.

Define
$$\mathcal{P} = (-\infty, x], \ x \in \mathbb{R}.$$

It can be easily verified that $\mathcal{P}$ is a π-system. And we show in the following that $\mathcal{P}$ generates $\mathcal{B}(\mathbb{R})$.

Proof: It is clear from
$$(-\infty, x] = \bigcap_n (-\infty, x + 1/n) \, , \ \forall x \in \mathbb{R}$$

that $\sigma(\mathcal{P}) \subset \mathcal{B}(\mathbb{R})$. To show $\sigma(\mathcal{P}) \supset \mathcal{B}(\mathbb{R})$, note that every open set of $\mathbb{R}$ is a countable union of open intervals. It therefore suffices to show that the open intervals of the form (a, b) are in $\sigma(\mathcal{P})$. This is indeed the case, since

$$(a, b) = (-\infty, a]^c \cap \left(\bigcup_n (-\infty, b - 1/n] \right).$$

Note that the above holds even when $b \leq a$, in which case $(a, b) = \emptyset$.

Theorem 1.4.6 (Extension Theorem) *Let $\mathcal{F}_0$ be a field on Ω, and let $\mathcal{F} = \sigma(\mathcal{F}_0)$. If $\mathbb{P}_0$ is a countably additive set function $\mathbb{P}_0 : \mathcal{F}_0 \to [0, 1]$ with $\mathbb{P}_0(\emptyset) = 0$ and $\mathbb{P}_0(\Omega) = 1$, then there exists a probability measure on $(\Omega, \mathcal{F})$ such that*

$$\mathbb{P} = \mathbb{P}_0 \textit{ on } \mathcal{F}_0.$$

Proof: We first define for any $E \subset \Omega$,

$$\mathbb{P}(E) = \inf_{\{A_n\}} \left\{ \sum_n \mathbb{P}_0(A_n) : A_n \in \mathcal{F}_0, E \subset \bigcup_n A_n \right\}.$$

We next prove that

(a) $\mathbb{P}$ is an outer measure.

(b) $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{M})$, where $\mathcal{M}$ is a σ-field of $\mathbb{P}$-measurable sets in $\mathcal{F}$.

(c) $\mathcal{F}_0 \subset \mathcal{M}$

(d) $\mathbb{P} = \mathbb{P}_0$ on $\mathcal{F}_0$.

Note that (c) immediately implies that $\mathcal{F} \subset \mathcal{M}$. If we restrict $\mathbb{P}$ to the domain $\mathcal{F}$, we obtain a probability measure on $(\Omega, \mathcal{F})$ that coincide with $\mathbb{P}_0$ on $\mathcal{F}_0$. The theorem is then proved. In the following we prove (a)-(d).

(a) We first define outer measure. A set function μ on $(\Omega, \mathcal{F})$ is an outer measure if

   (i) $\mu(\emptyset) = 0$.

   (ii) $E \subset F$ implies $\mu(E) \leq \mu(F)$. (monotonicity)

   (iii) $\mu\left(\bigcup_n E_n\right) \leq \sum_n \mu(E_n)$, where $E_1, E_2, \ldots \in \mathcal{F}$. (countable subadditivity)

- It is obvious that $\mathbb{P}(\emptyset) = 0$, since we may choose $E_n = \emptyset \ \forall n$.

- For $E \subset F$, choose $\{A_n\}$ such that $E \subset (\bigcup_n A_n)$ and $F \subset (\bigcup_n A_n) \cup (F - E)$. Monotonicity is now obvious.

- To show countable subadditivity, note that for each n, we can find a collection $\{C_{nk}\}_{k=1}^{\infty}$ such that $C_{nk} \in \mathcal{F}_0$, $E_n \subset \bigcup_k C_{nk}$, and $\sum_k \mathbb{P}_0(C_{nk}) \leq \mathbb{P}(E_n) + \epsilon 2^{-n}$, where $\epsilon > 0$. Since $\bigcup_n E_n \subset \bigcup_n \bigcup_k C_{nk}$, $\mathbb{P}(\bigcup_n E_n) \leq \sum_{n,k} \mathbb{P}_0(C_{nk}) \leq \sum_n \mathbb{P}(E_n) + \epsilon$. Since ϵ is arbitrarily chosen, the countable subadditivity is proved.

(b) Now we define $\mathcal{M}$ as

$$\mathcal{M} = \{A \subset \Omega | \mathbb{P}(A \cap E) + \mathbb{P}(A^c \cap E) = \mathbb{P}(E), \ \forall E \subset \Omega\}.$$

$\mathcal{M}$ contains sets that "split" every set $E \subset \Omega$ well. We call these sets $\mathbb{P}$-measurable. $\mathcal{M}$ has an equivalent definition,

$$\mathcal{M} = \{A \subset \Omega | \mathbb{P}(A \cap E) + \mathbb{P}(A^c \cap E) \leq \mathbb{P}(E), \ \forall E \subset \Omega\},$$

since $E = (A \cap E) \cup (A^c \cap E)$ and the countable subadditivity of $\mathbb{P}$ dictates that $\mathbb{P}(A \cap E) + \mathbb{P}(A^c \cap E) \geq \mathbb{P}(E)$. To prove that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{M})$, where $\mathcal{M}$ is a σ-field of $\mathbb{P}$-measurable sets in $\mathcal{F}$. We first establish:

- **Lemma 1.** If $A_1, A_2, \ldots \in \mathcal{M}$ are disjoint, then $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.
Proof: First note that

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1 \cap (A_1 \cup A_2)) + \mathbb{P}(A_1^c \cap (A_1 \cup A_2)) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

Induction thus obtains finite additivity. Now for any $m \in \mathbb{N}$, we have by monotonicity,

$$\sum_{n \leq m} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n \leq m} A_n\right) \leq \mathbb{P}\left(\bigcup_n A_n\right).$$

Since m is arbitrarily chosen, we have $\sum_n \mathbb{P}(A_n) \leq \mathbb{P}(\bigcup_n A_n)$. Combining this with subadditivity, we obtain Lemma 1. Next we prove that $\mathcal{M}$ is a field.

- **Lemma 2.** $\mathcal{M}$ is a field on Ω.
Proof: It is trivial that $\Omega \in \mathcal{M}$ and that $A \in \mathcal{M} \Rightarrow A^c \in \mathcal{M}$. It remains to prove that $A, B \in \mathcal{M} \Rightarrow A \cap B \in \mathcal{M}$. We first write,

$$(A \cap B)^c = (A^c \cap B) \cup (A \cap B^c) \cup (A^c \cap B^c).$$

Then

$$\begin{aligned}
& \mathbb{P}\left((A \cap B) \cap E\right) + \mathbb{P}\left((A \cap B)^c \cap E\right) \\
= \ & \mathbb{P}\left(A \cap B \cap E\right) + \mathbb{P}\left\{[(A^c \cap B) \cap E] \cup [(A \cap B^c) \cap E] \cup [(A^c \cap B^c) \cap E]\right\} \\
\leq \ & \mathbb{P}\left(A \cap (B \cap E)\right) + \mathbb{P}\left(A^c \cap (B \cap E)\right) + \mathbb{P}\left(A \cap (B^c \cap E)\right) + \mathbb{P}\left(A^c \cap (B^c \cap E)\right) \\
= \ & \mathbb{P}\left(B \cap E\right) + \mathbb{P}\left(B^c \cap E\right) = \mathbb{P}\left(E\right).
\end{aligned}$$

Using the second definition of $\mathcal{M}$, we have $A \cap B \in \mathcal{M}$. Hence $\mathcal{M}$ is a field. Next we establish that $\mathcal{M}$ is a σ-field. To show this we only need to show that $\mathcal{M}$ is closed to countable union. We first prove two technical lemmas.

- **Lemma 3.** Let $A_1, A_2, \ldots \in \mathcal{M}$ be disjoint. For each $m \in \mathbb{N}$, let $B_m = \bigcup_{n \leq m} A_n$. Then for all m and $E \subset \Omega$, we have

$$\mathbb{P}\left(E \cap B_m\right) = \sum_{n \leq m} \mathbb{P}\left(E \cap A_n\right).$$

Proof: We prove by induction. First, note that the lemma holds trivially when $m = 1$. Now suppose it holds for some m, we show that $\mathbb{P}\left(E \cap B_{m+1}\right) = \sum_{n \leq m+1} \mathbb{P}\left(E \cap A_n\right)$. Note that $B_m \cap B_{m+1} = B_m$ and $B_m^c \cap B_{m+1} = A_{m+1}$. So

$$\begin{aligned}
\mathbb{P}\left(E \cap B_{m+1}\right) &= \mathbb{P}\left(B_m \cap E \cap B_{m+1}\right) + \mathbb{P}\left(B_m^c \cap E \cap B_{m+1}\right) \\
&= \mathbb{P}\left(E \cap B_m\right) + \mathbb{P}\left(E \cap A_{m+1}\right) \\
&= \sum_{n \leq m+1} \mathbb{P}\left(E \cap A_n\right).
\end{aligned}$$

- **Lemma 4.** Let $A_1, A_2, \ldots \in \mathcal{M}$ be disjoint, then $\bigcup_n A_n \in \mathcal{M}$.
Proof: For any $m \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{P}\left(E\right) &= \mathbb{P}\left(E \cap B_m\right) + \mathbb{P}\left(E \cap B_m^c\right) \\
&= \sum_{n \leq m} \mathbb{P}\left(E \cap A_n\right) + \mathbb{P}\left(E \cap B_m^c\right) \\
&\geq \sum_{n \leq m} \mathbb{P}\left(E \cap A_n\right) + \mathbb{P}\left(E \cap \left(\bigcup_n A_n\right)^c\right),
\end{aligned}$$

since $\left(\bigcup_n A_n\right)^c \subset B_m^c$. Since m is arbitrary, we have

$$\begin{aligned}
\mathbb{P}\left(E\right) &\geq \sum_n \mathbb{P}\left(E \cap A_n\right) + \mathbb{P}\left(E \cap \left(\bigcup_n A_n\right)^c\right) \\
&\geq \mathbb{P}\left(E \cap \left(\bigcup_n A_n\right)\right) + \mathbb{P}\left(E \cap \left(\bigcup_n A_n\right)^c\right).
\end{aligned}$$

Hence $\bigcup_n A_n \in \mathcal{M}$. Now we are read to prove:

- **Lemma 5.** $\mathcal{M}$ is a σ-field of subsets of Ω.
  Proof: It suffices to show if $E_1, E_2, \ldots \in \mathcal{M}$, $\bigcup_n E_n \in \mathcal{M}$. Define $A_1 = E_1$, $A_i = E_i \cap E_1^c \cap E_2^c \cap \cdots \cap E_{i-1}^c$ for $i \geq 2$. Then $A_1, A_2, \ldots \in \mathcal{M}$ are disjoint and $\bigcup_n E_n = \bigcup_n A_n \in \mathcal{M}$ by Lemma 4.

(c) We now prove $\mathcal{F}_0 \subset \mathcal{M}$.
Proof: Let $A \in \mathcal{F}_0$, we need to show that $A \in \mathcal{M}$. For any $E \subset \Omega$ and any $\epsilon > 0$, we can find a sequence of $E_1, E_2, \ldots \in \mathcal{F}_0$ such that $E \subset \bigcup_n E_n$ such that,

$$\sum_n \mathbb{P}_0(E_n) \leq \mathbb{P}(E) + \epsilon.$$

By countable additivity of $\mathbb{P}_0$ on $\mathcal{F}_0$, we have $\mathbb{P}_0(E_n) = \mathbb{P}_0(E_n \cap A) + \mathbb{P}_0(E_n \cap A^c)$. Hence

$$\begin{aligned}\sum_n \mathbb{P}_0(E_n) &= \sum_n \mathbb{P}(E_n \cap A) + \sum_n \mathbb{P}(E_n \cap A^c) \\ &\geq \mathbb{P}\left(\left(\bigcup E_n\right) \cap A\right) + \mathbb{P}\left(\left(\bigcup E_n\right) \cap A^c\right) \\ &\geq \mathbb{P}(E \cap A) + \mathbb{P}(E \cap A^c).\end{aligned}$$

Since ϵ is arbitrarily chosen, we have $\mathbb{P}(E) \geq \mathbb{P}(E \cap A) + \mathbb{P}(E \cap A^c)$. Hence $A \in \mathcal{M}$.

(d) Finally, we prove that $\mathbb{P} = \mathbb{P}_0$ on $\mathcal{F}_0$.
Proof: Let $E \in \mathcal{F}_0$. It is obvious from the definition of $\mathbb{P}$ that $\mathbb{P}(E) \leq \mathbb{P}_0(E)$. Let $A_1, A_2, \ldots \in \mathcal{F}_0$ and $E \subset \bigcup_n A_n$. Define a disjoint sequence of subsets $\{B_n\}$ such that $B_1 = A_1$ and $B_i = A_i \cap A_1^c \cap A_2^c \cap \cdots \cap A_{i-1}^c$ for $i \geq 2$. We have $B_n \subset A_n$ for all n and $\bigcup_n A_n = \bigcup_n B_n$. Using countable additivity of $\mathbb{P}_0$,

$$\mathbb{P}_0(E) = \mathbb{P}_0\left(E \cap \left(\bigcup_n B_n\right)\right) = \sum_n \mathbb{P}_0(E \cap B_n).$$

Hence

$$\mathbb{P}_0(E) \leq \sum_n \mathbb{P}_0(B_n) \leq \sum_n \mathbb{P}_0(A_n).$$

Now it is obvious that $\mathbb{P}(E) \geq \mathbb{P}_0(E)$. The proof is now complete.

1.5 Exercises

1. Prove that an arbitrary intersection of σ-fields is a σ-field.

2. Show that

$$\lim_{n\to\infty}\left(-\frac{1}{n}, 1-\frac{1}{n}\right] = [0,1).$$

3. Let $\mathbb{R}$ be the sample space. We define a sequence E_n of subsets of $\mathbb{R}$ by

$$E_n = \begin{cases} \left(-\frac{1}{n}, \frac{1}{2}-\frac{1}{n}\right] & \text{if } n \text{ is odd,} \\ \left[\frac{1}{3}-\frac{1}{n}, \frac{2}{3}+\frac{1}{n}\right) & \text{if } n \text{ is even.} \end{cases}$$

Find $\liminf E_n$ and $\limsup E_n$. Let the probability $\mathbb{P}$ be given by the Lebesgue measure on the unit interval $[0,1]$ (that is, the length of interval). Compare $\mathbb{P}(\liminf E_n)$, $\liminf \mathbb{P}(E_n)$, $\mathbb{P}(\limsup E_n)$, and $\limsup \mathbb{P}(E_n)$.

4. Prove the following:
(a) If the events E and F are independent, then so are E^c and F^c.
(b) The events Ω and $\emptyset$ are independent of any event E.
(c) In addition to Ω and $\emptyset$, is there any event that is independent of itself?

5. Show that $\sigma(\{[a,b] | \forall a \leq b,\ a,b \in \mathbb{R}\}) = \mathcal{B}(\mathbb{R})$.

Chapter 2

Random Variable

2.1 Measurable Functions

Random variables are measurable functions from Ω to $\mathbb{R}$. We first define measurable functions and examine their properties. Let $(S, \mathcal{G})$ be a general measurable space, where $\mathcal{G}$ is a σ-field on a set S. For example, $(\Omega, \mathcal{F})$ is a measurable space, on which random variables are defined.

Definition 2.1.1 (Measurable function) *A function $f : S \to \mathbb{R}$ is $\mathcal{G}$-measurable if, for any $A \in \mathcal{B}(\mathbb{R})$,*

$$f^{-1}(A) \equiv \{s \in S | f(s) \in A\} \in \mathcal{G}.$$

We simply call a function *measurable* if there is no possibility for confusion.

Remarks:

- For a $\mathcal{G}$-measurable function f, f^{-1} is a mapping from $\mathcal{B}$ to $\mathcal{G}$, while f is a mapping from S to $\mathbb{R}$.
- For some set $E \in \mathcal{G}$, the indicator function I_E is $\mathcal{G}$-measurable.
- The mapping f^{-1} preserves all set operations:

$$f^{-1}\left(\bigcup_n A_n\right) = \bigcup_n f^{-1}(A_n), \quad f^{-1}(A^c) = \left(f^{-1}(A)\right)^c, \text{ etc.}$$

  $\{f^{-1}(A) | A \in \mathcal{B}\}$ is thus a σ-field. It may be called the σ-field generated by f.

Properties:

(a) If $\mathcal{C} \subset \mathcal{B}$ and $\sigma(\mathcal{C}) = \mathcal{B}$, then $f^{-1}(A) \in \mathcal{G}\ \forall A \in \mathcal{C}$ implies that f is $\mathcal{G}$-measurable.
Proof: Let $\mathcal{E} = \{B \in \mathcal{B} | f^{-1}(B) \in \mathcal{G}\}$. By definition $\mathcal{E} \subset \mathcal{B}$. Now it suffices to show that $\mathcal{B} \subset \mathcal{E}$. First, $\mathcal{E}$ is a σ-field, since inverse mapping preserves all set operations. And since $f^{-1}(A) \in \mathcal{G}\ \forall A \in \mathcal{C}$, we have $\mathcal{C} \subset \mathcal{E}$. Hence $\sigma(\mathcal{C}) = \mathcal{B} \subset \mathcal{E}$.

(b) f is $\mathcal{G}$-measurable if

$$\{s \in S | f(s) \leq c\} \in \mathcal{G} \ \ \forall c \ \in \mathbb{R}.$$

Proof: Let $\mathcal{C} = \{(-\infty, c]\}$, apply (a).

(c) (b) also holds if we replace $f(s) \leq c$ by $f(s) \geq c$, $f(s) > c$, etc.

(d) If f is measurable and a is a constant, then af and $f + a$ are measurable.

(e) If both f and g are measurable, then $f + g$ is also measurable.
Proof: Note that we can always find a rational number $r \in (f(s), c - g(s))$ if $f(s) + g(s) < c$. We can represent

$$\{f(s) + g(s) \leq c\} = \bigcup_r \left(\{f(s) < r\} \cap \{g(s) < c - r\}\right),$$

which is in $\mathcal{G}$ for all $c \in \mathbb{R}$, since the set of rational numbers is countable.

(f) If both f and g are measurable, then fg is also measurable.
Proof: It suffices to prove that if f is measurable, then f^2 is measurable, since $fg = \left((f+g)^2 - f^2 - g^2\right)/2$. But $\{f(s)^2 \leq c\} = \{f(s) \in [-\sqrt{c}, \sqrt{c}]\} \in \mathcal{G}$ for all $c \geq 0$ and $\{f(s)^2 \leq c\} = \emptyset \in \mathcal{G}$ for $c < 0$.

(g) Let $\{f_n\}$ be a sequence of measurable functions. Then $\sup f_n$, $\inf f_n$, $\liminf f_n$, and $\limsup f_n$ are all measurable ($\sup f_n$ and $\inf f_n$ may be infinite, though, hence we should consider Borel sets on the extended real line).
Proof: Note that $\{\sup f_n(s) \leq c\} = \bigcap_n \{f_n(s) \leq c\} \in \mathcal{G}$ and $\{\inf f_n(s) \geq c\} = \bigcap_n \{f_n(s) \geq c\} \in \mathcal{G}$. Now the rest is obvious.

(h) If $\{f_n\}$ are measurable, then $\{\lim f_n \text{ exists in } \mathbb{R}\} \in \mathcal{G}$.
Proof: Note that the set on which the limit exists is

$$\{\limsup f_n < \infty\} \cap \{\liminf f_n > -\infty\} \cap g^{-1}(0),$$

where $g = \limsup f_n - \liminf f_n$ is measurable.

(i) If $\{f_n\}$ are measurable and $f = \lim f_n$ exists, then f is measurable.
Proof: Note that for all $c \in \mathbb{R}$,

$$\{f \leq c\} = \bigcap_{m \geq 1} \bigcup_{k} \bigcap_{n \geq k} \left\{ f_n \leq c + \frac{1}{m} \right\}.$$

(j) A simple function f, which takes the form $f(s) = \sum_{i=1}^{n} c_i I_{A_i}$, where $(A_i \in \mathcal{G})$ are disjoint and (c_i) are constants, is measurable.
Proof: Use (d) and (e) and the fact that indicator functions are measurable.

Definition 2.1.2 (Borel Functions) *If f is $\mathcal{B}(\mathbb{R})$-measurable, it is called* Borel *function.*

Borel functions can be more general. For example, a $\mathcal{B}(S)$-measurable function, where S is a general topological space, may be referred to as a Borel function.

- If both f and g are $\mathcal{G}$-measurable, then the composition function $g \circ f$ is $\mathcal{G}$-measurable.

- If g is a continuous real function, then g is Borel. It is well known that a real function f is continuous if and only if the inverse image of every open set is an open set. By the definition of $\mathcal{B}(\mathbb{R})$, for every $A \in \mathcal{B}(\mathbb{R})$, A can be represented by a countable union of open intervals. It is then obvious that $f^{-1}(A)$ is also in $\mathcal{B}(\mathbb{R})$.

2.2 Random Variables

Definition 2.2.1 (Random Variable) *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define a* random variable *X as a $\mathcal{F}$-measurable function from Ω to $\mathbb{R}$, ie, $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$.*

Remarks:

- A random variable X is *degenerate* if $X(\omega) = c$, a constant for all ω. For all $B \in \mathcal{B}(\mathbb{R})$, if $c \in B$, then $X^{-1}(B) = \Omega \subset \mathcal{F}$, and if $c \notin B$, then $X^{-1}(B) = \emptyset \subset \mathcal{F}$.

- From Property (b) of measurable functions, if $\{\omega \in \Omega | X(\omega) \leq c\} \in \mathcal{F}\ \forall c\ \in \mathbb{R}$, then X is a random variable.

- If X and Y are random variables defined on a same probability space, then cX, $X+c$, X^2, $X+Y$, and XY are all random variables.

- If $\{X_n\}$ is a sequence of random variables, then $\sup X_n$, $\inf X_n$, $\limsup X_n$, $\liminf X_n$, and $\lim X_n$ (if it exists), are all random variables (possibly unbounded).

- If X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and f is a Borel function, then $f(X)$ is also a random variable on the same probability space.

- The concept of random variable may be more general. For example, X may be a mapping from Ω to a separable Banach space with an appropriate σ-field.

Example 2.2.2 *For the coin tossing experiments, we may define a random variable by $X(H) = 1$ and $X(T) = 0$, where H and T are the outcomes of the experiment, ie, head and tail, respectively. If we toss the coin for n times, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{\infty} X_i$ is also a random variable. As $n \to \infty$, $\bar{X}_n$ becomes a degenerate random variable as we know by the law of large numbers. $\lim \bar{X}_n$ is still a random variable since the following event is in $\mathcal{F}$,*

$$\left\{ \frac{\textit{number of heads}}{\textit{number of tosses}} \to \frac{1}{2} \right\} = \{\limsup \bar{X}_n = 1/2\} \cap \{\liminf \bar{X}_n = 1/2\}$$

Definition 2.2.3 (Distribution of Random Variable) *The distribution P_X of a random variable X is the probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induced by X. Specifically,*

$$P_X(A) = \mathbb{P}(X^{-1}(A)) \quad \textit{for all} \quad A \in \mathcal{B}(\mathbb{R}).$$

- We may write the distribution function as a composite function $P_X = \mathbb{P} \circ X^{-1}$. When there is no ambiguity about the underlying random variable, we write P in place of P_X for simplicity.

- P is indeed a probability measure (verify this). Hence all properties of the probability measure apply to P. P is often called the *law* of a random variable X.

Definition 2.2.4 (Distribution Function) *The distribution function F_X of a random variable is defined by*

$$F_X(x) = P_X\{(-\infty, x]\} \quad \textit{for all } x \ \in \mathbb{R}.$$

We may omit the subscript of F_X for simplicity. Note that since $\{(-\infty, x],\ x \in \mathbb{R}\}$ is a π-system that generates $\mathcal{B}(\mathbb{R})$, F uniquely determines P.

Properties:

(a) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(b) $F(x) \leq F(y)$ if $x \leq y$.

(c) F is right continuous.

Proof: (a) Let $x_n \to -\infty$. Since $(-\infty, x_n] \downarrow \emptyset$, we have $F(x_n) = P\{(-\infty, x_n]\} \to P(\emptyset) = 0$. The other statement is similarly established. (b) It follows from $(-\infty, x] \subset (-\infty, y]$ if $x \leq y$. (c) Fix an x, it suffices to show that $F(x_n) \to F(x)$ for any sequence $\{x_n\}$ such that $x_n \downarrow x$. It follows, however, from the fact that $(-\infty, x_n] \downarrow (-\infty, x]$ and the monotone convergence of probability measure.

Remark: If $P(\{x\}) = 0$, we say that P does not have *point probability mass* at x, in which case F is also left-continuous. For any sequence $\{x_n\}$ such that $x_n \uparrow x$, we have

$$F(x_n) = P((-\infty, x_n]) \to P((-\infty, x)) = F(x) - P(\{x\}) = F(x).$$

2.3 Random Vectors

An n-dimensional random vector is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. We may write a random vector X as $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))'$.

Example 2.3.1 *Consider the coin tossing experiment. Define a r.v. $X(H) = 1$ and $X(T) = 0$, and another r.v. $Y(H) = 1$ and $Y(T) = 0$. We may define a random vector $Z = (X, Y)'$. Z is obviously a mapping from $\Omega = \{H, T\}$ to $\mathbb{R}^2$. Specifically,*

$$Z(H) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \textit{and} \quad Z(T) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Example 2.3.2 *Consider tossing the coin twice. Let X_1 be a random variable that takes 1 if the first toss gives Head and 0 otherwise, and let X_2 be a random variable that takes 1 if the second toss gives Head and 0 otherwise. Then the random vector $X = (X_1, X_2)'$ is a function from $\Omega = \{HH, HT, TH, TT\}$ to $\mathbb{R}^2$:*

$$X(HH) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad X(HT) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad X(TH) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad X(TT) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Definition 2.3.3 (Distribution of Random Vector) *The distribution of an n-dimensional random vector $X = (X_1, \dots, X_n)'$ is a probability measure on $\mathbb{R}^n$,*

$$P_X(A) = \mathbb{P}\{\omega | X(\omega) \in A\} \quad \forall A \in \mathcal{B}(\mathbb{R}^n).$$

The distribution of a random vector $X = (X_1, \dots, X_n)$ is conventionally called the *joint distribution* of $X_1, \dots, X_n$. The distribution of a subvector of X is called the *marginal distribution.*

The marginal distribution is a projection of the joint distribution. Consider a random vector $Z = (X', Y')'$ with two subvectors $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$. Let $P_X(A)$ be the marginal distribution of X for $A \in \mathcal{B}(\mathbb{R}^m)$. We have

$$P_X(A) = P_Z(A \times \mathbb{R}^n) = \mathbb{P}\{\omega | Z(\omega) \in A \times \mathbb{R}^n\},$$

where the cylinder set $A \times \mathbb{R}^n$ is obviously an element in $\mathcal{B}(\mathbb{R}^{m+n})$.

Definition 2.3.4 (Joint Distribution Function) *The distribution function of a random vector $X = (X_1, \dots, X_n)'$ is defined by*

$$F_X(x_1, \dots, x_n) = \mathbb{P}\{\omega | X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\}.$$

The n-dimensional real function F_X is conventionally called the joint distribution function *of $X_1, \dots, X_2$.*

2.4 Density

Let μ be a measure on $(S, \mathcal{G})$, and let f_n be a simple function of the form $f_n(s) = \sum_{k=1}^n c_k I_{A_k}$, where $(A_k \in \mathcal{G})$ are disjoint and (c_k) are real nonnegative constants. We have

Definition 2.4.1 *The Lebesgue integral of f with respect to μ by*

$$\int f d\mu = \sum_{k=1}^{m} c_k \mu(A_k).$$

For a general nonnegative function f, we have

Definition 2.4.2 *The Lebesgue integral of f with respect to μ by*

$$\int f d\mu = \sup_{\{f_n \leq f\}} \int f_n d\mu,$$

where $\{f_n\}$ are simple functions.

In words, the Lebesgue integral of a general function f is the sup of the integrals of simple functions that are below f. For example, we may choose $f_n = \alpha_n \circ f$, where

$$\alpha_n(x) = \begin{cases} 0 & f(x) = 0 \\ 2^{-n}(k-1) \quad \text{if} & 2^{-n}(k-1) < f(x) \leq 2^{-n}k, \quad \text{for } k = 1, \ldots, n2^n \\ n & f(x) > n \end{cases}$$

For functions that are not necessarily nonnegative, we define

$$\begin{aligned} f^+(x) &= \max(f(x), 0) \\ f^-(x) &= \max(-f(x), 0). \end{aligned}$$

Then we have

$$f(x) = f^+ - f^-.$$

The Lebesgue integral of f is now defined by

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu.$$

If both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite, then we call f integrable with respect to μ.

Remarks:

- The function f is called integrand. The notation $\int f d\mu$ is a simplified form of $\int_S f(x)\mu(dx)$.
- The summation $\sum_n c_n$ is a special case of Lebesgue integral, which is taken with respect to the counting measure. The counting measure on $\mathbb{R}$ assigns 1 to each point in $\mathbb{Z}$.
- The Lebesgue integral generalizes the Riemann integral. It exists and coincides with the Riemann integral whenever that the latter exists.

Definition 2.4.3 (Absolute Continuity of Measures) *Let μ and ν be two measures on $(S, \mathcal{G})$. ν is* absolutely continuous *with respect to μ if*

$$\nu(A) = 0 \quad \textit{whenever} \quad \mu(A) = 0, \quad A \in G.$$

For example, given μ, we may construct a measure ν by

$$\nu(A) = \int_A f d\mu, \quad A \in \mathcal{G},$$

where f is nonnegative. It is obvious that ν, so constructed, is absolutely continuous with respect to μ.

Theorem 2.4.4 (Radon-Nikodym Theorem) *Let μ and ν be two measures on a measurable space $(S, \mathcal{G})$. If ν is absolutely continuous with respect to μ, then there exists a nonnegative measurable function f such that ν can be represented as*

$$\nu(A) = \int_A f d\mu, \quad A \in \mathcal{G}.$$

The function f is called the Radon-Nikodym derivative of ν with respect to μ. It is uniquely determined up to μ-null sets. We may denote $f = \partial\nu/\partial\mu$.

Density Recall that P_X is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If P_X is absolutely continuous with respect to a measure μ, then there exists a nonnegative function p_X such that

$$P_X(A) = \int_A p_X d\mu, \quad \forall A \in \mathcal{B}(\mathbb{R}). \tag{2.1}$$

- If the measure μ in (2.1) is a Lebesgue measure, the function p_X is conventionally called the *probability density function* of X. If such a pdf exists, we say that X is a continuous random variable.

- If P_X is absolutely continuous with respect to the counting measure μ, then p_X is conventionally called the discrete probabilities and X is called a discrete random variable.

2.5 Independence

The independence of random variables is defined in terms of σ-fields they generate. We first define

Definition 2.5.1 (σ-field Generated by Random Variable) *Let X be a random variable. The σ-field generated by X, denoted by $\sigma(X)$, is defined by*

$$\sigma(X) = \left\{X^{-1}(A) | A \in \mathcal{B}(\mathbb{R})\right\}.$$

- $\sigma(X)$ is the smallest σ-field to which X is measurable.

- The σ-field generated by a random vector $X = (X_1, \dots, X_n)'$ is similarly defined: $\sigma(X) = \sigma(X_1, \dots, X_n) = \{X^{-1}(A) | A \in \mathcal{B}(\mathbb{R}^n)\}$.

- $\sigma(X)$ may be understood as the set of information that the random variable X contains about the state of the world. Speaking differently, $\sigma(X)$ is the collection of events E such that, for a given outcome, we can tell whether the event E has happened based on the observance of X.

Definition 2.5.2 (Independence of Random Variables) *Random variables $X_1, \ldots, X_n$ are independent if the σ-fields, $\sigma(X_1), \ldots, \sigma(X_n)$, are independent.*

Let $p(x_{i_k})$ be the Radon-Nikodym density of the distribution of X_{i_k} with respect to Lebesgue or counting measure. And let, with some abuse of notation, $p(x_{i_1}, \ldots, x_{i_n})$ be the Radon-Nikodym density of the distribution of $X_{i_1}, \ldots, X_{i_n}$, with respect to the product of the measures to which the marginal densities $p(x_{i_1}), \ldots, p(x_{i_n})$ are defined. The density p may be pdf or discrete probabilities, depending on whether the corresponding random variable is continuous or discrete. We have the following theorem.

Theorem 2.5.3 *The random variables $X_1, X_2, \ldots$ are independent if and only if for any $(i_1, \ldots, i_n)$,*

$$p(x_{i_1}, \ldots, x_{i_n}) = \prod_{k=1}^{n} p(x_{i_k})$$

almost everywhere with respect to the measure for which p is defined.

Proof: It suffices to prove the case of two random variables. Let $Z = (X, Y)'$ be a two-dimensional random vector, and let $\mu(dx)$ and $\mu(dy)$ be measure to which $p(x)$ and $p(y)$ are defined. The joint density $p(x, y)$ is then defined with respect to the measure $\mu(dx)\mu(dy)$ on $\mathbb{R}^2$. For any $A, B \in \mathcal{R}$, we have

$$P_Z(A \times B) = \mathbb{P}\{Z^{-1}(A \times B)\} = \mathbb{P}\{X^{-1}(A) \cap Y^{-1}(B)\}.$$

X and Y are independent iff

$$P_Z(A \times B) = \mathbb{P}\{X^{-1}(A) \cap Y^{-1}(B)\} = \mathbb{P}\{X^{-1}(A)\}\mathbb{P}\{Y^{-1}(B)\} = P_X(A)P_Y(B).$$

And $P_Z(A \times B) = P_X(A)P_Y(B)$ holds iff

$$\begin{aligned}\int\int_{A\times B} p(x,y)\mu(dx)\mu(dy) &= \int_A p(x)\mu(x) \int_B p(y)\mu(dy) \\ &= \int\int_{A\times B} p(x)p(y)\mu(dx)\mu(dy),\end{aligned}$$

where the second equality follows from Fubini's theorem.

2.6 Exercises

1. Verify that $P_X(\cdot) = \mathbb{P}\left(X^{-1}(\cdot)\right)$ is a probability measure on $\mathcal{B}(\mathbb{R})$.

2. Let E and F be two events with probabilities $\mathbb{P}(E) = 1/2, \mathbb{P}(F) = 2/3$ and $\mathbb{P}(E \cap F) = 1/3$. Define random variables $X = \mathrm{I}(E)$ and $Y = \mathrm{I}(F)$. Find the joint distribution of X and Y. Also, obtain the conditional distribution of X given Y.

3. If a random variable X is endowed with the following density function,

$$p(x) = \frac{x^2}{18} \,\mathrm{I}\{-3 < x < 3\},$$

compute $\mathbb{P}\{\omega||X(\omega)| < 1\}$.

4. Suppose the joint probability density function of X and Y is given by

$$p(x, y) = 3(x + y)\,\mathrm{I}\{0 \leq x + y \leq 1, 0 \leq x, y \leq 1\}.$$

(a) Find the marginal density of X.
(b) Find $\mathbb{P}\{\omega|X(\omega) + Y(\omega) < 1/2\}$.

Chapter 3

Expectations

3.1 Integration

Expectation is integration. Before studying expectation, therefore, we first dig deeper into the theory of integration.

Notations Let μ be a measure on $(S, \mathcal{G})$.

- We denote $\mu(f) = \int f d\mu$ and $\mu(f; A) = \int_A f d\mu = \int f I_A d\mu$, where $A \in \mathcal{G}$.
- We say that f is μ-integrable if $\mu(|f|) = \mu(f^+) + \mu(f^-) < \infty$, in which case we write $f \in L^1(S, \mathcal{G}, \mu)$.
- If in addition, f is nonnegative, then we write $f \in L^1(S, \mathcal{G}, \mu)^+$.
- $E \in \mathcal{G}$ is μ-null if $\mu(E) = 0$.
- A statement is said to hold almost everywhere (a.e.) if the set E on which the statement is false is μ-null.

Properties of Integration

- If $f \in L^1(S, \mathcal{G}, \mu)$, then $|\mu(f)| \leq \mu(|f|)$.
- If $f, g \in L^1(S, \mathcal{G}, \mu)$, then $af + bg \in L^1(S, \mathcal{G}, \mu)$, where $a, b \in \mathbb{R}$. Furthermore, $\mu(af + bg) = a\mu(f) + b\mu(g)$.
- $\mu(f; A)$ is a measure on $(S, \mathcal{G})$.

Theorem 3.1.1 (Monotone Convergence Theorem) *If f_n is a sequence of nonnegative measurable functions such that, except on a μ-null set, $f_n \uparrow f$, then*

$$\mu(f_n) \uparrow \mu(f).$$

Note that the monotone convergence of probability is implied by the monotone convergence theorem. Take $f_n = I_{A_n}$ and $f = I_A$, where A_n is a monotone increasing sequence of sets in $\mathcal{G}$ that converge to A, and let $\mu = \mathbb{P}$ be a probability measure. Then $\mu(f_n) = \mathbb{P}(A_n) \uparrow \mathbb{P}(A) = \mu(f)$.

Theorem 3.1.2 (Fatou's Lemma) *For a sequence of nonnegative measurable functions f_n, we have*

$$\mu(\liminf f_n) \leq \liminf \mu(f_n).$$

Proof: Note that $\inf_{n\geq k} f_n$ is monotone increasing and $\inf_{n\geq k} f_n \uparrow \liminf f_n$. In addition, since $f_k \geq \inf_{n\geq k} f_n$ for all k, we have $\mu(f_k) \geq \mu(\inf_{n\geq k} f_n) \uparrow \mu(\liminf f_n)$ by Monotone Convergence Theorem.

Theorem 3.1.3 (Reverse Fatou's Lemma) *If a sequence of nonnegative measurable functions f_n are bounded by a measurable nonnegative function g for all n and $\mu(g) < \infty$, then*

$$\mu(\limsup f_n) \geq \limsup \mu(f_n).$$

Proof: Apply Fatou Lemma to $(g - f_n)$.

Theorem 3.1.4 (Dominated Convergence Theorem) *Suppose that f_n and f are measurable, that $f_n(s) \to f(s)$ for every $s \in S$, and that (f_n) is* dominated *by some $g \in L^1(S, \mathcal{G}, \mu)^+$, ie,*

$$|f_n(s)| \leq g(s), \quad \forall s \in S, \ \forall n,$$

then

$$\mu(|f_n - f|) \to 0,$$

so that

$$\mu(f_n) \to \mu(f).$$

In addition, $f \ \in L^1(S, \mathcal{G}, \mu)$.

Proof: It is obvious that $|f(s)| \leq g(s)$ $\forall s \in S$. Hence $|f_n - f| \leq 2g$, where $\mu(2g) < \infty$. We apply the reverse Fatou Lemma to $(|f_n - f|)$ and obtain

$$\limsup \mu(|f_n - f|) \leq \mu(\limsup |f_n - f|) = \mu(0) = 0.$$

Since $|\mu(f_n) - \mu(f)| = |\mu(f_n - f)| \leq \mu(|f_n - f|)$, we have

$$\lim_{n\to\infty} |\mu(f_n) - \mu(f)| \leq \limsup \mu(|f_n - f|) = 0.$$

3.2 Expectation

Definition 3.2.1 (Expectation) *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. The* expectation *of X, $\mathbb{E}X$, is defined by*

$$\mathbb{E}X = \int X d\mathbb{P}.$$

More generally, let f be a Borel function,

$$\mathbb{E}f(X) = \int f(X)\mathbb{P}.$$

$\mathbb{E}X$ is also called the mean of X, and $\mathbb{E}f(X)$ is called the f-moment of X.

Theorem 3.2.2 (Change of Variable) *We have*

$$\mathbb{E}f(X) = \int f dP_X = \int f p_X d\mu, \tag{3.1}$$

where p_X is the density of X with respect to measure μ.

Proof: First consider indicator functions of the form $f(X) = I_A(X)$, where $A \in \mathcal{B}$. We have $f(X)(\omega) = I_A \circ X(\omega) = I_{X^{-1}(A)}(\omega)$. Then

$$\mathbb{E}f(X) = \mathbb{E}I_A \circ X = \mathbb{P}(X^{-1}(A)) = P_X(A).$$

And we have

$$P_X(A) = \int I_A dP_X = \int f dP_X \quad \text{and} \quad P_X(A) = \int I_A p_X d\mu = \int f p_X d\mu.$$

Hence the theorem holds for indicator functions. Similarly we can show that it is true for simple functions. For a general nonnegative function f, we can choose a sequence of simple functions (f_n) such that $f_n \uparrow f$. The monotone convergence theorem is then applied to obtain the same result. For general functions, note that $f = f^+ - f^-$.

All properties of integration apply to the expectation. In addition, we have the following convergence theorems.

- (Monotone Convergence Theorem) If $0 \leq X_n \uparrow X$, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$.
- (Fatou's Lemma) If $X_n \geq 0$, then $\mathbb{E}(\liminf X_n) \leq \liminf \mathbb{E}(X_n)$.

- (Reverse Fatou's Lemma) If $X_n \leq X$ for all n and $\mathbb{E}X < \infty$, then $\mathbb{E}\limsup X_n \geq \limsup \mathbb{E}X_n$.

- (Dominated Convergence Theorem) If $|X_n(\omega)| \leq Y(\omega)$ $\forall(n, \omega)$, where $\mathbb{E}Y < \infty$, then
$$\mathbb{E}(|X_n - X|) \to 0,$$
which implies that
$$\mathbb{E}X_n \to \mathbb{E}X.$$

- (Bounded Convergence Theorem) If $|X_n(\omega)| \leq K$ $\forall(n, \omega)$, where $K < \infty$ is a constant, then
$$\mathbb{E}(|X_n - X|) \to 0.$$

3.3 Moment Inequalities

Definitions: Moments Let X and Y be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Recall that we call $\mathbb{E}f(X)$ the f-moment of X. In particular, if $f(x) = x^k$, $\mu_k \equiv \mathbb{E}X^k$ is called the k-th *moment* of X. If $f(x) = (x - \mu_1)^k$, we call $\mathbb{E}(X - \mu_1)^k$ the k-th *central moment* of X. Particularly, the second central moment is called the *variance*.

The *covariance* of X and Y is defined as
$$\text{cov}(X, Y) = \mathbb{E}(X - \mu_x)(Y - \mu_y),$$
where μ_x and μ_y are the means of X and Y, respectively. $\text{cov}(X, X)$ is of course the variance of X. Let σ_X^2 and σ_Y^2 denote the variances of X and Y, respectively, we define the *correlation* of X and Y by
$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

For a random vector $X = (X_1, \ldots, X_n)'$, the second moment is given by $\mathbb{E}XX'$, a symmetric matrix. Let $\mu = \mathbb{E}X$, then $\Sigma_X = \mathbb{E}(X - \mu)(X - \mu)'$ is called the variance-covariance matrix, or simply the covariance matrix. If $Y = AX$, where A is a conformable constant matrix, then $\Sigma_Y = A\Sigma_X A'$. This relation reduces to $\sigma_Y^2 = a^2\sigma_X^2$, if X and Y are scalar random variables and $Y = aX$, where a is a constant.

The moments of a random variable X contain the same information as the distribution (or the law) dose. We have

Theorem 3.3.1 *Let X and Y be two random variables (possibly defined on different probability spaces). Then $P_X = P_Y$ if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$ for all Borel functions whenever the expectation is finite.*

Proof: If $P_X = P_Y$, then we have $\mathbb{E}f(X) = \mathbb{E}f(Y)$ by (3.1). Conversely, set $f = I_B$, where B is any Borel set. Then $\mathbb{E}f(X) = \mathbb{E}f(Y)$ implies that $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$, ie, $P_X = P_Y$.

In the following, we prove a set of well-known inequalities.

Theorem 3.3.2 (Chebyshev Inequality) $\mathbb{P}\{|X| \geq \varepsilon\} \leq \frac{\mathbb{E}|X|^k}{\varepsilon^k}$, *for any $\varepsilon > 0$ and $k > 0$.*

Proof: It follows from the fact that $\varepsilon^k I_{|X| \geq \varepsilon} \leq |X|^k$.

Remarks:

- We have as a special case of the Chebyshev's inequality,

$$\mathbb{P}\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2},$$

where μ and σ^2 are the mean and the variance of X, respectively. If a random variable has a finite variance, this inequality states that it's tail probabilities are bounded.

- Another special case concerns nonnegative random variables. In this case, we have *Markov's Inequality*, which states that for a nonnegative random variable X,

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}X, \text{ for all } a > 0.$$

Theorem 3.3.3 (Cauchy-Schwartz Inequality) $(\mathbb{E}XY)^2 \leq (\mathbb{E}X^2)(\mathbb{E}Y^2)$

Proof: Without loss of generality, we consider the case when $X \geq 0$, $Y \geq 0$. Note first that if $\mathbb{E}(X^2) = 0$, then $X = 0$ a.s., in which case the inequality holds with equality. Now we consider the case when $\mathbb{E}(X^2) > 0$ and $\mathbb{E}(Y^2) > 0$. Let

$X_* = X/\left(\mathbb{E}(X^2)\right)^{1/2}$ and $Y_* = Y/\left(\mathbb{E}(Y^2)\right)^{1/2}$. Then we have $\mathbb{E}X_*^2 = \mathbb{E}Y_*^2 = 1$. Then we have

$$0 \leq \mathbb{E}(X_* - Y_*)^2 = \mathbb{E}(X_*^2 + Y_*^2 - 2X_*Y_*) = 1 + 1 - 2\mathbb{E}(X_*Y_*),$$

which results in $\mathbb{E}(X_*Y_*) \leq 1$. The Cauchy-Schwartz inequality then follows.

Remarks:

- It is obvious that equality holds only when Y is a linear function of X.
- If we apply Cauchy-Schwartz Inequality to $X - \mu_X$ and $Y - \mu_Y$, then we have

$$\text{cov}(X, Y)^2 \leq \text{var}(X)\text{var}(Y).$$

To introduce Jensen's inequality, recall that $f : \mathbb{R} \to \mathbb{R}$ is convex if $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$, where $\alpha \in [0, 1]$. If f is twice differentiable, then f is convex if and only if $f'' \geq 0$. Finally, if f is convex, it is automatically continuous.

Theorem 3.3.4 (Jensen's Inequality) *If f is convex, then $f(\mathbb{E}X) \leq \mathbb{E}f(X)$.*

Proof: Since f is convex, there exists a linear function ℓ such that

$$\ell \leq f \quad \text{and} \quad \ell(\mathbb{E}X) = f(\mathbb{E}X).$$

It follows that

$$\mathbb{E}f(X) \geq \mathbb{E}\ell(X) = \ell(\mathbb{E}X) = f(\mathbb{E}X).$$

Remarks:

- Functions such as $|x|$, x^2, and $\exp(\theta x)$ are all convex functions of x.
- The inequality is reversed for concave functions such as $log(x)$, $x^{1/2}$, etc.

Definition 3.3.5 (L^p Norm) *Let $1 \leq p < \infty$. The L^p norm of a random variable X is defined by*

$$\|X\|_p \equiv \left(\mathbb{E}|X|^p\right)^{1/p}.$$

Note that $L^p \equiv L^p(\Omega, \mathcal{F}, \mathbb{P})$ denotes a normed space of random variables that satisfies $\mathbb{E}|X|^p < \infty$.

Theorem 3.3.6 (Monotonicity of L^p Norms) *If $1 \leq p \leq q < \infty$ and $X \in L^q$, then $X \in L^p$, and*

$$\|X\|_p \leq \|X\|_q$$

Proof: Define $Y_n = \{\min(|X|, n)\}^p$. For any $n \in \mathbb{N}$, Y_n is bounded, hence both Y_n and $Y_n^{q/p}$ are in L^1. Since $x^{q/p}$ is a convex function of x, we use Jensen's inequality to obtain

$$(\mathbb{E}Y_n)^{q/p} \leq \mathbb{E}\left(Y_n^{q/p}\right) = \mathbb{E}\left(\{\min(|X|, n)\}^q\right) \leq \mathbb{E}\left(|X|^q\right).$$

Now the monotone convergence theorem obtains the desired result.

3.4 Conditional Expectation

Let X be a random variable on $L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subset \mathcal{F}$ be a sub-σ-field.

Definition 3.4.1 (Conditional Expectation) *The* conditional expectation *of X given $\mathcal{G}$, denoted by $\mathbb{E}(X|\mathcal{G})$, is a $\mathcal{G}$-measurable random variable such that for every $A \in \mathcal{G}$,*

$$\int_A \mathbb{E}(X|\mathcal{G}) d\mathbb{P} = \int_A X d\mathbb{P}. \tag{3.2}$$

In particular, if $\mathcal{G} = \sigma(Y)$, where Y is a random variable, we write $\mathbb{E}(X|\sigma(Y))$ simply as $\mathbb{E}(X|Y)$.

The conditional expectation is a local average. To see this, let $\{F_k\}$ be a partition of Ω with $\mathbb{P}(F_k) > 0$ for all k. Let $\mathcal{G} = \sigma(\{F_k\})$. According to the definition in (3.2), we have

$$\int_{F_k} \mathbb{E}(X|\mathcal{G}) d\mathbb{P} = \mathbb{E}(X|\mathcal{G})\mathbb{P}(F_k) = \int_{F_k} X d\mathbb{P}.$$

Thus $\mathbb{E}(X|\mathcal{G})$ can be written as

$$\mathbb{E}(X|\mathcal{G}) = \sum_k c_k I_{F_k},$$

where

$$c_k = \frac{\int_{F_k} X d\mathbb{P}}{\mathbb{P}(F_k)}.$$

The conditional expectation $\mathbb{E}(X|\mathcal{G})$ may be viewed as a random variable that takes values that are local averages of X over the partitions made by $\mathcal{G}$. If $\mathcal{G}_1 \subset \mathcal{G}$, $\mathcal{G}$ is said to be "finer" than $\mathcal{G}_1$. In other words, $\mathbb{E}(X|\mathcal{G})$ is more "random" than $\mathbb{E}(X|\mathcal{G}_1)$, since the former can take more values. Example 1 gives two extreme cases.

Example 3.4.2 *If $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$, which is a degenerate random variable. If $\mathcal{G} = \mathcal{F}$, then $\mathbb{E}(X|\mathcal{G}) = X$.*

Example 3.4.3 *Let E and F be two events that satisfy $\mathbb{P}(E) = \mathbb{P}(F) = 1/2$ and $\mathbb{P}(E \cap F) = 1/3$. E and F are obviously not independent. We define two random variables, $X = I_E$ and $Y = I_F$. It is obvious that $\{F, F^c\}$ is a partition of Ω and $\sigma(\{F, F^c\}) = \sigma(Y) = \{\emptyset, \Omega, F, F^c\}$. The conditional expectation of $\mathbb{E}(X|Y)$ may be written as*

$$\mathbb{E}(X|Y) = c_1^* I_F + c_2^* I_{F^c},$$

where $c_1^* = \mathbb{P}(F)^{-1} \int_F X\mathbb{P} = \mathbb{P}(F)^{-1}\mathbb{P}(F \cap E) = 2/3$, and $c_2^* = \mathbb{P}(F^c)^{-1} \int_{F^c} X\mathbb{P} = \mathbb{P}(F^c)^{-1}\mathbb{P}(F^c \cap E) = 1/3$.

Existence of Conditional Expectation Note that

$$\mu(A) = \int_A X d\mathbb{P}, \quad A \in \mathcal{G}$$

defines a measure on $(\Omega, \mathcal{G})$ and that μ is absolutely continuous with respect to $\mathbb{P}$. By the Radon-Nikodym theorem, there exists a $\mathcal{G}$-measurable random variable Y such that

$$\mu(A) = \int_A Y d\mathbb{P}.$$

The random variable Y is exactly $\mathbb{E}(X|\mathcal{G})$. It is unique up to $\mathbb{P}$-null sets.

Definition 3.4.4 (Conditional Probability) *The conditional probability may be defined as a random variable $\mathbb{P}(E|\mathcal{G})$ such that*

$$\int_A \mathbb{P}(E|\mathcal{G}) d\mathbb{P} = \mathbb{P}(A \cap E).$$

Check that the conditional probability behaves like ordinary probabilities, in that it satisfies the axioms of the probability, at least in a.s. sense.

Properties:

- (Linearity) $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$.
- (Law of Iterative Expectation) The definition of conditional expectation directly implies $\mathbb{E}X = \mathbb{E}\left[\mathbb{E}(X|\mathcal{G})\right]$.

- If X is $\mathcal{G}$-measurable, then $\mathbb{E}(XY|\mathcal{G}) = X\mathbb{E}(Y|\mathcal{G})$ with probability 1.

  Proof: First, $X\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable. Now let $X = I_F$, where $F \in \mathcal{G}$. For any $A \in \mathcal{G}$, we have

  $$\int_A \mathbb{E}(I_F Y|\mathcal{G}) d\mathbb{P} = \int_A I_F Y d\mathbb{P} = \int_{A\cap F} Y d\mathbb{P} = \int_{A\cap F} \mathbb{E}(Y|\mathcal{G}) d\mathbb{P} = \int_A I_F \mathbb{E}(Y|\mathcal{G}) d\mathbb{P}.$$

  Hence the statement holds for $X = I_F$. For general random variables, use linearity and monotone convergence theorem.

- Using the above two results, it is trivial to show that X and Y are independent if and only if $\mathbb{E}f(X)g(Y) = \mathbb{E}f(X)\mathbb{E}g(Y)$ for all Borel functions f and g.

- Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be sub-σ-fields and $\mathcal{G}_1 \subset \mathcal{G}_2$. Then, with probability 1,

  $$\mathbb{E}\left[\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1\right] = \mathbb{E}(X|\mathcal{G}_1).$$

  Proof: It follows from, for any $A \in \mathcal{G}_1 \subset \mathcal{G}_2$,

  $$\int_A \mathbb{E}\left[\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1\right] d\mathbb{P} = \int_A \mathbb{E}(X|\mathcal{G}_2) d\mathbb{P} = \int_A X d\mathbb{P} = \int_A \mathbb{E}(X|\mathcal{G}_1) d\mathbb{P}.$$

- (Doob-Dynkin) There exists a measurable function f such that $\mathbb{E}(X|Y) = f(Y)$.

Conditional Expectation as Projection The last property implies that

$$\mathbb{E}\left[\mathbb{E}(X|\mathcal{G})|\mathcal{G}\right] = \mathbb{E}(X|\mathcal{G}),$$

which suggest that the conditional expectation is a projection operator, projecting a random variable onto a sub-σ-field. This is indeed the case. It is well known that $H = L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with inner product defined by $\langle X, Y\rangle = \mathbb{E}XY$, where $X, Y \in L^2$. Consider a subspace $H_0 = L^2(\Omega, \mathcal{G}, \mathbb{P})$, where $\mathcal{G} \subset \mathcal{F}$. The projection theorem in functional analysis guarantees that for any random variable $X \in H$, there exists a $\mathcal{G}$-measurable random variable Y such that

$$\mathbb{E}(X - Y)W = 0 \text{ for all } W \in H_0. \tag{3.3}$$

Y is called the (orthogonal) projection of X on H_0. Write $W = I_A$ for any $A \in \mathcal{G}$, the equation (3.3) implies that

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P} \text{ for all } A \in \mathcal{G}.$$

It follows that Y is indeed a version of $\mathbb{E}(X|\mathcal{G})$.

Conditional Expectation as the Best Predictor Consider the problem of predicting Y given X. We call $\phi(X)$ a *predictor*, where ϕ is a Borel function. We have the following theorem,

Theorem 3.4.5 *If $Y \in L^2$, then $\mathbb{E}(Y|X)$ solves the following problem,*

$$\min_{\phi} \mathbb{E}(Y - \phi(X))^2.$$

Proof: We have

$$\begin{aligned} \mathbb{E}(Y - \phi(X))^2 &= \mathbb{E}([Y - \mathbb{E}(Y|X)] + [\mathbb{E}(Y|X) - \phi(X)])^2 \\ &= \mathbb{E}\left\{[Y - \mathbb{E}(Y|X)]^2 + [\mathbb{E}(Y|X) - \phi(X)]^2 \right. \\ &\qquad \left. +2[Y - \mathbb{E}(Y|X)][\mathbb{E}(Y|X) - \phi(X)]\right\}. \end{aligned}$$

By the law of iterative expectation, $\mathbb{E}[Y - \mathbb{E}(Y|X)][\mathbb{E}(Y|X) - \phi(X)] = 0$. Hence

$$\mathbb{E}(Y - \phi(X))^2 = \mathbb{E}[Y - \mathbb{E}(Y|X)]^2 + \mathbb{E}[\mathbb{E}(Y|X) - \phi(X)]^2.$$

Since ϕ only appears in the second term, the minimum of which is attained when $\mathbb{E}(Y|X) = \phi(X)$, it is now clear that $\mathbb{E}(Y|X)$ minimizes $\mathbb{E}(Y - \phi(X))^2$.

Hence the conditional expectation is the best predictor in the sense of minimizing mean squared forecast error (MSFE). This fact is the basis of regression analysis and time series forecasting.

3.5 Conditional Distribution

Suppose that X and Y are two random variables with joint density $p(x, y)$.

Definition 3.5.1 (Conditional Density) *The* conditional density *of X given $Y = y$ is obtained by*

$$p(x|y) = \frac{p(x, y)}{\int p(x, y)\mu(dx)}.$$

The conditional expectation $\mathbb{E}(X|Y = y)$ may then be represented by

$$\mathbb{E}(X|Y = y) = \int xp(x|y)\mu(dx).$$

For any Borel function f such that $f(X) \in L^1$, we may show that $\mathbb{E}(f(X)|Y = y) = \int f(x)p(x|y)\mu(dx)$ solves the following problem ,

$$\min_{\phi} \int\int (\phi(y) - f(x))^2 p(x,y)\mu(dx)\mu(dy).$$

It is clear that $\mathbb{E}(X|Y = y)$ is a deterministic function of y. Thus we write $g(y) = \mathbb{E}(X|Y = y)$. Recall that $\mathbb{E}(X|Y)$ is a Borel function of Y. Indeed, here we have

$$g(Y) = \mathbb{E}(X|Y).$$

To show this, first note that for all $F \in \sigma(Y)$, there exists $A \in \mathcal{B}$ such that $F = Y^{-1}(A)$. We now have

$$\begin{aligned}
\int_F g(Y)d\mathbb{P} &= \int_A g(y)p(y)\mu(dy) \\
&= \int_A \left(\int xp(x|y)\mu(dx)\right) p(y)\mu(dy) \\
&= \int_{\mathbb{R}} \int_{\times A} xp(x|y)p(y)\mu(dx)\mu(dy) \\
&= \int_F X d\mathbb{P} \\
&= \int_F \mathbb{E}(X|Y)d\mathbb{P}.
\end{aligned}$$

Example 3.5.2 *If $p(x,y) = (x+y)I_{\{0 \le x,y \le 1\}}$. To obtain $\mathbb{E}(X|Y)$, we calculate*

$$\mathbb{E}(X|Y = y) = \int xp(x|y)dx = \int_0^1 x\frac{x+y}{\frac{1}{2}+y}dx = \frac{\frac{1}{3}+y}{\frac{1}{2}+y}.$$

Then $\mathbb{E}(X|Y) = (1/3 + Y/2)/(1/2 + Y)$.

3.6 Exercises

1. Let the sample space $\Omega = \mathbb{R}$ and the probability $\mathbb{P}$ on Ω be given by

$$\mathbb{P}\left\{\frac{1}{3}\right\} = \frac{1}{3} \quad \text{and} \quad \mathbb{P}\left\{\frac{2}{3}\right\} = \frac{2}{3}.$$

Define a sequence of random variables by

$$X_n = \left(3 - \frac{1}{n}\right) \mathrm{I}(A_n) \quad \text{and} \quad \mathrm{X = 3\ I\left(\lim_{n\to\infty} A_n\right)},$$

where

$$A_n = \left[\frac{1}{3} + \frac{1}{n}, \frac{2}{3} + \frac{1}{n}\right)$$

for $n = 1, 2, \ldots$.

(a) Show that $\lim_{n\to\infty} A_n$ exists so that X is well defined.

(b) Compare $\lim_{n\to\infty} \mathbb{E}(X_n)$ with $\mathbb{E}(X)$.

(c) Is it true that $\lim_{n\to\infty} \mathbb{E}(X_n - X)^2 = 0$?

2. Let X_1 and X_2 be two zero-mean random variables with correlation ρ. Suppose the variances of X_1 and X_2 are the same, say σ^2. Prove that

$$\mathbb{P}\left(|X_1 + X_2| \geq k\sigma\right) \leq \frac{2(1+\rho)}{k^2}.$$

3. Prove *Cantelli's inequality*, which states that if a random variable X has mean μ and variance $\sigma^2 < \infty$, then for all $a > 0$,

$$\mathbb{P}(X - \mu \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

[Hint: You may first show $\mathbb{P}(X - \mu \geq a) \leq \mathbb{P}\left((X - \mu + y)^2 \geq (a + y)^2\right)$, use Markov's inequality, and then minimize the resulting bound over the choice of y.]

4. Let the sample space $\Omega = [0, 1]$ and the probability on Ω be given by the density

$$p(x) = 2x$$

over $[0, 1]$. We define random variables X and Y by

$$X(\omega) = \begin{cases} 1, & 0 \leq \omega < 1/4, \\ 0, & 1/4 \leq \omega < 1/2, \\ -1, & 1/2 \leq \omega < 3/4, \\ 0, & 3/4 \leq \omega \leq 1, \end{cases} \quad \text{and} \quad \mathrm{Y}(\omega) = \begin{cases} 1, & 0 \leq \omega < 1/2, \\ 0, & 1/2 \leq \omega \leq 1. \end{cases}$$

(a) Find the conditional expectation $\mathbb{E}(X^2|Y)$

(b) Show that $\mathbb{E}(\mathbb{E}(X^2|Y)) = \mathbb{E}(X^2)$.

Chapter 4

Distributions and Transformations

4.1 Alternative Characterizations of Distribution

4.1.1 Moment Generating Function

Let X be a random variable with density p. The moment generating function (MGF) of X is given by

$$m(t) = \mathbb{E}\exp(tX) = \int \exp(tx)p(x)d\mu(x).$$

Note that the moment generating function is the Laplace transform of the density. The name of MGF is due to the fact that

$$\frac{d^k m}{dt^k}(0) = \mathbb{E}X^k.$$

4.1.2 Characteristic Function

The MGF may not exist, but we can always define characteristic function, which is given by

$$\phi(t) = \mathbb{E}\exp(itX) = \int \exp(itx)p(x)d\mu(x).$$

Note that the characteristic function is the Fourier transform of the density. Since $|\exp(itx)|$ is bounded, $\phi(t)$ is always defined.

4.1.3 Quantile Function

We define the τ-quantile or fractile of X (with distribution function F) by

$$Q_\tau = \inf\{x|F(x) \geq \tau\}, \quad 0 < \tau < 1.$$

In particular, if $\tau = 1/2$, $Q_{1/2}$ is conventionally called the median of X.

4.2 Common Families of Distributions

In the following we get familiar with some families of distributions that are frequently used in practice. Given a family of distributions $\{P_\theta\}$ indexed by θ, we call the index θ *parameter*. If θ is finite dimensional, we call $\{P_\theta\}$ a parametric family of distributions.

Uniform The uniform distribution is a continuous distribution with the following density with respect to the Lebesgue measure,

$$p_{a,b}(x) = \frac{1}{b-a} I_{[a,b]}(x), \quad a < b.$$

We denote the uniform distribution with parameters a and b by Uniform(a, b).

Bernoulli The Bernoulli distribution is a discrete distribution with the following density with respect to the counting measure,

$$p_\theta(x) = \theta^x (1-\theta)^{1-x}, \quad x \in \{0, 1\}, \text{ and } \theta \in [0, 1].$$

The Bernoulli distribution, denoted by Bernoulli(θ), usually describes random experiments with binary outcomes such as success $(x = 1)$ or failure $(x = 0)$. The parameter θ is then interpreted as the probability of success, $\mathbb{P}\{x = 1\}$.

Binomial The Binomial distribution, corresponding to n-consecutive coin tossing, is a discrete distribution with the following density with respect to counting measure,

$$p_{n,\theta}(x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}, \ x \in \{0, 1, \dots, n\}.$$

We may use Binomial distribution, denoted by Binomial(n, θ), to describe the outcomes of repeated trials, in which case n is the number of trials and θ is the probability of success for each trial.

Note that if $X \sim$ Binomial(n, θ), it can be represented by a sum of n i.i.d. (independently and identically distributed) Bernoulli(θ) random variables.

Poisson The Poisson distribution is a discrete distribution with the following density,

$$p_\lambda(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad x \in \{0, 1, 2, \ldots\}.$$

The Poisson distribution typically describes the probability of the number of events occurring in a fixed period of time. For example, the number of phone calls in a given time interval may be modeled by a Poisson(λ) distribution, where the parameter λ is the expected number of calls. Note that the Poisson(λ) density is a limiting case of the Binomial($n, \lambda/n$) density,

$$\binom{n}{x}\left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{n!}{(n-x)!n^x}\left(1 - \frac{\lambda}{n}\right)^{-x}\left(1 - \frac{\lambda}{n}\right)^n \frac{\lambda^x}{x!} \to e^{-\lambda}\frac{\lambda^x}{x!}.$$

Normal The normal (or Gaussian) distribution, denoted by $N(\mu, \sigma^2)$ is a continuous distribution with the following density with respect to Lebesgue measure,

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The parameter μ and σ^2 are the mean and the variance of the distribution, respectively. In particular, $N(0, 1)$ is called standard normal. The normal distribution was invented for the modeling of observation error, and is now the most important distribution in probability and statistics.

Exponential The exponential distribution, denoted by Exponential(λ) is a continuous distribution with the following density with respect to Lebesgue measure,

$$p_\lambda(x) = \lambda e^{-\lambda x}.$$

The cdf of the Exponential(λ) distribution is given by

$$F(x) = 1 - e^{-\lambda x}.$$

The exponential distribution typically describes the waiting time before the arrival of next Poisson event.

Gamma The Gamma distribution, denoted by Gamma(k, λ) is a continuous distribution with the following density,

$$p_{k,\lambda} = \frac{1}{\Gamma(k)}(\lambda x)^{k-1} e^{-\lambda x}, \quad x \in [0, \infty),$$

where $\Gamma(\cdot)$ is gamma function defined as follows,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

The parameter k is called shape parameter and $\lambda > 0$ is called scale parameter.

- Special cases
  - Let $k = 1$, then Gamma$(1, \lambda)$ reduces to Exponential(λ).
  - If k is an integer, Gamma$(1, \lambda)$ reduces to an Erlang distribution, i.e., the sum of k independent exponentially distributed random variables, each of which has a mean of λ.
  - Let ℓ be an integer and $\lambda = 1/2$, then Gamma$(\ell/2, 1/2)$ reduces to χ^2_ℓ, chi-square distribution with ℓ degrees of freedom.
- The gamma function generalizes the factorial function. To see this, note that $\Gamma(1) = 1$ and that by integration by parts, we have

  $$\Gamma(z+1) = z\Gamma(z).$$

  Hence for positive integer n, we have $\Gamma(n+1) = n!$.

Beta The Beta distribution, denoted by Beta(a, b), is a continuous distribution on $[0, 1]$ with the following density,

$$p_{a,b}(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad x \in [0,1],$$

where $B(a, b)$ is the beta function defined by

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx, \quad a, b > 0.$$

Both $a > 0$ and $b > 0$ are shape parameters. Since the support of Beta distributions is $[0, 1]$, it is often used to describe unknown probability value such as the probability of success in a Bernoulli distribution.

- The beta function is related to the gamma function by

  $$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

- Beta(a, b) reduces to Uniform$[0, 1]$ if $a = b = 1$.

Table 4.1: Mean, Variance, and Moment Generating Function

Distribution	Mean	Variance	MGF
Uniform$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t}$
Bernoulli(θ)	θ	$\theta(1-\theta)$	$(1-\theta)+\theta e^t$
Poisson(λ)	λ	λ	$\exp(\lambda(e^t-1))$
Normal(μ,σ^2)	μ	σ^2	$\exp\left(\mu t+\frac{1}{2}\sigma^2t^2\right)$
Exponential(λ)	λ^{-1}	λ^{-2}	$(1-t/\lambda)^{-1}$
Gamma(k,λ)	k/λ	k/λ^2	$(\lambda/(\lambda-t))^k$
Beta(a,b)	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{a+r}{a+b+r}\right)\frac{t^k}{k!}$

Cauchy The Cauchy distribution, denoted by Cauchy(a,b), is a continuous distribution with the following density,

$$p_{a,b}(x)=\frac{1}{\pi b\left(1+\left(\frac{x-a}{b}\right)^2\right)},\quad b>0.$$

The parameter a is called the location parameter and b is called the scale parameter. Cauchy$(0,1)$ is called the standard Cauchy distribution, which coincides with Student's t-distribution with one degree of freedom.

- The Cauchy distribution is a heavy-tail distribution. It does not have any finite moment. Its mode and median are well defined and are both equal to a.
- When U and V are two independent standard normal random variables, then the ratio U/V has the standard Cauchy distribution.
- Like normal distribution, Cauchy distribution is (strictly) stable, ie, if X_1, X_2, X are i.i.d. Cauchy, then for any constants a_1 and a_2, the random variable $a_1X_1+a_2X_2$ has the same distribution as cX with some constants c.

Multinomial The multinomial distribution generalizes the binomial distribution to describe more than two categories. Let $X=(X_1,\ldots,X_m)$. For the experiment of tossing a coin for n times, X would take $(k,n-k)'$, ie, there are k heads and $n-k$ tails. For the experiment of rolling a die for n times, X would take $(x_1,...,x_m)$, where $\sum_{k=1}^m x_k=n$. The multinomial density is given by

$$p(x_1,\ldots,x_m;p_1,...,p_m)=\frac{n!}{x_1!\cdots x_m!}p_1^{x_1}\cdots p_m^{x_m},\quad x\in\{0,1,\ldots,n\},\ \sum_{k=1}^m x_k=n,$$

where parameter $(p_k, k = 1, \ldots, m)$ is the probability of getting $k-$th outcome in each coin tossing or die rolling. When $m = 2$, the multinomial distribution reduces to binomial distribution. The continuous analogue of multinomial distribution is multivariate normal distribution.

4.3 Transformed Random Variables

In this section, we study three commonly used techniques to derive the distributions of transformed random variables $Y = g(X)$, given the distribution of X. We denote by F_X the distribution function of X.

4.3.1 Distribution Function Technique

By the definition of distribution function, we may directly calculate $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.

Example 4.3.1 *Let $X \sim$ Uniform$[0,1]$ and $Y = -\log(1-X)$. It is obvious that, in a.s. sense, $Y \geq 0$ and $1 - \exp(-Y) \in [0,1]$. Thus, for $y \geq 0$, the distribution function of Y is given by*

$$\begin{aligned} F_Y(y) &= \mathbb{P}(1 - \log(1-X) \leq y) \\ &= \mathbb{P}(X \leq 1 - e^{-y}) \\ &= 1 - e^{-y}, \end{aligned}$$

since $F_X(x) = x$ for $x \in [0,1]$. $F_Y(y) = 0$ for $y < 0$. Note that $Y \sim$ Exponential(1).

Example 4.3.2 *Let X_i be independent random variables with distribution function F_i, $i = 1, \ldots, n$. Then the distribution of $Y = \max\{X_1, \ldots, X_n\}$ is given by*

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_i \leq y\}\right) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leq y) \\ &= \prod_{i=1}^{n} F_i(y). \end{aligned}$$

Example 4.3.3 *Let $X = (X_1, X_2)'$ be a random vector with distribution P and density $p(x_1, x_2)$ with respect to measure μ. Then the distribution of $Y = X_1 + X_2$ is given by*

$$\begin{aligned} F_Y(y) &= \mathbb{P}\{X_1 + X_2 \le y\} \\ &= P\{(x_1, x_2)|x_1 + x_2 \le y\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y - x_2} p(x_1, x_2)\mu(dx_1)\mu(dx_2). \end{aligned}$$

4.3.2 MGF Technique

The moment generating function (MGF) uniquely determines distributions. When MGF of $Y = g(X)$ is easily obtained, we may identify the distribution of Y by writing the MGF into a form that corresponds to some particular distribution. For example, if (X_i) are independent random variables with MGF m_i, then the MGF of $Y = \sum_{i=1}^n X_i$ is given by

$$m(t) = \mathbb{E}e^{t(X_1 + \cdots + X_n)} = \prod_{i=1}^n m_i(t).$$

Example 4.3.4 *Let $X_i \sim$ Poisson(λ_i) be independent over i. Then the MGF of $Y = \sum_{i=1}^n X_i$ is*

$$m(t) = \prod_{i=1}^n \exp\left(\lambda_i\left(e^t - 1\right)\right) = \exp\left(\left(e^t - 1\right)\sum_{i=1}^n \lambda_i\right).$$

This suggests that $Y \sim$ Poisson$(\sum_i \lambda_i)$.

Example 4.3.5 *Let $X_i \sim N(\mu_i, \sigma_i^2)$ be independent over i. Then the MGF of $Y = \sum_{i=1}^n c_i X_i$ is*

$$m(t) = \prod_{i=1}^n \exp\left(c_i\mu_i t + \frac{1}{2}c_i^2\sigma_i^2 t^2\right) = \exp\left(t\sum_{i=1}^n c_i\mu_i + \frac{t^2}{2}\sum_{i=1}^n c_i^2\sigma_i^2\right).$$

This suggests that $Y \sim N(\sum_i c_i\mu_i, \sum_{i=1}^n c_i^2\sigma_i^2)$.

4.3.3 Change-of-Variable Transformation

If the transformation function g is one-to-one, we may find the density of $Y = g(X)$ from that of X by the change-of-variable transformation. Let $g = (g_1, \ldots, g_n)'$

and $x = (x_1, \ldots, x_n)'$. And let P_X and P_Y denote the distributions of X and Y, respectively. Assume P_X and P_Y admit density p_X and p_Y with respect to μ, the counting or the Lebesgue measure on $\mathbb{R}^n$.

For any $B \in \mathcal{B}(\mathbb{R})$, we define $A = g^{-1}(B)$. We have $A \in \mathcal{B}(\mathbb{R})$ since g is measurable. It is clear that $\{X \in A\} = \{Y \in B\}$. We therefore have

$$P_Y(B) = P_X(A) = \int_A p_X(x)\mu(dx).$$

If μ is counting measure, we have

$$\int_A p_X(x)\mu(dx) = \sum_{x\in A} p_X(x) = \sum_{y\in B} p_X(g^{-1}(y)).$$

Hence the density p_Y of Y is given by

$$p_Y(y) = p_X(g^{-1}(y)).$$

If μ is Lebesgue measure and g is differentiable, we use the change-of-variable formula to obtain,

$$\int_A p_X(x)\mu(dx) = \int_A p_X(x)dx = \int_B p_X(g^{-1}(y)) \left|\text{det}\dot{g}\left(g^{-1}(y)\right)\right|^{-1} dy,$$

where $\dot{g}$ is the Jacobian matrix of g, ie, the matrix of the first partial derivatives of f, $[\partial g_i/\partial x_j]$. Then we obtain the density of Y,

$$p_Y(y) = p_X(g^{-1}(y)) \left|\text{det}\dot{g}\left(g^{-1}(y)\right)\right|^{-1}.$$

Example 4.3.6 *Suppose we have two random variables X_1 and X_2 with joint density*

$$\begin{aligned} p(x_1, x_2) &= 4x_1x_2 \quad if \quad 0 < x_1,\ x_2 < 1, \\ &= 0 \quad otherwise \end{aligned}$$

Define $Y_1 = X_1/X_2$ and $Y_2 = X_1X_2$. The problem is to obtain the joint density of (Y_1, Y_2) from that of (X_1, X_2). First note that the inverse transformation is

$$x_1 = (y_1y_2)^{1/2} \quad and \quad x_2 = (y_2/y_1)^{1/2}.$$

Let $\mathcal{X} = \{(x_1, x_2)|0 < x_1,\ x_2 < 1\}$ denote the support of the joint density of (X_1, X_2). Then the support of the joint density of (Y_1, Y_2) is given by $\mathcal{Y} = \{(y_1, y_2)|y_1, y_2 > 0, y_1y_2 < 1, y_2 < y_1\}$. Then

$$|det\dot{g}(x)| = \left| det\begin{pmatrix} \frac{1}{x_2} & -\frac{x_1}{x_2^2} \\ x_2 & x_1 \end{pmatrix} \right| = \left| det\begin{pmatrix} \sqrt{\frac{y_1}{y_2}} & \frac{\sqrt{y_1y_2}}{y_2/y_1} \\ \sqrt{y_2/y_1} & \sqrt{y_1y_2} \end{pmatrix} \right| = 2y_1.$$

Hence the joint density of (Y_1, Y_2) *is given by*

$$p(y_1, y_2) = \frac{4(y_1y_2)^{1/2}(y_2/y_1)^{1/2}}{2y_1} = \frac{2y_2}{y_1}.$$

4.4 Multivariate Normal Distribution

4.4.1 Introduction

Definition 4.4.1 (Multivariate Normal) *A random vector* $X = (X_1, \ldots, X_n)'$ *is said to be multivariate normally distributed if for all* $a \in \mathbb{R}^n$, $a'X$ *has a univariate normal distribution.*

Let $Z = (Z_1, \ldots, Z_n)'$ be a n-dimensional random vector, where (Z_i) are i.i.d. $N(0, 1)$. We have $\mathbb{E}Z = 0$ and $\mathrm{var}(Z) = I_n$. For all $a \in \mathbb{R}^n$, we have

$$\mathbb{E}e^{it(a'z)} = \prod_{k=1}^{n} \mathbb{E}e^{ita_kz_k} = \prod_{k=1}^{n} \phi_Z(a_kt) = \prod_{k=1}^{n} e^{-\frac{1}{2}a_k^2t^2} = e^{-\frac{1}{2}\sum_{k=1}^n a_k^2},$$

which is the characteristic function of a $N(0, \sum_{k=1}^n a_k^2)$ random variable. Hence Z is multivariate normal. We may write $Z \sim N(0, I_n)$, and call it *standard multivariate normal.*

Using similar argument, we can show that X is multivariate normal if it can be written as

$$X = \mu + \Sigma^{1/2}Z,$$

where Z is standard multivariate normal, μ is an n-vector, and Σ is a symmetric and positive definite matrix. It is easy to see that $\mathbb{E}X = \mu$ and $\mathrm{var}(X) = \Sigma$. We write $X \sim N(\mu, \Sigma)$.

Characteristic Function for Random Vectors For a random vector X, the characteristic function may be defined as $\phi_X(t) = \mathbb{E}\exp(it'X)$, where $t \in \mathbb{R}^n$. The characteristic function of Z (defined above) is obviously

$$\phi_Z(t) = \exp\left(-\frac{1}{2}t't\right).$$

Let $X \sim N(\mu, \Sigma)$. It follows that

$$\phi_X(t) = \mathbb{E}e^{it'X} = e^{it'\mu}\phi_Z\left(\Sigma^{1/2}t\right) = \exp\left(it'\mu - \frac{1}{2}t'\Sigma t\right).$$

Joint Density The joint density of Z is given by,

$$p(z) = \prod_{i=1}^{n} p(z_i) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}z'z\right).$$

The Jacobian matrix of of the affine transformation $X = \mu + \Sigma^{1/2}Z$ is $\Sigma^{1/2}$, hence

$$p(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

Remarks:

- A vector of univariate normal random variables is not necessarily a multivariate normal random vector. A counter example is $(X, Y)'$, where $X \sim N(0, 1)$ and $Y = X$ if $|X| > c$ and $Y = -X$ if $|X| < c$, where c is about 1.54.
- If Σ is singular, then there exists some $a \in \mathbb{R}^n$ such that $\mathrm{var}(a'X) = a'\Sigma a = 0$. This implies that X is random only on a subspace of $\mathbb{R}^n$. We may say that the joint distribution of X is degenerate in this case.

4.4.2 Marginals and Conditionals

Throughout this section, let $X \sim N(\mu, \Sigma)$.

Lemma 4.4.2 (Affine Transformation) *If $Y = AX + b$, then $Y \sim N(A\mu + b, A\Sigma A')$.*

Proof: Exercise. (Hint: use c.f. arguments.)

To introduce marginal distributions, we partition X conformably into

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right),$$

where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$.

Marginal Distribution Apply Lemma 1 with $A = (I_{n_1}, 0)$ and $b = 0$, we have $X_1 \sim N(\mu_1, \Sigma_{11})$. In other words, the marginal distributions of a multivariate normal distribution are also multivariate normal.

Lemma 4.4.3 (Independence) *X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.*

Proof: The "only if" part is obvious. If $\Sigma_{12} = 0$, then Σ is a block diagonal,

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

Hence

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix},$$

and

$$|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22}|.$$

Then the joint density of x_1 and x_2, can be factored as

$$\begin{aligned} p(x) = p(x_1, x_2) &= (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\ &= (2\pi)^{-n_1/2}|\Sigma_{11}|^{-1/2} \exp\left(-\frac{1}{2}(x_1-\mu_1)'\Sigma_{11}^{-1}(x_1-\mu_1)\right) \\ &\quad \cdot (2\pi)^{-n_2/2}|\Sigma_{22}|^{-1/2} \exp\left(-\frac{1}{2}(x_2-\mu_2)'\Sigma_{22}^{-1}(x_2-\mu_2)\right) \\ &= p(x_1)p(x_2). \end{aligned}$$

Hence X_1 and X_2 are independent.

Theorem 4.4.4 (Conditional Distribution) *The conditional distribution of X_1 given X_2 is $N(\mu_{1|2}, \Sigma_{11|2})$, where*

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma_{11|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Proof: First note that

$$\begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Since

$$\begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I & 0 \\ -\Sigma_{12}\Sigma_{22}^{-1} & I \end{pmatrix} = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{pmatrix},$$

$X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ and X_2 are independent. We write

$$X_1 = A_1 + A_2,$$

where

$$A_1 = \left(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\right), \quad A_2 = \Sigma_{12}\Sigma_{22}^{-1}X_2.$$

Since A_1 is independent of X_2, the conditional distribution of A_1 given X_2 is the unconditional distribution of A_1, which is

$$N\left(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

A_2 may be treated as a constant given X_2, which only shifts the mean of the condtional distribution of X_1 given X_2. We have thus obtained the desired result.

From the above result, we may see that the conditional mean of X_1 given X_2 is linear in X_2, and that the conditional variance of X_1 given X_2 does not depend on X_2. Of course the conditional variance of X_1 given X_2 is less than the unconditional variance of X_1, in the sense that $\Sigma_{11} - \Sigma_{11|2}$ is a positive semi-definite matrix.

4.4.3 Quadratic Forms

Let X be an n-by-1 random vector and A be an n-by-n deterministic matrix, the quantity $X'AX$ is called the quadratic form of X with respect to A. In this section we consider the distribution of the quadratic forms of X when X is multivariate normal. First we introduce a few important distributions that are related with the quadratic forms of normal vectors.

chi-square distribution If $Z = (Z_1, \ldots, Z_n)' \sim N(0, I_n)$, it is well known that

$$Z'Z = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2,$$

which is called chi-square distribution with n degrees of freedom.

Student t distribution Let $T = \frac{Z}{\sqrt{V/m}}$, where $Z \sim N(0,1)$ and $V \sim \chi_m^2$ and Z and V are independent, then $T \sim t_m$, the Student t distribution with m degrees of freedom.

F distribution Let $F = \frac{V_1/m_1}{V_2/m_2}$, where V_1 and V_2 are independent $\chi_{m_1}^2$ and $\chi_{m_2}^2$, respectively. Then $F \sim F_{m_1,m_2}$, the F distribution with degrees of freedom m_1 and m_2.

Theorem 4.4.5 *Let $X \sim N(0, \Sigma)$, where Σ is nonsingular. Then*

$$X'\Sigma^{-1}X \sim \chi^2_n.$$

Proof: Note that $\Sigma^{-1/2}X \sim N(0, I_n)$.

To get to the next theorem, recall that a square matrix is a *projection* if and only if $P^2 = P$.[1] If, in addition, P is symmetric, then P is an orthogonal projection.

Theorem 4.4.6 *Let $Z \sim N(0, I_n)$ and P be an m-dimensional orthogonal projection in $\mathbb{R}^n$, then we have*

$$Z'PZ \sim \chi^2_m.$$

Proof: It is well known that P may be decomposed into

$$P = H_m H_m',$$

where H_m is an $n \times m$ orthogonal matrix such that $H_m' H_m = I_m$. Note that $H_m' Z \sim N(0, I_m)$ and $Z'PZ = (H_m' Z)'(H_m' Z)$.

Theorem 4.4.7 *Let $Z \sim N(0, I_n)$, and let A and B be deterministic matrices, then $A'Z$ and $B'Z$ are independent if and only if $A'B = 0$.*

Proof: Let $C = (A, B)$. Without loss of generality, we assume that C is full rank (if it is not, then throw away linearly dependent columns). We have

$$C'Z = \left(\begin{array}{c} A'Z \\ B'Z \end{array} \right) \sim N\left(0, \left(\begin{array}{cc} A'A & A'B \\ B'A & B'B \end{array} \right)\right).$$

It is now clear that $A'Z$ and $B'Z$ are independent if and only if the covariance $A'B$ is null.

It is immediate that we have

Corollary 4.4.8 *Let $Z \sim N(0, I_n)$, and let P and Q be orthogonal projections such that $PQ = 0$, then $Z'PZ$ and $Z'QZ$ are independent.*

Proof: Note that since $PQ = 0$, then PZ and QZ are independent. Hence the independence of $Z'PZ = (PZ)'(PZ)$ and $Z'QZ = (QZ)'(QZ)$.

[1] Matrices that satisfy this property is said to be idempotent.

Using the above results, we can easily prove

Theorem 4.4.9 *Let $Z \sim N(0, I_n)$, and let P and Q be orthogonal projections of dimensions m_1 and m_2, respectively. If $PQ = 0$, then*

$$\frac{Z'PZ/m_1}{Z'QZ/m_2} \sim F_{m_1,m_2}.$$

Finally, we prove a useful theorem.

Theorem 4.4.10 *Let (X_i) be i.i.d. $N(\mu, \sigma^2)$, and define*

$$\begin{aligned} \bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \end{aligned}$$

We have

(a) $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

(b) $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

(c) $\bar{X}_n$ *and* S_n^2 *are independent*

(d) $\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \sim t_{n-1}$

Proof: Let $X = (X_1, \ldots, X_n')$ and ι be an $n \times 1$ vector of ones, then $X \sim N(\mu\iota, I_n)$. (a) follows from $\bar{X}_n = \frac{1}{n}\iota' X$. Define $P_\iota = \iota\iota'/n = \iota(\iota'\iota)^{-1}\iota$, which is the orthogonal projection on the span of ι. Then we have

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = (X - \iota\iota' X/n)'(X - \iota\iota' X/n) = X'(I - P_\iota)X.$$

Hence

$$\frac{(n-1)S_n^2}{\sigma^2} = \left(\frac{X - \mu\iota}{\sigma}\right)'(I_n - P_\iota)\left(\frac{X - \mu\iota}{\sigma}\right).$$

(b) follows from the fact that $\frac{X-\mu\iota}{\sigma} \sim N(0, I_n)$ and that $(I_n - P_\iota)$ is an $(n-1)$-dimensional orthogonal projection. To prove (c), we note that $\bar{X}_n = \frac{\iota'}{n}P_\iota X$ and $S_n^2 = \frac{1}{n-1}\left((I - P_\iota)X\right)'\left((I - P_\iota)X\right)$, and that $P_\iota X$ and $(I - P_\iota)X$ are independent by Theorem 4.4.7. Finally, (d) follows from

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}{\sqrt{\frac{\frac{(n-1)S_n^2}{\sigma^2}}{n-1}}}.$$

4.5 Exercises

1. Derive the characteristic function of the distribution with density
$$p(x) = \exp(-|x|)/2.$$

2. Let X and Y be independent standard normal variables. Find the density of a random variable defined by
$$U = \frac{X}{Y}.$$
[Hint: Let $V = Y$ and first find the joint density of U and V.]

3. Let X and Y have bivariate normal distribution with mean and variance
$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$
(a) Find a constant α^* such that $Y - \alpha^* X$ is independent of X. Show that $\mathrm{var(Y - \alpha X) \geq var(Y - \alpha^* X)}$ for any constant α.
(b) Find the conditional distribution of $X + Y$ given $X - Y$.
(c) Obtain $\mathbb{E}(X|X + Y)$.

4. Let $X = (X_1, \ldots, X_n)'$ be a random vector with mean $\mu\iota$ and variance Σ, where μ is a scalar, ι is the n-vector of ones and Σ is an n by n symmetric matrix. We define
$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S_n^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}.$$
Consider the following assumptions:
(A1) X has multivariate normal distribution,
(A2) $\Sigma = \sigma^2 I$,
(A3) $\mu = 0$.
We claim:
(a) $\overline{X}_n$ and S_n^2 are uncorrelated.
(b) $\mathbb{E}(\overline{X}_n) = \mu$.
(c) $\mathbb{E}(S_n^2) = \sigma^2$.
(d) $\overline{X}_n \sim N(\mu, \sigma^2/n)$.
(e) $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$.
(f) $\sqrt{n}(\overline{X}_n - \mu)/S_n \sim t_{n-1}$.
What assumptions in (A1), (A2), and (A3) are needed for each of (a) – (f) to hold. Prove (a) – (f) using the assumptions you specified.

Chapter 5

Introduction to Statistics

5.1 General Settings

The fundamental postulate of statistical analysis is that the observed data are realized values of a vector of random variables defined on a common probability space. This postulate is not verifiable. It is a philosophical view of the world that we choose to take, and we call it the *probabilistic view*. An alternative view would be that the seemingly random data are generated from a deterministic but chaotic law. We only consider the probabilistic view, which is main stream among economists.

Let $X = (X_1, \ldots, X_n)$ be variables of interest, where for each i, X_i may be a vector. The objective of statistical inference, is to study the joint distribution of X based on the observed sample.

The First Example: For example, we may study the relationship between individual income (*income*) and the characteristics of the individual such as education level (*edu*), work experience (*expr*), gender, etc. The variables of interest may then be $X_i = (income_i, edu_i, expr_i, gender_i)$. We may reasonably postulate that (X_i) are independently and identically distributed (i.i.d.). Hence the study of the joint distribution of X reduces to that of the joint distribution of X_i. To achieve this, we take a sample of the whole population, and observe $(X_i, i = 1, \ldots, n)$, where i denotes individuals. In this example in particular, we may focus on the conditional distribution of *income* given *edu*, *expr*, and *gender*.

The Second Example: For another example, in macroeconomics, we may be interested in the relationship among government expenditure (g_t), GDP growth (y_t), inflation (π_t), and unemployment (u_t). The variables of interest may be $X_t =$

(g_t, y_t, π_t, u_t). One of the objective of empirical analysis, in this example, may be to study the conditional distribution of unemployment given past observations on government expenditure, GDP growth, inflation, as well as itself. The problem of this example lies with, first, the i.i.d. assumption on X_t is untenable, and second, the fact that we can observe X_t only once. In other words, an economic data generating process is nonindependent and time-irreversible. It is clear that the statistical study would go nowhere unless we impose (sometimes strong) assumptions on the evolution of X_t, stationarity for example.

In this chapter, for simplicity, we have the first example in mind. In most cases, we assume that $X_1, \ldots, X_n$ are i.i.d. with a distribution P_θ that belongs to a family of distributions $\{P_\theta | \theta \in \Theta\}$ where θ is called parameter and Θ a parameter set. In this course we restrict θ to be finite-dimensional. This is called the *parametric approach* to statistical analysis. The nonparametric approach refers to the case where we do not restrict the distribution to any family of distributions, which is in a sense to allow θ to be infinite-dimensional. In this course we mainly consider the parametric approach.

Definition 5.1.1 (Statistic) *A statistic is a real-valued (or vector-valued) measurable function $\tau(X)$ of a random sample $X = (X_1, \ldots, X_n)$.*

Note that the statistic is a random variable (or vector) itself.

Statistical inference consists of two procedures: estimation of and hypothesis testing on θ. For the purpose of estimating θ, we need to construct a vector-valued statistic called *estimator*, $\hat{\theta}(X) : \mathcal{X} \to \mathcal{T}$, where $\mathcal{X}$ is called the state space (the range of X), and where $\mathcal{T}$ includes Θ. It is customary to omit X in $\hat{\theta}(X)$ and to write $\hat{\theta}$.

For the purpose of hypothesis testing on θ, we need to construct a statistic called *test statistic*, $\tau(X) : \mathcal{X} \to \mathcal{T}$, where $\mathcal{T}$ is a subset of $\mathbb{R}$. A hypothesis divides Θ into two disjoint and exhaustive subsets. We rely on the value of τ to decide whether θ_0, the true parameter, is in one of them.

Sufficient Statistic Let $\tau = \tau(X)$ be a statistic, and $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$ be a family of distributions of X.

Definition 5.1.2 (Sufficient Statistic) *We define that τ is* sufficient *for $\mathcal{P}$ (or more precisely θ) if the conditional distribution of X given τ does not depend on θ.*

The distribution of X can be any member of the family $\mathcal{P}$. Therefore, the conditional distribution of X given τ would depend on θ in general. τ is sufficient in the sense that the distribution of X is uniquely determined by the value of τ.

Sufficient statistics are useful in data reduction. It is less costly to infer θ from a statistic τ than from X, since the former, being a function of the latter, is of lower dimension. The sufficiency of τ guarantees that τ contains all information about θ in X.

Example 5.1.3 *Suppose that $X \sim N(0, \sigma^2)$ and $\tau = |X|$. Conditional on $\tau = t$, X can take t or $-t$. Since the distribution of X is symmetric about the origin, each has a conditional probability of 1/2, regardless of the value of σ^2. The statistic τ is thus sufficient.*

Example 5.1.4 *Let X_1 and X_2 be independent Poisson(λ). $\tau = X_1 + X_2$ is a sufficient statistic. First, the joint density of X_1 and X_2 is*

$$p_\lambda(x_1, x_2) = \exp(-2\lambda)\frac{\lambda^{x_1+x_2}}{x_1!x_2!}, \ x_1, x_2 = 0, 1, 2, \ldots.$$

We may show that $p(x_1|\tau = t) = p_\lambda(x_1, t)/p_\lambda(t)$ is λ-free. The same is of course true for $p(x_2|\tau = t)$. Hence τ is sufficient.

Theorem 5.1.5 (Fisher-Neyman Factorization) *A statistic $\tau = \tau(X)$ is sufficient if and only if there exist two functions f and g such that the density of X is factorized as*

$$p_\theta(x) = f(\tau(x), \theta)g(x).$$

This theorem implies that if two samples give the same value for a sufficient statistic, then the MLE based on the two samples yield the same estimate of the parameters.

Example 5.1.6 *Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ). We may write the joint distribution of $X = (X_1, \ldots, X_n)$ as*

$$p_\lambda(x) = e^{n\lambda}\frac{\lambda^{x_1+\cdots+x_n}}{\prod_{i=1}^n x_i!} = f(\tau(x), \lambda)g(x),$$

where $\tau(x) = \sum_{i=1}^n x_i$, $f(t, \lambda) = \exp(-n\lambda)\lambda^t$, and $g(x) = \left(\prod_{i=1}^n x_i!\right)^{-1}$. Hence $\tau(x)$ is sufficient for λ.

Example 5.1.7 *Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. The joint density is*

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - n\frac{\mu^2}{2\sigma^2}\right). \end{aligned}$$

It is clear that $\tau(x) = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right)$ is sufficient for $(\mu, \sigma^2)'$.

Minimal Sufficient Statistic Sufficient statistic is by no means unique. $\tau(x) = (x_1, \ldots, x_n)'$, for example, is always sufficient. Let τ and κ be two statistics and κ is sufficient. It follows immediately from the Fisher-Neyman factorization theorem that if $\tau = h(\kappa)$ for some function h, then τ is also sufficient. If h is a many-to-one function, then τ provides further data reduction than κ. We call a sufficient statistic *minimal* if it is a function of every sufficient statistic. A minimal sufficient statistic thus achieves data reduction to the best extent.

Definition 5.1.8 *Exponential Family The exponential family refers to the family of distributions that have densities of the form*

$$p_\theta(x) = \exp\left[\sum_{i=1}^{m} a_i(\theta)\tau_i(x) + b(\theta)\right] g(x),$$

where m is a positive integer.

To emphasize the dependence on m, we may call the above family m-parameter exponential family.

- Note that for the m-parameter exponential family, by the factorization theorem, $\tau(x) = (\tau_1(x), \ldots, \tau_m(x))'$ is a sufficient statistic.
- If $X_1, \ldots, X_n$ are i.i.d. with density

  $$p_\theta(x_i) = \exp\left[a(\theta)\tau_i(x_i) + b(\theta)\right] g(x_i),$$

  then the joint density of $X = (X_1, \ldots, X_n)'$ is

  $$p_\theta(x) = \exp\left[a(\theta)\sum_{i=1}^{n} \tau_i(x_i) + nb(\theta)\right] \prod_{i=1}^{n} g(x_i).$$

  This implies that $\sum_i \tau(x_i)$ is a sufficient statistic.

The exponential family includes many distributions that are in frequent use.

Example 5.1.9 (One-parameter exponential family)

- *Poisson(λ)*

  $$p_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!} = e^{x\log\lambda - \lambda}\frac{1}{x!}.$$

- *Bernoulli(θ)*

  $$p_\theta(x) = \theta^x(1-\theta)^{1-x} = \exp\left(x\log(\theta/(1-\theta)) + \log(1-\theta)\right).$$

Example 5.1.10 (Two-parameter exponential family) • *$N(\mu, \sigma^2)$*

$$\begin{aligned} p_{\mu,\sigma^2} &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2} - \left(\frac{\mu^2}{2\sigma^2} + \log\sigma\right)\right) \end{aligned}$$

- *Gamma(α, β)*

$$\begin{aligned} p_{\alpha,\beta} &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \\ &= \exp\left((\alpha-1)\log x - \frac{1}{\beta}x - (\log\Gamma(\alpha) + \alpha\log\beta)\right). \end{aligned}$$

Remark on Bayesian Approach The Bayesian approach to probability is one of the different interpretations of the concept of probability. Bayesians view probability as an extension of logic that enables reasoning with uncertainty. Bayesians do not reject or accept a hypothesis, but evaluate the probability of a hypothesis. To achieve this, Bayesians specify some *prior* distribution $p(\theta)$, which is then updated in the light of new relevant data by the Bayes' rule,

$$p(\theta|x) = p(\theta)\frac{p(x|\theta)}{p(x)},$$

where $p(x) = \int p(x|\theta)p(\theta)d\theta$. Note that Bayesians treat θ as random, hence the conditional-density notation of $p(\theta|x)$, which is called *posterior* density.

5.2 Estimation

5.2.1 Method of Moment

Let $X_1, \dots, X_n$ be i.i.d. random variables with a common distribution P_θ, where the parameter vector θ is to be estimated. And let $x_1, \dots, x_n$ be a realized sample. We call the underlying distribution P_θ the *population*, the moments of which we call *population moments.* Let $\mathbf{f}$ be a vector of measurable functions $\mathbf{f}(x) = (f_1(x), \dots, f_m(k))'$, the $\mathbf{f}$-population moments of P_θ are given by

$$\mathbb{E}_\theta \mathbf{f} = \int f dP_\theta.$$

In contrast, we call the sample average of $(\mathbf{f}(x_i))$ the *sample moments*. Note that the sample average may be regarded as the moment of the distribution that assigns probability mass $1/n$ to each realization x_i. This distribution is called the *empirical distribution*, which we denote P_n. Obviously, the moments of the empirical distribution equal the corresponding sample moments

$$\mathbb{E}_n\mathbf{f} = \int \mathbf{f}dP_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}(x_i).$$

The method of moment (MM) equates population moment to sample moment so that the parameter vector θ may be solved. In other words, the MM estimation solves the following set of equations for the parameter vector θ,

$$\mathbb{E}_\theta\mathbf{f} = \mathbb{E}_n\mathbf{f}. \tag{5.1}$$

This set of equations are called the moment conditions.

Example 5.2.1 *Let X_i be i.i.d. Poisson(λ). To estimate λ, we may solve the following equation,*

$$\mathbb{E}_\lambda X_i = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

It is immediate that the MM estimator of λ is exactly $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Example 5.2.2 *Let X_i be i.i.d. N(μ, σ^2). To estimate μ and σ^2, we may solve the following system of equations*

$$\begin{aligned}\mathbb{E}_{\mu,\sigma^2}X &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \mathbb{E}_{\mu,\sigma^2}(X-\mu)^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

This would obtain

$$\hat{\mu} = \bar{x}, \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}).$$

A Remark on GMM If the number of equations (moment conditions) in (5.1) exceeds the number of parameters to be estimated, then the parameter θ is over-identified. In such cases, we may use the generalized method of moments (GMM) to estimate θ. The basic idea of GMM is to minimize some distance measure between the population moments and their corresponding sample moments. A popular approach is to solve the following quadratic programming problem,

$$\min_{\theta \in \Theta} \mathbf{d}(\theta; x)' W \mathbf{d}(\theta; x),$$

where $\mathbf{d}(\theta; x) = \mathbb{E}_\theta \mathbf{f} - \mathbb{E}_n \mathbf{f}$ and W is a positive definite weighting matrix. The detailed properties of GMM is out of the scope of this text.

5.2.2 Maximum Likelihood

Let $p(x, \theta)$ be the density of the distribution P_θ. We write $p(x, \theta)$, instead of $p_\theta(x)$, to emphasize that the density is a function of θ as well as that of x. We define *likelihood function* as

$$p(\theta; x) = p(x, \theta).$$

The likelihood function is a function of the parameter θ given a sample x. Obviously, it is intuitively appealing to assume that if $\theta = \theta_0$, the true parameter, then the likelihood function $p(\theta; x)$ achieves the maximum. This is indeed the fundamental assumption of the maximum likelihood estimation (MLE), which is defined as follows,

Definition 5.2.3 (MLE) *The maximum likelihood estimator (MLE) of θ is given by*

$$\hat{\theta}_{ML} = arg \max_{\theta \in \Theta} p(\theta; x).$$

Remark: Let τ be any sufficient statistic for the parameter θ. According the factorization theorem, we have $p(x, \theta) = f(\tau(x), \theta) g(x)$. Then $\hat{\theta}_{ML}$ maximizes $f(\tau(x), \theta)$ with respect to θ. Therefore, $\hat{\theta}_{ML}$ is always a function of $\tau(X)$. This implies that if MLE is a sufficient statistic, then it is always minimal.

Log Likelihood It is often easier to maximize the logarithm of the likelihood function,

$$\ell(\theta; x) = \log(p(\theta; x)).$$

Since the log function is monotone increasing, maximizing log likelihood yields the same estimates.

First Order Condition If the log likelihood function $\ell(\theta; x)$ is differentiable and globally concave for all x, then the ML estimator can be obtained by solving the first order condition (FOC),

$$\frac{\partial \ell}{\partial \theta}(\theta; x) = 0$$

Note that $s(\theta; x) = \frac{\partial \ell}{\partial \theta}(\theta; x)$ is called score functions.

Theorem 5.2.4 (Invariance Theorem) *If $\hat{\theta}$ is an ML estimator of θ and $\pi = g(\theta)$ be a function of θ, then $g(\hat{\theta})$ is an ML estimator of π.*

Proof: If g is one-to-one, then

$$p(\theta; x) = p(g^{-1}g(\theta); x) = p^*(g(\theta); x).$$

Both ML estimators, $\hat{\theta}$ and $\widehat{g(\theta)}$, maximize the likelihood function and it is obvious that

$$\hat{\theta} = g^{-1}\left(\widehat{g(\theta)}\right).$$

This implies $g(\hat{\theta}) = \widehat{g(\theta)} = \hat{\pi}$. If g is many-to-one, $\hat{\pi} = g(\hat{\theta})$ still corresponds to $\hat{\theta}$ that maximizes $p(\theta; x)$. Any other value of π would correspond to θ that results in lower likelihood. Q.E.D.

Example 5.2.5 (Bernoulli(θ)) *Let $(X_i, i = 1, \ldots, n)$ be i.i.d. Bernoulli(θ), then the log likelihood function is given by*

$$\ell(\theta; x) = \left(\sum_{i=1}^{n} x_i\right) \log \theta + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \theta).$$

The FOC yields

$$\hat{\theta}^{-1} \sum_{i=1}^{n} x_i - (1 - \hat{\theta})^{-1}(n - \sum_{i=1}^{n} x_i) = 0,$$

which is solved to obtain $\hat{\theta} = \bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. Note that to estimate the variance of X_i, we need to estimate $v = \theta(1 - \theta)$, a function of θ. By the invariance theorem, we obtain $\hat{v} = \hat{\theta}(1 - \hat{\theta})$.

Example 5.2.6 **($N(\mu, \sigma^2)$)** *Let X_i be i.i.d. $N(\mu, \sigma^2)$, then the log-likelihood function is given by*

$$\ell(\mu, \sigma^2; x) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Solving the FOC gives

$$\begin{aligned}\hat{\mu} &= \bar{x} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.\end{aligned}$$

Note that the ML estimators are identical to the MM estimators.

Example 5.2.7 (Uniform([0,θ) *)] Let X_i be i.i.d. Uniform($[0, \theta]$). Then*

$$\begin{aligned}p(\theta; x) &= \frac{1}{\theta^n}\prod_{i=1}^{n} I_{0 \le x_i \le \theta} \\ &= \frac{1}{\theta^n} I_{\{\min_{1\le i\le n} x_i \ge 0\}} I_{\{\max_{1\le i\le n} x_i \le \theta\}}.\end{aligned}$$

It follows that $\hat{\theta} = \max\{x_1, \ldots, x_n\}$.

5.2.3 Unbiasedness and Efficiency

Let $\mathbb{P}_\theta$ denote the probability measure in Ω corresponding to P_θ in $\mathcal{X}$, and let $\mathbb{E}_\theta$ denote the expectation taken with respect to $\mathbb{P}_\theta$.

Definition 5.2.8 (Unbiasedness) *An estimator $\hat{\theta}$ is unbiased if for all $\theta \in \Theta$,*

$$\mathbb{E}_\theta \hat{\theta} = \theta.$$

Unbiasedness is a desirable property. Loosely speaking, it refers to the description that “the estimation is correct in average”. To describe how “varied” an estimator would be, we often use the mean squared error, which is defined as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \theta)^2.$$

We may decompose the MSE as

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2 + (\mathbb{E}_\theta\hat{\theta} - \theta)^2.$$

For an unbiased estimator $\hat{\theta}$, the second term vanishes, then the MSE is equal to the variance.

In general, MSE is a function of the unknown parameter θ and it is impossible to find an estimator that has the smallest MSE for all $\theta \in \Theta$. However, if we restrict our attention to the class of unbiased estimators, we may find an estimator that enjoys the smallest variance (hence MSE) for all $\theta \in \Theta$. This property is known as *uniformly minimum variance unbiasedness* (UMVU). More precisely, we have

Definition 5.2.9 (UMVU Estimator) *An estimator $\hat{\theta}$ is called an UMVU estimator if it satisfies*

(1) $\hat{\theta}$ is unbiased,

(2) $\mathbb{E}_\theta(\hat{\theta}-\theta)^2 \le \mathbb{E}_\theta(\tilde{\theta}-\theta)^2$ for any unbiased estimator $\tilde{\theta}$.

5.2.4 Lehmann-Scheffé Theorem

The prominent Lehmann-Scheffé Theorem helps to find UMVU estimators. First, we introduce some basic concepts in the decision-theoretic approach of statistical estimation.

Definition 5.2.10 (Loss Function) *Loss function is any function $\ell(t,\theta)$ that assigns disutility to each pair of estimate t and parameter value θ.*

Examples of Loss Function

- $\ell(t,\theta) = (t-\theta)^2$, squared error.
- $\ell(t,\theta) = |t-\theta|$, absolute error.
- $\ell(t,\theta) = c\mathrm{I}\{|t-\theta| > \epsilon\}$, fixed loss out of bound.

Definition 5.2.11 (Risk Function) *For an estimator $T = \tau(X)$, the risk function is defined by*

$$r(\tau,\theta) = \mathbb{E}_\theta \ell(T,\theta).$$

It can be observed that risk function is the expected loss of an estimator for each value of θ. Risk functions corresponding to the loss functions in the above examples are

Examples of Risk Function

- $r(\tau,\theta) = \mathbb{E}_\theta(\tau(X)-\theta)^2$, mean squared error.
- $r(\tau,\theta) = \mathbb{E}_\theta|\tau(X)-\theta|$, mean absolute error.
- $r(\tau,\theta) = c\mathbb{P}_\theta\{|\tau-\theta| > \epsilon\}$

In the decision-theoretic approach of statistical inference, estimators are constructed by minimizing some appropriate loss or risk functions.

Definition 5.2.12 (Minimax Estimator) *An estimator τ_* is called minimax if*

$$\sup_{\theta\in\Theta} r(\tau_*, \theta) \leq \sup_{\theta\in\Theta} r(\tau, \theta)$$

for every other estimator τ.

Note that $\sup_{\theta\in\Theta} r(\tau, \theta)$ measures the maximum risk of an estimator τ.

Theorem 5.2.13 (Rao-Blackwell Theorem) *Suppose that the loss function $\ell(t, \theta)$ is convex in t and that S is a sufficient statistic. Let $T = \tau(X)$ be an estimator for θ with finite mean and risk. If we define $T_* = \mathbb{E}_\theta(T|S)$ and write $T_* = \tau_*(X)$, then we have*

$$r(\tau_*, \theta) \leq r(\tau, \theta).$$

Proof: Since $\ell(t, \theta)$ is convex in t, Jensen's inequality gives

$$\ell(T_*, \theta) = \ell(\mathbb{E}_\theta(T|S), \theta) \leq \mathbb{E}_\theta(\ell(T, \theta)|S).$$

We conclude by taking expectations on both sides and applying the law of iterative expectations.

Note that $\mathbb{E}_\theta(T|S)$ is not a function of θ, since S is sufficient.

Definition 5.2.14 (Complete Statistic) *A statistic T is complete if $\mathbb{E}_\theta f(T) = 0$ for all $\theta \in \Theta$ implies $f = 0$ a.s. P_θ.*

Theorem 5.2.15 (Lehmann-Scheffé Theorem) *If S is complete and sufficient and $T = \tau(X)$ is an unbiased estimator of $g(\theta)$, then $f(S) = \mathbb{E}_\theta(T|S)$ is a UMVU estimator.*

Proof: Apply Rao-Blackwell Theorem with the squared loss function $\ell(t, \theta) = (t - \theta)^2$.

Note that $f(S)$ is also a unique unbiased estimator. Suppose there exists another unbiased estimator $\tilde{f}(S)$, then $\mathbb{E}_\theta(f(S) - \tilde{f}(S)) = 0$. But the completeness of S guarantees that $f = \tilde{f}$.

Given a complete and sufficient statistic, it is then straightforward to obtain a UMVU estimator. What we have to do is to take any unbiased estimator T and obtain the desired UMVU estimator as $T^* = \mathbb{E}_\theta(T|S)$.

Example 5.2.16 *Let $(X_i,\ i = 1, \ldots, n)$ be i.i.d. Uniform$(0, \theta)$, and let $S = \max_i X_i$. S is sufficient and complete. To see the completeness, note that*

$$\mathbb{P}_\theta(S \leq s) = (\mathbb{P}_\theta(X_i \leq s))^n = \left(\frac{s}{\theta}\right)^n .$$

The density of S is thus

$$p_\theta(s) = \frac{ns^{n-1}}{\theta^n} I\{0 \leq s \leq \theta\}.$$

$\mathbb{E}_\theta f(T) = 0$ for all θ implies

$$\int_0^\theta s^{n-1} f(s) ds = 0, \text{ for all } \theta.$$

This is only possible when $f = 0$.

Now we proceed to find a UMVU estimator. Let $T = 2X_1$, which is an unbiased estimator for θ. Suppose $S = s$, then X_1 can take s with probability $1/n$, since every member of $(X_i, i = 1, \ldots, n)$ is equally likely to be the maximum. When $X_1 \neq s$, which is of probability $(n-1)/n$, X_1 is uniformly distributed on $(0, s)$. Thus we have

$$\begin{aligned} \mathbb{E}_\theta(T|S = s) &= 2\mathbb{E}_\theta(X_1|S = s) \\ &= 2\left(\frac{1}{n}s + \frac{n-1}{n}\frac{s}{2}\right) \\ &= \frac{n+1}{n}s \end{aligned}$$

The UMVU estimator of θ is thus obtained as

$$T^* = \frac{n+1}{n} \max_{1 \leq i \leq n} X_i.$$

5.2.5 Efficiency Bound

It is generally not possible to construct an UMVU estimator. However, we show in this section that there exists a lower bound for the variance of unbiased estimators, which we call *efficiency bound*. If an unbiased estimator achieves the efficiency bound, we say that it is an efficient estimator.

Let $\ell(\theta; x)$ be the log-likelihood function. Recall that we have defined score function $s(\theta; x) = \partial\ell/\partial\theta(\theta; x)$. We further define:

(a) Hessian: $h(\theta; x) = \frac{\partial^2 \ell}{\partial\theta\partial\theta'}(\theta; x)$.

(b) Fisher Information: $I(\theta) = \mathbb{E}_\theta s(\theta; X)s(\theta; X)'$.

(c) Expected Hessian: $H(\theta) = \mathbb{E}_\theta h(\theta; X)$.

Note that for a vector of independent variables, the scores and Hessians are additive. Specifically, let X_1 and X_2 be independent random vectors, let $X = (X_1', X_2')'$. Denote the scores and the Hessians of X_i, $i = 1, 2$, by $s(\theta; x_i)$ and $H(\theta; x_i)$ respectively, and denote the score and the Hessian of X by $s(\theta; x)$ and $H(\theta; x)$, respectively. Then it is clear that

$$\begin{aligned} s(\theta; x) &= s(\theta; x_1) + s(\theta; x_2) \\ h(\theta; x) &= h(\theta; x_1) + h(\theta; x_2). \end{aligned}$$

We can also show that

$$\begin{aligned} I(\theta) &= I_1(\theta) + I_2(\theta) \\ H(\theta) &= H_1(\theta) + H_2(\theta), \end{aligned}$$

where $I(\theta)$, $I_1(\theta)$, and $I_2(\theta)$ denote the information matrix of X, X_1, X_2, respectively, and the notations of H, H_1, and H_2 are analogous.

From now on, we assume that a random vector X has joint density $p(x, \theta)$ with respect to Lebesque measure μ. Note that the notation $p(x, \theta)$ emphasizes the fact that the joint density of X is a function of both x and θ. We let $\hat{\theta}$ (or more precisely, $\hat{\theta}(X)$) be an unbiased estimator for θ. And we impose the following regularity conditions on $p(x, \theta)$,

Regularity Conditions

(a) $\frac{\partial}{\partial\theta} \int p(x, \theta) d\mu(x) = \int \frac{\partial}{\partial\theta} p(x, \theta) d\mu(x)$

(b) $\frac{\partial^2}{\partial\theta\partial\theta'} \int p(x, \theta) d\mu(x) = \int \frac{\partial^2}{\partial\theta\partial\theta'} p(x, \theta) d\mu(x)$

(c) $\int \hat{\theta}(x) \frac{\partial}{\partial\theta'} p(x, \theta) d\mu(x) = \frac{\partial}{\partial\theta'} \int \hat{\theta}(x) p(x, \theta) d\mu(x)$.

Under these regularity conditions, we have a few results that are both useful in proving subsequent theorems and interesting in themselves.

Lemma 5.2.17 *Suppose that Condition (a) holds, then*

$$\mathbb{E}_\theta s(\theta; X) = 0.$$

Proof: We have

$$\begin{aligned}
\mathbb{E}_\theta s(\theta;X) &= \int s(\theta;x)p(x,\theta)d\mu(x) \\
&= \int \frac{\partial}{\partial\theta}\ell(\theta;x)p(x,\theta)d\mu(x) \\
&= \int \frac{\frac{\partial}{\partial\theta}p(x,\theta)}{p(x,\theta)}p(x,\theta)d\mu(x) \\
&= \frac{\partial}{\partial\theta}\int p(x,\theta)d\mu(x) \\
&= 0
\end{aligned}$$

Lemma 5.2.18 *Suppose that Condition (b) holds, then*

$$I(\theta) = -H(\theta).$$

Proof: We have

$$\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta;x) = \frac{\frac{\partial^2}{\partial\theta\partial\theta'}p(x,\theta)}{p(x,\theta)} - \frac{\partial}{\partial\theta}\log p(x,\theta)\frac{\partial}{\partial\theta'}\log p(x,\theta).$$

Then

$$\begin{aligned}
H(\theta) &= \int\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta;x)\right)p(x,\theta)d\mu(x) \\
&= \frac{\partial^2}{\partial\theta\partial\theta'}\int p(x,\theta)d\mu(x) - I(\theta) \\
&= -I(\theta).
\end{aligned}$$

Lemma 5.2.19 *Let $\hat\theta(X)$ be an unbiased estimator for θ, and suppose the Condition (c) holds, then*

$$\mathbb{E}_\theta\hat\theta(X)s(\theta;X)' = I.$$

Proof: We have

$$\begin{aligned}
\mathbb{E}_\theta\hat\theta(X)s(\theta;X)' &= \int \hat\theta(x)\frac{\frac{\partial p}{\partial\theta'}(x,\theta)}{p(x,\theta)}p(x,\theta)d\mu(x) \\
&= \frac{\partial}{\partial\theta'}\int\hat\theta(x)p(x,\theta)d\mu(x) \\
&= I.
\end{aligned}$$

Theorem 5.2.20 (Cramer-Rao Bound) *Let $\hat{\theta}(X)$ be an unbiased estimator of θ, and if Conditions (a) and (c) hold, then,*

$$var_\theta\left(\hat{\theta}(X)\right) \geq I(\theta)^{-1}.$$

Proof: Using the above lemmas, we have

$$\mathrm{var}_\theta \begin{pmatrix} \hat{\theta}(X) \\ s(\theta; X) \end{pmatrix} = \begin{pmatrix} \mathrm{var}_\theta\left(\hat{\theta}(X)\right) & I \\ I & I(\theta) \end{pmatrix} \equiv A.$$

Recall that the covariance matrix A must be positive definite. We choose $B' = (I, -I(\theta)^{-1})$, then we must have $B'AB \geq 0$. The conclusion follows.

Example 5.2.21 *Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ). The the log-likelihood, the score, and the Fisher's information of each X_i are given by*

$$\begin{aligned} \ell(\lambda; x_i) &= -\lambda + x_i \log\lambda - \log x_i! \\ s(\lambda; x_i) &= -1 + x_i/\lambda \\ I_i(\lambda) &= 1/\lambda. \end{aligned}$$

Then the information matrix $I(\lambda)$ of $X = (X_1, \ldots, X_n)'$ is $I(\lambda) = nI_1(\lambda) = n/\lambda$. Recall $\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator for λ. And we have

$$var_\lambda(\bar{X}) = var_\lambda(X_1)/n = \lambda/n.$$

Hence the estimator $\bar{X}$ is an UMVU estimator.

5.3 Hypothesis Testing

5.3.1 Basic Concepts

Suppose a random sample $X = (X_1, \ldots, X_n)'$ is drawn from a population characterized by a parametric family $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$. We partition the parameter set Θ as

$$\Theta = \Theta_0 \cup \Theta_1.$$

A statistical hypothesis is of the following form:

$$\mathrm{H}_0 : \ \theta \in \Theta_0 \quad \mathrm{H}_1 : \ \theta \in \Theta_1,$$

where H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis.*

A test statistic, say τ, is used to partition the state space $\mathcal{X}$ into the disjoint union of the *critical region* C and the *acceptance region* A,

$$\mathcal{X} = C \cup A.$$

The critical region is conventionally given as

$$C = \{x \in \mathcal{X} | \tau(x) \geq c\},$$

where c is a constant that is called *critical value.* If the observed sample is within the critical region, we reject the null hypothesis. Otherwise, we say that we fail to reject the null and thus accept the alternative hypothesis. Note that different tests differ in their critical regions. In the following, we denote tests using their critical regions.

For $\theta \in \Theta_0$, $P_\theta(C)$ is the probability of rejecting H_0 when it is true. We thus define

Definition 5.3.1 (Size) *The* size *of a test C is*

$$\max_{\theta \in \Theta_0} P_\theta(C).$$

Obviously, it is desirable to have a small size. For $\theta \in \Theta_1$, $P_\theta(C)$ is the probability of rejecting H_0 when it is false. If this probability is large, we say that the test is powerful. Conventionally, we call $\pi(\theta) = P_\theta(C)$ the *power function.* The power function restricted to the domain Θ_1 characterizes the *power* of the test.

Given two tests with a same size, C_1 and C_2, if $P_\theta(C_1) > P_\theta(C_2)$ at $\theta \in \Theta_1$, we say that C_1 is *more powerful* than C_2. If there is a test C_* that satisfies $P_\theta(C_*) \geq P_\theta(C)$ at $\theta \in \Theta_1$ for any test C of the same size, then we say that C_* is the *most powerful* test. Furthermore, if the test C_* is such that $P_\theta(C_*) \geq P_\theta(C)$ for all $\theta \in \Theta_1$ for any test C of the same size, then we say that C_* is the *uniformly most powerful.*

If Θ_0 (or Θ_1) is a singleton set, ie, $\Theta_0 = \{\theta_0\}$, we call the hypothesis $\mathrm{H}_0 : \ \theta = \theta_0$ *simple.* Otherwise, we call it *composite* hypothesis.

In particular, when both H_0 and H_1 are simple hypotheses, say, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, $\mathcal{P}$ consists of two distributions P_{θ_0} and P_{θ_1}, which we denote as P_0 and P_1, respectively. It is clear that $P_0(C)$ and $P_1(C)$ are the size and the power of the test C, respectively. Note that both $P_0(C)$ and $P_1(A)$ are probabilities of making mistakes. $P_0(C)$ is the probability of rejecting the true null, and $P_1(A)$ is the probability of accepting the false null. Rejecting the true null is often called the type-I error, and accepting the false null is called the type-II error.

5.3.2 Likelihood Ratio Tests

Assume that both the null and the alternative hypotheses are simple, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$. Let $p(x, \theta_0)$ and $p(x, \theta_1)$ be the densities of P_0 and P_1, respectively. We have

Theorem 5.3.2 (Neyman-Pearson Lemma) *Let c be a constant. The test*

$$C_* = \left\{ x \,\middle|\, \lambda(x) = \frac{p(x,\theta_1)}{p(x,\theta_0)} \geq c \right\}$$

is the most powerful test.

Proof: Suppose C is any test with the same size as C_*. Assume without loss of generality that C and C_* are disjoint. It follows that

$$\begin{array}{rcll} p(x,\theta_1) & \geq & cp(x,\theta_0) & \text{on } C_* \\ p(x,\theta_1) & < & cp(x,\theta_0) & \text{on } C. \end{array}$$

Hence we have

$$P_1(C_*) = \int_{C_*} p(x,\theta_1)d\mu(x) \geq c \int_{C_*} p(x,\theta_0)d\mu(x) = cP_0(C_*),$$

and

$$P_1(C) = \int_{C} p(x,\theta_1)d\mu(x) < c \int_{C} p(x,\theta_0)d\mu(x) = cP_0(C).$$

Since $P_0(C_*) = P_0(C)$ (the same size), we have $P_1(C_*) \geq P_1(C)$. Q.E.D.

Remarks:

- For obvious reasons, test of the same form as C_* is also called *likelihood ratio* (LR) test. The constant c is to be determined by pre-specifying a size, ie, by solving for c the equation $P_0(C) = \alpha$, where α is prescribed small number.

- We may view $p(x, \theta_1)$ (or $p(x, \theta_0)$) as marginal increases of power (size) when the point x is added to the critical region C. The Neyman-Pearson Lemma shows that those points contributing more power increase per unit increase in size should be included in C for an optimal test.

- For any monotone increasing function f, the test $\{x \in \mathcal{X} | (f \circ \lambda)(x) \geq c'\}$ is identical to that is based on $\lambda(x)$. It is hence also an LR test. Indeed, the LR tests are often based on monotone increasing transformations of λ whose null distributions are easier to obtain.

For composite hypotheses, we have the generalized LR test based on the ratio

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_1} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)}.$$

The Neyman-Pearson Lemma does not apply to the generalized LR test. However, it performs well in many contexts.

Example 1: Simple Student-t Test

First consider a simple example. Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, 1)$, and we test

$$\mathrm{H}_0: \ \mu = 0 \quad \text{against} \quad \mathrm{H}_1: \mu = 1$$

Since both the null and the alternative are simple, Neyman-Pearson Lemma ensures that the likelihood ratio test is the best test. The likelihood ratio is

$$\begin{aligned} \lambda(x) &= \frac{p(x,1)}{p(x,0)} \\ &= \frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - 1)^2\right)}{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i - 0)^2\right)} \\ &= \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i - \frac{n}{2}\right). \end{aligned}$$

We know that $\tau(X) = n^{-1/2}\sum_{i=1}^n X_i$ is distributed as $\mathrm{N}(0,1)$ under the null. We may use this construct a test. Note that we can write $\tau(x) = f \circ \lambda(x)$, where $f(z) = n^{-1/2}(\log z + n/2)$ is a monotone increasing function. The test

$$C = \{x | \tau(x) \geq c\}$$

is then an LR test. It remains to determine c. Suppose we allow the probability of type-I error to be 5%, that is a size of 0.05, we may solve for c the equation $P_0(C) = 0.05$. Since $\tau(X) \sim N(0,1)$ under the null, we can look up the $N(0,1)$ table and find that

$$P_0(x | \tau(x) \geq 1.645) = 0.05.$$

This implies $c = 1.645$.

Example 2: One-Sided Student-t Test

Now we test

$$\mathrm{H}_0: \ \mu = 0 \quad \text{against} \quad \mathrm{H}_1: \mu > 0. \tag{5.2}$$

The alternative hypothesis is now composite. From the preceding analysis, however, it is clear that for any $\mu_1 > 0$, C is the most powerful test for

$$\mathrm{H}_0 : \ \mu = 0 \quad \text{against} \quad \mathrm{H}_1 : \mu = \mu_1.$$

We conclude that C is the uniformly most powerful test.

Example 3: Two-Sided F Test

Next we let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, and test

$$\mathrm{H}_0 : \ \mu = \mu_0 \quad \text{against} \quad \mathrm{H}_1 : \mu \neq \mu_0.$$

Here we have two unknown parameters, μ and σ^2, but the null and the alternative hypotheses are concerned with the parameter μ only. We consider the generalized LR test with the following generalized likelihood ratio

$$\lambda(x) = \frac{\sup_{\mu,\sigma^2} (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right)}{\sup_{\sigma^2} (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\right)}.$$

Recall that the ML estimator of μ and σ^2 are

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

Hence $\hat{\mu}$ and $\hat{\sigma}^2$ achieve the sup on the numerator. On the denominator,

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2$$

achieves the sup. Then we have

$$\begin{aligned}\lambda(x) &= \frac{(2\pi\hat{\sigma}^2)^{-n/2} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i - \hat{\mu})^2\right)}{(2\pi\tilde{\sigma}^2)^{-n/2} \exp\left(-\frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^n (x_i - \mu_0)^2\right)} \\ &= \left(\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{n/2} \\ &= \left(1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{n/2}.\end{aligned}$$

We define

$$\tau(x) = (n-1)\frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

It is clear that τ is a monotone increasing transformation of λ. Hence the generalized LR test is given by $C = \{x|\tau(x) \geq c\}$ for a constant c. Note that

$$\tau(X) = \frac{V_1/1}{V_2/(n-1)},$$

where

$$V_1 = \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}\right)^2 \quad \text{and} \quad V_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}.$$

Under H_0, we can show that $V_1 \sim \chi_1^2$, $V_2 \sim \chi_{n-1}^2$, and V_1 and V_2 are independent. Hence, under H_0,

$$\tau(X) \sim F_{1,n-1}.$$

To find the critical value c for a size-α test, we look up the F table and find constants $F_{1,n-1}(\alpha)$ such that

$$P_0\{x|\tau(x) \leq F_{1,n-1}(\alpha)\} = \alpha.$$

From the preceding examples, we may see that the hypothesis testing problem consists of three steps in practice: first, forming an appropriate test statistic, second, finding the distribution of this statistic under H_0, and finally making a decision. If the outcome of the test statistic is deemed as unlikely under H_0, the null hypothesis H_0 is rejected, in which case we *accept* H_1. The Neyman-Peason Lemma gives important insights on how to form a test statistic that leads to a powerful test. In the following example, we illustrate a direct approach that is not built on likelihood ratio.

Example 4: Two-Sided Student-t Test

For the testing problem of Example 3, we may construct a Student-t test statistic as follows,

$$\tilde{\tau}(x) = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)}}.$$

However, $\tilde{\tau}$ is *not* a monotone increasing transformation of λ. Hence the test based on $\tilde{\tau}$ is *not* a generalized LR test any more. However, we can easily derive the distribution of $\tilde{\tau}$ if the null hypothesis is true. Indeed, we have

$$\tilde{\tau}(X) = \frac{Z}{\sqrt{V/(n-1)}},$$

where

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \quad \text{and} \quad V = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}.$$

Under H_0, we can show that $Z \sim N(0,1)$, $V \sim \chi^2_{n-1}$, and Z and V are independent. Hence, under H_0,

$$\tilde{\tau}(X) \sim t_{n-1}.$$

To find the critical value c for a size-α test, we look up the t table and find a constant $t_{n-1}(1-\alpha/2) > 0$ such that

$$P_0\{x| - t_{n-1}(1-\alpha/2) \leq \tilde{\tau}(x) \leq t_{n-1}(1-\alpha/2)\} = 1-\alpha.$$

Finally, to see the connection between this test and the F test in Example 3, note that $F_{1,n-1} \equiv t^2_{n-1}$.

5.4 Exercises

1. Let X_1 and X_2 be independent Poisson(λ). Show that $\tau = X_1 + X_2$ is a sufficient statistic.

2. Let $(X_i, i = 1, \ldots, n)$ be a random sample from the underlying distribution given by the density
$$p(x,\theta) = \frac{2x}{\theta^2} \,\mathrm{I}\{0 \leq x \leq \theta\}.$$
(a) Find the MLE of θ.
(b) Show that $T = \max\{X_1, \ldots, X_n\}$ is sufficient.
(c) Let
$$\begin{aligned} S_1 &= (\max\{X_1, \ldots, X_m\}, \max\{X_{m+1}, \ldots, X_n\}), \\ S_2 &= (\max\{X_1, \ldots, X_m\}, \min\{X_{m+1}, \ldots, X_n\}), \end{aligned}$$
where $1 < m < n$. Discuss the sufficiency of S_1 and S_2.

3. Let $(X_i, i = 1, \ldots, n)$ be i.i.d. Uniform($\alpha - \beta, \alpha + \beta$), where $\beta > 0$, and let $\theta = (\alpha, \beta)$.
(a) Find a minimal sufficient statistic τ for θ.
(b) Find the ML estimator $\hat{\theta}_{\mathrm{ML}}$ of θ. (Hint: Graph the region for θ such that the joint density $p(x,\theta) > 0$.)
(c) Given the fact that τ in (a) is complete, find the UMVU estimator of α. (Hint: Note that $\mathbb{E}_\theta(X_1) = \alpha$.)

4. Let $(X_i, i = 1, \ldots, n)$ be a random sample from a normal distribution with mean μ and variance σ^2. Define
$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad S_n^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}.$$

(a) Obtain the Cramer-Rao lower bound.
(b) See whether $\overline{X}_n$ and S_n^2 attain the lower bound.
(c) Show that $\overline{X}_n$ and S_n^2 are jointly sufficient for μ and σ^2.
(d) Are $\overline{X}_n$ and S_n^2 the UMVU estimators?

5. Let X_1 and X_2 be independent and uniformly distributed on $(\theta, \theta+1)$. Consider the two tests with critical regions C_1 and C_2 given by

$$\begin{aligned} C_1 &= \{(x_1, x_2) | x_1 \geq 0.95\}, \\ C_2 &= \{(x_1, x_2) | x_1 + x_2 \geq c\}, \end{aligned}$$

to test $H_0 : \theta = 0$ versus $H_1 : \theta = 1/2$.
(a) Find the value of c so that C_2 has the same size as C_1.
(b) Find and compare the powers of C_1 and C_2.
(c) Show how to get a test that has the same size, but is more powerful than C_2.

Chapter 6

Asymptotic Theory

6.1 Introduction

Let $X_1, \ldots, X_n$ be a sequence of random variables, and let $\hat{\beta}_n = \hat{\beta}(X_1, \ldots, X_n)$ be an estimator for the population parameter β. For $\hat{\beta}_n$ to be a good estimator, it must be *asymptotically consistent*, ie, $\hat{\beta}_n$ converges to β in some sense as $n \to \infty$. Furthermore, it is desirable to have an *asymptotic distribution* of β_n, if properly standardized. That is, there may be a sequence of number a_n such that $a_n(\hat{\beta}_n - \beta)$ converges in some sense to a random variable Z with a known distribution. If in particular Z is normal (or Gaussian), we say $\hat{\beta}_n$ is asymptotically normal.

Asymptotic distribution is also important for hypothesis testing. If we can show that a test statistic has an asymptotic distribution, then we may relax assumptions on the finite sample distribution of $X_1, \ldots, X_n$. This would make our test more robust to mis-specifications of the model.

We study basic asymptotic theories in this chapter. They are essential tools for proving asymptotic consistency and deriving asymptotic distributions. In this section we first study the convergence of a sequence of random variables. As a sequence of measurable functions, the converging behavior of random variables is much richer than that of real numbers.

6.1.1 Modes of Convergence

Let (X_n) and X be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 6.1.1 (a.s. Convergence) *X_n converges almost surely (a.s.) to X,* *written as $X_n \to_{a.s.} X$, if*

$$\mathbb{P}\{\omega | X_n(\omega) \to X(\omega)\} = 1.$$

Equivalently, the a.s. convergence can be defined as

$$\mathbb{P}\{\omega | \, |X_n(\omega) - X(\omega)| > \epsilon \; i.o.\} = 0.$$

or

$$\mathbb{P}\{\omega | \, |X_n(\omega) - X(\omega)| < \epsilon \; e.v.\} = 1.$$

Definition 6.1.2 (Convergence in Probability) *X_n converges in probability to X, written as $X_n \to_p X$, if*

$$\mathbb{P}\{\omega | \, |X_n(\omega) - X(\omega)| > \epsilon\} \to 0.$$

Remarks:

- The convergence in probability may be equivalently defined as

$$\mathbb{P}\{\omega | \, |X_n(\omega) - X(\omega)| \leq \epsilon\} \to 1.$$

- Most commonly, X in the definition is a degenerate random variable (or simply, a constant).
- The definition carries over to the case where X_n is a sequence of random vectors. In this case the distance measure $| \cdot |$ should be replaced by the Euclidian norm.

Definition 6.1.3 (L^p Convergence) *X_n converges in L^p to X, written as $X_n \to_{L^p} X$, if*

$$\mathbb{E}\, |X_n(\omega) - X(\omega)|^p \to 0, \quad p > 0.$$

In particular, if $p = 2$, L^2 convergence is also called the mean squared error convergence.

Definition 6.1.4 (Convergence in Distribution) *X_n converges in distribution to X, written as $X_n \to_d X$, if for every function f that is bounded and continuous a.s. in P_X,*

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X).$$

Remarks:

- Note that for the convergence in distribution, (X_n) and X need not be defined on a common probability space. It is not a convergence of X_n, but that of probability measure induced by X_n, ie, $P_{X_n}(B) = \mathbb{P} \circ X_n(B)$, $B \in \mathcal{B}(\mathbb{R})$.
- Recall that we may also call P_{X_n} the law of X_n. Thus the convergence in distribution is also called convergence in law. More technically, we may call convergence in distribution as weak convergence, as opposed to strong convergence in the set of probability measures. Strong convergence refers to convergence in the distance metric of probability measure (e.g., total variation metric).
- In the definition of convergence in distribution, the function f need not be continuous at every point. The requirement of a.s. continuity allows f to be discontinuous on a set $S \subset \mathbb{R}$ that $P_X(S) = 0$.

Without proof, we give the following three lemmas, each of which supplies an equivalent definition of convergence in distribution.

Lemma 6.1.5 *Let F_n and F be the distribution function of X_n and X, respectively. $X_n \to_d X$ if and only if*

$$F_n(x) \to F(x) \text{ for every continuous point } x \text{ of } F.$$

Lemma 6.1.6 *Let ϕ_n and ϕ be the characteristic function of X_n and X, respectively. $X_n \to_d X$ if and only if*

$$\phi_n(t) \to \phi(t) \text{ for all } t.$$

Lemma 6.1.7 *$X_n \to_d X$ if and only if $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every bounded and uniformly continuous function f.* [1]

We have

Theorem 6.1.8 *Both a.s. convergence and L^p convergence imply convergence in probability, which implies convergence in distribution.*

[1] A function $f : D \to \mathbb{R}$ is uniformly continuous on D if for every $\epsilon > 0$, there exists $\delta > 0$ such that $|f(x_1) - f(x_2)| < \epsilon$ for $x_1, x_2 \in D$ that satisfy $|x_1 - x_2| < \delta$.

Proof: (a) To show that a.s. convergence implies convergence in probability, we let $E_n = \{|X_n - X| > \epsilon\}$. By Fatou's lemma,

$$\lim_{n\to\infty} \mathbb{P}\{E_n\} = \limsup \mathbb{P}\{E_n\} \le \mathbb{P}\{\limsup E_n\} = \mathbb{P}\{E_n \ i.o.\}.$$

The conclusion follows.

(b) The fact that L^p convergence implies convergence in probability follows from the Chebysheve inequality

$$\mathbb{P}\{|X_n - X| > \epsilon\} \le \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p}.$$

(c) To show that convergence in probability implies convergence in distribution, we first note that for any $\epsilon > 0$, if $X > z + \epsilon$ and $|X_n - X| < \epsilon$, then we must have $X_n > z$. That is to say, $\{X_n > z\} \supset \{X > z + \epsilon\} \cap \{|X_n - X| < \epsilon\}$. Taking complements, we have

$$\{X_n \le z\} \subset \{X \le z + \epsilon\} \cup \{|X_n - X| \ge \epsilon\}.$$

Then we have

$$\mathbb{P}\{X_n \le z\} \le \mathbb{P}\{X \le z + \epsilon\} + \mathbb{P}\{|X_n - X| \ge \epsilon\}.$$

Since $X_n \to_p X$, $\limsup \mathbb{P}\{X_n \le z\} \le \limsup \mathbb{P}\{X \le z + \epsilon\}$. Let $\epsilon \downarrow 0$, we have

$$\limsup \mathbb{P}\{X_n \le z\} \le \mathbb{P}\{X \le z\}.$$

Similarly, using the fact that $X < z - \epsilon$ and $|X_n - X| < \epsilon$ imply $X_n < z$, we can show that

$$\liminf \mathbb{P}\{X_n \le z\} \ge \mathbb{P}\{X < z\}.$$

If $\mathbb{P}\{X = z\} = 0$, then $\mathbb{P}\{X \le z\} = \mathbb{P}\{X < z\}$. Hence

$$\limsup \mathbb{P}\{X_n \le z\} = \limsup \mathbb{P}\{X_n \le z\} = \mathbb{P}\{X \le z\}.$$

This establishes

$$\lim_{n\to\infty} F_n(z) = F(z) \text{ for every continuous point of } F.$$

Other directions of the theorem do not hold. And a.s. convergence does not imply L^p convergence, nor does the latter imply the former. Here are a couple of counter examples:

Counter Examples Consider the probability space $([0,1], \mathcal{B}([0,1]), \mu)$, where μ is Lebesgue measure and $\mathcal{B}([0,1])$ is the Borel field on $[0,1]$. Define X_n by

$$X_n(\omega) = n^{1/p} I_{0 \leq \omega \leq 1/n}, \quad p > 0,$$

and define Y_n by

$$Y_n = I_{(b-1)/a \leq \omega \leq b/a}, \quad n = a(a-1)/2 + b, \quad 1 \leq b \leq a, \quad a = 1, 2, \ldots.$$

It can be shown that $X_n \to 0$ a.s., but $\mathbb{E}X_n^p = 1$ for all n. On the contrary, $\mathbb{E}Y_n^p = 1/a \to 0$, but $Y_n(\omega)$ does not converge for any $\omega \in [0,1]$.

It also follows from the above counter examples that convergence in probability does not imply a.s. convergence. Suppose it does, we would have $\to_{L^p} \Rightarrow \to_p \Rightarrow \to_{\text{a.s.}}$. But we have

Theorem 6.1.9 *If $X_n \to_p X$, then there exists a subsequence X_{n_k} such that $X_{n_k} \to_{a.s.} X$.*

Proof: For any $\epsilon > 0$, we may choose n_k such that

$$\mathbb{P}\left\{|X_{n_k} - X| > \epsilon\right\} \leq 2^{-k}.$$

Since

$$\sum_{k=1}^{\infty} \mathbb{P}\left\{|X_{n_k} - X| > \epsilon\right\} \leq \sum_{k=1}^{\infty} 2^{-k} < \infty,$$

Borel-Cantelli Lemma dictates that

$$\mathbb{P} \limsup_{n \to \infty} \{|X_{n_k} - X| > \epsilon\} = \mathbb{P}\{|X_{n_k} - X| > \epsilon \;\; i.o.\} = 0.$$

It is clear that convergence in distribution does not imply convergence in probability, since the former does not even require that X_n be defined on a common probability space. However, we have

Theorem 6.1.10 *Let X_n be defined on a common probability space and let c be constant. If $X_n \to_d c$, then $X_n \to_p c$.*

Proof: Let $f(x) = I_{|x-c|>\epsilon}$ for any $\epsilon > 0$. Since f is continuous at c and $X_n \to_d c$, we have

$$\mathbb{E}f(X_n) = \mathbb{P}\{|X_n - c| > \epsilon\} \to \mathbb{E}f(c) = 0.$$

Theorem 6.1.11 *Let f be a continuous function. We have,*

(a) if $X_n \to_{a.s.} X$, then $f(X_n) \to_{a.s.} f(X)$,

(b) if $X_n \to_p X$, then $f(X_n) \to_p f(X)$,

(c) if $X_n \to_d X$, then $f(X_n) \to_d f(X)$. (Continuous Mapping Theorem)

Proof: (a) Omitted.
(b) For any $\epsilon > 0$, there exists $\delta > 0$ such that $|x - c| \leq \delta$ implies $|f(x) - f(c)| \leq \epsilon$. So we have

$$\{|X_n - X| \leq \delta\} \subset \{|f(X_n) - f(X)| \leq \epsilon\},$$

which implies

$$\{|X_n - X| > \delta\} \supset \{|f(X_n) - f(X)| \leq \epsilon\}.$$

Hence

$$\mathbb{P}\{|X_n - X| > \delta\} \geq \mathbb{P}\{|f(X_n) - f(X)| > \epsilon\}.$$

The theorem follows.
(c) It suffices to show that for any bounded and continuous function g,

$$\mathbb{E}g(f(X_n)) \to \mathbb{E}g(f(X)).$$

But this is guaranteed by $X_n \to_d X$, since $g \circ f$ is also bounded and continuous.

Using the above results, we easily obtain,

Theorem 6.1.12 (Slutsky Theorem) *If $X_n \to_d c$ and $Y_n \to_p Y$, where c is a constant, then*

(a) $X_n Y_n \to_d cY$,

(b) $X_n + Y_n \to_d c + Y$.

6.1.2 Small o and Big O Notations

We first introduce small o and big O notations for sequences of real numbers.

Definition 6.1.13 (Small o and Big O) *Let (a_n) and (b_n) be sequences of real numbers. We write $x_n = o(a_n)$ and $y_n = O(b_n)$, respectively, when*

$$\frac{x_n}{a_n} \to 0 \quad \text{and} \quad \left|\frac{y_n}{b_n}\right| < M$$

for some constant $M > 0$.

Remarks:

- In particular, if we take $a_n = b_n = 1$ for all n, the sequence $x_n = o(1)$ converges to zero and sequence $y_n = O(1)$ is bounded.
- We may write $o(a_n) = a_n o(1)$ and $O(b_n) = b_n O(1)$. However, these are *not* equalities in the usual sense. It is understood that $o(1) = O(1)$ but $O(1) \neq o(1)$.
- For $y_n = O(1)$, if suffices to have $|y_n| < M$ for large n. If $|y_n| < M$ for all $n > N$, then we would have $|y_n| < M^*$ for all n, where $M^* = \max\{y_1, y_2, \ldots, y_n, M\}$.
- $O(o(1)) = o(1)$
  Proof: Let $x_n = o(1)$ and $y_n = O(x_n)$. It follows from $|y_n/x_n| < M$ that $|y_n| < M|x_n| \to 0$.
- $o(O(1)) = o(1)$
  Proof: Let $x_n = O(1)$ and $y_n = o(x_n)$. It follows from $|y_n| < \frac{M}{|x_n|}|y_n| = M\frac{|y_n|}{|x_n|} \to 0$.
- $o(1)O(1) = o(1)$
  Proof: Let $x_n = o(1)$ and $y_n = O(x_n)$. It follows from $|x_n y_n| < M|x_n| \to 0$.
- In general, we have

$$O(o(a_n)) = O(a_n o(1)) = a_n O(o(1)) = a_n o(1) = o(a_n).$$

In probability, we have

Definition 6.1.14 (Small o_p and Big O_p) *Let X_n and Y_n be sequences of random variables. We say $X_n = o_p(a_n)$ if $X_n/a_n \to_p 0$, and $Y_n = O_p(b_n)$ if for any $\epsilon > 0$, there exists $M > 0$ such that $\mathbb{P}(|Y_n/b_n| > M) < \epsilon$.*

If we take $a_n = b_n = 1$ for all n, then $X_n = o_p(1) \to_p 0$, and for any $\epsilon > 0$, there exists $M > 0$ such that $\mathbb{P}(|Y_n| > M) < \epsilon$. In the latter case, we say that Y_n is stochastically bounded.

Analogous to the real series, we have the following results.

Lemma 6.1.15 *We have*

(a) $O_p(o_p(1)) = o_p(1)$,

(b) $o_p(O_p(1)) = o_p(1)$,

(c) $o_p(1)O_p(1) = o_p(1)$.

Proof: (a) Let $X_n = o_p(1)$ and $Y_n = O_p(X_n)$, we show that $Y_n = o_p(1)$. For any $\epsilon > 0$, since $|Y_n|/|X_n| \leq M$ and $|X_n| \leq M^{-1}\epsilon$ imply $|Y_n| \leq \epsilon$, we have $\{|Y_n| \leq \epsilon\} \supset \{|Y_n| \leq |X_n|M\} \cap \{|X_n| \leq M^{-1}\epsilon\}$. Taking complements, we have

$$\{|Y_n| > \epsilon\} \subset \{|Y_n| > |X_n|M\} \cup \{|X_n| > M^{-1}\epsilon\}.$$

Thus

$$\mathbb{P}\{|Y_n| > \epsilon\} \leq \mathbb{P}\{|Y_n|/|X_n| > M\} + \mathbb{P}\{|X_n| > M^{-1}\epsilon\}.$$

This holds for any $M > 0$. We can choose M such that the first term on the right be made arbitrarily small. And since M is a constant, the second term goes to zero. Thus $\mathbb{P}\{|Y_n| > \epsilon\} \to 0$, i.e., $Y_n = o_p(1)$.
(b) Let $X_n = O_p(1)$ and $Y_n = o_p(X_n)$, we show that $Y_n = o_p(1)$. For any $\epsilon > 0$ and $M > 0$, we have

$$\mathbb{P}\{|Y_n| > M\epsilon\} \leq \mathbb{P}\{|Y_n|/|X_n| > \epsilon\} + \mathbb{P}\{|X_n| > M\}.$$

The first term on the right goes to zero, and the second term can be made arbitrarily small by choosing a large M.
(c) Left for exercise.

In addition, we have

Theorem 6.1.16 *If $X_n \to_d X$, then*

(a) $X_n = O_p(1)$, *and*

(b) $X_n + o_p(1) \to_d X$.

Proof: (a) For any $\epsilon > 0$, we have sufficiently large M such that $\mathbb{P}(|X| > M) < \epsilon$, since $\{|X| > M\} \downarrow \emptyset$ as $M \uparrow \infty$. Let $f(x) = I_{|x|>M}$. Since $X_n \to_d X$ and f is bounded and continuous a.s., we have $\mathbb{E}(f(X_n)) = \mathbb{P}(|X_n| > M) \to \mathbb{E}f(X) = \mathbb{P}(|X| > M) < \epsilon$. Therefore, $\mathbb{P}(|X_n| > M) < \epsilon$ for large n.
(b) Let $Y_n = o_p(1)$. And let f be any uniformly continuous and bounded function and let $M = \sup |f(x)|$. For any $\epsilon > 0$, there exists a δ such that $|Y_n| \leq \delta$ implies $|f(X_n + Y_n) - f(X_n)| \leq \epsilon$. Hence

$$\begin{aligned}
& |f(X_n + Y_n) - f(X_n)| \\
= \; & |f(X_n + Y_n) - f(X_n)| \cdot I_{|Y_n| \leq \delta} + |f(X_n + Y_n) - f(X_n)| \cdot I_{|Y_n| > \delta} \\
\leq \; & \epsilon + 2MI_{|Y_n| > \delta}
\end{aligned}$$

Hence

$$\mathbb{E}|f(X_n + Y_n) - f(X_n)| \leq \epsilon + 2M\mathbb{P}\{|Y_n| > \delta\}.$$

Then we have

$$\begin{aligned} |\mathbb{E}f(X_n + Y_n) - \mathbb{E}f(X)| &= |\mathbb{E}[f(X_n + Y_n) - f(X_n) + f(X_n) - f(X)]| \\ &\leq \mathbb{E}|f(X_n + Y_n) - f(X_n)| + |\mathbb{E}f(X_n) - \mathbb{E}f(X)| \\ &\leq \epsilon + 2M\mathbb{P}\{|Y_n| > \delta\} + |\mathbb{E}f(X_n) - \mathbb{E}f(X)|. \end{aligned}$$

The third term goes to zero since $X_n \to_d X$, the second term goes to zero since $Y_n = o_p(1)$, and $\epsilon > 0$ is arbitrary. Hence $\mathbb{E}f(X_n + Y_n) \to \mathbb{E}f(X)$.

Corollary 6.1.17 *If $X_n \to_d X$ and $Y_n \to_p c$, then $X_nY_n \to_d cX$.*

Proof: We have

$$X_nY_n = X_n(c + o_p(1)) = cX_n + O_p(1)o_p(1) = cX_n + o_p(1).$$

Then the conclusion follows from CMT.

6.1.3 Delta Method

Let $\hat{\theta}_n$ be an estimator of the parameter θ with true value θ_0. If $\hat{\theta}_n$ is consistent, then we may write

$$\hat{\theta}_n = \theta_0 + o_p(1).$$

If, in addition, $\hat{\theta}_n$ has an asymptotic distribution with a_n convergence rate, then

$$\hat{\theta}_n = \theta_0 + O_p(1/a_n).$$

The delta method is used to derive the asymptotic distribution of $f(\hat{\theta}_n)$, when f is differentiable and $\hat{\theta}_n$ is asymptotically normal,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \to_d N(0, \Sigma).$$

Let $\Delta(\theta) = \partial f(\theta)/\partial\theta'$. The Taylor expansion of $f(\theta)$ around θ_0 gives

$$\begin{aligned} f(\hat{\theta}_n) &= f(\theta_0) + \Delta(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + o\left(\|\hat{\theta}_n - \theta_0\|\right) \\ &= f(\theta_0) + \Delta(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + o\left(O_p\left(1/\sqrt{n}\right)\right) \\ &= f(\theta_0) + \Delta(\theta_0)\left(\hat{\theta}_n - \theta_0\right) + o_p\left(1/\sqrt{n}\right). \end{aligned}$$

This implies

$$\sqrt{n}\left(f(\hat{\theta}_n) - f(\theta_0)\right) = \Delta(\theta_0)\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) + o_p(1) \to_d N(0, \Delta(\theta_0)\Sigma\Delta(\theta_0)').$$

Example 6.1.18 *Let $\theta = (\alpha, \beta)'$. If $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to_d N(0, \Sigma)$, then the asymptotic distribution of $\hat{\alpha}_n/\hat{\beta}_n$ is $N(0, \Delta(\theta_0)\Sigma\Delta(\theta_0)')$ with*

$$\Delta(\theta) = \left(\frac{\partial f}{\partial \alpha}, \frac{\partial f}{\partial \beta}\right) = \left(\frac{1}{\beta}, -\frac{\alpha}{\beta^2}\right).$$

6.2 Limit Theorems

6.2.1 Law of Large Numbers

The law of large numbers (LLN) states that sample average converges in some sense to the population mean. In this section we state three LLN's for independent random variables. It is more difficult to establish LLN's for sequences of random variables with dependence. Intuitively, every additional observation of dependent sequence brings less information to the sample mean than that of independent sequence.

Theorem 6.2.1 (Weak LLN (Khinchin)) *If $X_1, \ldots, X_n$ are i.i.d. with mean $\mu < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_p \mu.$$

Proof: We only prove the case when $\mathrm{var}(X_i) < \infty$. The general proof is more involved. The theorem follows easily from

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right)^2 = \frac{1}{n}\mathbb{E}(X_i - \mu)^2 \to 0,$$

since L^2 convergence implies convergence in probability.

Theorem 6.2.2 (Strong LLN) *If $X_1, \ldots, X_n$ are i.i.d. with mean $\mu < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} \mu.$$

Proof: Since the mean exists, we may assume $\mu = 0$ and prove

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} 0.$$

The general proof is involved. Here we prove the case when $\mathbb{E}X_i^4 < \infty$. We have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^4 &= \frac{1}{n^4}\left(\sum_{i=1}^n \mathbb{E}X_i^4 + 6\sum_{i\neq j}\mathbb{E}X_i^2X_j^2\right) \\
&= n^{-3}\mathbb{E}X_i^4 + 3\frac{n(n-1)}{n^4}\mathbb{E}X_i^2\mathbb{E}X_j^2 \\
&= O(n^{-2}).
\end{aligned}$$

This implies $\mathbb{E}\sum_{n=1}^\infty \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^4 < \infty$, which further implies $\sum_{n=1}^\infty \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^4 < \infty$ a.s. Then we have

$$\frac{1}{n}\sum_{i=1}^n X_i \to_{a.s} 0.$$

Without proof, we also give a strong LLN that only requires independence,

Theorem 6.2.3 (Kolmogorov's Strong LLN) *If $X_1, \ldots, X_n$ are independent with $\mathbb{E}X_i = \mu_i$ and var$(X_i) = \sigma_i^2$, and if $\sum_{i=1}^\infty \sigma_i^2/i^2 < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^n X_i \to_{a.s.} \frac{1}{n}\sum_{i=1}^n \mu_i.$$

The first application of LLN is in deducing the probability p of getting head in the coin-tossing experiment. If we define $X_i = 0$ when we get tail in the i-th tossing and $X_i = 1$ when we get head. Then the LLN guarantees that $\frac{1}{n}\sum_{i=1}^n X_i$ converges to $\mathbb{E}X_i = p \cdot 1 + (1-p)\cdot 0 = p$. This converge to a probability, indeed, is the basis of the "frequentist" interpretation of probability.

Sometimes we need LLN for measurable functions of random variables, say, $g(X_i, \theta)$, where θ is a non-random parameter vector taken values in Θ. The Uniform LLN's establishe that $\frac{1}{n}\sum_{i=1}^n g(X_i,\theta)$ converges in some sense uniformly in $\theta \in \Theta$. More precisely, we have

Theorem 6.2.4 (Uniform Weak LLN) *Let $X_1, \ldots, X_n$ be i.i.d., Θ be compact, and $g(x,\theta)$ be a measurable function that is continuous in x for every $\theta \in \Theta$. If $\mathbb{E}\sup_{\theta\in\Theta}|g(X,\theta)| < \infty$, then*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n g(X_i,\theta) - \mathbb{E}g(X_1,\theta)\right| \to_p 0.$$

6.2.2 Central Limit Theorem

The central limit theorem states that sample average, under suitable scaling, converges in distribution to a normal (Gaussian) random variable.

We consider the sequence $\{X_{in}\}$, $i = 1, \ldots, n$. Note that the sequence has double subscript in, with n denotes sample size and i the index within sample. We call such data structure as *double array.* We first state without proof the celebrated

Theorem 6.2.5 (Lindberg-Feller CLT) *Let $X_{1n}, \ldots, X_{nn}$ be independent with $\mathbb{E}X_i = \mu_i$ and $var(X_i) = \sigma_i^2 < \infty$. Define $\sigma_n^2 = \sum_{i=1}^n \sigma_i^2$. If for any $\epsilon > 0$,*

$$\frac{1}{\sigma_n^2} \sum_{i=1}^n \mathbb{E}(X_{in} - \mu_i)^2 I_{|X_{in}-\mu_i| > \epsilon\sigma_n} \to 0, \tag{6.1}$$

then

$$\frac{\sum_{i=1}^n (X_{in} - \mu_i)}{\sigma_n} \to_d N(0, 1).$$

The condition in (6.1) is called the Lindberg condition. As it is often difficult to check, we often use the Liapounov condition, which implies the Lindberg condition. The Liapounov condition states that if for some $\delta > 0$,

$$\sum_{i=1}^n \mathbb{E} \left| \frac{X_{in} - \mu_i}{\sigma_n} \right|^{2+\delta} \to 0. \tag{6.2}$$

To see that Liapounov is stronger than Lindberg, let $\xi_{ni} = \frac{X_{in}-\mu_i}{\sigma_n}$. We have

$$\sum_{i=1}^n \mathbb{E}\xi_{in}^2 I_{|\xi_{in}|>\epsilon} \le \frac{\sum_{i=1}^n \mathbb{E}|\xi_{in}|^3}{\epsilon}.$$

Using the Lindberg-Feller CLT, we obtain

Theorem 6.2.6 (Lindberg-Levy CLT) *If $X_1, \ldots, X_n$ are i.i.d. with mean zero and variance $\sigma^2 < \infty$, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \to_d N(0, \sigma^2).$$

Proof: Let $Y_{in} = X_i/\sqrt{n}$. Y_{in} is thus an independent double array with $\mu_i = 0$, $\sigma_i^2 = \sigma^2/n$, and $\sigma_n^2 = \sigma^2$. It suffices to check the Lindberg condition

$$\frac{1}{\sigma_n^2} \sum_{i=1}^n \mathbb{E}Y_{in}^2 I_{|Y_{in}|>\epsilon\sigma_n} = \frac{1}{\sigma^2}\mathbb{E}X_i^2 I_{|X_i|>\epsilon\sigma\sqrt{n}} \to 0$$

by dominated convergence theorem. Note that $Z_n = X_i^2 I_{|X_i|>\epsilon\sigma\sqrt{n}} \leq X_i^2 < \infty$ and $Z_n(\omega) \to 0$ for all $\omega \in \Omega$.

6.3 Asymptotics for Maximum Likelihood Estimation

As an application of the asymptotic theory we have learned, we present in this section the asymptotic properties of Maximum Likelihood Estimator (MLE). The tests based on MLE, such as likelihood ratio (LR), Wald test, and Lagrange multiplier (LM) test, are also discussed.

Throughout the section, we assume that $X_1, \ldots, X_n$ are i.i.d. random variables with a common distribution that belongs to a parametric family. We assume that each distribution in the parametric family admits a density $p(x, \theta)$ with respect to a measure μ. Let $\theta_0 \in \Theta$ denote the true value of θ, let P_0 the distribution with density $p(x, \theta_0)$, and let $E_0(\cdot) \equiv \int \cdot p(x, \theta_0) d\mu(x)$, an integral operator with respect to P_0.

6.3.1 Consistency of MLE

We first show that the expected log likelihood with respect to P_0 is maximized at θ_0. Let $p(x_i, \theta)$ and $\ell(x_i, \theta)$ denote the likelihood and the log likelihood, respectively. We consider the function of θ,

$$E_0 \ell(\cdot, \theta) = \int \ell(x, \theta) p(x, \theta_0) d\mu(x).$$

Lemma 6.3.1 *We have for all $\theta \in \Theta$,*

$$E_0 \ell(\cdot, \theta_0) \geq E_0 \ell(\cdot, \theta).$$

Proof: Note that $\log(\cdot)$ is a concave function. Hence by Jensen's inequality,

$$\begin{aligned} E_0 \ell(\cdot, \theta) - E_0 \ell(\cdot, \theta_0) &= E_0 \log \frac{p(\cdot, \theta)}{p(\cdot, \theta_0)} \\ &\leq \log E_0 \frac{p(\cdot, \theta)}{p(\cdot, \theta_0)} \\ &= \log \int \frac{p(\cdot, \theta)}{p(\cdot, \theta_0)} p(\cdot, \theta_0) d\mu(x) = 0. \end{aligned}$$

Under our assumptions, the MLE of θ_0 is defined by

$$\hat{\theta} = \text{argmax}_\theta \frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta).$$

We have

Theorem 6.3.2 (Consistency of MLE) *Under certain regularity conditions, we have*

$$\hat{\theta} \to_p \theta_0.$$

Proof: The regularity conditions ensure that the uniform weak LLN applies to $\ell(X_i,\theta)$,

$$\frac{1}{n}\sum_{i=1}^n \ell(X_i,\theta) \to_p E_0(\cdot,\theta)$$

uniformly in $\theta \in \Theta$. The conclusion then follows.

6.3.2 Asymptotic Normality of MLE

Theorem 6.3.3 *Under certain regularity conditions, we have*

$$\sqrt{n}(\hat{\theta}-\theta_0) \to_d N(0, I(\theta_0)^{-1}),$$

where $I(\cdot)$ is the Fisher's information.

Proof: The regularity conditions are to ensure:

(a) $n^{-1/2}\sum_{i=1}^n s(X_i,\theta_0) \to_d N(0,I(\theta_0))$.

(b) $n^{-1}\sum_{i=1}^n h(X_i,\theta_0) \to_p E_0 h(\cdot,\theta_0) = H(\theta_0) = -I(\theta_0)$.

(c) $\bar{s}(x,\theta) \equiv n^{-1}\sum_{i=1}^n s(x_i,\theta)$ is differentiable at θ_0 for all x.

(d) $\hat{\theta} = \theta_0 + O_p(n^{-1/2})$.

By Taylor's expansion,

$$\bar{s}(x,\theta) = \bar{s}(x,\theta_0) + \bar{h}(x,\theta_0)(\theta-\theta_0) + o(\|\theta-\theta_0\|).$$

We have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\hat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta_0) + \left(\frac{1}{n}\sum_{i=1}^n h(x_i,\theta_0)\right)\sqrt{n}(\hat{\theta}-\theta_0) + o_p(1).$$

Then

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= -\left(\frac{1}{n}\sum_{i=1}^n h(x_i,\theta_0)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta_0)+o_p(1)\\ &\to_d N(0,I(\theta_0)^{-1}).\end{aligned}$$

6.3.3 MLE-Based Tests

Suppose $\theta \in \mathbb{R}^m$. For simplicity, let the hypothesis be

$$\mathrm{H}_0 : \theta = \theta_0 \quad \mathrm{H}_1 : \theta \neq \theta_0.$$

We consider the following three celebrated test statistics:

$$\begin{aligned}\mathrm{LR} &= 2\left(\sum_{i=1}^n \ell(x_i,\hat{\theta})-\sum_{i=1}^n \ell(x_i,\theta_0)\right)\\ \mathrm{Wald} &= \sqrt{n}(\hat{\theta}-\theta_0)'I(\hat{\theta})\sqrt{n}(\hat{\theta}-\theta_0)\\ \mathrm{LM} &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s(x_i,\theta_0)\right)' I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s(x_i,\theta_0)\right).\end{aligned}$$

LR measures the difference between restricted likelihood and unrestricted likelihood. Wald measures the difference between estimated and hypothesized values of the parameter. And LM measures the first derivative of the log likelihood at the hypothesized value of the parameter. Intuitively, if the null hypothesis holds, all three quantities should be small.

For the Wald statistic, we may replace $I(\hat{\theta})$ by $\frac{1}{n}\sum_{i=1}^n s(X_i,\hat{\theta})s(X_i,\hat{\theta})'$, $-H(\hat{\theta})$, or $-\frac{1}{n}\sum_{i=1}^n h(X_i,\hat{\theta})$. The asymptotic distribution of Wald would not be affected.

Theorem 6.3.4 *Suppose the conditions in Theorem 6.3.3 hold. We have*

$$LR,\ Wald,\ LM \to_d \chi^2_m.$$

Proof: Using Taylor's expansion,

$$\begin{aligned}\bar{\ell}(x,\theta) &= \bar{\ell}(x,\theta_0)+\bar{s}(x,\theta_0)'(\theta-\theta_0)+\frac{1}{2}(\theta-\theta_0)'\bar{h}(x,\theta_0)(\theta-\theta_0)+o(\|\theta-\theta_0\|^2)\\ \bar{s}(x,\theta) &= \bar{s}(x,\theta_0)+\bar{h}(x,\theta_0)(\theta-\theta_0)+o(\|\theta-\theta_0\|).\end{aligned}$$

Plugging $\bar{s}(x,\theta_0)=\bar{s}(x,\theta)-\bar{h}(x,\theta_0)(\theta-\theta_0)-o(\|\theta-\theta_0\|)$ in the first equation above, we obtain

$$\bar{\ell}(x,\theta)=\bar{\ell}(x,\theta_0)+\bar{s}(x,\theta)(\theta-\theta_0)-\frac{1}{2}(\theta-\theta_0)'\bar{h}(x,\theta_0)(\theta-\theta_0)+o(\|\theta-\theta_0\|^2).$$

We then have

$$\sum_{i=1}^{n} \ell(X_i, \hat{\theta}) - \sum_{i=1}^{n} \ell(X_i, \theta_0) = -\frac{1}{2}\sqrt{n}(\hat{\theta} - \theta)' \left(\frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta) \right) \sqrt{n}(\hat{\theta} - \theta) + o_p(1),$$

since $\frac{1}{n}\sum_{i=1}^{n} s(X_i, \hat{\theta}) = 0$. The asymptotic distribution of LR then follows.

For the Wald statistic, we have under regularity conditions that $I(\theta)$ is continuous at $\theta = \theta_0$ so that $I(\hat{\theta}) = I(\theta_0) + o_p(1)$. Then the asymptotic distribution follows from $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, I(\theta_0)^{-1})$.

The asymptotic distribution of the LM statistic follows from $\frac{1}{n}\sum_{i=1}^{n} s(X_i, \theta_0) \to_d N(0, I(\theta_0))$.

6.4 Exercises

1. Suppose $X_1, \ldots, X_n$ are i.i.d. Exponential(1), and define $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$.
(a) Find the characteristic function of X_1.
(b) Find the characteristic function of $Y_n = \sqrt{n}(\overline{X}_n - 1)$.
(c) Find the limiting distribution of Y_n.

2. Prove the following statements from the definition of convergence in probability,
(a) $o_p(1)o_p(1) = o_p(1)$
(b) $o_p(1)O_p(1) = o_p(1)$.

3. Let $X_1, \ldots, X_n$ be a random sample from a $N(0, \sigma^2)$ distribution. Let $\overline{X}$ be the sample mean and let S_n be the second sample moment $\sum_{i=1}^{n} X_i^2/n$. Using the asymptotic theory, find an approximation to the distribution of each of the following statistics:
(a) S_n.
(b) $\log S_n$.
(c) $\overline{X}_n/S_n$.
(d) $\log(1 + \overline{X}_n)$.
(e) $\overline{X}_n^2/S_n$.

4. A random sample of size n is drawn from a normal population with mean θ and variance θ, i.e., the mean and variance are known to be equal but the common value is not known. Let $\overline{X}_n = \sum_{i=1}^{n} X_i/n$, $S_n^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. and $T_n = \sum_{i=1}^{n} X_i^2/n$.

(a) Calculate $\pi = \text{plim}_{n\to\infty} T_n$.
(b) Find the maximum-likelihood estimator of θ and show that it is a differentiable function of T_n.
(c) Find the asymptotic distribution of T_n, i.e., find the limit distribution of $\sqrt{n}(T_n - \pi)$.
(d) Derive the asymptotic distribution of the ML estimator by using the delta method.
(e) Check your answer to part (d) by using the information to calculate the asymptotic variance of the ML estimator.
(f) Compare the asymptotic efficiencies of the ML estimator, the sample mean $\overline{X}_n$, and the sample variance S_n^2.

References

Bierens, Herman J. (2005), Introduction to the Mathematical and Statiscal Foundations of Econometrics, Cambridge University Process.

Chang, Yoosoon & Park, Joon Y. (1997), Advanced Probability and Statistics for Economists, Lecture Notes.

Dudley, R.M (2003), Real Analysis and Probability (2nd Ed.), Cambridge University Process.

Rosenthal, Jeffrey S. (2006), A First Look at Rigorous Probability Theory (2nd Ed.) World Scientific.

Williams, David (2001), Probability with Martingales, Cambridge University Process.

Su, Liangjun (2007), Advanced Mathematical Statistics (in Chinese), Peking University Press.

Index