An Intermediate Course in Probability and Statistics

Shiu-Sheng Chen[†]

First Version: September 2008
Current Version: November 2020

Department of Economics, National Taiwan University. email: sschen@ntu.edu.tw

Introduction

This lecture note aims at equipping students with advanced probability theory and statistical tools. It provides an intermediate level coverage of material suitable for students having taken introductory statistics.

References

1. Fraser, D.A.S. (1976). Probability and Statistics–Theory and Applications, Duxbury Press.
2. Gut, Allan (2009). An Intermediate Course in Probability, Springer-Verlag (2nd ed).
3. Gut, Allan (2013). Probability: A Graduate Course, Springer-Verlag (2nd ed).
4. Hansen, Bruce (2008). Econometrics, manuscript.
5. Lehmann, Erich L. (2001). Elements of Large-Sample Theory, Springer-Verlag.
6. Mittelhammer, Ron C. (1995). Mathematical Statistics for Economics and Business, Springer-Verlag.
7. Ramanathan, Ramu (1993). Statistical Methods in Econometrics, Academic Press.
8. Resnick, Sideny I. (2001). A Probability Path, Birkhäuser.
9. Roussas, George G. (2002). A course in Mathematical Statistics, Elsevier.
10. White, Halbert (2001). Asymptotic Theory for Econometricians, Academic Press.

Contents

Chapter 1

Probability Model and Random Variables

1.1 Sets

We provide a brief review of the notations and operations on sets.

Set Operations

1. Union: $A \cup B = \{x : x \in A \ \text{ or } \ x \in B\}$;
2. Intersection: $A \cap B = \{x : x \in A \ \text{ and } \ x \in B\}$;
3. Complement: $A^c = \{x : x \notin A\}$;
4. Difference: $A - B = A \cap B^c = \{x : x \in A \ \text{ and } \ x \notin B\}$;
5. Symmetric difference: $A \triangle B = (A - B) \cup (B - A)$.

Some Additional Terminology

1. The empty set: $\varnothing = \{\}$;
2. Subset: A is a subset of B, $A \subset B$, if $x \in A$ implies $x \in B$;
3. Disjoint: A and B are disjoint if $A \cap B = \varnothing$;
4. Power set associated with Ω: $2^{\Omega} = \{A : A \subset \Omega\}$, which is the set of all subsets of Ω.

Theorem 1. (de Morgan Formulas)

1. $\left(\bigcup_{k=1}^{n} A_k\right)^c = \bigcap_{k=1}^{n} A_k^c$;
2. $\left(\bigcap_{k=1}^{n} A_k\right)^c = \bigcup_{k=1}^{n} A_k^c$

Exercise 1. Prove the de Morgan Formulas.

The de Morgan Formulas can be extended to infinite unions and infinite intersections.

1. $\left(\bigcup_{k=1}^{\infty} A_k\right)^c = \bigcap_{k=1}^{\infty} A_k^c$;
2. $\left(\bigcap_{k=1}^{\infty} A_k\right)^c = \bigcup_{k=1}^{\infty} A_k^c$

1.2 Monotone Sequences of Sets

Our first discussion deals with sequences of sets and various types of limits of such sequences. The limits are also sets. We start with two simple definitions.

Definition 1 (Sequences of Sets). Suppose that $(A_1, A_2, \cdots)$ is a sequence of sets.

1. The sequence is increasing if $A_n \subset A_{n+1}$ for every $n \in \mathbb{N}_+$.
2. The sequence is decreasing if $A_{n+1} \subset A_n$ for every $n \in \mathbb{N}_+$.

If a sequence of sets is either increasing or decreasing, we can define the limit of the sequence in a way that turns out to be quite natural.

Definition 2 (Limit of Sets). Suppose that $(A_1, A_2, \cdots)$ is a sequence of sets.

1. If the sequence is increasing, we define
$$\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n$$
2. If the sequence is decreasing, we define
$$\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n$$

1.3 Probability Model

Definition 3 (Random Experiment). Random experiment is an action whose outcome is uncertain in advance (ex ante) of its occurrence.

For instance, tossing a coin, or throwing a die.

Definition 4 (Sample Space/State Space). The totality of all possible outcomes of a random experiment is referred to as sample space (state space), which is denoted by Ω. The distinct individual elements of Ω are called sample points or elementary events (denoted by ω).

1. Tossing a coin
$$\Omega = \{H, T\}$$
2. Throwing a die
$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

We now define a special set called σ-algebra.

Definition 5 (**σ-algebra/σ-field).** A σ-algebra $\mathcal{F}$ is a non-empty collection of subsets of Ω, which satisfies

1. $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$.
2. $A_i \in \mathcal{F}\ \forall i \geq 1$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

In words, a σ-algebra $\mathcal{F}$ is simply a nonempty collection of subsets of Ω that is closed under complement and taking countable unions.

Theorem 2. According to the definition of the σ-algebra, we have

1. $\Omega \in \mathcal{F}$.
2. $\varnothing \in \mathcal{F}$.
3. $A_i \in \mathcal{F}\ \forall i \geq 1$ implies $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Proof. The first two are trivial. For the third one,
$$A_i \in \mathcal{F} \Rightarrow A_i^c \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{F} \Rightarrow \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c \in \mathcal{F} \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}.$$
□

Example 1. Consider tossing a coin twice,
$$\Omega = \{HH, HT, TH, TT\}$$

1. $A = \{\Omega, \varnothing, HH, (HT, TH, TT)\}$ is a σ-algebra.
2. $B = \{\Omega, \varnothing, (HH, TT)\}$ is not a σ-algebra.

Example 2. We can further provide some special σ-algebras.

1. The Power set associated with Ω, 2^{Ω} (i.e., the collection of *all* subsets of Ω) is a σ-algebra.
2. $\{\varnothing, \Omega\}$ is a trivial σ-algebra consisting of only two types of events: "nothing happens" and "something happens."

We now need a *mathematical model* of a random experiment. Before introducing the probability model, I would recommend you to read pages 1–2 in Gut (2009), which provide a very intuitive discussion to motivate the probability model.[1]

Definition 6 (Probability Model). Random experiment (or random phenomenon) can be represented by a probability space $(\Omega, \mathcal{F}, P)$, where

- Ω is the sample space (state space) including all possible outcomes of the experiment.
- $\mathcal{F}$ is the σ-algebra of subsets of Ω (event space).
- $P(\cdot) : \mathcal{F} \mapsto [0,1]$ is the probability measure assigned to any element of $\mathcal{F}$, and satisfies the following axioms:
  1. $0 \leq P(A), \forall A \in \mathcal{F}$.
  2. $P(\Omega) = 1$.
  3. $P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j)$, $A_j \in \mathcal{F}$ and A_j's are disjoint.

Axiom 3 is called *countable additivity*. A probability measure $P(\cdot)$ is also called a probability function. It is worth noting that the probability model only specifies what qualification a function has to have in order to be entitled to be called a probability. It does not tell us *what probability really is*. In general, for a countable sample space Ω, it will be possible to define more than one probability function. For example, let $\Omega = \{H, T\}$. Consider the function $f : 2^{\Omega} \mapsto [0,1]$, and $g : 2^{\Omega} \mapsto [0,1]$ as in Table 1.1.

Table 1.1: Alternative Probability Function

A	$f(A)$	g(A)
$\varnothing$	0	0
$\{H\}$	1/2	1/3
$\{T\}$	1/2	2/3
$\{H, T\}$	1	1

The reason we would like to impose some mathematical structures on the set of all events (i.e., the σ-algebra) is to make sure that we may construct new events from old ones without trouble assigning probabilities. For instance, given that we know the probability of event A, it may be of interest to know the probability that event A does not happen, $P(A^c)$. Moreover, suppose that

[1] Gut, Allan (2009). An Intermediate Course in Probability, Springer-Verlag.

we know the probabilities of events A and B, we may also want to know the probability of the event that either A or B happens, $P(A \cup B)$.

The σ-algebra is just a definition of which sets may be considered as events. Elements not in $\mathcal{F}$ simply have no defined probability measure. Basically, σ-algebras are the "patch" that lets us avoid some pathological behaviors of mathematics, namely non-measurable sets. If we restrict ourselves to countable sets, then we can take $\mathcal{F} = 2^{\Omega}$ the power set of Ω, and we won't have any of these problems because for countable Ω, 2^{Ω} consists only of measurable sets.

Note that under the context of a probability model,

1. Any subset A is called an event if and only if $A \in \mathcal{F}$.

2. A σ-algebra on sample space Ω is also called an *event space*.

The choice of σ-algebra depends on what we would like to model. Consider rolling a fair die once, $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we want to model the beliefs of a person who will be told after the experiment only whether or not 1 has come up, then a proper σ-algebra would be $\{\varnothing, \Omega, \{1\}, \{2, 3, 4, 5, 6\}\}$. On the other hand, the power set associated with Ω is also a candidate but it is not a good choice. For instance, the event $\{1, 2\}$ is not a conceivable event for the individual knowing only whether or not 1 has come up.

1.3.1 Continuity Theorem

Generally speaking, a function is continuous if it preserves limits. Thus, the following results are the continuity theorems of probability. Part 1 is the continuity theorem for increasing events and part 2 the continuity theorem for decreasing events.

Theorem 3 (Continuity Theorem). Suppose that $(A_1, A_2, \cdots)$ is a sequence of events.

1. If the sequence is increasing then

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\cup_{n=1}^{\infty} A_n)$$

2. If the sequence is decreasing then

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\cap_{n=1}^{\infty} A_n)$$

Proof.

1. Let $B_1 = A_1$ and let $B_i = A_i - A_{i-1} = A_i \cap A_{i-1}^c$ for $i = 2, 3, \ldots$. Note that the collection of events $\{B_1, B_2, \ldots\}$ is pairwise disjoint and

$$\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$$

Then

$$P\left(\cup_{i=1}^{\infty} A_i\right) = P\left(\cup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i)$$

But

$$P(B_1) = P(A_1)$$

and

$$P(B_i) = P(A_i) - P(A_{i-1}) \text{ for } i = 2, 3, \ldots.$$

Therefore,

$$\sum_{i=1}^{n} P(B_i) = P(A_n)$$

and hence we have

$$P\left(\cup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n)$$

2. Since the sequence $\{A_1, A_2, \ldots, \}$ is decreasing, the sequence of complements $\{A_1^c, A_2^c, \ldots\}$ is increasing. Hence using the result in Part 1 with DeMorgan's law, we have

$$\begin{aligned} P\left(\cap_{i=1}^{\infty} A_i\right) &= 1 - P\left(\left(\cap_{i=1}^{\infty} A_i\right)^c\right) = 1 - P\left(\cup_{i=1}^{\infty} A_i^c\right) \\ &= 1 - \lim_{n\to\infty} P(A_n^c) = \lim_{n\to\infty} \left[1 - P(A_n^c)\right] = \lim_{n\to\infty} P(A_n) \end{aligned}$$

□

1.4 Random Variables

Definition 7 (Random Variable). Random variable is simply a *measurable* function, $X(\omega) : \Omega \mapsto \mathbb{R}$.

Given a probability space $(\Omega, \mathcal{F}, P)$, a random variable X is measurable if for every x,

$$\{\omega : X(\omega) \leq x\} \in \mathcal{F},$$

In plain English, "measurable" just means "nice."

So a random variable $X(\cdot)$ is a function whose value is determined by the outcome of an experiment. Note that X is a function whose domain is the sample space Ω, and whose codomain is the set of real numbers $\mathbb{R}$.

For instance, given $\Omega = \{H, T\}$,

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = H, \\ 0 & \text{if } \omega = T. \end{cases}$$

Definition 8 (Cumulative Distribution Function, CDF). A random variable is described by a CDF

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

We will from now on denote $F_X(x)$ as $F(x)$ to prevent global warming. CDF has following properties.

Proposition 1. (**CDF**) Let X be a random variable with CDF, $F(x) = P(X \leq x)$.

1. If X and Y have the same CDF, they are said to be *identically distributed*. We denote it as $X \overset{d}{=} Y$.
2. $F(-\infty) = 0$.
3. $F(\infty) = 1$.
4. $F(x) \geq 0 \; \forall x \in \mathbb{R}$.
5. $F(\cdot)$ is non-decreasing (weakly increasing).
6. $F(\cdot)$ is right continuous
$$\lim_{h \to 0^+} F(x+h) = F(x).$$

Proof.

1. Omit. Beyond the scope of this note.
2. Let $x_1 > x_2 > \cdots$ be a decreasing sequence with $x_n \to -\infty$ as $n \to \infty$. The intervals $(-\infty, x_n]$ are decreasing in n and have intersection $\varnothing$. The result now follows from Theorem 3 for decreasing events.
3. Let $x_1 < x_2 < \cdots$ be an increasing sequence with $x_n \to \infty$ as $n \to \infty$. The intervals $(-\infty, x_n]$ are increasing in n and have union $\mathbb{R}$. The result now follows from Theorem 3 for increasing events.
4. Trivial as $F(x) = P(X \leq x) \geq 0$.
5. $F(\cdot)$ is non-decreasing (weakly increasing). Clearly, for $a \leq b$
$$F(b) = P(X \leq b) = P(X \leq a) + P(a < X \leq b) = F(a) + P(a < X \leq b)$$
Hence,
$$F(a) \leq F(b).$$

6. Fix $x \in \mathbb{R}$. Let $x_1 > x_2 > \cdots$ be a decreasing sequence with $x_n \to x$ as $n \to \infty$. The intervals $(-\infty, x_n]$ are decreasing in n and have intersection $(-\infty, x]$. The result now follows from Theorem 3 for decreasing events.

□

1.4.1 Two Types of Random Variables

Recall the definition of random variables,

$$X(\omega) : \Omega \mapsto \mathbb{R}$$

According to the *range* of a random variable X, we have different types of random variables: *discrete* and *continuous*. The range of a discrete random variable is countable, while the range of a continuous random variable is uncountable

If the domain Ω is countable, then the range of X, dented by $X(\Omega)$, is countable as well. Hence, X is a discrete random variable. It is worth noting that the codomain of a random variable is $\mathbb{R}$, which is uncountable. On the other hand, if both the domain and range are uncountable, X is called a *continuous* random variable. However, it is possible to define a discrete random variable on a continuous (uncountable) sample space. For example, for a continuous sample space, $\Omega = (0, 1)$, the random variable defined by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in (0, 1/2] \\ 0, & \text{if } \omega \in (1/2, 1) \end{cases}$$

is discrete.

Table 1.2 shows four combinations.

Table 1.2: Domain and Range of Random Variables

Domain of X Ω	Range of X $X(\Omega)$	Random Variable $X(\omega)$
Countable	Countable	Discrete random variable on a discrete sample space
Countable	Uncountable	This combination cannot happen
Uncountable	Countable	Discrete random variable on a continuous sample space
Uncountable	Uncountable	Continuous random variable on a continuous sample space

1.4.2 Discrete Random Variables

For discrete random variables, probability is assigned using the probability mass function.

Definition 9 (Probability Mass Function, pmf). Suppose that X is a discrete random variable, taking values on some countable sample space $B \subseteq \mathbb{R}$. Then the probability mass function $f(x)$ for X is given by $f(x) : \mathbb{R} \mapsto [0,1]$

$$f(x) = \begin{cases} P(X = x), & x \in B \\ 0, & x \in \mathbb{R} - B \end{cases}$$

so that

1. $f(x) > 0, \forall x \in B$.
2. $\sum_{x \in B} f(x) = 1$.
3. Given that $A \subseteq B$, $P(X \in A) = \sum_{x \in A} f(x)$.

We would like to introduce the support of a random variable here.

Definition 10 (Support). The support of a random variable X, denoted by $\text{supp}(X)$, is the set of points where its density is positive.

$$\text{supp}(X) = \{x \in \mathbb{R} : f(x) > 0\}.$$

Clearly, set B is the support of the discrete random variable X.

Example 3 (Bernoulli Distribution). Random variable $X \sim \text{Bernoulli}(p)$ if the pmf is

$$f(x) = p^x(1-p)^{1-x}, \quad \text{supp}(X) = \{0,1\}$$

Example 4 (Binomial Distribution). The Binomial arises when we repeat Bernoulli trials n times. Let $\{X_i\}_{i=1}^n \sim^{i.i.d.} \text{Bernoulli}(p)$ and $Y = \sum_{i=1}^n X_i$, then

$$Y \sim \text{Binomial}(n,p)$$

with pmf

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad \text{supp}(Y) = \{y | y = 0,1,2,3,\ldots,n\}$$

Example 5 (Geometric Distribution). Let X denote the number of trails *until first success*. Then $X \sim \text{Geo}(p)$ with pmf

$$f(x) = (1-p)^x p, \quad \text{supp}(X) = \{x | x = 0,1,2,3,\ldots\}$$

Example 6 (Poisson Distribution). $X \sim \text{Poisson}(\lambda)$ with pmf

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad \text{supp}(X) = \{x|x = 0, 1, 2, 3, \ldots\}$$

The applications of the Poisson distribution includes

1. The number of soldiers of the Prussian army killed accidentally by horse kick per year (von Bortkewitsch, 1898, p. 25).[2]

2. The number of bankruptcies that are filed in a month (Jaggia, Kelly, 2012 p. 158).[3]

3. The number of arrivals at a car wash in one hour (Anderson et al., 2012, p. 236).[4]

4. The number of file server virus infection at a data center during a 24-hour period . The number of Airbus 330 aircraft engine shutdowns per 100,000 flight hours. The number of asthma patient arrivals in a given hour at a walk-in clinic (Doane, Seward, 2010, p. 232).[5]

5. The number of hungry persons entering McDonald's restaurant. The number of work-related accidents over a given production time, The number of birth, deaths, marriages, divorces, suicides, and homicides over a given per iod of time (Weiers, 2008, p. 187).[6]

6. The number of customers who call to complain about a service problem per month (Donnelly, Jr., 2012, p. 215).[7]

7. The number of visitors to a Web site per minute (Sharpie, De Veaux, Velleman, 2010, p. 654).[8]

8. The number of calls to consumer hot line in a 5-minute period (Pelosi, Sandifer, 2003, p. D1).[9]

9. The number of telephone calls per minute in a small business. The number of arrivals at a turnpike tollbooth par minute between 3 A.M. and 4 A.M. in January on the Kansas Turnpike (Black, 2012, p. 161).[10]

[2] Bortkewitsch, L. (1898). Das Gesetz der Kleinen Zah len. Leipzig, Germay: Teubner.

[3] Jaggia, S., Kelly, A. (2012) Business Statistics - Communicating with Numbers. New York, NY: McGraw-Hill Irvin.

[4] Anderson, D. R., Sweeney, D. J., Williams, T. A., (2012), Essentials of Modern Business Statistics with Microsoft Excel. Mason, OH: South-Western, Cengage Learning.

[5] Doane, D., Seward, L. (2010) Applied Statistics in Business and Economics, 3rd Edition, Mcgraw-Hill, 2010.

[6] Weiers, R. M. (2008) Introduction to Business Statistics. Mason, OH: South-Western, Cengage Learning.

[7] Donnelly, Jr., R. A. (2012) Business Statistics. Upper Saddle River, NJ: Pearson Education, Inc.

[8] Sharpie, N. R., De Veaux, R. D., Velleman, P. F. (2010) Business Statistics. Boston, MA: Addison Wesley.

[9] Pelosi, M. K., Sandifer, T.M. (2003) Elementary Statistics. New York, NY: John Wiley and Sons, Inc.

[10] Black, K. (2012) Business Statistics For Contemporary Decision Making. New York, NY: John Wiley and Sons, Inc.

1.4.3 Continuous Random Variables

If Ω is uncountable, and $F(x)$ is continuous on $\mathbb{R}$, the random variable is continuous. We use probability density function to assign probability.

Definition 11 (Probability Density Function, pdf). A probability density function is *any* function, $f : \mathbb{R} \mapsto \mathbb{R}$ such that

1. $f(x) > 0, \forall x \in \text{supp}(X)$.
2. $\int_{x \in \text{supp}(X)} f(x)dx = 1$.

If X is a continuous random variable, then

1. $P(a \leq X \leq b) = \int_a^b f(z)dz$.
2. $F(x) = \int_{-\infty}^x f(z)dz$.
3. $P(X = a) = 0$.
4. $\int_a^b f(z)dz = P(a \leq X < b) = P(a \leq X \leq b) = P(a < X < b) = P(a < X \leq b)$.

Example 7 (Uniform Distribution). $X \sim U[l, h]$, if the pdf is

$$f(x) = \frac{1}{h - l}, \quad \text{supp}(X) = \{x | l \leq x \leq h\}$$

Note that pdf is not unique! For instance, Figure 1.1 shows two possible pdfs of the $U(0, 1)$ random variable.

Theorem 4. Let $F(\cdot)$ be any distribution function and define

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}$$

to be its inverse function (quantile function) for $0 < t < 1$.

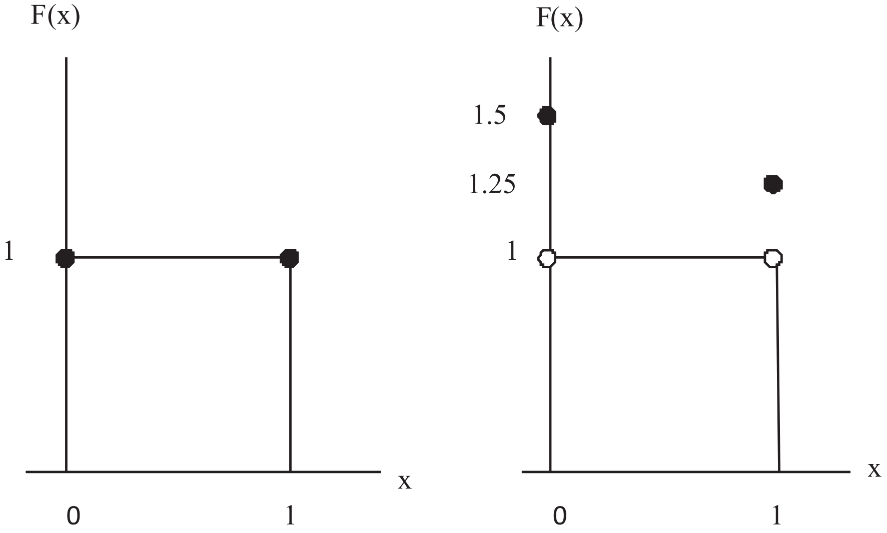
If $U \sim U[0, 1]$, and $X = F^{-1}(U)$, then the distribution function of X is $F(\cdot)$.

Proof. Since $F^{-1}(t) \leq x$ iff. $t \leq F(x)$,

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

□

Figure 1.1: Two Possible pdfs of the $U(0,1)$ Random Variable



A common application of Theorem 4 is the simulation of random variables with a particular distribution. Once we obtain a random variable $U \sim U[0,1]$ via simulation, we can then obtain a random variable $X = F^{-1}(U)$, which has distribution function $F(\cdot)$. For example, consider a $\exp(\beta)$ random variable with distribution function

$$F(x) = 1 - e^{-\frac{1}{\beta}x}$$

Since $F(x)$ is strictly increasing over the set where $F(x) > 0$, we can solve the inverse $F^{1}(t)$ via the equation

$$1 - e^{-\frac{1}{\beta}F^{-1}(t)} = t,$$

which gives us

$$F^{-1}(t) = -\beta \log(1-t).$$

Hence, let u be one simulated realization drawn $U[0,1]$, then

$$x = F^{-1}(u) = -\beta \log(1-u)$$

is a simulated realization drawn from $\exp(\beta)$.

However it is worth noting that $F^{-1}(U)$ is not necessarily easily computable.

Example 8 (**Normal Distribution).** A random variable X has the normal distribution with two parameters μ and σ^2 if X has a continuous distribution with the following pdf:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad \text{supp}(X) = \{x| - \infty < x < \infty\}$$

Let $z = \sqrt{2}x$, we thus have $dz = \sqrt{2}dx$, and

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}2x^2} \sqrt{2}dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-x^2} dx$$
$$= \frac{1}{\sqrt{\pi}}\sqrt{\pi} = 1, \quad \text{by the Gaussian Integral (see Theorem 79)}$$

i.e., the pdf of a $N(0,1)$ random variable is integrated to 1. Hence, let $y = \sigma z + \mu$, and thus $dy = \sigma dz$, then we have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} dy = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)^2} \sigma dz$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 1$$

i.e., the pdf of a $N(\mu, \sigma^2)$ random variable is integrated to 1.

Example 9 (**Gamma Distribution).** A continuous random variable X follows a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function is:

$$f(x) = \frac{x^{\alpha-1}e^{-\frac{1}{\beta}x}}{\beta^\alpha \Gamma(\alpha)}, \quad \text{supp}(X) = \{x|0 < x < \infty\},$$

where $\Gamma(\alpha)$ is the Gamma function (see Definition 67).

1. It is denoted by $X \sim \text{Gamma}(\alpha, \beta)$.

2. Given $\alpha = 1$, we obtain an Exponential distribution:

$$\exp(\beta) \overset{d}{=} \text{Gamma}(1, \beta)$$

3. Given $\alpha = \frac{k}{2}$ and $\beta = 2$, we obtain a χ^2 distribution with degree of freedom k:

$$\chi^2(k) \overset{d}{=} \text{Gamma}\left(\frac{k}{2}, 2\right)$$

Example 10 (**Student's *t* Distribution).** If a random variable X has the following pdf

$$\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{\sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

with support $\text{supp}(X) = \{x| - \infty < x < \infty\}$ and a parameter k, then it is called a Student's t distribution, and denoted by

$$X \sim t(k)$$

Example 11 (Log-Normal Distribution). A random variable X has a log-normal distribution if

$$\log X \sim N(\mu, \sigma^2).$$

1.4.4 Mixed Distribution Random Variables

We now show you a *mixed distribution* random variable for fun!

Exercise 2. Pick any $p \in (0,1)$. Let X be a random variable which has the following CDF:

$$F(x) = pI_{\{x \geq 0\}} + (1-p)\Phi(x),$$

where

$$I_{\{x \geq 0\}} = \begin{cases} 1 & \text{if} \quad x \geq 0, \\ 0 & \text{if} \quad \text{otherwise.} \end{cases}$$

an indicator function, and $\Phi(x) = P(N(0,1) \leq x)$ the CDF of the standard normal random variable. As a practical exercise, see if you can show the follows.

1. Check if $F(x)$ is indeed a CDF.

2. Plot $F(x)$.

3. Find out the pdf $f(x)$.

1.5 Moments

1.5.1 Expectation

- X a discrete random variable:

$$E(X) = \sum_{x_i \in \text{supp}(X)} x_i f(x_i).$$

- X a continuous random variable:

$$E(X) = \int_{x \in \text{supp}(X)} x f(x) dx.$$

Given $h(x)$ a function of random variable X,

$$E(h(X)) = \begin{cases} \sum_{x_i \in \text{supp}(X)} h(x_i) f(x_i), \\ \int_{x \in \text{supp}(X)} h(x) f(x) dx. \end{cases}$$

1.5.2 r-th Moment

$$E(X^r) = \begin{cases} \sum_{x_i \in \text{supp}(X)} x_i^r f(x_i), \\ \int_{x \in \text{supp}(X)} x^r f(x) dx. \end{cases}$$

is the r-th moment of X.

$$Var(X) = E[(X - E(X))^2] = \begin{cases} \sum_{x_i \in \text{supp}(X)} (x_i - E(X))^2 f(x_i), \\ \int_{x \in \text{supp}(X)} (x - E(X))^2 f(x) dx. \end{cases}$$

is the variance of X.

Let X be a random variable, we distinguish 3 cases.

1. $E(X)$ exists and is finite.
2. $E(X)$ exists and is infinite.
3. $E(X)$ does not exist.

Case 1 is a normal case, so we focus on examples of cases 2 and 3.

Example 12. ($E(X)$ **exists and is infinite**)

$$f(x) = \begin{cases} \frac{1}{x^2} & \text{if } x \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that

1. $f(x) \geq 0$ and $\int_1^\infty x^{-2} = 1$ is a pdf.
2. $E(X) = \int_1^\infty x x^{-2} dx = \int_1^\infty x^{-1} dx = \ln x]_1^\infty = \infty - 0 = \infty$.

Example 13. ($E(X)$ **does not exist**) A standard Cauchy random variable.

$$f(x) = \frac{1}{\pi} \frac{1}{1 + x^2}, \quad x \in \mathbb{R}.$$

Thus,

$$E(X) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{1 + x^2} = \frac{1}{2\pi} \log(1 + x^2)]_{x=-\infty}^{\infty} = \infty - \infty.$$

1.6 Quantile

Definition 12 (p–**th Quantile).** Let X be a random variable with CDF, $F(\cdot)$. Pick any $p \in (0, 1)$, a p-th quantile of X is a number x_p such that

$$P(X \leq x_p) \geq p \text{ and } P(X \geq x_p) \geq 1 - p.$$

That is,

$$p \leq F(x_p) \leq p + P(X = x_p).$$

Now suppose that X is a continuous variable, $P(X = x_p) = 0$, so the p-th quantile of X is a number x_p such that

$$F(x_p) = p.$$

If $p = 0.25$, x_p is called quartile. If $p = 0.5$, x_p is called median.

Example 14. Let X ~Bernoulli($\frac{1}{2}$). Then $x_{0.5} = \{x : 0 \leq x < 1\}$. This provides an example that quantile is not unique!

However, if we restrict restrict the quantiles to the range of X, then the quantile function can be defined as

$$x_p = F^{-1}(p) = \begin{cases} 0, & p \leq 0.5 \\ 1, & 0.5 < p \end{cases}$$

1.7 Some Useful Inequalities

Theorem 5. (**Markov Inequality**) Let $\varepsilon > 0$ and $p > 0$,

$$P(|X| \geq \varepsilon) \leq \frac{E|X|^p}{\varepsilon^p}.$$

Proof.

$$\begin{aligned} E|X|^p &= \int_{\text{supp}(X)} |x|^p f(x)dx \\ &= \int_{|x|\geq\varepsilon} |x|^p f(x)dx + \int_{|x|<\varepsilon} |x|^p f(x)dx \\ &\geq \int_{|x|\geq\varepsilon} |x|^p f(x)dx \\ &\geq \int_{|x|\geq\varepsilon} \varepsilon^p f(x)dx \\ &= \varepsilon^p \int_{|x|\geq\varepsilon} f(x)dx \\ &= \varepsilon^p P(|X| \geq \varepsilon). \end{aligned}$$

Thus,

$$P(|X| \geq \varepsilon) \leq \frac{E|X|^p}{\varepsilon^p}.$$

□

Note that, let $Y = X - E(X)$, $\varepsilon = k\sigma$ and $p = 2$, we can obtain Chebyshev's Inequality:

Theorem 6. (**Chebyshev's Inequality**)

$$P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}.$$

Theorem 7. (**Jensen's Inequality**) Let $\phi(x)$ be a smooth convex function. Then

$$\phi(E(X)) \leq E(\phi(X)).$$

Proof. Let $\mu = E(X)$. Since $\phi(x)$ is convex,

$$\phi(X) \geq \phi(\mu) + \phi'(\mu)(X - \mu).$$

Take expectation in both side,

$$E(\phi(X)) \geq \phi(\mu) + 0 = \phi(\mu).$$

That is,

$$E(\phi(X)) \geq \phi(E(X)).$$

□

Chapter 2

Multivariate Random Variables

In this chapter, we focus mostly on bivariate random variables. Let X and Y be jointly distributed random variables. We will denote the pair of random variables as (X, Y), and call this random vector as a random variable.

2.1 Bivariate Probability Distribution

Random experiment outcome is a pair of random variables (X, Y). The joint CDF is

$$F(x, y) = P(X \leq x, Y \leq y).$$

If (X, Y) is discrete, the distribution of (X, Y) is given by the joint pmf

$$f(x, y) = P(X = x, Y = y),$$

with properties that

1. $f(x, y) > 0$.
2. $\sum_i \sum_j f(x_i, y_i) = 1$.

If (X, Y) is continuous, the joint CDF is given by

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f(u, v) du dv,$$

and a joint pdf of (X, Y) is given by

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

with properties that

1. $f(x, y) \geq 0$.

2. $\int_{supp(X)} \int_{supp(Y)} f(x, y) dx dy = 1$.

3. $P(X = x, Y = y) = 0$.

The marginal pdf of X is given by

$$f(x) = \int_{y \in supp(Y)} f(x, y) dy.$$

Theorem 8 (Cauchy-Schwarz Inequality). For any random variables X and Y, we have

$$[E(XY)]^2 \leq E(X^2)E(Y^2),$$

or

$$|E(XY)| \leq \sqrt{E(X^2)E(Y^2)}.$$

Proof. Let c be a real number and define $Z = cX + Y$. Then

$$0 \leq E(Z^2) = c^2 E(X^2) + 2cE(XY) + E(Y^2).$$

The RHS can be seen as a quadratic function in the variable c. Since this quadratic expression is apparently non-negative, and $E(X^2) > 0$, it follows that the corresponding discriminant is non-positive. That is,

$$D = (2E(XY))^2 - 4E(X^2)E(Y^2) \leq 0,$$

which is what we want to prove.

□

Clearly, according to Cauchy-Schwarz inequality, we can easily derive that the correlation coefficient is between ±1. Simply define two new random variables as $X = U - EU$, $Y = W - EW$. Then by Cauchy-Schwarz inequality,

$$|E(U - EU)(W - EW)| \leq \sqrt{E(U - EU)^2 E(W - EW)^2},$$

or

$$|Cov(U, W)| \leq \sqrt{Var(U)Var(W)}.$$

2.2 Conditioning

Definition 13 (Conditional Distribution of Discrete Random Variables). Let (X, Y) be a discrete random variable. If $P(X = x) > 0$, the conditional pmf of $Y|X = x$ can be derived by

$$\begin{aligned} f_{Y|X=x}(y) &= P(Y = y|X = x), \\ &= \frac{P(Y = y, X = x)}{P(X = x)}, \\ &= \frac{f_{XY}(x, y)}{f_X(x)}. \end{aligned}$$

Note that $f_{Y|X=x}(y)$ is itself a true pmf. That is, it satisfies the following properties:

1. $f_{Y|X=x}(y) \geq 0 \;\; \forall y$.
2. $\sum_y f_{Y|X=x}(y) = 1$.
3. $P(Y \leq y|X = x) = \sum_{t \leq y} f_{Y|X=x}(t)$.

In analogy with the discrete case, we have the following definition for continuous random variables.

Definition 14 (Conditional Distribution of Continuous Random Variables). Let (X, Y) be a continuous random variable. The conditional pdf of $Y|X = x$ is defined as

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f(x)}$$

1. Therefore, the conditional CDF is

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(u)du$$

2. The conditional probability can be calculated by

$$P(a < Y < b|X = x) = \int_{a}^{b} f_{Y|X=x}(y)dy$$

Exercise 3. Figure out what the conditional pdf

$$g_{X|X\geq a}(x)$$

would be, given a continuous random variable X with pdf $f(x)$.

[Hint:]

$$g_{X|X\geq a}(x) = \frac{d}{dx}P(X \leq x|X \geq a).$$

2.3 Expectation and Conditional Expectation

Definition 15 (Expectation). Suppose (X,Y) is a random variable with joint pmf/pdf $f(x,y)$, then for the discrete case,

$$E(h(X,Y)) = \sum_{x\in supp(X)} \sum_{y\in supp(Y)} h(x,y)f(x,y),$$

and for the continuous case,

$$E(h(X,Y)) = \int_{x\in supp(X)} \int_{y\in supp(Y)} h(x,y)f(x,y)dxdy.$$

For instance, the covariance between X and Y is given that $h(x,y) = (x - EX)(y - EY)$:

$$Cov(X,Y) = E(X - EX)(Y - EY) = E(XY) - E(X)E(Y).$$

Proposition 2. The following properties have been already introduced in the course for elementary statistics.

- $Cov(aX + bY, Z) = aCov(X,Z) + bCov(Y,Z)$.
- $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$.

Now we define the conditional expectation.

Definition 16 (Conditional Expectation). Suppose (X,Y) is a random variable with joint pmf/pdf $f(x,y)$, then for the discrete case,

$$E(Y|X = x) = \sum_{y\in supp(Y)} y f_{Y|X=x}(y),$$

and for the continuous case,

$$E(Y|X = x) = \int_{y\in supp(Y)} y f_{Y|X=x}(y)dy.$$

Moreover, by definition,

$$E(g(X,Y)|X = x) = \int g(x,y)f_{Y|X=x}(y)dy.$$

Theorem 9 (Important Theorems for Conditional Expectation).

1. $E(c|X) = c$.
2. $E(Y + Z|X) = E(Y|X) + E(Z|X)$.
3. $E(cY|X) = cE(Y|X)$.
4. $E(g(X,Y)|X = x) = E(g(x,Y)|X = x)$.

It is worth noting that expectation $E(Y)$ is a constant. On the other hand, the conditional expectation $E(Y|X)$ is a function of random variable X, and thus it is a random variable. Since $E(Y|X)$ is a random variable, we would be interested in its expected value, which is shown in the following theorem.

Theorem 10 (**Law of Iterated Mathematical Expectation, LIME).**

$$E(E(Y|X)) = E(Y).$$

Proof. To be more precise, the above expectations are based on different probability distributions:

$$E_X(E_{Y|X}(Y|X)) = E_Y(Y).$$

For short, we denote $f_{Y|X=x}(y)$ as $f(y|x)$.

That is,

$$\begin{aligned} E[E(Y|X)] &= \int_x h(x)f(x)dx = \int_x \left[\int_y yf(y|x)dy \right] f(x)dx \\ &= \int_x \int_y y\frac{f(x,y)}{f(x)} f(x)dydx = \int_x \int_y yf(x,y)dydx \\ &= \int_y y \left[\int_x f(x,y)dx \right] dy = \int_y yf(y)dy = E(Y) \end{aligned}$$

□

Again, in order to save the Earth, we will also use the following notation for LIME:

$$EE(Y|X) = E(Y).$$

Theorem 11 (Useful Rule).

$$E(g(X)Y|X) = g(X)E(Y|X).$$

Proof. For any x,

$$\begin{aligned} E(g(X)Y|X = x) &= E(g(x)Y|X = x) \\ &= g(x)E(Y|X = x). \end{aligned}$$

This holds for any realization x, thus

$$E(g(X)Y|X) = g(X)E(Y|X).$$

□

This theorem is called "Useful Rule" by Gautam Tripathi.[1]

[1] Professor of Econometrics at the University of Luxembourg.

Theorem 12. If $G(X)$ is a monotonic (one-to-one increasing) transformation function of X, then

$$E(Y|X) = E(Y|G(X)).$$

Proof. By definition,

$$E(Y|G(X) = g) = \int_y y f_{Y|G(X)=g}(y)dy,$$

where

$$f_{Y|G(X)=g} = \frac{f_{GY}(g,y)}{f_G(g)}.$$

So we need to figure out the numerator and the denominator first.

$$\begin{aligned} F_{GY}(g,y) &= P(G(X) \leq g, Y \leq y), \\ &= P(X \leq G^{-1}(g), Y \leq y), \\ &= F_{XY}(G^{-1}(g), y). \end{aligned}$$

Therefore,

$$\begin{aligned} f_{GY}(g,y) &= \frac{\partial^2}{\partial g \partial y} F_{GY}(g,y), \\ &= \frac{\partial^2}{\partial g \partial y} F_{XY}(G^{-1}(g), y), \\ &= \frac{\partial}{\partial y}\left\{\frac{\partial}{\partial x} F_{XY}(G^{-1}(g), y)\frac{dG^{-1}(g)}{dg}\right\}, \\ &= \left[\frac{\partial}{\partial y}\frac{\partial}{\partial x} F_{XY}(x,y)\right]\frac{dG^{-1}(g)}{dg}, \\ &= f_{XY}(x,y)\frac{dG^{-1}(g)}{dg}. \end{aligned}$$

Moreover, since

$$F_G(g) = P(G(X) \leq g) = P(X \leq G^{-1}(g)) = F_X(G^{-1}(g))$$

we have

$$f_G(g) = \frac{d}{dg}F_G(g) = \frac{d}{dx}F_X(x)\frac{dG^{-1}(g)}{dg} = f_X(x)\frac{dG^{-1}(g)}{dg}.$$

Hence,

$$f_{Y|G(X)=g} = \frac{f_{GY}(g,y)}{f_G(g)} = \frac{f_{XY}(x,y)\frac{dG^{-1}(g)}{dg}}{f_X(x)\frac{dG^{-1}(g)}{dg}} = \frac{f_{XY}(x,y)}{f_X(x)} = f_{Y|X=x}(y).$$

Therefore,

$$\begin{aligned} E(Y|G(X)=g) &= \int_y y f_{Y|G(X)=g}(y)dy, \\ &= \int_y y f_{Y|X=x}(y)dy, \\ &= E(Y|X=x). \end{aligned}$$

That is,

$$E(Y|G(X)) = E(Y|X).$$

□

For instance, $E(Y|2X) = E(Y|X)$, or $E(Y|e^X) = E(Y|X)$. However, $E(Y|X^2) \neq E(Y|X)$ because X^2 is not a monotonic function of X.

Theorem 13 (Small Conditioning Set Wins Rule, SCSWR).

1. $E(E[Y|X,Z]|X) = E(Y|X)$.
2. $E(E[Y|X]|X,Z) = E(Y|X)$.

Proof. For the first case, given any x,

$$E(Y|X=x,Z) = h(Z)$$

is a function of Z. That is, given any z,

$$h(z) = E(Y|X=x, Z=z) = \int_y y f(y|x,z)dy$$

Therefore,

$$\begin{aligned}
E[E(Y|X=x,Z)|X=x) = E[h(Z)|X=x] &= \int_z h(z)f(z|x)dz \\
&= \int_z \left[\int_y yf(y|x,z)dy\right] f(z|x)dz \\
&= \int_z \int_y yf(y|x,z)f(z|x)dydz \\
&= \int_z \int_y y\frac{f(y,x,z)}{f(x,z)}\frac{f(x,z)}{f(x)}dydz \\
&= \int_y \int_z y\frac{f(y,x,z)}{f(x)}dzdy \\
&= \int_y y\frac{1}{f(x)}\int_z f(y,x,z)dzdy \\
&= \int_y y\frac{1}{f(x)}f(y,x)dy \\
&= \int_y yf(y|x)dy \\
&= E(Y|X=x)
\end{aligned}$$

The above result holds for all x, and hence

$$E(E[Y|X,Z]|X) = E(Y|X)$$

For the second case, we can simply apply the useful rule.

$$E(E[Y|X=x]|X=x,Z) = E(A(x)|X=x,Z) = A(x) = E(Y|X=x).$$

This holds for any x, so

$$E(E[Y|X]|X,Z) = E(Y|X).$$

□

2.4 Conditional Variance

Definition 17 (Conditional Variance). Let (X,Y) be a random variable. The conditional variance of Y given $X=x$ is

$$Var(Y|X=x) = E\left[(Y-E[Y|X])^2|X=x\right].$$

Some simple algebras can give us the following theorem (try it!).

Theorem 14.

$$Var(Y|X=x) = E(Y^2|X=x) - [E(Y|X=x)]^2.$$

Indeed, this is an analogy to the unconditional variance: $Var(Y) = E(Y^2) - [E(Y)]^2$.

Theorem 15. Let (X, Y) be a random variable, and let $g(\cdot)$ denote *any* function of X. Then

$$E([Y - g(X)]^2) = E(Var[Y|X]) + E([E(Y|X) - g(X)]^2).$$

Proof.

$$\begin{aligned} E([Y - g(X)]^2) &= E([Y - E(Y|X) + E(Y|X) - g(X)]^2), \\ &= \underbrace{E([Y - E(Y|X)]^2)}_{(i)} + E([E(Y|X) - g(X)]^2) \\ &\quad + 2\underbrace{E([Y - E(Y|X)][E(Y|X) - g(X)])}_{(ii)}. \end{aligned}$$

$$\begin{aligned} (i) &= E([Y - E(Y|X)]^2), \\ &= EE([Y - E(Y|X)]^2|X), \quad \text{by LIME}, \\ &= E[Var(Y|X)], \quad \text{by definition}. \end{aligned}$$

$$\begin{aligned} (ii) &= E\left([Y - E(Y|X)]\underbrace{[E(Y|X) - g(X)]}_{A(X)}\right), \\ &= E(A(X)Y - A(X)E[Y|X]), \\ &= E(A(X)Y) - E(A(X)E[Y|X]), \\ &= E(E(A(X)Y|X)) - E(A(X)E[Y|X]), \quad \text{by LIME}, \\ &= E(A(X)E[Y|X]) - E(A(X)E[Y|X]), \quad \text{by useful rule}, \\ &= 0. \end{aligned}$$

Therefore, we have shown that

$$\begin{aligned} E([Y - g(X)]^2) &= (i) + E([E(Y|X) - g(X)]^2) + 2(ii), \\ &= E(Var[Y|X]) + E([E(Y|X) - g(X)]^2). \end{aligned}$$

□

According to Theorem 15, we can obtain the following two lemmas.

Lemma 1. (**Analysis of Variance**) Suppose that $g(X)$ = constant = $E(Y)$, then

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)].$$

Proof.

$$\begin{aligned} E([Y - E(Y)]^2) &= E(Var[Y|X]) + E([E(Y|X) - E(Y)]^2), \\ &= E(Var[Y|X]) + E([E(Y|X) - E(E(Y|X))]^2), \\ &= E(Var[Y|X]) + Var(E[Y|X]). \end{aligned}$$

That is,

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)].$$

□

Lemma 2. (**Best Mean Squared Error Predictor**) Define the prediction error as

$$Y - g(X),$$

then

$$E(Y|X) = \arg\min_g E([Y - g(X)]^2).$$

That is, $E(Y|X)$ is the best predictor of Y via the minimum mean square error (MSE) criterion.

Proof. Since $E([Y - g(X)]^2) = E(Var[Y|X]) + E([E(Y|X) - g(X)]^2)$, and $E(Var[Y|X]) > 0$, it is trivial that $E(Y|X)$ minimizes the MSE, $E([Y - g(X)]^2)$. □

2.5 Applications to the Regression Models

Given that the best (conditional) predictor of Y is $E(Y|X)$, we can thus define the *prediction error* as follows:

$$\varepsilon \equiv Y - E(Y|X).$$

Rearrange the equation, we have the canonical regression model:

$$Y = E(Y|X) + \varepsilon.$$

Therefore, ε is also called the *regression error*. The regression error has following important properties.

Theorem 16 (**Important Properties of the Regression Error**)**.**

1. $E(\varepsilon|X) = 0$.

2. $E(\varepsilon) = 0$.

3. $Var(\varepsilon|X) = Var(Y|X)$.

4. $Cov(\varepsilon, h(X)) = 0$.

5. $Cov(\varepsilon, X) = 0$.

6. $Var(\varepsilon) = E(Var[Y|X])$.

Proof.

1. $$E(\varepsilon|X) = E[Y - E(Y|X)|X] = E(Y|X) - E[E(Y|X)|X] = E(Y|X) - E(Y|X) = 0.$$

2. By LIME, $E(\varepsilon) = E[E(\varepsilon|X)] = E[0] = 0$.

3. $$\begin{aligned} Var(\varepsilon|X) &= E\Big([\varepsilon - E(\varepsilon|X)]^2 \Big| X\Big) = E(\varepsilon^2|X), \text{ since } E(\varepsilon|X) = 0, \\ &= E\Big([Y - E(Y|X)]^2 \Big| X\Big) = Var(Y|X), \text{ by definition.} \end{aligned}$$

4. $$\begin{aligned} Cov(\varepsilon, h(X)) &= E(\varepsilon h(X)) - E(\varepsilon)E(h(X)) = E(\varepsilon h(X)), \\ &= E[E(\varepsilon h(X)|X)], \text{ by LIME,} \\ &= E[h(X)E(\varepsilon|X)], \text{ by useful rule,} \\ &= E[0] = 0. \end{aligned}$$

5. Simply let $h(X) = X$.

6. By Lemma 1,

$$\begin{aligned} Var(\varepsilon) &= Var[E(\varepsilon|X)] + E[Var(\varepsilon|X)], \\ &= Var[E(\varepsilon|X)] + E[Var(Y|X)] = E[Var(Y|X)]. \end{aligned}$$

□

From the course of elementary statistics, we have already learned that

$$a^* + b^* X = \arg\min_{a,b} E[(Y - a - bX)^2],$$

where

$$b^* = \frac{Cov(Y,X)}{Var(X)}, \quad a^* = E(Y) - b^* E(X).$$

Therefore,

- The best predictor of Y is $BP(Y|X) = E(Y|X)$.
- The best linear predictor of Y is $BLP(Y|X) = a^* + b^* X$.

Theorem 17. Suppose that $E(Y|X)$ is a linear function of X, then

$$BLP(Y|X) = E(Y|X) = BP(Y|X).$$

Proof. In general, since $E(Y|X)$ is the best predictor,

$$E\Big([Y - E(Y|X)]^2\Big) \leq E\Big([Y - BLP(Y|X)]^2\Big).$$

But since $E(Y|X)$ is a linear function of X, we have

$$E\Big([Y - BLP(Y|X)]^2\Big) \leq E\Big([Y - E(Y|X)]^2\Big).$$

Therefore, we have

$$E\Big([Y - E(Y|X)]^2\Big) = E\Big([Y - BLP(Y|X)]^2\Big),$$

so

$$E(Y|X) = BLP(Y|X).$$

□

2.6 Independence

Definition 18 (Independence of Two Events). Given a probability space $(\Omega, \mathcal{F}, P)$, events $A, B \in \mathcal{F}$ are (stochastically or statistically) independent if

$$P(A \bigcap B) = P(A)P(B).$$

Definition 19 (Independence of a Finite Number of Events I). The events $A_1, A_2, \ldots, A_n$ are independent if

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i), \quad \text{for all } I \subset \{1, 2, \ldots, n\}.$$

Note that we need to check $\sum_{k=2}^{n} \binom{n}{k} = 2^n - n - 1$ relationships for independence. Moreover, this condition can be rephrased as follows.

Definition 20 (Independence of a Finite Number of Events II). The events $A_1, A_2, \ldots, A_n$ are independent if

$$P\left(B_1 \bigcap B_2 \bigcap \cdots \bigcap B_n\right) = \prod_{i=1}^{n} P(B_i),$$

where B_i equals A_i or Ω.

We now define the independent two random variables.

Definition 21 (Independence of Two Random Variables). Two random variables (X, Y) are said to be independent if for all sets $A, B \subseteq \mathbb{R}$, we have

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

and we denote it as $X \perp Y$.

We now extend the independence concept to the n-variate case.

Definition 22 (Independence of Random Variables). Random variables (X_1, X_2, ..., X_n) are said to be independent if for all sets $A_j \subseteq \mathbb{R}$, $j = 1, 2, \ldots, n$, we have

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n).$$

And this definition gives us the following theorem.

Theorem 18 (Factorization Theorem I). Let $\boldsymbol{X} = (X_1 \quad X_2 \quad \cdots \quad X_n)'$. Random variables $(X_1, X_2, \ldots, X_n)$ are independent if and only if

$$f_{\boldsymbol{X}}(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \quad \forall x_i \in \mathbb{R},$$

where $f_{\boldsymbol{X}}$ and f_{X_i} are joint pdf(pmf) and marginal pdf(pmf), respectively.

Proof.

1. The "⇒" part.

$$\begin{aligned}
&P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n)\\
&\quad = P(X_1 \in A_1)P(X_2 \in A_2)\cdots P(X_n \in A_n),\\
&\quad = \int_{x_1 \in A_1} f_{X_1}(x_1)dx_1 \int_{x_2 \in A_2} f_{X_2}(x_2)dx_2 \cdots \int_{x_n \in A_n} f_{X_n}(x_n)dx_n,\\
&\quad = \int_{x_1 \in A_1}\int_{x_2 \in A_2}\cdots\int_{x_n \in A_n} f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)dx_1dx_2\cdots dx_n.
\end{aligned}$$

That is, $f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n)$ is indeed the joint pdf of $(X_1, X_2, \ldots, X_n)$, i.e., $f_{\boldsymbol{X}}(x_1, x_2 \ldots, x_n)$.

2. The "⇐" part.

$$
\begin{aligned}
&P(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) \\
&= \int_{x_1 \in A_1} \int_{x_2 \in A_2} \cdots \int_{x_n \in A_n} f_{\mathbf{X}}(x_1, x_2 \dots, x_n) dx_1 dx_2 \cdots dx_n, \\
&= \int_{x_1 \in A_1} \int_{x_2 \in A_2} \cdots \int_{x_n \in A_n} f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) dx_1 dx_2 \cdots dx_n, \\
&= \int_{x_1 \in A_1} f_{X_1}(x_1) dx_1 \int_{x_2 \in A_2} f_{X_2}(x_2) dx_2 \cdots \int_{x_n \in A_n} f_{X_n}(x_n) dx_n, \\
&= P(X_1 \in A_1) P(X_2 \in A_2) \cdots P(X_n \in A_n).
\end{aligned}
$$

□

Clearly, since Definition 22 is defined on all possible sets, we can define $B_j = (-\infty, x_j]$ such that independence implies

$$P(X_1 \in B_1, X_2 \in B_2, \dots, X_n \in B_n) = P(X_1 \in B_1) P(X_2 \in B_2) \cdots P(X_n \in B_n).$$

That is, $(X_1, X_2, \dots, X_n)$ are independent if

$$P\Big(\{X_1 \le x_1\} \bigcap \{X_2 \le x_2\} \bigcap \cdots \bigcap \{X_n \le x_n\}\Big) = P(\{X_1 \le x_1\}) P(\{X_2 \le x_2\}) \cdots P(\{X_n \le x_n\}).$$

Therefore, it follows that independence implies the following theorem.

Theorem 19 (**Factorization Theorem II).** Let $\mathbf{X} = (X_1 \;\; X_2 \;\; \cdots \;\; X_n)'$. Random variables $(X_1, X_2, \dots, X_n)$ are independent if and only if

$$F_{\mathbf{X}}(x_1, x_2 \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \quad \forall x_i \in \mathbb{R},$$

where $F_{\mathbf{X}}$ and F_{X_i} are joint CDF and marginal CDF, respectively.

Theorem 20. If $X \perp Y$, then

$$g(X) \perp h(Y).$$

Proof. Pick any $B_1, B_2 \subseteq \mathbb{R}$ and define $f^{-1}(A) = \{x \in \mathbb{R} : f(x) \in A\}$,

$$
\begin{aligned}
P(g(X) \in B_1, h(Y) \in B_2) &= P(X \in g^{-1}(B_1), Y \in h^{-1}(B_2)), \\
&= P(X \in g^{-1}(B_1)) P(Y \in h^{-1}(B_2)), \quad \text{by independence}, \\
&= P(g(X) \in B_1) P(h(Y) \in B_2).
\end{aligned}
$$

□

In this theorem, what is the requirement for functions $g(\cdot)$ and $h(\cdot)$? At least, we would require $g(\cdot)$ and $h(\cdot)$ to be continuous functions.

Theorem 21. If $X \perp Y$, then

$$f_{Y|X}(y) = f_Y(y).$$

We here introduce a weaker concept of independence: mean independence.

Definition 23 (Mean Independence). A random variable Y is said to be mean independent of X if

$$E(Y|X) = \text{constant} = C.$$

In other words, the conditional expectation of Y given X is the same for all values of X. Note that Y is mean independent of X, and by LIME

$$E(Y) = E[E(Y|X)] = E[C] = C.$$

That is, if the conditional expectation of Y is the same for all values of X, then the unconditional expectation of Y coincides with that common conditional expectation.

Theorem 22. Let Y be mean independent of X. Suppose $h(X)$ is any function of X, then Y is mean independent of $h(X)$.

Proof.

$$\begin{aligned} LHS = E(Y|h(X)) &= E\Big(E(Y|h(X))\Big|h(X), X\Big), \quad \text{by useful rule,} \\ &= E\Big(E(Y|h(X), X)|h(X)\Big), \quad \text{by SCSWR.} \end{aligned}$$

However,

$$\begin{aligned} E(Y|h(X), X) &= E(E[Y|h(X), X]|X), \quad \text{by useful rule,} \\ &= E[Y|X], \quad \text{by SCSWR,} \\ &= C. \end{aligned}$$

Hence,

$$LHS = E(Y|h(X)) = E(C|h(X)) = C.$$

□

Some remarks are worth addressing.

1. Stochastic Independence ⇒ Mean Independence

2. Mean Independence $\nRightarrow$ Stochastic Independence

3. X is stochastically independent of $Y \Rightarrow Y$ is stochastically independent of X

4. X is mean independent of $Y \nRightarrow Y$ is mean independent of X

That is, mean independence is NOT symmetric.

Example 15 (Three-Point Distribution). Let $X \in \{0,1\}$ and $Y \in \{-1,0,1\}$. The joint pmf is

$$f(x,y) = \begin{cases} \frac{1}{3} & \text{for}(1,-1),(0,0),(1,1), \\ 0 & \text{otherwise} \end{cases}$$

It can be easily shown that $E(Y|X) = 0$ a constant, but $E(X|Y)$ is not a constant. That is, X is NOT mean independent of Y. Moreover, this example also shows that $P(X = 0, Y = 0) \neq P(X = 0)P(Y = 0)$. That is, mean independence does NOT imply stochastic independence.

Theorem 23. Suppose Y is mean independent of X, then X and Y are uncorrelated.

Proof. Since Y is mean independent of X,

$$E(Y|X) = C,$$

and

$$E(Y) = E(E(Y|X)) = E(C) = C.$$

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y), \\ &= E(E[XY|X]) - E(X)C, \\ &= E(XE[Y|X]) - E(X)C, \\ &= E(XC) - E(X)C, \\ &= E(X)C - E(X)C = 0. \end{aligned}$$

□

Let us consider some applications of mean independence. Define

$$\varepsilon = Y - BP(Y|X) = Y - E(Y|X),$$

then

$$E(\varepsilon|X) = 0.$$

That is, ε is mean independent of X. In contrast, define

$$v = Y - BLP(Y|X) = Y - a^* - b^* X,$$

where $b^* = \frac{Cov(X,Y)}{Var(X)}$, and $a^* = E(Y) - b^* E(X)$. Then

$$E(v|X) = E(Y|X) - a^* - b^* X.$$

So v is in general NOT mean independent of X. However, we can easily show that v is uncorrelated with X: $Cov(v, X) = 0$.

Chapter 3

Transforms

3.1 Univariate Transformation

We have already learned how to do univariate transformation in the elementary statistics course. Now we simply provide an example to refresh your memory.

Example 16. Let $X \overset{d}{=} U[0,1]$ and $Y = X^2$. We would like to find the pdf of Y.
First of all, we need to know the support of Y. Since $X \in [0,1]$, $Y \in [0,1]$ as well.

So pick any $y \in [0,1]$, $F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) = F_X(\sqrt{y}) - 0 = F_X(\sqrt{y})$. Hence,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(\sqrt{y})}{dx}\frac{dx}{dy} = \underbrace{f_X(\sqrt{y})}_{1} \cdot \frac{1}{2} y^{-\frac{1}{2}} = \frac{1}{2} y^{-\frac{1}{2}}.$$

That is,

$$f_Y(y) = \begin{cases} \frac{1}{2} y^{-\frac{1}{2}} & \text{if } y \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\int_0^1 f_Y(y)dy = 1.$$

(Check it!)

One more example is provided.

Example 17. Let $X \overset{d}{=} U[0,1]$ and $Y = -\log X$. We would like to find the pdf of Y.
Again, note that the support of Y is $\text{supp}(Y) = \{y : 0 \le y < \infty\}$.

Pick any $y \in [0,\infty)$, $F_Y(y) = P(Y \le y) = P(-\log X \le y) = P(\log X \ge -y) = P(e^{\log X} \ge e^{-y}) = P(X \ge e^{-y}) = 1 - P(X \le e^{-y}) = 1 - e^{-y}$. Hence,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} e^{-y} & \text{if } y \in [0,\infty), \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we present the general result.

Theorem 24 (Univariate Transformation Theorem). Suppose X is a continuous random variable, and $Y = g(X)$, where $g(\cdot)$ is a differentiable monotonic function, then the pdf of Y is

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|.$$

We omit the proof since it has already been shown in your elementary statistics course.

3.2 Multivariate Transformation

Now we turn to the multivariate case.

Let $\mathbf{X}$ be an $n \times 1$ random vector with pdf $f_{\mathbf{X}}(\mathbf{x})$ and support $S_{\mathbf{X}} \subseteq \mathbb{R}^n$. Moreover, let $g(g_1, g_2, \ldots, g_n)$ be one-to-one onto from $S_{\mathbf{X}}$ to some set $T \subseteq \mathbb{R}^n$. Now define

$$\mathbf{Y} = \begin{pmatrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \vdots \\ g_n(\mathbf{X}) \end{pmatrix} = g(\mathbf{X}).$$

Finally, assume that $g(\cdot)$ and its inverse are both continuously differentiable.

Theorem 25 (Multivariate Transformation Theorem). The pdf of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} f_{\mathbf{X}}(h_1(\mathbf{y}), h_2(\mathbf{y}), \ldots, h_n(\mathbf{y})) \cdot |\mathbf{J}|, & \text{for } y \in T, \\ 0 & \text{otherwise,} \end{cases}$$

where h is the inverse of g, i.e., $h(\mathbf{y}) = g^{-1}(\mathbf{y})$, and where

$$\mathbf{J} = \left| \frac{d\mathbf{X}}{d\mathbf{Y}} \right| = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

is the Jacobian.

Proof. Pick any set $B \subseteq \mathbb{R}^n$, and define

$$h(B) = g^{-1}(B) = \{\mathbf{x} \in \mathbb{R}^n : g(\mathbf{x}) \in B\}.$$

Then

$$P(\mathbf{Y} \in B) = P(g(\mathbf{X}) \in B) = P(\mathbf{X} \in h(B)) = \int_{h(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.$$

According to the formula for changing variables in multiple integrals, (for instance, see Pages 407–408 in Apostol (1969)).[1] we have

$$P(\mathbf{Y} \in B) = \int_B f_{\mathbf{X}}(h_1(\mathbf{y}), h_2(\mathbf{y}), \ldots, h_n(\mathbf{y})) \cdot |\mathbf{J}| d\mathbf{y}.$$

□

Let's see some examples.

Example 18. Given $\{X, Y\} \sim^{i.i.d.} N(0,1)$, show that

$$X + Y \overset{d}{=} N(0,2),$$

and

$$X - Y \overset{d}{=} N(0,2).$$

Let $U = X + Y$ and $V = X - Y$. Inversion yields

$$X = \frac{U+V}{2}, \quad Y = \frac{U-V}{2}.$$

The Jacobian is

$$\mathbf{J} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Therefore,

$$\begin{aligned} f_{UV}(u,v) &= f_{XY}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot \frac{1}{2}, \\ &= f_X\left(\frac{u+v}{2}\right) f_Y\left(\frac{u-v}{2}\right) \cdot \frac{1}{2}, \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u+v}{2}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u-v}{2}\right)^2} \cdot \frac{1}{2}, \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\left(\frac{u}{\sqrt{2}}\right)^2} \cdot \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\left(\frac{v}{\sqrt{2}}\right)^2}. \end{aligned}$$

The marginal pdf can be obtained as

$$f_U(u) = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\left(\frac{u}{\sqrt{2}}\right)^2}, \quad u \in \mathbb{R},$$

and

$$f_V(v) = \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\left(\frac{v}{\sqrt{2}}\right)^2}, \quad v \in \mathbb{R}.$$

[1] Apostol, Tom M. (1969). Calculus, Vol. 2: Multi-Variable Calculus and Linear Algebra with Applications.

Example 19 (Convolution Formula). Let X and Y be independent random variables with pdf $f_X(x)$ and $f_Y(y)$. Find the pdf of $X + Y$.

Clearly, we start with two variables but seek the distribution of just a new one. The trick is to set $U = X + Y$ and to introduce an *auxiliary variable* V, which may be arbitrarily defined. For instance, set $V = X$. Then the Jacobian is

$$\mathbf{J} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1.$$

Hence,

$$f_{UV}(u, v) = f_{XY}(v, u - v) \cdot 1 = f_X(v) f_Y(u - v),$$

and

$$f_U(u) = \int_{-\infty}^{\infty} f_X(v) f_Y(u - v) dv, \quad u \in \mathbb{R}.$$

This is called the *convolution formula*.

Example 20. $\{X_1, X_2\} \sim^{i.i.d.} U(0, 1)$. Find the pdf of $X_1 - X_2$.

Let $Y_1 = X_1 - X_2$, and $Y_2 = X_2$. That is, $X_1 = Y_1 + Y_2$, and $X_2 = Y_2$. Clearly,

$$\mathbf{J} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1.$$

$$f_{X_1 X_2}(x_1, x_2) = I_{\{0 \le x_1 \le 1\}}(x_1) \cdot I_{\{0 \le x_2 \le 1\}}(x_2),$$

where $I_A(x)$ is an indicator function. Hence,

$$f_{Y_1 Y_2}(y_1, y_2) = I_{\{0 \le y_1 + y_2 \le 1\}}(y_1 + y_2) \cdot I_{\{0 \le y_2 \le 1\}}(y_2).$$

However, the support of Y_1 is $supp(Y_1) = \{y_1 : -1 \le y_1 \le 1\}$, we have

$$f_{Y_1 Y_2}(y_1, y_2) = \begin{cases} 1 & \text{if } -1 \le y_1 \le 0, 0 \le y_2 \le 1, \\ 1 & \text{if } 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$f_{Y_1}(y_1) = \begin{cases} \int_{-y_1}^{1} 1 dy_2 & \text{if } -1 \le y_1 \le 0, 0 \le y_2 \le 1, \\ \int_{0}^{1-y_1} 1 dy_2 & \text{if } 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

That is,

$$f_{Y_1}(y_1) = \begin{cases} 1 + y_1 & \text{if } -1 \le y_1 \le 0, 0 \le y_2 \le 1, \\ 1 - y_1 & \text{if } 0 \le y_1 \le 1, 0 \le y_2 \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, y_1 is a Triangular distribution random variable.

3.3 Many to One

What if the transformation is not one-to-one? For instance, $Y = X^2$. Then

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= P(X^2 \leq y), \\ &= P(-\sqrt{y} \leq x \leq \sqrt{y}), \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Therefore,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}, \quad y \in \mathbb{R}_+.$$

Note that the function is 2 to 1 and that we obtain two terms. Now consider the general case. Let $\mathbf{X} \in S$ is a random variable with pdf $f_{\mathbf{X}}(\mathbf{x})$. Let $\mathbf{Y} = g(\mathbf{X})$, where $g : S \mapsto T$ is NOT one-to-one. But suppose S can be partitioned into m disjoint subsets $S_1, S_2, \ldots, S_m$ such that $g : S_k \mapsto T$ is one-to-one on each partition. Then

$$\begin{aligned} P(\mathbf{Y} \in B) &= P(g(\mathbf{X}) \in B), \\ &= P(\mathbf{X} \in g^{-1}(B)), \\ &= \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \\ &= \int_{g^{-1}(B) \cap S} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \\ &= \int_{g^{-1}(B) \cap (\cup_{i=1}^{m} S_i)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \\ &= \int_{\cup_{i=1}^{m} [g^{-1}(B) \cap S_i]} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}, \\ &= \sum_{i=1}^{m} \int_{g^{-1}(B) \cap S_i} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Hence, by Theorem 25, applied m times, yields

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} \sum_{k=1}^{m} f_{\mathbf{X}}(h_{1k}(\mathbf{y}), h_{2k}(\mathbf{y}), \ldots, h_{nk}(\mathbf{y})) \left|\mathbf{J_k}\right|, & \mathbf{y} \in T, \\ 0 & \text{otherwise.} \end{cases}$$

Where $(h_{1k}, h_{2k}, \ldots, h_{nk})$ is the inverse function corresponding to the mapping from S_k to T and $\mathbf{J_k}$ is the k-th Jacobian.

Here is an example.

Example 21. Let $\{X, Y\} \sim^{i.i.d.} N(0, 1)$. Consider the polar coordinate transformation.

$$R = \sqrt{X^2 + Y^2},$$

and

$$\Theta = \tan^{-1}\left(\frac{Y}{X}\right).$$

Note that $R \geq 0$, and $\tan\theta = \frac{Y}{X} \in (-\infty, \infty)$. That is, $\Theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Clearly, $(X, Y) \mapsto (R, \Theta)$ is not one-to-one transformation since (X, Y) and $(-X, -Y)$ are mapping to the same point.

Now partition the support of (X, Y) into S_1, S_2 and S_3:

$$supp(X, Y) = \underbrace{\{(x, y) : x > 0, y \in \mathbb{R}\}}_{S_1} \bigcup \underbrace{\{(x, y) : x < 0, y \in \mathbb{R}\}}_{S_2} \bigcup \underbrace{\{(x, y) : x = 0, y \in \mathbb{R}\}}_{S_3}$$

1. On S_1, $x > 0$, $y \in \mathbb{R}$. Thus $x = r\cos\theta$, $y = r\sin\theta$, where $r > 0$ and $\theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

$$J_1 = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

So on S_1,

$$f_{R,\Theta}(r, \theta) = f_{XY}(r\cos\theta, r\sin\theta)r = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}, \quad r > 0, \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

2. On S_2, $x < 0$, $y \in \mathbb{R}$. Thus $x = -r\cos\theta$, $y = r\sin\theta$.

$$J_2 = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} -\cos\theta & r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = -r.$$

So on S_2,

$$f_{R\Theta}(r, \theta) = f_{XY}(-r\cos\theta, r\sin\theta)r = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}, \quad r > 0, \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

3. On S_3, $f_{R\Theta}(r, \theta) = 0$.

Therefore,

$$f_{R\Theta}(r, \theta) = [f_{R\Theta}]_{S_1} + [f_{R\Theta}]_{S_2} + [f_{R\Theta}]_{S_3} = \frac{r}{\pi} e^{-\frac{r^2}{2}} \quad r > 0, \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right).$$

3.4 Sums of Independent Random Variables

In previous sections, we have learned how to handle transformation in order to find the distribution of new random variables. Now we are going to focus on sums of independent variables since the average of a set of random variables is a very important object in probability and statistics.

Two types of transformation are introduced: the moment generating function, and the characteristic function. Two common features of these transformations are that

1. Summation of independent random variables corresponds to *multiplication* of the transformation.

2. The transformation is 1 to 1.

3.4.1 The Moment Generating Function

In this section, some important theorems are presented without proofs. Interested readers may refer to advanced texts such as Fraser (1976), Resnick (2001) or Roussas (2002).

Definition 24. Let X be a random variable. The moment generating function (MGF) of X is

$$M_X(t) = E(e^{tX}),$$

provided there exists $h > 0$, such that the expectation exists and is finite for $|t| < h$.

We state the following important theorem without proof since proving this is far beyond the scope of this lecture. Intuitively, the proof relies on the fact that the moment generating function is a two-sided Laplace transformation of the pdf $f(x)$, and there is a unique association between a Laplace transformation and the function being transformed.

Theorem 26 (Uniqueness Theorem). Let X and Y be random variables. If there exists $h > 0$ such that

$$M_X(t) = M_Y(t), \quad \text{for} |t| < h,$$

then

$$X \overset{d}{=} Y.$$

Moreover, we have the following two theorems without proofs since the proofs are simply followed via definition and have already been shown in the elementary statistics course.

Theorem 27 (Multiplication Theorem). Let $\{X_i\}_{i=1}^n$ be independent random variables whose MGF $M_{X_i}(t)$ exist for $|t| < h$, for some $h > 0$, and let $Y = \sum_{i=1}^n X_i$, then

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t), \quad |t| < h.$$

Theorem 28. Let $\{X_i\}_{i=1}^n$ be independent random variables whose MGF $M_{X_i}(t)$ exist for $|t| < h$, for some $h > 0$, and let $Y = aX + b$ for constants a, b, then

$$M_Y(t) = e^{bt} M_X(at).$$

The following theorem shows that the derivatives at 0 of the MGF produce the moments (hence the name of the transformation).

Theorem 29. Let X be a random variable whose MGF $M_X(t)$ exists for $|t| < h$, for some $h > 0$. if all the moments exist, that is, $E(X^r) < \infty$ for all r, then

$$E(X^r) = M_X^{(r)}(0), \quad \text{for } r = 1, 2, \dots.$$

Proof. We show the case of a continuous random variable while the discrete case can be applied analogously. By definition,

$$M_X(t) = \int_{supp(X)} e^{tx} f_X(x)dx.$$

It can be obtained by differentiating under the integral sign,

$$M_X^{(r)}(t) = \int_{supp(X)} x^r e^{tx} f_X(x)dx.$$

Therefore,

$$M_X^{(r)}(t)|_{t=0} = M_X^{(r)}(0) = \int_{supp(X)} x^r f_X(x)dx = E(X^r).$$

□

Moreover, taking a Taylor expansion of the exponential function yields

$$e^{tX} = e^{t\cdot 0} + \frac{1}{1!}t \cdot e^{t\cdot 0}X + \frac{1}{2!}t^2 \cdot e^{t\cdot 0}X^2 + \cdots = 1 + \sum_{n=1}^{\infty} \frac{t^n X^n}{n!}.$$

Thus we have

$$M_X(t) = E(e^{tX}) = 1 + \sum_{n=1}^{\infty} \frac{t^n E(X^n)}{n!}.$$

By taking termwise differentiation yields the result in Theorem 29.

Finally, we define the MGF for a random vector.

Definition 25. Let **X** be a random n-vector. The (joint) moment generating function of **X** is

$$M_{\mathbf{X}}(t_1, t_2, \ldots, t_n) = E(e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}),$$

provided there exists $h_1, h_2, \ldots, h_n > 0$ such that the expectation exists for $|t_k| < h_k, k = 1, 2, \ldots, n$.

In vector notation,

$$M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{X}}),$$

provided there exists $\mathbf{h} > \mathbf{0}$, such that the expectation exists for $|\mathbf{t}| < \mathbf{h}$ (the inequalities being interpreted componentwise).

Theorem 30. Let $M_{\mathbf{X}}(t_1, t_2, \ldots, t_n)$ be the joint MGF of a random n-vector $\mathbf{X} = (X_1\ X_2 \cdots X_n)$. Then $(X_1, X_2, \ldots, X_n)$ are independent if and only if

$$M_{\mathbf{X}}(t_1, t_2, \ldots, t_n) = M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n)$$

3.4.2 Moment Generating Functions for Particular Distributions

1. Bernoulli: $X \sim \text{Bernoulli}(p)$
$$M_X(t) = (1-p) + pe^t$$

2. Binomial: $X \sim \text{Binomial}(n,p)$
$$M_X(t) = \left[(1-p) + pe^t\right]^n$$

3. Geometric: $X \sim \text{Geo}(p)$
$$M_X(t) = \frac{p}{1-(1-p)e^t}$$

4. Poisson: $X \sim \text{Poisson}(\lambda)$
$$M_X(t) = \exp\left[\lambda(e^t - 1)\right]$$

5. Uniform: $X \sim U[l,h]$
$$M_X(t) = \frac{e^{ht} - e^{lt}}{(h-l)t}$$

6. Normal: $X \sim N(\mu, \sigma^2)$
$$M_X(t) = \exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$$

7. Gamma: $X \sim \text{Gamma}(\alpha, \beta)$
$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$$

8. Exponential: $X \sim \exp(\beta)$
$$M_X(t) = \left(\frac{1}{1-\beta t}\right)$$

9. Chi-square: $\chi^2(k)$
$$M_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{k}{2}}$$

3.4.3 The Characteristic Function

A problem with the moment generating function is that it does not necessarily exist for all distributions. For instance, the Cauchy and the log-normal distributions are two such examples. We now introduce a new transformation that is technically complicate but exists for *all* distributions.

Definition 26 (The Characteristic Function). Let X be a random variable. The characteristic function of X is
$$\phi_X(t) = E(e^{itX}) = E(\cos tX + i \sin tX).$$

We present some important features of the characteristic function.

1. $\phi_X(t)$ always exists.
2. $\phi_X(0) = 1$

Theorem 31 (Uniqueness Theorem). Let X and Y be random variables. If

$$\phi_X(t) = \phi_Y(t),$$

then

$$X \overset{d}{=} Y.$$

Indeed, the above theorem follows from the Inversion Formula. See pages 141–145 in Roussas (2002).

Theorem 32 (Multiplication Theorem). Let $\{X_i\}_{i=1}^n$ be independent random variables with characteristics functions $\phi_{X_i}(t)$, and let $Y = \sum_{i=1}^n X_i$, then

$$\phi_Y(t) = \prod_{i=1}^n \phi_{X_i}(t).$$

Therefore, if, in addition, $\{X_i\}_{i=1}^n$ are i.i.d. random variables, then

$$\phi_Y(t) = [\phi_X(t)]^n.$$

Theorem 33. Let X be a random variable and a and b be real numbers. Then

$$\phi_{aX+b}(t) = e^{ibt}\phi_X(at).$$

We have shown a series expansion of the moment generating function in previous section. Following is the counterpart for characteristic functions:

Theorem 34. Let X be a random variable. If $E|X|^n < \infty$, then

(a) $\phi_X^{(k)}(0) = i^k \cdot E(X^k)$, $k = 1, 2, \ldots, n$.

(b) $\phi_X(t) = 1 + \sum_{k=1}^{\infty} E(X^k) \cdot \frac{(it)^k}{k!}$.

Finally, we close this section by introducing the characteristic function for a random vector.

Definition 27. The characteristic function of a random n-vector $\mathbf{X}$ is defined as

$$\phi_{\mathbf{X}}(t) = E(e^{it'\mathbf{X}}).$$

Chapter 4

Multivariate Normal Distribution

4.1 Random Vector and Variance-Covariance Matrix

Consider the random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$$

with expectation

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix},$$

and variance-covariance matrix (we will call it variance for short):

$$\begin{aligned} \Lambda = Var(\mathbf{X}) &= E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})', \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_n) \\ \vdots & \cdots & \ddots & \vdots \\ Cov(X_n, X_1) & Cov(X_n, X_2) & \cdots & Var(X_n) \end{pmatrix}. \end{aligned}$$

Note that

1. Λ is symmetric.
2. Λ is positive-semidefinite. That is, for all $d \in \mathbb{R}^n$,

$$d'\Lambda d = d'E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'d = E\Big([d'(\mathbf{X}-\boldsymbol{\mu})]^2\Big) \geq 0.$$

Theorem 35. Let $\mathbf{X}$ be a random n-vector with mean $\boldsymbol{\mu}$ and variance Λ. Further, let $\mathbf{B}$ be an $m \times n$ matrix, let $\mathbf{b}$ be a constant m-vector, and let $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. Then

$$E(\mathbf{Y}) = \mathbf{B}\boldsymbol{\mu} + \mathbf{b},$$

$$Var(\mathbf{Y}) = \mathbf{B}\Lambda\mathbf{B}'.$$

Proof. First we have

$$E(\mathbf{Y}) = E(\mathbf{B}\mathbf{X} + \mathbf{b}) = \mathbf{B}E(\mathbf{X}) + \mathbf{b} = \mathbf{B}\boldsymbol{\mu} + \mathbf{b}.$$

Second,

$$\begin{aligned} Var(\mathbf{Y}) &= Var(\mathbf{B}\mathbf{X} + \mathbf{b}), \\ &= E\big(\mathbf{B}\mathbf{X} + \mathbf{b} - [\mathbf{B}\boldsymbol{\mu} + \mathbf{b}]\big)\big(\mathbf{B}\mathbf{X} + \mathbf{b} - [\mathbf{B}\boldsymbol{\mu} + \mathbf{b}]\big)', \\ &= E(\mathbf{B}(\mathbf{X} - \boldsymbol{\mu}))(\mathbf{B}(\mathbf{X} - \boldsymbol{\mu}))', \\ &= \mathbf{B}E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\mathbf{B}', \\ &= \mathbf{B}Var(\mathbf{X})\mathbf{B}', \\ &= \mathbf{B}\Lambda\mathbf{B}'. \end{aligned}$$

□

4.2 Multivariate Normal Distribution

We will introduce three different definitions of the multivariate normal distribution.

Definition 28 (Multivariate Normal Distribution I). A random n-vector $\mathbf{X}$ is said to be normal (multivariate normal) iff $\forall \mathbf{a} \in \mathbb{R}^n$, the random variable $\mathbf{a}'\mathbf{X}$ is normal. We often write $\mathbf{X} \overset{d}{=} N(\boldsymbol{\mu}, \Lambda)$.

The definition states that a random vector is normal if and only if *every* linear combination of its components is normal. Note that no assumption whatsoever is made about independence between the components of $\mathbf{X}$.

Here are some consequences of this definition.

1. Every component of $\mathbf{X}$ is Gaussian (but the reversed statement is not true).

2. $\sum_{i=1}^{n} X_i$ is also Gaussian.

3. $X_1 + X_3 + X_{27} + X_{670} + 2X_{401}$ is also Gaussian.

4. If $\mathbf{X}$ consists of independent Gaussian components, then $\mathbf{X}$ is normal. However, just stacking up normal random variables will NOT yield a multivariate normal random vector.

Example 22. Suppose $X \overset{d}{=} N(0,1)$, and Z is independent of X such that $P(Z=1) = P(Z=-1) = \frac{1}{2}$. Now let $Y = ZX$.

1. Find the distribution of Y.
2. Is $\begin{pmatrix} X \\ Y \end{pmatrix}$ a multivariate normal random vector?

First of all, let $\Phi(\cdot)$ denotes the CDF of a $N(0,1)$ random variable,

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(ZX \le y), \\ &= P(ZX \le y, Z = 1) + P(ZX \le y, Z = -1), \\ &= P(X \le y, Z = 1) + P(-X \le y, Z = -1), \\ &= P(X \le y)P(Z = 1) + P(-X \le y)P(Z = -1), \quad \text{since } X \perp Z, \\ &= \phi(y) \cdot \frac{1}{2} + \phi(y) \cdot \frac{1}{2} \quad \text{by symmetric}, \\ &= \Phi(y). \end{aligned}$$

However,

$$P(X + Y = 0) = P(X + ZX = 0) = P(Z = -1) = \frac{1}{2} \neq 0.$$

That is, $X + Y$ is not normally distributed, otherwise we expect $P(X + Y = 0) = 0$. It is worth noting that in this case that X and Y are NOT independent since

$$|Y| = |X|.$$

So if you stack up “dependent” normal variables, you will NOT get a multivariate normal random vector.

Theorem 36. Let $\mathbf{X}$ be a multivariate normal n-vector. Suppose that

$$U_1 = \sum_{i=1}^{n} b_i X_i,$$

and

$$U_2 = \sum_{i=1}^{n} c_i X_i.$$

Then U_1 and U_2 have a bivariate normal distribution.

Proof. For any constants γ and β

$$\gamma U_1 + \beta U_2 = \gamma \sum_i b_i X_i + \beta \sum_i c_i X_i = \sum_i (\gamma b_i + \beta c_i) X_i = \sum_i a_i X_i$$

is normal by construction. Hence, U_1 and U_2 have a multivariate (bivariate) normal distribution.

□

Theorem 37. Let $\mathbf{X} \overset{d}{=} N(\boldsymbol{\mu}, \Lambda)$. Further, let $\mathbf{B}$ be an $m \times n$ matrix, let $\mathbf{b}$ be a constant m-vector, and let $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. Then

$$\mathbf{Y} \overset{d}{=} N(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\Lambda\mathbf{B}').$$

Proof. We have shown in Theorem 35 that the expectation and variance of $\mathbf{Y}$ are $\mathbf{B}\boldsymbol{\mu}+\mathbf{b}$ and $\mathbf{B}\Lambda\mathbf{B}'$, respectively. So all we need to show is that $\mathbf{Y}$ is a multivariate normal random vector.

Pick any vector $\mathbf{a}$

$$\begin{aligned} \mathbf{a}'\mathbf{Y} &= \mathbf{a}'\mathbf{B}\mathbf{X} + \mathbf{a}'\mathbf{b}, \\ &= (\mathbf{B}'\mathbf{a})'\mathbf{X} + \mathbf{a}'\mathbf{b}, \\ &= \underbrace{\mathbf{c}'\mathbf{X}}_{\overset{d}{=}\text{Normal}} + \mathbf{a}'\mathbf{b}, \\ &\overset{d}{=} \text{Normal}. \end{aligned}$$

□

Example 23. Suppose

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{d}{=} N\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & -2 \\ -2 & 7 \end{bmatrix} \right).$$

Let $Y_1 = X_1 + X_2$ and $Y_2 = 2X_1 - 3X_2$, find the joint distribution of $(Y_1, Y_2)'$.

Clearly,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ 2X_1 - 3X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{B}\mathbf{X}.$$

Hence, we know that

$$\mathbf{Y} \overset{d}{=} N(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Lambda\mathbf{B}') = \left(\begin{bmatrix} 3 \\ -4 \end{bmatrix}, \begin{bmatrix} 4 & -17 \\ -17 & 91 \end{bmatrix} \right).$$

Now we turn to define the normal distribution by the characteristic function.

Definition 29 (Multivariate Normal Distribution II). A random n-vector $\mathbf{X}$ is multivariate normal if its characteristic function is

$$\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Lambda\mathbf{t}},$$

where $\boldsymbol{\mu} = E(\mathbf{X})$, $\Lambda = Var(\mathbf{X})$.

Hence, the MGF of multivariate normal random vector is

$$M_{\mathbf{X}}(\mathbf{t}) = e^{\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Lambda\mathbf{t}}.$$

Finally we provide the definition of the multivariate normal distribution via probability density function.

Definition 30 (Multivariate Normal Distribution III). A random n-vector $\mathbf{X}$ with $\boldsymbol{\mu} = E(\mathbf{X})$, $\Lambda = Var(\mathbf{X})$, such that $\Lambda > 0$, is $N(\boldsymbol{\mu}, \Lambda)$ distributed if the density function is

$$f_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{\det(\Lambda)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \mathbb{R}^n.$$

Where $\det(\Lambda)$ is the determinant of Λ.

Note that under the assumption that $\det(\Lambda) > 0$ (non-singular), Definitions I, II and III are equivalent.

Definition 31 (Singular Distribution). A continuous random variable X for which pdf does not exist, we call that X has a singular distribution.

Let's see a bivariate case as an example.

Example 24 (Bivariate Normal Distribution). For $i = 1, 2$, let $\mu_i = E(X_i)$, $\sigma_i^2 = Var(X_i)$, $\sigma_{12} = Cov(X_1, X_2)$ and $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$. Thus,

$$\Lambda = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

and

$$\Lambda^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix} = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} & f_{X_1X_2}(x_1, x_2) \\ &= \left(\frac{1}{2\pi}\right) \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \end{aligned}$$

The following theorem is a very important one for multivariate normal random vector. In general we know that "independent" implies "uncorrelated", but the reverse is not true. However, if random variables are multivariate normal, then "uncorrelated" implies "independent"!

Theorem 38. Let $\mathbf{X}$ be a normal random vector. The components of $\mathbf{X}$ are *independent* **if and only if** they are *uncorrelated.*

Proof. We show only the $\Leftarrow$ part since the converse always being true. Since $X_1, X_2, \ldots, X_n$ are uncorrelated,

$$\Lambda = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \vdots \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix},$$

hence,

$$\Lambda^{-1} = \begin{pmatrix} \sigma_1^{-2} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-2} & 0 & \vdots \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^{-2} \end{pmatrix}.$$

The pdf is

$$\begin{aligned} f_X(\mathbf{x}) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{\det(\Lambda)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Lambda^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\prod_{i=1}^n \sigma_i} \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{X_i-\mu_i}{\sigma_i}\right)^2\right\}, \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2\right\}, \\ &= \prod_{i=1}^n f_{X_i}(x_i). \end{aligned}$$

□

Let's see an example.

Example 25. Let X_1 and X_2 be independent $N(0,1)$ random variables. Show that $X_1 + X_2$ and $X_1 - X_2$ are independent.

First of all, since X_1 and X_2 are independent, then stacking X_1 and X_2 gives us multivariate normal.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \overset{d}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

Clearly,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{BX}.$$

So

$$\mathbf{Y} \overset{d}{=} N(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\Lambda\mathbf{B}'),$$

or

$$\mathbf{Y} \overset{d}{=} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}\right).$$

That is, Y_1 and Y_2 are jointly normal and uncorrelated. Thus they are independent.

Theorem 39. Let $\mathbf{X} \overset{d}{=} N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, where $\sigma^2 > 0$. Let $\mathbf{C}$ be an arbitrary orthogonal matrix, and set $\mathbf{Y} = \mathbf{CX}$. Then

$$\mathbf{Y} \overset{d}{=} N(\mathbf{C}\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

Proof.

$$\mathbf{Y} \overset{d}{=} N(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}(\sigma^2\mathbf{I})\mathbf{C}') \overset{d}{=} N(\mathbf{C}\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

□

Clearly, $Y_1, Y_2, \ldots, Y_n$ are independent since $\mathbf{Y}$ is multivariate normal and $Y_1, Y_2, \ldots, Y_n$ are uncorrelated.

We now state a theorem from linear algebra.

Theorem 40 (Gram-Schmidt Process). Given variable $X_1, X_2, \ldots, X_n$, and constant $a_{11}, a_{12}, \ldots, a_{1n}$ and

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1n}X_n.$$

If $\sum_{j=1}^{n} a_{1j}^2 = 1$, then there exist a_{ij}, $i = 2, 3, \ldots, n$; $j = 1, 2, \ldots, n$ such that

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{in}X_n, \quad i = 2, 3, \ldots, n$$

and the transformation from $X_1, X_2, \ldots, X_n$ to $Y_1, Y_2, \ldots, Y_n$ ($\mathbf{Y} = \mathbf{AX}$) is an orthogonal transformation. That is, $\mathbf{A} = [a_{ij}]$ is an orthogonal matrix.

Note that a matrix $\mathbf{A}$ is said to be orthogonal if $\mathbf{A}^{-1} = \mathbf{A}'$. Furthermore, we know that

$$\sum_i Y_i^2 = \mathbf{Y}'\mathbf{Y} = (\mathbf{AX})'(\mathbf{AX}) = \mathbf{X}'\mathbf{A}'\mathbf{AX} = \mathbf{X}'\mathbf{X} = \sum_i X_i^2.$$

We now provide examples to show how useful Theorems 39 and 40 are. The following theorems are presented in the elementary statistics without a solid proof.

Theorem 41. Given $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$, and

$$\bar{X}_n = \frac{1}{n}\sum_i X_i, \quad S_n^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2.$$

1. $$\frac{(n-1)S_n^2}{\sigma^2} \overset{d}{=} \chi^2(n-1)$$

2. The sample mean $\bar{X}_n$ and sample variance S_n^2 are independent.

Proof. First of all, since $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$, we know that

$$\mathbf{X} \overset{d}{=} N(\boldsymbol{\mu}, \sigma^2\mathbf{I}).$$

According to Theorem 40, there exists an orthogonal matrix $\mathbf{C}$ such that the first row has all elements equal to $1/\sqrt{n}$. For instance,

$$\mathbf{C} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{2\cdot 3}} & \frac{1}{\sqrt{2\cdot 3}} & \frac{-2}{\sqrt{2\cdot 3}} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{(n-1)\cdot n}} & \frac{1}{\sqrt{(n-1)\cdot n}} & \cdots & \cdots & \cdots & \frac{1}{\sqrt{(n-1)\cdot n}} & \frac{-(n-1)}{\sqrt{(n-1)\cdot n}} \end{pmatrix},$$

Then we set $\mathbf{Y} = \mathbf{CX}$. By Theorem 39, we have

$$\mathbf{Y} \overset{d}{=} N(\mathbf{C}\boldsymbol{\mu}, \sigma^2 \mathbf{I}).$$

That is, $\mathbf{Y}$ has multivariate normal distribution with a diagonal variance-covariance matrix. Hence, $Y_1, Y_2, \ldots, Y_n$ are independent. By construction, we have

$$Y_1 = \sqrt{n}\bar{X}_n, \quad E(Y_1) = \sqrt{n}\mu. \quad Var(Y_1) = \sigma^2,$$

and for $i = 2, 3, \ldots, n$,

$$E(Y_i) = 0, \quad Var(Y_i) = \sigma^2.$$

Moreover,

$$\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = Y_2^2 + Y_3^2 + \cdots + Y_n^2 = \sum_{i=2}^{n} Y_i^2.$$

Therefore,

1. we can show that
$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{\sigma^2} = \frac{\sum_{i=2}^{n} Y_i^2}{\sigma^2} = \sum_{i=2}^{n}\left(\frac{Y_i - 0}{\sigma}\right)^2 \overset{d}{=} \chi^2(n-1)$$
2. $\bar{X}_n = \frac{1}{\sqrt{n}} Y_1$ and $S_n^2 = \frac{1}{n-1}(Y_2^2 + Y_3^2 + \cdots + Y_n^2)$ are independent since Y_1 and $(Y_2, \ldots, Y_n)$ are independent.

□

Finally, we present a theorem called the Daly's Theorem, which is a generalization of above example.

Theorem 42 (Daly's Theorem). Given a random n-vector $\mathbf{X}$. Let $\mathbf{X} \overset{d}{=} N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ and set $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Suppose that $g(\mathbf{X}) = g(X_1, X_2, \ldots, X_n)$ is *translation invariant*, that is, for all $\mathbf{X} \in \mathbb{R}^n$, we have

$$g(\mathbf{X} + a \cdot \mathbf{1}) = g(X_1 + a, X_2 + a, \ldots, X_n + a) = g(X_1, X_2, \ldots, X_n) = g(\mathbf{X})$$

for all constant a. Then $\bar{X}_n$ and $g(\mathbf{X})$ are independent.

Proof. Note that the variance-covariance matrix is $\sigma^2\mathbf{I}$, which implies that $X_1, X_2, \ldots, X_n$ are independent normal random variables with mean $\mu_1, \mu_2, \ldots, \mu_n$ and variance σ^2. Hence the joint density is

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu_i)^2}$$

Hence, the joint MGF of the vector $\begin{pmatrix} \bar{X} \\ g(\mathbf{X}) \end{pmatrix}$ is

$$E\left(e^{t_1\bar{X}+t_2 g(X_1,X_2,\ldots,X_n)}\right) = \frac{1}{(\sqrt{2\pi})^n\sigma^n}\int\cdots\int e^{\frac{t_1}{n}\sum_i x_i + t_2 g(x_1,x_2,\ldots,x_n)} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\mu_i)^2} dx_1 \cdots dx_n$$

The exponent is

$$\frac{t_1}{n}\sum_i x_i + t_2 g(x_1, x_2, \ldots, x_n) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu_i)^2 = \left[\frac{\sigma^2 t_1^2}{2n} + t_1\bar{\mu}\right] - \frac{1}{2\sigma^2}\sum_i\left(x_i - \mu_i - \frac{\sigma^2 t_1}{n}\right)^2 + t_2 g(x_1, x_2, \ldots, x_n)$$

where $\bar{\mu} = \frac{1}{n}\sum_i \mu_i$. Hence,

$$E\left(e^{t_1\bar{X}+t_2 g(X_1,X_2,\ldots,X_n)}\right) = e^{\frac{\sigma^2 t_1^2}{2n}+t_1\bar{\mu}} \frac{1}{(\sqrt{2\pi})^n\sigma^n}\int\cdots\int e^{-\frac{1}{2\sigma^2}\sum_i\left(x_i-\mu_i-\frac{\sigma^2 t_1}{n}\right)^2 + t_2 g(x_1,x_2,\ldots,x_n)} dx_1\cdots dx_n$$

It is worth noting that as $\bar{X} \sim N(\bar{\mu}, \frac{\sigma^2}{n})$, $e^{\frac{\sigma^2 t_1^2}{2n}+t_1\bar{\mu}}$ is the MGF of $\bar{X}$, i.e.,

$$M_{\bar{X}}(t_1) = E(e^{t_1\bar{X}}) = e^{\bar{\mu}t_1 + \frac{1}{2}\frac{\sigma^2}{n}t_1^2}$$

Moreover, since $g(\mathbf{X})$ is translation invariant

$$E\left(e^{t_1\bar{X}+t_2 g(X_1,X_2,\ldots,X_n)}\right) = M_{\bar{X}}(t_1)\frac{1}{(\sqrt{2\pi})^n\sigma^n}\int\cdots\int e^{-\frac{1}{2\sigma^2}\sum_i\left(x_i-\mu_i-\frac{\sigma^2 t_1}{n}\right)^2 + t_2 g\left(x_1-\frac{\sigma^2 t_1}{n}, x_2-\frac{\sigma^2 t_1}{n}, \ldots, x_n - \frac{\sigma^2 t_1}{n}\right)} dx_1\cdots dx_n$$

Let

$$(y_1, y_2, \ldots, y_n) = \left(x_1 - \frac{\sigma^2 t_1}{n}, x_2 - \frac{\sigma^2 t_1}{n}, \ldots, x_n - \frac{\sigma^2 t_1}{n}\right)$$

we thus have

$$E\left(e^{t_1\bar{X}+t_2 g(X_1,X_2,\ldots,X_n)}\right) = M_{\bar{X}}(t_1)\frac{1}{(\sqrt{2\pi})^n\sigma^2}\int\cdots\int e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu_i)^2 + t_2 g(y_1,y_2,\ldots,y_n)} dy_1\cdots dy_n$$

That is,

$$\begin{aligned} E\left(e^{t_1\bar{X}+t_2 g(X_1,X_2,\ldots,X_n)}\right) &= E(e^{t_1\bar{X}})\int\cdots\int e^{t_2 g(y_1,y_2,\ldots,y_n)}\frac{1}{(\sqrt{2\pi})^n\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_i (y_i-\mu_i)^2} dy_1\cdots dy_n \\ &= M_{\bar{X}}(t_1)E(e^{t_2 g(\mathbf{X})}) = M_{\bar{X}}(t_1)M_{g(\mathbf{X})}(t_2) \end{aligned}$$

By Theorem 30, $\bar{X}_n$ and $g(\mathbf{X})$ are independent.

□

That is, sample mean $\bar{X}_n$ is independent with any translation-invariant function of $\mathbf{X}$. Clearly, $S_n^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X}_n)^2$ is a translation invariant function such that the theorem can be applied.

Chapter 5

Sampling Theory

Definition 32 (**Random Sample).** A random sample $\{X_i\}_{i=1}^n$ from a population is a collection of i.i.d. random variables.

Definition 33 (**Statistics).** Any function of the random sample is called a statistic:

$$T_n = T(X_1, X_2, \dots, X_n).$$

Definition 34 (**Sampling Distribution).** If $T_n = T(X_1, X_2, \dots, X_n)$ is a statistic, then the distribution of T_n is called the sampling distribution.

For instance, if $\{X_i\}_{i=1}^n$ is a random sample from Bernoulli(p), then

$$T_n = \sum_{i=1}^n X_i \overset{d}{=} \text{Binomial}(n, p).$$

That is, $\sum_{i=1}^n X_i$ is a statistic with Binomial(n,p) as its sampling distribution.

Most properties of sampling distributions will be presented without further derivations since they have already been shown in the elementary statistics course.

5.1 Sample Mean

Let $\{X_i\}_{i=1}^n \sim^{i.i.d.} (\mu, \sigma^2)$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1. $E(\bar{X}_n) = \mu$.

2. $Var(\bar{X}_n) = \frac{\sigma^2}{n}$.

5.2 Sample Variance

Let $\{X_i\}_{i=1}^n \sim^{i.i.d.} (\mu, \sigma^2)$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

1. $S_n^2 = \frac{1}{n}\sum_i X_i^2 - \bar{X}_n^2 = \left[\frac{1}{n}\sum_i (X_i - \mu)^2\right] - (\bar{X}_n - \mu)^2$.

2. $E(S_n^2) = \left(1 - \frac{1}{n}\right)\sigma^2$.

3. $Var(S_n^2) = \frac{(n-1)^2}{n^3}\left(\mu_4 - \left[\frac{n-3}{n-1}\right]\mu_2^2\right)$, where $\mu_4 = E(X-\mu)^4$, $\mu_2 = E(X-\mu)^2$.

5.3 χ^2 Distribution

Theorem 43. Given that $Y \sim \chi^2(n)$.

1. $E(Y) = n$.

2. $Var(Y) = 2n$.

3. $M_Y(t) = (1-2t)^{-n/2}$.

Proof. Since $\chi^2(n) \overset{d}{=} \text{Gamma}\left(\frac{n}{2}, 2\right)$, then

$$\begin{aligned} E(Y) &= \frac{n}{2} \times 2 = n \\ Var(Y) &= \frac{n}{2} \times 2^2 = 2n \\ M_Y(t) &= \left(\frac{1}{1-2t}\right)^{\frac{n}{2}} \end{aligned}$$

□

Theorem 44. If $Z \sim N(0,1)$, and

$$Y = Z^2,$$

then $Y \overset{d}{=} \chi^2(1)$.

Proof.

$$\begin{aligned} M_{Z^2}(t) = E(e^{tZ^2}) &= \int_{-\infty}^{\infty} e^{tz^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(1-2t)z^2}dz \\ &= \frac{1}{\sqrt{1-2t}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{(1-2t)}}}e^{-\frac{1}{2}\left(\frac{z}{\sqrt{\frac{1}{(1-2t)}}}\right)^2}dz = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}} \end{aligned}$$

□

Theorem 45. If $\{Z_i\}_{i=1}^n \sim^{i.i.d.} N(0,1)$, and

$$Y = \sum_{i=1}^n Z_i^2,$$

then $Y \overset{d}{=} \chi^2(n)$.

Proof. Since $Z_i^2 \sim \chi^2(1)$,

$$M_{Z_i^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}$$

$$M_Y(t) = M_{\sum Z_i^2}(t) = \prod_{i=1}^n M_{Z_i^2}(t) = \left[\left(\frac{1}{1-2t}\right)^{\frac{1}{2}}\right]^n = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$$

□

5.4 Student's *t* Distribution

Theorem 46. Let $Z \overset{d}{=} N(0,1)$ and $W \overset{d}{=} \chi_k^2$, then

$$U = \frac{Z}{\sqrt{\frac{W}{k}}} \overset{d}{=} t_k.$$

Proof. Let $U = \frac{Z}{\sqrt{\frac{W}{k}}}$, and $V = W$. Hence, inversion yields

$$Z = U\left(\frac{V}{k}\right)^{\frac{1}{2}}$$
$$W = V$$

The Jacobian is

$$\mathbf{J} = \begin{vmatrix} \frac{\partial Z}{\partial U} & \frac{\partial Z}{\partial V} \\ \frac{\partial W}{\partial U} & \frac{\partial W}{\partial V} \end{vmatrix} = \begin{vmatrix} \left(\frac{V}{k}\right)^{\frac{1}{2}} & 0 \\ 0 & 1 \end{vmatrix} = \left(\frac{V}{k}\right)^{\frac{1}{2}}$$

Since Z and W are independent,

$$f_{ZW}(z,w) = f_Z(z)f_W(w).$$

Therefore,

$$\begin{aligned} f_{UV}(u,v) &= f_Z\left(u\left(\frac{v}{k}\right)^{\frac{1}{2}}\right) f_W(v)\left(\frac{v}{k}\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{\pi}} e^{-\frac{1}{2}\left(u\left(\frac{v}{k}\right)^{\frac{1}{2}}\right)^2} \frac{v^{\frac{k}{2}-1}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} e^{-\frac{1}{2}v}\left(\frac{V}{k}\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{2\pi}2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)k^{\frac{1}{2}}} v^{\frac{k+1}{2}-1} \exp\left[-\frac{1}{2}\left(1+\frac{u^2}{k}\right)v\right] \\ &= \frac{1}{\sqrt{k\pi}2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)} v^{\frac{k+1}{2}-1} \exp\left[-\frac{1}{2}\left(1+\frac{u^2}{k}\right)v\right] \end{aligned}$$

Let

$$c = \frac{1}{\sqrt{k\pi}2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)},$$

and

$$h(u) = \frac{1}{2}\left(1+\frac{u^2}{k}\right)$$

we can obtain

$$f_{UV}(u,v) = cv^{\frac{k+1}{2}-1}\exp\left[-h(u)v\right]$$

Hence,

$$f_U(u) = c\int_0^\infty v^{\frac{k+1}{2}-1}e^{-h(u)v}dv$$

According to the property of Gamma function (see Theorem 80),

$$\begin{aligned} f_U(u) &= c\left[\frac{1}{h(u)}\right]\Gamma\left(\frac{k+1}{2}\right) \\ &= \frac{1}{\sqrt{k\pi}2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)}\left[\frac{1}{\frac{1}{2}\left(1+\frac{u^2}{k}\right)}\right]^{\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right) \\ &= \frac{1}{\sqrt{k\pi}2^{\frac{k+1}{2}}\Gamma\left(\frac{k}{2}\right)}2^{\frac{k+1}{2}}\left(1+\frac{u^2}{k}\right)^{-\frac{k+1}{2}}\Gamma\left(\frac{k+1}{2}\right) \\ &= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}\frac{1}{\sqrt{k\pi}}\left(1+\frac{u^2}{k}\right)^{-\frac{k+1}{2}} \end{aligned}$$

This is exactly the pdf of the student's t distribution. □

5.5 Sampling from a Normal Distribution

Let $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu,\sigma^2)$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X}_n)^2$.

1.
$$\bar{X}_n \overset{d}{=} N\left(\mu, \frac{\sigma^2}{n}\right).$$

2. By Theorem 44,
$$\left[\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\right]^2 \overset{d}{=} \chi^2(1).$$

3. By Theorem 41,
$$\frac{(n-1)S_n^2}{\sigma^2} \overset{d}{=} \chi^2_{n-1}.$$

4. By Theorem 41, $\bar{X}_n$ and S_n^2 are independent.

5. By Theorem 45,
$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \overset{d}{=} \chi^2(n).$$

6. By Theorem 46,
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \overset{d}{=} t(n-1).$$

Chapter 6

Asymptotic Theory

6.1 Convergence Concepts

Let $\{X_1, X_2, \ldots\} = \{X_n\}_{n=1}^{\infty}$ is a sequence of random variables, and let X be another random variable.

Definition 35 (Convergence in Quadratic Mean). X_n converges in quadratic mean to the random variable X if

$$E|X_n - X|^2 \to 0 \text{ as } n \to \infty.$$

We denote it as

$$X_n \xrightarrow{L^2} X.$$

It is also called converges in square mean (or mean-square convergence).

An alternative notation for mean-square convergence is

$$X_n \xrightarrow{m.s.} X.$$

Definition 36 (Convergence in Distribution). X_n converges in distribution to the random variable X if $\forall x \in C(F)$,

$$F_n(x) \to F(x) \text{ as } n \to \infty,$$

where $C(F) = \{x : F(x) \text{ is continuous at } x\}$. That is, for every x which is a continuous point of F. We denote it as

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

Convergence in distribution means that X_n is approximately distributed with distribution function $F(x)$ for large n. That is, probabilities regarding X_n may be well approximated using probabilities regarding X. The approximation error decreases to zero as n increases to infinity.

However, the statement $X_n \xrightarrow{d} X$ does not say how large n must be in order for the approximation to be practically useful. To answer this question, we need a further result dealing explicitly with the approximation error as a function of n.

We will kind of "abuse" the notations such as $X_n \xrightarrow{d} N(0,1)$ as $n \longrightarrow \infty$ instead of the formally more correct, but lengthier $X_n \xrightarrow{d} X$ as $n \longrightarrow \infty$, where $X \overset{d}{=} N(0,1)$.

Note that the convergence specified by this definition is ***pointwise***, and only has to occur at points x where F is continuous. Note that the sequence $\{F_n(x)\}$ is said to ***pointwise converge*** to $F(x)$ if and only if for every $\varepsilon > 0$, there is a natural number $N(x)$ such that all $n \geq N(x)$, $|Fn(x) - F(x)| < \varepsilon$.[1]

Definition 37 (Convergence in Probability). X_n converges in probability to the random variable X if one of the following equivalent conditions holds:

(a) $\forall\ \varepsilon > 0$,
$$P(|X_n - X| > \varepsilon) \to 0 \ \text{ as } \ n \to \infty.$$

(b) $\forall\ \varepsilon > 0$,
$$P(|X_n - X| < \varepsilon) \to 1 \ \text{ as } \ n \to \infty.$$

(c) Given $\varepsilon > 0$, $\delta > 0$, $\exists N(\varepsilon, \delta)$ such that
$$P(|X_n - X| > \varepsilon) < \delta, \forall\ n > N.$$

(d) Given $\varepsilon > 0$, $\delta > 0$, $\exists N(\varepsilon, \delta)$ such that
$$P(|X_n - X| < \varepsilon) > 1 - \delta, \forall\ n > N.$$

That is, $P(|X_{N+1} - X| < \varepsilon) > 1 - \delta$, $P(|X_{N+2} - X| < \varepsilon) > 1 - \delta$, and so on.

We denote it as
$$X_n \xrightarrow{p} X.$$

Clearly, convergence in probability means that X_n is close to X with high probability.

We now consider a stronger mode of convergence: *almost sure convergence*.

[1] On the other hand, The sequence $\{F_n(x)\}$ ***uniformly converges*** to $F(x)$ if and only if for every $\varepsilon > 0$, there is a natural number N such that for all x and all $n \geq N$, $|Fn(x) - F(x)| < \varepsilon$.

Reference https://www.physicsforums.com/threads/convergence-of-random-variables.167343/

Definition 38 (**Almost Sure Convergence**). X_n converges almost surely (a.s.) to the random variable X if

$$P(\{\omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1.$$

More formally, given $\varepsilon > 0$, $\delta > 0$, $\exists N(\varepsilon, \delta)$ such that

$$P(|X_{N+1} - X| < \varepsilon, |X_{N+2} - X| < \varepsilon, \ldots) > 1 - \delta.$$

We denote it as

$$X_n \xrightarrow{a.s.} X.$$

It is also called that X_n *almost everywhere* or *with probability 1* or *strongly* towards X.

It is worth noting that the difference between convergence in probability and convergence almost surely is that convergence in probability requires all of the *individual* probabilities $P(|X_{N+1} - X| < \varepsilon)$, $P(|X_{N+2} - X| < \varepsilon)$,... to be larger than $1 - \delta$, while convergence almost surely requires that the *joint* probability $P(|X_{N+1} - X| < \varepsilon, |X_{N+2} - X| < \varepsilon, \ldots)$ should be larger than $1 - \delta$.

6.2 Relations Between the Convergence Concepts

We first outline the relations between the convergence concepts. Details will be provided in numerous theorems.

1. $X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{p} X.$
2. $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X.$
3. $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X.$
4. $X_n \xrightarrow{d} X \not\Rightarrow X_n \xrightarrow{p} X.$
5. For any constant c, $X_n \xrightarrow{d} c \Leftrightarrow X_n \xrightarrow{p} c$

Theorem 47. Convergence in quadratic mean implies convergence in probability, that is

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

Proof. By Markov inequality with $p = 2$ (see Theorem 5),

$$P(|X_n - X| \geq \varepsilon) \leq \frac{E|X_n - X|^2}{\varepsilon^2} \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

□

Theorem 48. Almost sure convergence implies convergence in probability, that is

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

Proof. Let $A = \{|X_{N+1} - X| < \varepsilon, |X_{N+2} - X| < \varepsilon, \ldots\}$, and $B_i = \{|X_{N+i} - X| < \varepsilon\}$, for $i = 1, 2, \ldots$. Because $A \subset B_i$, $P(B_i) > P(A)$, for $i = 1, 2, \ldots$. If X_n converges almost surely to X then $P(A) > 1-\delta$, and hence $P(B_i) > 1-\delta$, for $i = 1, 2, \ldots$, which thus implying that X_n converges in probability to X □

Theorem 49. Convergence in probability implies convergence in distribution, that is

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

Proof. For any $\varepsilon > 0$,

$$\begin{aligned} F_n(x) &= P(X_n \leq x), \\ &= P(X_n \leq x, X > x + \varepsilon) + P(X_n \leq x, X \leq x + \varepsilon), \\ &\leq P(X_n \leq x, X > x + \varepsilon) + P(X \leq x + \varepsilon), \\ &= P(X_n - X < -\varepsilon) + P(X \leq x + \varepsilon), \\ &\leq P(|X_n - X| > \varepsilon) + P(X \leq x + \varepsilon), \end{aligned}$$

that is,

$$F_n(x) \leq F(x + \varepsilon) + P(|X_n - X| > \varepsilon). \tag{6.1}$$

Also,

$$\begin{aligned} F(x - \varepsilon) &= P(\{X \leq x - \varepsilon\} \cap \{X_n > x\}) + P(\{X \leq x - \varepsilon\} \cap \{X_n \leq x\}), \\ &\leq P(\{X \leq x - \varepsilon\} \cap \{X_n > x\}) + P(X_n \leq x), \\ &= P(X_n - X > \varepsilon) + P(X_n \leq x), \\ &\leq P(|X_n - X| > \varepsilon) + F_n(x), \end{aligned}$$

that is,

$$F(x - \varepsilon) \leq F_n(x) + P(|X_n - X| > \varepsilon). \tag{6.2}$$

Since $X_n \xrightarrow{p} X$, by letting $n \to \infty$ in equations (6.1) and (6.2), we have

$$F(x - \varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x + \varepsilon).$$

This relation holds for all x and for all $\varepsilon > 0$. Finally, suppose that $x \in C(F)$ and let $\varepsilon \to 0$. It follows that

$$F(x) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x).$$

that is,

$$\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = \lim_{n\to\infty} F_n(x) = F(x).$$

□

Theorem 50. For any constant c, $X_n \xrightarrow{d} c \;\Leftrightarrow\; X_n \xrightarrow{p} c$, as $n \longrightarrow \infty$.

Proof. We only need to show "⇒."

Given $X_n \xrightarrow{d} c$, we know that

$$F_n(x) \longrightarrow F(x) = 1_{\{x \geq c\}}.$$

That is, if we treat the constant c as a random variable, then for $\delta > 0$, $F(c+\delta) = P(c \leq c+\delta) = 1$ and $F(c-\delta) = P(c \leq c-\delta) = 0$. Figure 6.1 shows the distribution function of a constant c.

Let $\varepsilon > 0$. Then

$$\begin{aligned} P(|X_n - c| > \varepsilon) &= 1 - P(|X_n - c| \leq \varepsilon), \\ &= 1 - P(c - \varepsilon \leq X_n \leq c + \varepsilon), \\ &= 1 - \left[F_n(c+\varepsilon) - F_n(c-\varepsilon) + P(X_n = c - \varepsilon)\right], \\ &\leq 1 - \left[F_n(c+\varepsilon) - F_n(c-\varepsilon)\right], \\ &\longrightarrow 1 - \left[F(c+\varepsilon) - F(c-\varepsilon)\right], \\ &= 1 - \left[1 - 0\right] = 0. \end{aligned}$$

□

Let's use an example to show how to apply Theorem 50.

Example 26. Given that $X_n \sim N\left(0, \frac{1}{n}\right)$. Then show that

$$X_n \xrightarrow{p} 0.$$

Clearly, $\sqrt{n}X_n \sim N(0,1)$. For $x < 0$,

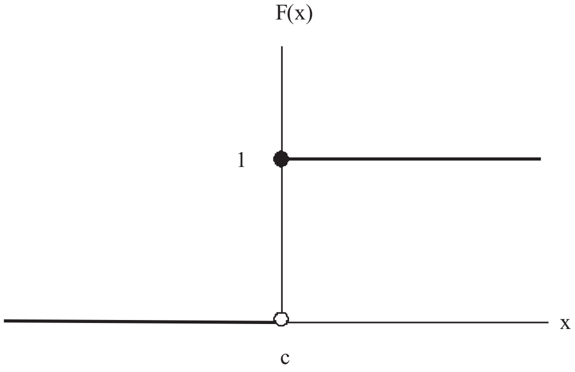
$$\begin{aligned} F_n(x) &= P(X_n \leq x), \\ &= P(\sqrt{n}X_n \leq \sqrt{n}x), \\ &= P(N(0,1) \leq \sqrt{n}x) \to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Note that $\sqrt{n}x \to -\infty$ since $x < 0$.

For $x > 0$,

$$F_n(x) = P(N(0,1) \leq \sqrt{n}x) \to 1 \quad \text{as} \quad n \to \infty.$$

Figure 6.1: Distribution Function of A Constant



That is, for $x \neq 0$ (note that the discontinuous point is $x = 0$),

$$\lim_{n\to\infty} F_n(x) = F(x) = 1_{\{x\geq 0\}},$$

so $X_n \xrightarrow{d} 0$, which implies $X_n \xrightarrow{p} 0$ by Theorem 50.

According to this example, we can also realize the motivation for considering only points of continuity of $F(x)$. That is, the concept of convergence of distribution can be applied in such a case since we require $F(x)$ to be a CDF (not any function), which is a right continuous function.

Of course, this result can be shown by using the Chebyshev's Inequality.

Theorem 51 (Continuous Mapping Theorem). Let $\{X_n\}$ be a sequence of random variables, and X be another random variable. Suppose $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function. Then

$$X_n \xrightarrow{p} X \;\;\Rightarrow\;\; g(X_n) \xrightarrow{p} g(X).$$

Proof. By the continuity of $g(\cdot)$ we have , for given $\varepsilon, \delta > 0$,

$$|X_n - X| < \delta \;\text{ implies }\; |g(X_n) - g(X)| < \varepsilon.$$

That is, $\{|X_n - X| < \delta\} \subseteq \{|g(X_n) - g(X)| < \varepsilon\}$, and

$$P(|g(X_n) - g(X)| < \varepsilon) \geq P(|X_n - X| < \delta) \longrightarrow 1, \;\text{ as }\; n \longrightarrow \infty.$$

Thus,

$$P(|g(X_n) - g(X)| < \varepsilon) \longrightarrow 1 \;\text{ as }\; n \longrightarrow \infty.$$

□

Theorem 52. Let $X_n \xrightarrow{p} X$, and $Y_n \xrightarrow{p} Y$. Then

(a) $X_n + Y_n \xrightarrow{p} X + Y$.

(b) $X_n Y_n \xrightarrow{p} XY$.

(c) $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{Y}$, where $Y_n \neq 0$ and $Y \neq 0$.

We need the following lemma to prove Theorem 52.

Lemma 3. If E_n and F_n are two sequences of events, then

$$P(E_n) \longrightarrow 1, P(F_n) \longrightarrow 1 \Rightarrow P(E_n \bigcap F_n) \longrightarrow 1.$$

Proof. Since

$$P((E_n \bigcap F_n)^c) = P(E_n^c \bigcup F_n^c) \leq P(E_n^c) + P(F_n^c) \longrightarrow 0,$$

we have

$$P(E_n \bigcap F_n) = 1 - P((E_n \bigcap F_n)^c) \longrightarrow 1.$$

□

We now prove Theorem 52 (a), and leave (b) and (c) as exercises.

Proof. From

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

it follows that[2]

$$\left\{|X_n - X| < \frac{\varepsilon}{2} \text{ and } |Y_n - Y| < \frac{\varepsilon}{2}\right\} \subseteq \{|X_n - X| + |Y_n - Y| < \varepsilon\} \subseteq \{|(X_n + Y_n) - (X + Y)| < \varepsilon\}.$$

Therefore,

$$P(|(X_n + Y_n) - (X + Y)| < \varepsilon) \geq P\left(|X_n - X| < \frac{\varepsilon}{2} \text{ and } |Y_n - Y| < \frac{\varepsilon}{2}\right).$$

Now since $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, which imply

$$P\left(|X_n - X| < \frac{\varepsilon}{2}\right) \longrightarrow 1,$$

and

$$P\left(|Y_n - Y| < \frac{\varepsilon}{2}\right) \longrightarrow 1.$$

[2] Clearly, $|X_n - X| < \frac{\varepsilon}{2}$ and $|Y_n - Y| < \frac{\varepsilon}{2}$ implies $|X_n - X| + |Y_n - Y| < \varepsilon$ but the converse is not necessary true. Moreover, $|X_n - X| + |Y_n - Y| < \varepsilon$ implies $|(X_n + Y_n) - (X + Y)| < \varepsilon$ but the converse is not necessary true.

Hence Lemma 3 proves

$$P(|(X_n + Y_n) - (X + Y)| < \varepsilon) \geq P\left(|X_n - X| < \frac{\varepsilon}{2} \text{ and } |Y_n - Y| < \frac{\varepsilon}{2}\right) \longrightarrow 1.$$

□

Indeed, if we extend Theorem 51 to random vector.

Theorem 53. (Continuous Mapping Theorem for Random Vectors) Suppose $\mathbf{h}(\cdot)$ is a vector-valued continuous function. If

$$\mathbf{Z_n} \xrightarrow{p} \mathbf{Z},$$

then

$$\mathbf{h}(\mathbf{Z_n}) \xrightarrow{p} \mathbf{h}(\mathbf{Z}).$$

Hence, Theorem 52 can be easily proved by setting $\mathbf{Z_n} = (X_n \quad Y_n)'$.

The next theorem about combinations of convergence in probability and in distribution will be very useful to derive the asymptotic distribution of estimators.

Theorem 54 (Slutsky's Theorem). Let $X_n \xrightarrow{d} X$, and $Y_n \xrightarrow{p} c$. Then

(a) $X_n + Y_n \xrightarrow{d} X + c$.

(b) $X_n Y_n \xrightarrow{d} Xc$.

(c) $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$, where $Y_n \neq 0$ and $c \neq 0$.

Proof. We prove Theorem 54 (a), and leave (b) and (c) as exercises.

$$\begin{aligned} F_{X_n+Y_n}(a) &= P(X_n + Y_n \leq a), \\ &= P\left(\{X_n + Y_n \leq a\} \bigcap \{|Y_n - c| \leq \varepsilon\}\right) + P\left(\{X_n + Y_n \leq a\} \bigcap \{|Y_n - c| > \varepsilon\}\right) \end{aligned}$$

Let

$$S_1 = \{X_n + Y_n \leq a\} \bigcap \{|Y_n - c| \leq \varepsilon\},$$

$$S_2 = \{X_n + Y_n \leq a\} \bigcap \{Y_n - c \geq -\varepsilon\},$$

$$S_3 = \{X_n \leq a + \varepsilon - c\}.$$

Note that S_1 implies S_2, and S_2 implies S_3, hence, $S_1 \subset S_2 \subset S_3$, which suggests that $P(S_1) \le P(S_2) \le P(S_3)$. We thus have

$$\begin{aligned} F_{X_n+Y_n}(a) &= P(X_n + Y_n \le a), \\ &= P(S_1) + P\left(\{X_n + Y_n \le a\} \bigcap \{|Y_n - c| > \varepsilon\}\right), \\ &\le P(S_3) + P\left(\{X_n + Y_n \le a\} \bigcap \{|Y_n - c| > \varepsilon\}\right), \\ &\le P(S_3) + P(|Y_n - c| > \varepsilon), \\ &= P(X_n \le a + \varepsilon - c) + P(|Y_n - c| > \varepsilon), \\ &= F_n(a + \varepsilon - c) + P(|Y_n - c| > \varepsilon). \end{aligned}$$

Since $X_n \xrightarrow{d} X$, we have $F_n(a + \varepsilon - c) \longrightarrow F_X(a + \varepsilon - c)$. Moreover, since $Y_n \xrightarrow{p} c$, we have $P(|Y_n - c| > \varepsilon) \longrightarrow 0$. That is,

$$\limsup_{n\to\infty} F_{X_n+Y_n}(a) \le F_X(a + \varepsilon - c).$$

Using a similar argument, we have

$$F_X(a - \varepsilon - c) \le \liminf_{n\to\infty} F_{X_n+Y_n}(a).$$

Thus, we have

$$F_X(a - \varepsilon - c) \le \liminf_{n\to\infty} F_{X_n+Y_n}(a) \le \limsup_{n\to\infty} F_{X_n+Y_n}(a) \le F_X(a + \varepsilon - c).$$

Let $\varepsilon \to 0$,

$$F_X(a - c) \le \liminf_{n\to\infty} F_{X_n+Y_n}(a) \le \limsup_{n\to\infty} F_{X_n+Y_n}(a) \le F_X(a - c).$$

That is,

$$\lim_{n\to\infty} F_{X_n+Y_n}(a) = F_X(a - c),$$

or

$$F_{Y_n+X_n}(a) \longrightarrow F_X(a - c) = F_{X+c}(a),$$

since

$$P(X \le a - c) = P(X + c \le a).$$

□

6.3 Convergence via Transforms

Theorem 55 (Continuity Theorem for Moment Generating Functions). Let $X_1, X_2, \ldots$ be random variables such that $M_{X_n}(t)$ exists for $|t| < h$, for some $h > 0$, and for all n. Suppose that X is a random variable whose moment generating function, $M_X(t)$ exists for $|t| \leq h_1 < h$ for some $h_1 > 0$ and that

$$M_{X_n}(t) \longrightarrow M_X(t), \text{ as } n \longrightarrow \infty.$$

Then

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

Theorem 56 (Continuity Theorem for Characteristic Functions). Let $X_1, X_2, \ldots$ be random variables, and suppose that

$$\phi_{X_n}(t) \longrightarrow \phi_X(t), \text{ as } n \longrightarrow \infty.$$

Then

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

The above theorems for convergence via transforms are useful to prove the central limit theorem. Moreover, the converse of Theorem 56 is also of interest. Namely, if $X_1, X_2, \ldots$ are random variables such that

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

then

$$\phi_{X_n}(t) \longrightarrow \phi_X(t), \text{ as } n \longrightarrow \infty.$$

According to Theorem 56, we can derive the following theorem (try it!).

Theorem 57. Let $X_1, X_2, \ldots$ be random variables, and suppose that, for some real number c,

$$\phi_{X_n}(t) \longrightarrow e^{itc}, \text{ as } n \longrightarrow \infty.$$

Then

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

6.4 Landau Symbol (Order Notation)

Definition 39 (Landau Symbols to Real Numbers). Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers.

1. $a_n = O(b_n)$ if for some finite real number $\Delta > 0$, there exists a finite integer N such that for all $n \geq N$,
$$\left|\frac{a_n}{b_n}\right| < \Delta,$$
i.e., $\lim_{n\to\infty}\left|\frac{a_n}{b_n}\right| < \infty$.
2. $a_n = o(b_n)$ if for every real number $\delta > 0$ there exists a finite integer $N(\delta)$ such that for all $n \geq N(\delta)$,
$$\left|\frac{a_n}{b_n}\right| < \delta,$$
i.e., $\lim_{n\to\infty}\frac{a_n}{b_n} = 0$.

When a_n and b_n both tend to infinity, $a_n = o(b_n)$ states that a_n tends to infinity more slowly than b_n; when both tend to 0, it states that a_n tends to zero faster than b_n. Obviously, if $a_n = o(b_n)$, then $a_n = O(b_n)$.

Example 27. Let $a_n = \frac{1}{n^2}$ and $b_n = \frac{1}{n}$. Clearly, as $n \to \infty$
$$\frac{a_n}{b_n} = \frac{1/n^2}{1/n} = \frac{1}{n} \to 0.$$
That is,
$$\frac{1}{n^2} = o\left(\frac{1}{n}\right)$$

We will show an example of the use of the small o notation.

Example 28. Consider the following sequence
$$a_n = \frac{1}{n} - \frac{2}{n^2} + \frac{4}{n^3}.$$
Then we have first-order approximation as
$$a_n \approx \frac{1}{n},$$
i.e.,
$$a_n = \frac{1}{n} + o\left(\frac{1}{n}\right).$$
Moreover, we have the second-order approximation as
$$a_n \approx \frac{1}{n} - \frac{2}{n^2}.$$

That is,

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

Note that we have the following simple properties of the small o relations.

Lemma 4. (Small o Relations)

1. $a_n = o(1)$ iff. $a_n \longrightarrow 0$.
2. If $a_n = o(b_n)$, then $\frac{a_n}{b_n} = o(1)$. Thus $o(b_n) = b_n o(1)$.
3. $a_n = o(b_n)$ implies $ca_n = o(b_n)$. Thus $o(b_n) = ko(b_n)$, $k = 1/c$.

Most of the time, we are interested in the order relationship of power of n.

Definition 40. Let $\{a_n\}$ be a sequence of real number.

1. The sequence $\{a_n\}$ is at most of order n^k, denoted $a_n = O(n^k)$, if for some finite real number $\Delta > 0$, there exists a finite integer N such that for all $n > N$,

$$|n^{-k} a_n| < \Delta.$$

2. The sequence $\{a_n\}$ is of order smaller than n^k, denoted $a_n = o(n^k)$, if for every $\delta > 0$ there exists a finite integer $N(\delta)$ such that for all $n \geq N(\delta)$,

$$|n^{-k} a_n| < \delta,$$

i.e., $n^{-k} a_n \to 0$.

Now we turn our focus on the notion of orders of magnitude for convergence in probability.

Definition 41 (Landau Symbols to Random Variables). Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables.

1. $X_n = O_p(Y_n)$ if $\forall \varepsilon > 0$, $\exists$ a positive real number $M_\varepsilon < \infty$, such that

$$P\left(|X_n| \leq M_\varepsilon |Y_n|\right) > 1 - \varepsilon.$$

That is, $\{X_n\}$ is bounded in probability.

2. $X_n = o_p(Y_n)$ if

$$\frac{X_n}{Y_n} \xrightarrow{p} 0.$$

That is, $\{X_n\}$ is vanish in probability.

Note that

1. $X_n = o_p(1)$ iff. $X_n \xrightarrow{p} 0$.

2. $c_n o_p(Y_n) = o_p(c_n Y_n)$ for constant c_n, and $C_n o_p(Y_n) = o_p(C_n Y_n)$ for random variable C_n.

3. If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$. That is, if X_n converge in distribution, then it is bounded in probability.

4. $O_p(n^a) o_p(n^b) = o_p(n^{a+b})$.

For more details about Landau Symbols, see chapter 2 in White (2001) or section 1.4/2.1 in Lehmann (2001).

6.5 Weak Law of Large Number

Theorem 58 (Weak Law of Large Numbers (WLLN) I). Let $\{X_n\}$ be a sequence of i.i.d. random variables such that $E(X_1) = \mu$, and $Var(X_1) = \sigma^2 < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_i X_i \xrightarrow{p} \mu.$$

Equivalently, we can also denote WLLN as

$$\bar{X}_n - \mu \xrightarrow{p} 0,$$

$$\begin{aligned} \bar{X}_n - \mu &= o_p(1), \\ \bar{X}_n &= \mu + o_p(1). \end{aligned}$$

Proof. Pick any ε > 0, by Markov Inequality,

$$\begin{aligned} P(|\bar{X}_n - \mu| > \varepsilon) &\leq \frac{E|\bar{X}_n - \mu|^2}{\varepsilon^2}, \\ &= \frac{E(\bar{X}_n - \mu)^2}{\varepsilon^2}, \\ &= \frac{Var(\bar{X}_n)}{\varepsilon^2}, \\ &= \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0 \text{ as } n \longrightarrow \infty. \end{aligned}$$

□

The above version of WLLN has been shown in the elementary statistics course. However, the assumption of finite second moment can be relaxed. We thus have the following version of WLLN assuming finite first moment only.

Theorem 59 (Weak Law of Large Numbers (WLLN) II). Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables such that $E(X_1) = \mu < \infty$, then

$$\bar{X}_n = \frac{1}{n}\sum_i X_i \xrightarrow{p} \mu.$$

Proof. According to Theorem 57, it is sufficient to show that

$$\phi_{\bar{X}_n}(t) \longrightarrow e^{it\mu}, \quad \text{as} \quad n \longrightarrow \infty.$$

Let $Y_n = \sum_i X_i$. By Theorem 33, we have

$$\phi_{\bar{X}_n}(t) = \phi_{Y_n}\left(\frac{t}{n}\right).$$

Moreover, by Theorem 32,

$$\phi_{\bar{X}_n}(t) = \phi_{Y_n}\left(\frac{t}{n}\right) = \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n.$$

Moreover, after introducing the Landau Symbol, we can rewrite Theorem 34 as

$$\begin{aligned}
\phi_X(t) &= 1 + \sum_{k=1}^{\infty} E(X^k) \cdot \frac{(it)^k}{k!}, \\
&= 1 + \sum_{k=1}^{n} E(X^k) \cdot \frac{(it)^k}{k!} + o(|t|^n), \quad \text{[see Theorem 4.2 in Chapter 4 of Gut (2013)]} \\
&= 1 + E(X)\frac{it}{1!} + o(t). \quad [n = 1]
\end{aligned}$$

Hence,

$$\phi_{X_1}\left(\frac{t}{n}\right) = 1 + E(X_1)\frac{i\frac{t}{n}}{1!} + o\left(\frac{t}{n}\right),$$

and

$$\begin{aligned}
\phi_{\bar{X}_n}(t) &= \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n, \\
&= \left[1 + E(X_1)\frac{i\frac{t}{n}}{1!} + o\left(\frac{t}{n}\right)\right]^n, \\
&= \left[1 + i\frac{t}{n}\mu + o\left(\frac{t}{n}\right)\right]^n, \\
&= \left[1 + \frac{it\mu[1 + o(1)]}{n}\right]^n \longrightarrow e^{it\mu} \quad \text{as} \quad n \longrightarrow \infty.
\end{aligned}$$

□

Note that in the above proof, we have applied the fact that given $a_n \longrightarrow a$,

$$\left(1 + \frac{a_n}{n}\right)^n \longrightarrow e^a \text{ as } n \longrightarrow \infty.$$

6.6 Central Limit Theorem

Theorem 60 (The Central Limit Theorem, CLT). Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Then

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1).$$

Proof. Let $Z_i = \frac{X_i - \mu}{\sigma}$. Hence we have $E(Z_1) = 0$, $Var(Z_1) = E(Z_1^2) = 1$ and

$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma}.$$

Let $Y_n = \sum_i Z_i$ and

$$W_n = \sqrt{n}\bar{Z}_n = \frac{Y_n}{\sqrt{n}}.$$

Hence,

$$\begin{aligned}
\phi_{W_n}(t) = \phi_{\frac{Y_n}{\sqrt{n}}}(t) = \phi_{Y_n}\left(\frac{t}{\sqrt{n}}\right) &= \left[\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n, \\
&= \left[1 + E(Z_1)\frac{i\frac{t}{\sqrt{n}}}{1!} + E(Z_1^2)\frac{\left(i\frac{t}{\sqrt{n}}\right)^2}{2!} + o\left(\frac{t^2}{n}\right)\right]^n, \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n, \\
&= \left[1 - \frac{\frac{t^2}{2}[1 - o(1)]}{n}\right]^n \longrightarrow e^{-t^2/2} \text{ as } n \longrightarrow \infty.
\end{aligned}$$

That is,

$$\sqrt{n}\bar{Z}_n \xrightarrow{d} N(0,1),$$

and thus

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1).$$

□

6.7 Delta Method and Cramer-Wold Theorem

We now introduce some theorems that is useful to derive asymptotic distributions.

Theorem 61 (Cramer-Wold Device). Let $\mathbf{X_n} \in \mathbb{R}^k$ be a sequence of i.i.d. random k-vector, and $\mathbf{a} \in \mathbb{R}^k$ be a constant vector.

$$\mathbf{X_n} \xrightarrow{d} \mathbf{X} \text{ iff. } \mathbf{a}'\mathbf{X_n} \xrightarrow{d} \mathbf{a}'\mathbf{X}$$

Proof. See pages 282–283 in Mittelhammer (1995). □

According to the Cramer-Wold Device, we can easily obtain the following multivariate CLT.

Theorem 62 (Multivariate CLT). Let $\{\mathbf{X_n}\}$ be a sequence of i.i.d. random k-vector with mean vector $E(\mathbf{X_i}) = \boldsymbol{\theta}$ and variance-covariance matrix $Var(\mathbf{X_i}) = \Sigma$, for all i, and Σ is positively definite. Then

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X_i} - \boldsymbol{\theta}\right) \xrightarrow{d} N(\mathbf{0}, \Sigma).$$

Proof. Let $Z_i = \mathbf{a}'\mathbf{X_i}$. Thus

$$E(Z_i) = E(\mathbf{a}'\mathbf{X_i}) = \mathbf{a}'E(\mathbf{X_i}) = \mathbf{a}'\boldsymbol{\theta},$$

$$Var(Z_i) = Var(\mathbf{a}'\mathbf{X_i}) = \mathbf{a}'Var(\mathbf{X_i})\mathbf{a} = \mathbf{a}'\Sigma\mathbf{a}.$$

Note that

$$\frac{\sqrt{n}\,(n^{-1}\sum_i Z_i - E(Z_i))}{\sqrt{Var(Z_i)}} = \frac{\sqrt{n}(n^{-1}\mathbf{a}'(\sum_i(\mathbf{X_i} - \boldsymbol{\theta})))}{\sqrt{\mathbf{a}'\Sigma\mathbf{a}}}.$$

By univariate CLT, we have

$$\frac{\sqrt{n}\,(n^{-1}\sum_i Z_i - E(Z_i))}{\sqrt{Var(Z_i)}} \xrightarrow{d} N(0, 1),$$

hence

$$\frac{\sqrt{n}(n^{-1}\mathbf{a}'(\sum_i(\mathbf{X_i} - \boldsymbol{\theta})))}{\sqrt{\mathbf{a}'\Sigma\mathbf{a}}} \xrightarrow{d} N(0, 1).$$

That is,

$$\sqrt{n}(n^{-1}\mathbf{a}'(\sum_i(\mathbf{X_i} - \boldsymbol{\theta}))) \xrightarrow{d} N(0, \mathbf{a}'\Sigma\mathbf{a}).$$

By Cramer-Wold Device,

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X_i} - \boldsymbol{\theta}\right) \xrightarrow{d} N(\mathbf{0}, \Sigma).$$

□

Theorem 63 (Univariate Delta Method). Given $g(\cdot)$ a continuous and differentiable function. If

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)N(0, \sigma^2).$$

Proof. By Taylor theorem, we expand $g(X_n)$ around $g(\theta)$,

$$g(X_n) = g(\theta) + (X_n - \theta)g'(X_n^*),$$

where $X_n^* = \lambda X_n + (1 - \lambda)\theta$, $\lambda \in [0, 1]$.

Since $\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ by construction, we have

$$\sqrt{n}(X_n - \theta) = O_p(1).$$

That implies

$$X_n - \theta = o_p(1),$$

or

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

It follows that

$$X_n^* = \lambda X_n + (1 - \lambda)\theta \xrightarrow{p} \theta,$$

and by CMT, we have

$$g'(X_n^*) \xrightarrow{p} g'(\theta).$$

That is,

$$g'(X_n^*) = g'(\theta) + o_p(1)$$

Therefore,

$$\begin{aligned}\sqrt{n}(g(X_n) - g(\theta)) &= \sqrt{n}(X_n - \theta)[g'(\theta) + o_p(1)],\\ &= \sqrt{n}(X_n - \theta)g'(\theta) + \sqrt{n}(X_n - \theta)o_p(1),\\ &\xrightarrow{d} g'(\theta)N(0, \sigma^2).\end{aligned}$$

□

Well, do not forget that

$$g'(\theta)N(0, \sigma^2) \overset{d}{=} N\left(0, \left[g'(\theta)\right]^2 \sigma^2\right).$$

Let's see an example.

Example 29. Suppose that $\{X_i\}_{i=1}^n \sim^{i.i.d.}$Bernoulli$(p)$, where $p \neq 1/2$. Find the asymptotic distribution of $\bar{X}(1-\bar{X})$.

You should figure it out by yourself that the asymptotic distribution of $\bar{X}(1-\bar{X})$ is

$$\bar{X}(1-\bar{X}) \sim^A N\left(p(1-p), \frac{p(1-p)(1-2p)^2}{n}\right).$$

Theorem 64 (**Multivariate Delta Method**)**.** Suppose that $\mathbf{W_n}$ is a sequence of random k-vector such that

$$\sqrt{n}(\mathbf{W_n} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \Sigma),$$

where Σ is positively definite. Let $g(\cdot) : \mathbb{R}^k \mapsto \mathbb{R}$ be a function such that

$$\nabla g(\mathbf{y}) = \frac{dg(\mathbf{y})}{d\mathbf{y}} = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \frac{\partial g}{\partial y_2} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix},$$

and $\nabla g(\boldsymbol{\theta})$ is non-zero and continuous at $\boldsymbol{\theta}$. Then

$$\sqrt{n}(g(\mathbf{W_n}) - g(\boldsymbol{\theta})) \xrightarrow{d} N(0, \nabla g(\boldsymbol{\theta})'\Sigma\nabla g(\boldsymbol{\theta})).$$

Here is an application of the Delta Method.

Example 30. Consider the following regression model

$$\{y_i, x_{1i}, x_{2i}\}_{i=1}^n,$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i = \mathbf{x_i}'\beta + e_i,$$
$$e_i|\mathbf{x_i} \sim (0, \sigma^2),$$

The parameter of interest is

$$\theta = \frac{\beta_1}{\beta_2}.$$

Find the asymptotic distribution of the analog estimator

$$\hat{\theta} = \frac{\hat{\beta}_1}{\hat{\beta}_2},$$

where

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left(\sum_i \mathbf{x_i}\mathbf{x_i}'\right)^{-1}\left(\sum_i \mathbf{x_i} y_i\right).$$

Since we know that

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, V\right), \quad V = \sigma^2[E(\mathbf{x_i x_i}')]^{-1}.$$

Hence, by Delta Method,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(\mathbf{0}, H_\beta' V H_\beta)$$

where

$$H_\beta = \begin{pmatrix} 0 \\ 1/\beta_2 \\ -\beta_1/\beta_2^2 \end{pmatrix}.$$

Chapter 7

Estimation

Given random sample $\{X_i\}_{i=1}^n \sim F(\mathbf{x}, \theta)$. Suppose that we use a statistic $\hat{T}_n(X_1, X_2, \ldots, X_n) = \hat{T}_n(\mathbf{X})$ to estimate population features denoted by $q(\theta)$, $\theta \in \Theta$, where Θ is called the parameter space. We call $\hat{T}_n$ an estimator of $q(\theta)$.

7.1 Small Sample Criteria for Estimators

Definition 42 (Unbiased Estimator). An estimator $\hat{T}_n$ is said to be unbiased if

$$E(\hat{T}_n) = q(\theta).$$

Hence, if $\hat{T}_n$ is a biased estimator, then

$$B(\theta) = E(\hat{T}_n) - q(\theta)$$

is called *bias*.

Definition 43 (Mean Square Error). The mean square error is defined by

$$MSE(q(\theta), \hat{T}_n) = E(\hat{T}_n - q(\theta))^2.$$

Mean square error is a most popular measure of distance between $\hat{T}_n$ and $q(\theta)$. It has been shown in your elementary statistics that

$$MSE(q(\theta), \hat{T}_n) = Var(\hat{T}_n) + [E(\hat{T}_n) - q(\theta)]^2 = Var(\hat{T}_n) + [B(\theta)]^2.$$

Definition 44 (Minimum Variance Unbiased Estimator, MVUE). $\hat{T}_n^*$ is said to be an MVUE if for all $\theta \in \Theta$

$$\hat{T}_n^* = \arg\min_{\{\hat{T}_n:\ \hat{T}_n \text{ is unbiased for } q(\theta)\}} E(\hat{T}_n - q(\theta))^2$$

To compare two estimator, we have the following criterion.

Definition 45 (Relative Efficiency). Let T_1 and T_2 be estimators of $q(\theta)$. We say that T_1 is more efficient than T_2 if

$$MSE(q(\theta), T_1) \leq MSE(q(\theta), T_2).$$

7.2 Large Sample Properties of Estimators

Definition 46 (Consistency). $\hat{\theta}_n$ is consistent for θ if

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

That is, an $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta}$ converges to θ in probability.

Definition 47 (Best Asymptotically Normal (BAN) Estimator). $\hat{\theta}_n$ is a BAN estimator of θ if

1. $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$.
2. $\sigma^2 \leq r^2$ for $\tilde{\theta}_n$ such that $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, r^2)$.

Notice that some books would add one more condition for a BAN estimator: $\hat{\theta}_n \xrightarrow{p} \theta$. However, it is clear that condition 1 (asymptotically normal) in the theorem has already implied consistency (why?).

7.3 Interval Estimation

Definition 48 (Confidence Interval). Let $L(\mathbf{X})$, $U(\mathbf{X})$ be two statistics such that $L(\mathbf{X}) \leq U(\mathbf{X})$. We say that the random interval

$$[L(\mathbf{X}), U(\mathbf{X})]$$

is a $(1 - \alpha) \cdot 100\%$ confidence interval for θ if

$$P_\theta(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha.$$

We have already learned how to construct confidence intervals and approximate confidence intervals in elementary statistics. For instance, consider the following two cases:

Case I $\{X_i\}_{i=1}^n \sim^{i.i.d.} N(\mu, \sigma^2)$.

Let $\bar{X}_n = \frac{1}{n}\sum_i X_i$, $S_n^2 = \frac{1}{n}\sum_i (X_i - \bar{X}_n)^2$. In Case I, the statistic

$$\varphi = \frac{\sqrt{n-1}(\bar{X}_n - \mu)}{S_n} = \frac{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}{\sqrt{\frac{nS_n^2}{\sigma^2}/(n-1)}} \overset{d}{=} t(n-1).$$

That is, the statistic φ has an exact t distribution, and is therefore exactly free of unknown parameters. Hence we say that φ is an exactly *pivotal* statistic. The $(1-\alpha)\cdot 100\%$ exact confidence interval for μ is

$$\left[\bar{X}_n - t_{\alpha/2}(n-1)\frac{S_n}{\sqrt{n-1}}, \quad \bar{X}_n + t_{\alpha/2}(n-1)\frac{S_n}{\sqrt{n-1}}\right].$$

Case II $\{X_i\}_{i=1}^n \sim^{i.i.d.} (\mu, \sigma^2)$.

In Case II, without the assumption of normality, the same statistic

$$\varphi = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$$

does not have any known distribution. However, it can be shown that

$$\varphi = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

Thus the asymptotic distribution of φ is the standard normal, which does not depend on the parameters. We say that φ is *asymptotically pivotal*. Then the approximate confidence interval is given by

$$\left[\bar{X}_n - Z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \quad \bar{X}_n + Z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right],$$

where

$$P\left(-Z_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \le Z_{\alpha/2}\right) = 1 - \alpha + o(1).$$

We now illustrate how to construct a *confidence region* for two estimators.

Example 31. Given $\{X_i\} \sim^{i.i.d.} N(\mu, \sigma^2)$, where μ and σ^2 are both unknown. Then find a $(1-\alpha)\cdot 100\%$ confidence region for (μ, σ^2).

Let $\bar{X}_n = n^{-1}\sum_i X_i$ and $S_n^2 = n^{-1}\sum_i (X_i - \bar{X}_n)^2$. Since we know that

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

and

$$\frac{nS_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

By Theorem 42, $\bar{X}_n$ and S_n^2 are independent.

Now choose $c > 0$ such that

$$P\left(-c \le N(0,1) \le c\right) = \sqrt{1-\alpha},$$

and choose a, b such that

$$P(a \le \chi^2_{n-1} \le b) = \sqrt{1-\alpha}.$$

Therefore,

$$P\left(-c \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le c, a \le \frac{nS_n^2}{\sigma^2} \le b\right) = P\left(-c \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le c\right) P\left(a \le \frac{nS_n^2}{\sigma^2} \le b\right) = 1 - \alpha.$$

That is,

$$P\left((\mu - \bar{X}_n)^2 \le \frac{c^2\sigma^2}{n}, \frac{nS_n^2}{b} \le \sigma^2 \le \frac{nS_n^2}{a}\right) = 1 - \alpha.$$

7.4 Maximum Likelihood Estimation

Let X be a random variable with pdf $f(x,\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Suppose that $\{X_1, X_2, \dots, X_n\}$ is a random sample, then the joint density of the random sample is

$$f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

The likelihood of the sample is thus a function of θ:

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n; \theta).$$

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

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x_1, x_2, \dots, x_n),$$

or equivalently

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta; x_1, x_2, \dots, x_n).$$

Theorem 65 (Invariance Property). If $\hat{\theta}$ is the MLE of θ, and let $G(\theta)$ be any function of θ, then the MLE of $G(\theta)$ is $G(\hat{\theta})$.

Proof. Let $G(\theta) = g$. Define an induced likelihood function L^*:

$$L^*(g, \mathbf{x}) = \max_{\{\theta : G(\theta) = g\}} L(\theta, \mathbf{x}).$$

Let $\hat{g}$ be the MLE of L^*,

$$\hat{g} = \arg\max_g L^*(g, \mathbf{x}).$$

Hence,

$$\begin{aligned} L^*(\hat{g}, \mathbf{x}) &= \max_g \max_{\{\theta : G(\theta) = g\}} L(\theta, \mathbf{x}), \quad \text{(by definition)} \\ &= \max_\theta L(\theta, \mathbf{x}), \quad \text{(by integrated maximization)} \\ &= L(\hat{\theta}, \mathbf{x}). \quad \text{(since } \hat{\theta} \text{ is an MLE)} \end{aligned}$$

On the other hand,

$$\begin{aligned} L(\hat{\theta}, \mathbf{x}) &= \max_{\{\theta : G(\theta) = G(\hat{\theta})\}} L(\theta, \mathbf{x}) \quad \text{(since } \hat{\theta} \text{ is an MLE)} \\ &= L^*(G(\hat{\theta}), \mathbf{x}). \quad \text{(by definition of } L^*\text{)} \end{aligned}$$

Therefore, we have just shown that

$$L^*(\hat{g}, \mathbf{x}) = L^*(G(\hat{\theta}), \mathbf{x}).$$

□

Definition 49 (Score Function). The function

$$S_\theta = \frac{\partial \ln L}{\partial \theta}$$

is called the score for estimating θ.

Theorem 66. The score function has zero expectation:

$$E(S_\theta) = 0.$$

Proof. Note that since

$$\int L(\theta, \mathbf{x}) dx = 1,$$

differentiating both sides partially with respect to θ gives us

$$0 = \int \frac{\partial L}{\partial \theta} d\mathbf{x} = \int \frac{1}{L(\theta, \mathbf{x})} \frac{\partial L}{\partial \theta} L(\theta, \mathbf{x}) d\mathbf{x} = \int \frac{\partial \ln L}{\partial \theta} L d\mathbf{x} = E\left(\frac{\partial \ln L}{\partial \theta}\right).$$

That is,

$$E(S_\theta) = E\left(\frac{\partial \ln L}{\partial \theta}\right) = 0.$$

□

Definition 50 (Fisher Information). The function

$$I(\theta) = E(S_\theta^2) = Var(S_\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right].$$

is called the information for estimating θ.

Theorem 67 (Information Matrix Equality).

$$I(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right],$$

where $-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$ is called the *Hessian*.

From previous proof, we know that

$$0 = \int \frac{\partial \ln L}{\partial \theta} L d\mathbf{x}.$$

Differentiating both sides partially with respect to θ gives us

$$\begin{aligned} 0 &= \int \frac{\partial \ln L}{\partial \theta}\frac{\partial L}{\partial \theta} d\mathbf{x} + \int \frac{\partial^2 \ln L}{\partial \theta^2} L d\mathbf{x}, \\ &= \int \left[\frac{\partial \ln L}{\partial \theta}\right]^2 L(\theta, \mathbf{x}) d\mathbf{x} + \int \frac{\partial^2 \ln L}{\partial \theta^2} L d\mathbf{x}. \end{aligned}$$

That is,

$$I(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right].$$

Theorem 68 (Cramer-Rao Lower Bound, CRLB). Let $\hat{\theta}$ be an estimator of θ such that $E(\hat{\theta}) = u(\theta)$. Under some regularity conditions (see page 179 in Ramanathan (1993)), we have the Cramer-Rao inequality:

$$Var(\hat{\theta}) \geq \frac{[u'(\theta)]^2}{I(\theta)} = \frac{[1 + B'(\theta)]^2}{I(\theta)},$$

where $I(\theta)$ is the Fisher information, and $B(\theta) = E(\hat{\theta}) - \theta = u(\theta) - \theta$ is the bias. We then called

$$\underline{CR} = \frac{[1 + B'(\theta)]^2}{I(\theta)}$$

the Cramer-Rao Lower Bound.

Proof. It is well-known that the correlation coefficient of any two random variables Z, W is between -1 and 1. Hence, $[Cov(Z, W)]^2 \leq Var(Z)Var(W)$. Let

$$Z = \hat{\theta},$$

$$W = S_\theta,$$

we thus have $E(Z) = u(\theta)$, $E(W) = E(S_\theta) = 0$, $Var(W) = Var(S_\theta) = I(\theta)$.

Since

$$u(\theta) = E(\hat{\theta}) = \int \hat{\theta} L(\theta, \mathbf{x}) d\mathbf{x},$$

we have

$$u'(\theta) = \int \hat{\theta} \frac{\partial L}{\partial \theta} d\mathbf{x} = \int \hat{\theta} \left(\frac{1}{L} \frac{\partial L}{\partial \theta} \right) L d\mathbf{x} = \int \hat{\theta} S_\theta L d\mathbf{x} = \int Z W L d\mathbf{x} = E(ZW) = Cov(Z, W).$$

Therefore,

$$[u'(\theta)]^2 = [Cov(Z, W)]^2 \leq Var(Z) Var(W) = Var(\hat{\theta}) I(\theta).$$

□

Clearly, if $\tilde{\theta}$ is an unbiased estimator of θ, then the Cramer-Rao inequality becomes

$$Var(\tilde{\theta}) \geq \frac{1}{I(\theta)}.$$

Thus, suppose that $Var(\tilde{\theta}) = \frac{1}{I(\theta)} = \underline{CR}$, then $\tilde{\theta}$ is an MVUE (see Definition 44).

Finally, we present the following properties of MLE without proof. Interesting readers should refer to pages 192–198 in Ramanathan (1993).

Theorem 69 (Asymptotic Properties of MLE). Given that $\hat{\theta}$ is an MLE of θ. Then

1. $\hat{\theta} \xrightarrow{p} \theta$.

2. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [\Sigma(\theta)]^{-1})$, where $\Sigma(\theta) = \lim_{n \to \infty} \frac{I(\theta)}{n}$.

3. $\left[\frac{I(\hat{\theta})}{n} \right]^{-1} \xrightarrow{p} [\Sigma(\theta)]^{-1}$. That is, $\frac{I(\hat{\theta})}{n}$ is a consistent estimator of $\Sigma(\theta)$.

4. $\hat{\theta}$ is a BAN estimator of θ. (see Definition 47).

Chapter 8

Hypothesis Testing

In elementary statistics, we have learned how to conduct hypothesis tests using appropriate pivotal test statistics. However, no rationale was provided to suggest that they are best in any sense. We now consider a method for deriving rejection region corresponding to tests that are most powerful tests of a given size for testing simple hypothesis.

Let's review some useful definitions and notations.

Hypothesis Suppose $\{X_i\}_{i=1}^n \sim^{i.i.d.} f(x,\theta)$, where $\theta \in \Theta$. Any statement about θ is a hypothesis. Suppose that

$$\Theta_0 \bigcup \Theta_1 = \Theta, \quad \Theta_0 \bigcap \Theta_1 = \varnothing,$$

then statement that $H_0 : \theta \in \Theta_0$ is called a null hypothesis. Moreover, statement that $H_1 : \theta \in \Theta_1$ is called an alternative hypothesis.

Test A statistical test represents a rule of action. Given some data, we use this rule to decide whether we are able to reject or fail to reject the null hypothesis.

Simple and Composite Hypotheses If $card(\Theta_0) = 1$, it is called a simple hypothesis. On the other hand, if $card(\Theta_0) > 1$, it is called a composite hypothesis. Note that $card(A)$ denotes the cardinality, which gives the number of members of set A.

Critical Region It is also called a *rejection region*. It can be represented as

$$C(\mathbf{x}) = \{\mathbf{x}'s \text{ for which we reject } H_0\}$$

Type I and Type II Errors

1. Type I error: Rejecting H_0 when it is true.

2. Type II error: Failing to reject H_0 when it is false.

Power of a Test

$$\pi(\theta) = P(\text{reject } H_0 | \theta)$$

is called the power of the test. Hence,

$$P(\text{Type I error}) = P(\text{reject } H_0 | \theta \in \Theta_0),$$

$$P(\text{Type II error}) = P(\text{fail to reject } H_0 | \theta \in \Theta_1) = 1 - P(\text{reject } H_0 | \theta \in \Theta_1).$$

That is,

1. If $\theta \in \Theta_0$, $\pi(\theta)$ is the probability of making a type I error.
2. If $\theta \in \Theta_1$, $\pi(\theta)$ is the power.

Clearly, when $\theta \in \Theta_1$, we define

$$\beta(\theta) = P(\text{Type II error}) = 1 - \pi(\theta).$$

Size of a Test

$$\alpha = \max_{\theta \in \Theta_0} P(\text{Type I error}) = \max_{\theta \in \Theta_0} \pi(\theta)$$

Hence, the size of a test is the largest probability of making a type I error. α is also called *the level of significance*.

8.1 Most Powerful Tests

Definition 51 (Most Powerful (MP) Test). A test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ based on a critical region C is said to be a most powerful test of size α if

1. $\pi_C(\theta_0) = \alpha$
2. $\pi_C(\theta_1) \geq \pi_A(\theta_1)$ for any other critical region A where $\pi_A(\theta_0) = \alpha$.

The following theorem shows how to derive a most powerful test.

Theorem 70 (Neyman-Pearson Lemma). Suppose that a random sample $\{X_i\}_{i=1}^n$ has joint pdf $f(\mathbf{x}, \theta)$, where $\theta \in \Theta$. If there exists a constant $k \in (0, \infty)$ such that

$$P\left(\frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_0)} \geq k \,\middle|\, \theta = \theta_0 \right) = \alpha,$$

then the most powerful size-α test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ is given by

$$C(\mathbf{x}) = \left\{ \mathbf{x} : \frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_0)} \geq k \right\}.$$

Proof. Let A be another critical region of the same size α. Then

$$\int_C f(\theta_0, \mathbf{x}) d\mathbf{x} = \alpha = \int_A f(\theta_0, \mathbf{x}) d\mathbf{x}.$$

Now, let $\bar{C}$ and $\bar{A}$ denote the complement of C and A, respectively,

$$\begin{aligned}
&\int_C f(\theta_1, \mathbf{x}) d\mathbf{x} - \int_A f(\theta_1, \mathbf{x}) d\mathbf{x}, \\
&\quad = \left[\int_{C \cap A} f(\theta_1, \mathbf{x}) d\mathbf{x} + \int_{C \cap \bar{A}} f(\theta_1, \mathbf{x}) d\mathbf{x} \right] - \left[\int_{A \cap C} f(\theta_1, \mathbf{x}) d\mathbf{x} + \int_{A \cap \bar{C}} f(\theta_1, \mathbf{x}) d\mathbf{x} \right], \\
&\quad = \int_{C \cap \bar{A}} f(\theta_1, \mathbf{x}) d\mathbf{x} - \int_{A \cap \bar{C}} f(\theta_1, \mathbf{x}) d\mathbf{x}.
\end{aligned}$$

Note that in C, $f(\theta_1, \mathbf{x}) \geq kf(\theta_0, \mathbf{x})$, and in $\bar{C}$, $f(\theta_1, \mathbf{x}) < kf(\theta_0, \mathbf{x})$, which implies $-f(\theta_1, \mathbf{x}) > -kf(\theta_0, \mathbf{x})$.

Therefore,

$$\begin{aligned}
LHS &= \int_{C \cap \bar{A}} f(\theta_1, \mathbf{x}) d\mathbf{x} - \int_{A \cap \bar{C}} f(\theta_1, \mathbf{x}) d\mathbf{x}, \\
&> \int_{C \cap \bar{A}} kf(\theta_0, \mathbf{x}) d\mathbf{x} - \int_{A \cap \bar{C}} kf(\theta_0, \mathbf{x}) d\mathbf{x}, \\
&= \left[\int_{C \cap \bar{A}} kf(\theta_0, \mathbf{x}) d\mathbf{x} + \int_{A \cap C} f(\theta_0, \mathbf{x}) d\mathbf{x} \right] - \left[\int_{A \cap \bar{C}} kf(\theta_0, \mathbf{x}) d\mathbf{x} + \int_{A \cap C} kf(\theta_0, \mathbf{x}) d\mathbf{x} \right], \\
&= \int_C kf(\theta_0, \mathbf{x}) d\mathbf{x} - \int_A kf(\theta_0, \mathbf{x}) d\mathbf{x}, \\
&= k\alpha - k\alpha = 0.
\end{aligned}$$

That is,

$$\int_C f(\theta_1, \mathbf{x}) d\mathbf{x} > \int_A f(\theta_1, \mathbf{x}) d\mathbf{x}.$$

□

8.2 Uniformly Most Powerful Tests

Definition 52 (Uniformly Most Powerful (UMP) Test). A test of $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ based on a critical region C is said to be a uniformly most powerful test of size α if

$$\max_{\theta \in \Theta_0} \pi_C(\theta) = \alpha,$$

and for any other test based on critical region A that satisfies $\max_{\theta \in \Theta_0} \pi_A(\theta) = \alpha$, we have

$$\pi_C(\theta) \geq \pi_A(\theta)$$

for all $\theta \in \Theta_1$.

A UMP test often exists in the case of a one-sided composite alternative such as $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. A possible technique for determining a UMP test is first to derive the Neyman-Pearson test for a particular alternative value θ_1, and then show that the test does not depend on the specific alternative value. On the other hand, for the two-sided composite alternative $H_1 : \theta \neq \theta_0$, it is not possible to find a test that is UMP for every alternative value.

Example 32. Let $\{X_i\} \sim^{i.i.d.} N(\theta, 1)$. We want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where $\theta_0 < \theta_1$.

Clearly, it can be shown that

$$\frac{f(\theta_1, \mathbf{x})}{f(\theta_0, \mathbf{x})} = \exp\left[\sum_{i=1}^{n} \frac{(x_i - \theta_0)^2 - (x_i - \theta_1)^2}{2}\right].$$

By Neyman-Pearson Lemma, the MP test critical region for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ is given by

$$\begin{aligned} C(\mathbf{x}) &= \left\{\mathbf{x} : \exp\left[\sum_{i=1}^{n} \frac{(x_i - \theta_0)^2 - (x_i - \theta_1)^2}{2}\right] \geq k\right\}, \\ &= \left\{\mathbf{x} : \sum_{i=1}^{n} \frac{(x_i - \theta_0)^2 - (x_i - \theta_1)^2}{2} \geq \log k\right\}, \\ &= \left\{\mathbf{x} : \sum_{i=1}^{n} x_i \geq \frac{\log k}{\theta_1 - \theta_0} + \frac{n(\theta_1 + \theta_0)}{2}\right\}. \end{aligned}$$

Now we need to find out k. Given the size of the test is α, we have

$$\begin{aligned} \alpha = P(\mathbf{x} \in C \mid \theta = \theta_0) &= P\left(\bar{X}_n \geq \frac{\log k}{n(\theta_1 - \theta_0)} + \frac{(\theta_1 + \theta_0)}{2} \,\middle|\, \theta = \theta_0\right), \\ &= P\left(\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \geq \frac{\left[\frac{\log k}{n(\theta_1 - \theta_0)} + \frac{(\theta_1 + \theta_0)}{2}\right] - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \,\middle|\, \theta = \theta_0\right), \\ &= P\left(\frac{\bar{X}_n - \theta}{\sqrt{\frac{1}{n}}} \geq \frac{\left[\frac{\log k}{n(\theta_1 - \theta_0)} + \frac{(\theta_1 + \theta_0)}{2}\right] - \theta}{\sqrt{\frac{1}{n}}} \,\middle|\, \theta = \theta_0\right), \\ &= P\left(\sqrt{n}(\bar{X}_n - \theta_0) \geq \frac{\log k}{\sqrt{n}(\theta_1 - \theta_0)} + \frac{\sqrt{n}(\theta_1 - \theta_0)}{2}\right), \\ &= P\left(N(0,1) \geq \frac{\log k}{\sqrt{n}(\theta_1 - \theta_0)} + \frac{\sqrt{n}(\theta_1 - \theta_0)}{2}\right). \end{aligned}$$

Thus,

$$Z_\alpha = \frac{\log k}{\sqrt{n}(\theta_1 - \theta_0)} + \frac{\sqrt{n}(\theta_1 - \theta_0)}{2}.$$

So

$$\begin{aligned} C(\mathbf{x}) &= \left\{ \mathbf{x} : \sum_{i=1}^{n} x_i \geq \frac{\log k}{\theta_1 - \theta_0} + \frac{n(\theta_1 + \theta_0)}{2} \right\}, \\ &= \left\{ \mathbf{x} : \sum_{i=1}^{n} x_i \geq \sqrt{n} Z_\alpha + n\theta_0 \right\}, \\ &= \left\{ \mathbf{x} : \bar{X}_n \geq \theta_0 + \frac{Z_\alpha}{\sqrt{n}} \right\}. \end{aligned}$$

Let's see an example that UMP test exists for one-sided composite hypothesis.

Example 33. Let $\{X_i\} \sim^{i.i.d.} N(\theta, 1)$. We want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$.

Now pick any $\theta_1 > \theta_0$. So following the previous example, Neyman-Pearson Lemma gives us the MP test critical region for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ as

$$C(\mathbf{x}) = \left\{ \mathbf{x} : \bar{X}_n \geq \theta_0 + \frac{Z_\alpha}{\sqrt{n}} \right\}.$$

Note that $C(\mathbf{x})$ does not depend on θ_1, so it is also an UMP test.

8.3 Likelihood Ratio Test

Definition 53. Suppose that a random sample $\{X_i\}_{i=1}^n$ has joint pdf $f(\mathbf{x}, \theta)$, where $\theta \in \Theta$. Let $L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta)$ denotes the likelihood function. A test statistic for testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ is given by

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta} L(\theta, \mathbf{x})}{\sup_{\theta \in \Theta_0} L(\theta, \mathbf{x})},$$

which is called a likelihood ratio test.

Clearly, the critical region is

$$C(\mathbf{x}) = \{\mathbf{x} : \lambda \geq k\},$$

such that

$$\max_{\theta \in \Theta_0} P(\lambda \geq k) = \alpha.$$

Theorem 71 (Asymptotic Distribution of LR test). Let random sample $\{X_i\}_{i=1}^n$ has joint pdf $f(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$. Suppose that we would like to test the null hypothesis that

$$H_0 : \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_r \end{pmatrix} = \begin{pmatrix} \theta_1^0 \\ \theta_2^0 \\ \vdots \\ \theta_r^0 \end{pmatrix}$$

with the alternative hypothesis $H_1 : H_0$ is not true. Define

$$\lambda = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}, \mathbf{x})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}, \mathbf{x})} = \frac{L(\hat{\boldsymbol{\theta}})}{L(\hat{\boldsymbol{\theta}}_0)},$$

Then under H_0,

$$2 \log \lambda \xrightarrow{d} \chi_r^2.$$

Proof. Proof is based on the asymptotic properties of MLEs. See page 228–229 in Ramanathan (1993). □

8.4 Wald Test

Another popular test of $\theta = \theta_0$ is the Wald test. Given that an MLE estimator $\hat{\theta}$ is asymptotically normally distributed (see Theorem 69):

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [\Sigma(\theta)]^{-1}),$$

where $\Sigma(\theta) = \lim_{n\to\infty} \frac{I(\theta)}{n}$.

Then under $H_0 : \theta = \theta_0$,

$$\begin{aligned} W &= (\hat{\theta} - \theta_0)' I(\hat{\theta})(\hat{\theta} - \theta_0), \\ &= \underbrace{\sqrt{n}(\hat{\theta} - \theta_0)'}_{\xrightarrow{d} N(0,\Sigma(\theta_0)]^{-1})} \underbrace{\frac{I(\hat{\theta})}{n}}_{\xrightarrow{p} \Sigma(\theta_0)} \underbrace{\sqrt{n}(\hat{\theta} - \theta_0)}_{\xrightarrow{d} N(0,\Sigma(\theta_0)]^{-1})}, \\ &\xrightarrow{d} \chi_r^2. \end{aligned}$$

Chapter 9

The Bootstrap

9.1 The Empirical Distribution Function

Definition 54 (Empirical Distribution Function, EDF). Given a random sample $\{X_i\}_{i=1}^n \sim^{i.i.d.} F(x)$. The empirical distribution function $\hat{F}_n$ is the CDF that puts mass $\frac{1}{n}$ at each data point X_i. Formally,

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n 1(X_i \leq x)}{n},$$

where

$$1(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases}$$

is an indicator function.

Clearly,

$$E(1(X_i \leq x)) = 1 \cdot F(x) + 0 \cdot [1 - F(x)] = F(x),$$
$$E((1(X_i \leq x))^2) = 1^2 \cdot F(x) + 0^2 \cdot [1 - F(x)] = F(x),$$

and

$$Var(1(X_i \leq x)) = E((1(X_i \leq x))^2) - [E(1(X_i \leq x))]^2 = F(x) - [F(x)]^2 = F(x)[1 - F(x)].$$

Hence, we have

$$E[\hat{F}_n(x)] = \frac{\sum_{i=1}^n E(1(X_i \leq x))}{n} = \frac{nF(x)}{n} = F(x),$$

and

$$Var[\hat{F}_n(x)] = \frac{1}{n^2} \sum_i Var(1(X_i \leq x)) = \frac{nF(x)[1 - F(x)]}{n^2} = \frac{F(x)[1 - F(x)]}{n}.$$

Furthermore, by WLLN,

$$\hat{F}_n(x) \xrightarrow{p} F(x).$$

Finally, by CLT,

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} N(0, F(x)[1 - F(x)]).$$

9.2 Monte Carlo Simulations

Monte Carlo simulation is an important computer-aided tool in econometrics. The word "simulation" indicates that an artificial model of a real system is builded to study and understand the system. The name "Monte Carlo" is referred to the randomness inherent in the analysis. The term Monte Carlo Method was coined by S. Ulam and Nicholas Metropolis in reference to games of chance, a popular attraction in Monte Carlo, Monaco. Many years ago, some gamblers studied how they could maximize their chances of winning by using simulations to check the probability of occurrence for each possible case. In summary, Monte Carlo simulation is a method of analysis based on artificially recreating a chance process (usually with a computer), running it many times, and directly observing the results.

Let $\{x_i\}_{i=1}^n$ be the data (observations) randomly drawn from a population distribution F. Let $T_n = T_n(x_1, \dots, x_n, \theta)$ be a statistic of interest, where θ is a parameter which is in general assumed to represent the distribution, F. Thus, the exact distribution of T_n is

$$G_n(\tau, F) = P(T_n \leq \tau | F).$$

Since F (or θ) is unknown, G_n is in general unknown. Monte Carlo simulation uses numerical simulation to compute $G_n(\tau, F)$ for selected choices of F. The Monte Carlo simulation is conducted as follows.

1. The researcher chooses θ and sample size n to construct a hypothetical data generating process (DGP).

2. A random sample $\{x_i^*\}_{i=1}^n$ is drawn from distribution F characterized by θ using the computer's random number generater. (To be more precise, it should be called a "pseudo random number generater" since the number generated is not truly random. However, modern pseudo random generators are accurate enough that we can ignore this fact.) The generator generates sequences of values that appear to be drawn from a specified probability distribution.

3. Calculate the statistic $T_n = T_n(x_1^*, \dots, x_n^*, \theta)$ from the pseudo data.

4. Repeat steps 2 and 3 B times and store the results. Typically, $B = 1000$ or $B = 5000$. These results constitute a random sample of size B from the distribution of T_{nb}, where T_{nb}

denotes the experiment result in the b-th draw. We then have an *empirical distribution function* (EDF):

$$\hat{G}_n(\tau) = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb} \leq \tau),$$

where $1(\cdot)$ is the indicator function.

The theoretical justification for using simulation is the Fundamental Theorem of Statistics (FTS). According to FTS, the EDF of a set of independent drawings of a random variable generated by some DGP converges to the true CDF of the random variable under that DGP. That is, since

$$G_n(\tau) = P(T_n \leq \tau) = E(1(T_n \leq \tau)),$$

by WLLN, for any τ we have

$$\hat{G}_n(\tau) = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb} \leq \tau) \xrightarrow{p} E(1(T_n \leq \tau)) = G_n(\tau).$$

A Remark on Simulation For a random number generator, we use a number called *seed* to determine the probability space. For example, in GAUSS, we use

```
RNDSEED 123587;
```

In RATS, we use

```
SEED 123587
```

In R, we use

```
set.seed(123587)
```

When programming your own simulation, it is important to set up a seed such that the simulation results can be replicated.

9.2.1 Applications of Monte Carlo Simulations

The typical purpose of a Monte Carlo is to investigate the performance of a statistical procedure such as an estimator or a test. The performance in general depends on sample size n and the true data generating process F. For example, it is of interest to know the size and power of a particular test. Furthermore, we may use Monte carlo simulation to approximate the small-sample distribution of a particular estimate, and then compute its standard error or confidence interval.

9.2.2 Example: Empirical Power and Size of a t-test

Consider a simple regression model

$$y_t = \alpha + \beta x_t + e_t.$$

When the sample is large, we know that the t-ratio is asymptotically $N(0,1)$ distributed,

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \xrightarrow{d} N(0,1).$$

under null hypothesis that $H_0 : \beta = 0$. In this example, we will investigate the empirical size and empirical power of the t-statistic for the regression coefficient.

The DGP under H_0 is as follows:

$$y_t = 2.5 + e_t,$$

where e_t is drawn from the t distribution with 3 degrees of freedom. On the other hand, the DGP under H_1 is

$$y_t = 2.5 + \beta x_t + e_t,$$

where $x_t \sim N(0,1)$. We consider $\beta = 0.1, 0.5, 1.0$ and -0.1, and then conduct the hypothesis test of $\beta = 0$ under the 5% significance level.

Recall that the size is defined as

$$P(|t| > 1.96 \mid H_0 \text{ is true}),$$

while the power is defined as

$$P(|t| > 1.96 \mid H_1 \text{ is true}).$$

We can also compute the size-adjusted power. Using the DGP under null, we can find the critical value (cv) t^* so that

$$P(|t| > t^* \mid H_0) = 0.05.$$

Then the size-adjusted power is obtained by

$$P(|t| > t^* \mid H_1).$$

Table 9.1 reports the simulation results for different values of β with sample size 25. Number of replications is set to be 1000. The empirical power is higher when the true parameter β is far more from zero.

We can also investigate how sample size affects the empirical size and power of a test. Table 9.2 reports the case that under alternative hypothesis, $\beta = 0.1$. Clearly, the power improves when sample size gets larger.

Table 9.1: Empirical Size and Power of the t-test

β	Size	True 5% cv (t^*)	Power	Size-adjusted Power
0.1	0.060	2.045	0.082	0.067
0.5	0.060	2.045	0.648	0.615
1.0	0.060	2.045	0.993	0.991
−0.5	0.060	2.045	0.632	0.595

Table 9.2: Effects of Sample Size

Sample Size	Size	True 5% cv (t^*)	Power	Size-adjusted Power
25	0.060	2.045	0.082	0.067
50	0.052	1.966	0.099	0.099
100	0.059	2.060	0.156	0.132
1000	0.059	2.005	0.900	0.889

9.3 Bootstrap

This technique was invented by Bradley Efron (1979, 1981, 1982) and further developed by Efron and Tibshirani (1993). "Bootstrap" means that one available sample gives rise to many others by resampling (a concept reminiscent of pulling yourself up by your own bootstrap). In general, bootstrap is developed for inferential purposes (Efron, 1981, 1982).

Confidence intervals, hypothesis testing, and standard errors are all based on the idea of the sampling distribution (or asymptotic distribution) of a statistic. In many settings, we have no model for the population to construct exact sampling distribution. Furthermore, we cannot obtain enough sample to ensure the large sample theory works (the small sample problem). The bootstrap may help us out of these troubles. The idea of bootstrap is simple: use the one sample we have as though it were the population, taking many resamples from it to construct the bootstrap distribution. Then in statistical inference, use the bootstrap distribution in place of the sampling distribution.

Before discussing what bootstrap is, let's look at an example showing the poor performance of asymptotic approximation.

An Example of the Poor Performance of Asymptotic Approximation Consider the following regression model

$$\{y_i, x_{1i}, x_{2i}\}_{i=1}^{n}$$

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

$$\begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \sim N(0, I_2),$$

$$e_i \sim N(0, 3^2),$$

and $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 0.5$, n=300.

The parameter of interest is

$$\theta = \frac{\beta_1}{\beta_2}.$$

Hence, the true value of θ is $\theta_0 = 2$.

We can estimate θ by

$$\hat{\theta} = \frac{\hat{\beta}_1}{\hat{\beta}_2}.$$

According to Delta Method,

$$t(\hat{\theta}) = \frac{\hat{\theta} - \theta}{S_n(\hat{\theta})} \xrightarrow{d} N(0,1)$$

where

$$S_n(\hat{\theta}) = \sqrt{n^{-1}(\hat{H}'_\beta \hat{V} \hat{H}_\beta)},$$

$$\hat{H}_\beta = \begin{pmatrix} 0 \\ 1/\hat{\beta}_2 \\ -\hat{\beta}_1/\hat{\beta}_2^2 \end{pmatrix},$$

and $\hat{V}$ is the estimated variance-covariance matrix.

The exact distribution of $t(\hat{\theta})$ can be calculated by simulation with 10000 replications. We plot the exact distribution of $t(\hat{\theta})$ accompany with the standard normal distribution in Figure 9.1. Clearly, there is a dramatic divergence between the exact and asymptotic distributions. The exact distribution is skewed and not symmetric. The probability $P(|t| > 1.96) = 0.084 = 8.4\%$, which suggests an empirical size larger than 5%. That is, the asymptotic test over-reject in finite sample. **This simple simulation result presents that even the sample size is large ($n = 300$), asymptotic approximation is poor. The bootstrap may rescue us.**

9.4 Definition of the Bootstrap

Assume data $\{x_i\}_{i=1}^n$ come from an unknown distribution function F. Let

$$T_n = T_n(x_1, \ldots, x_n, F)$$

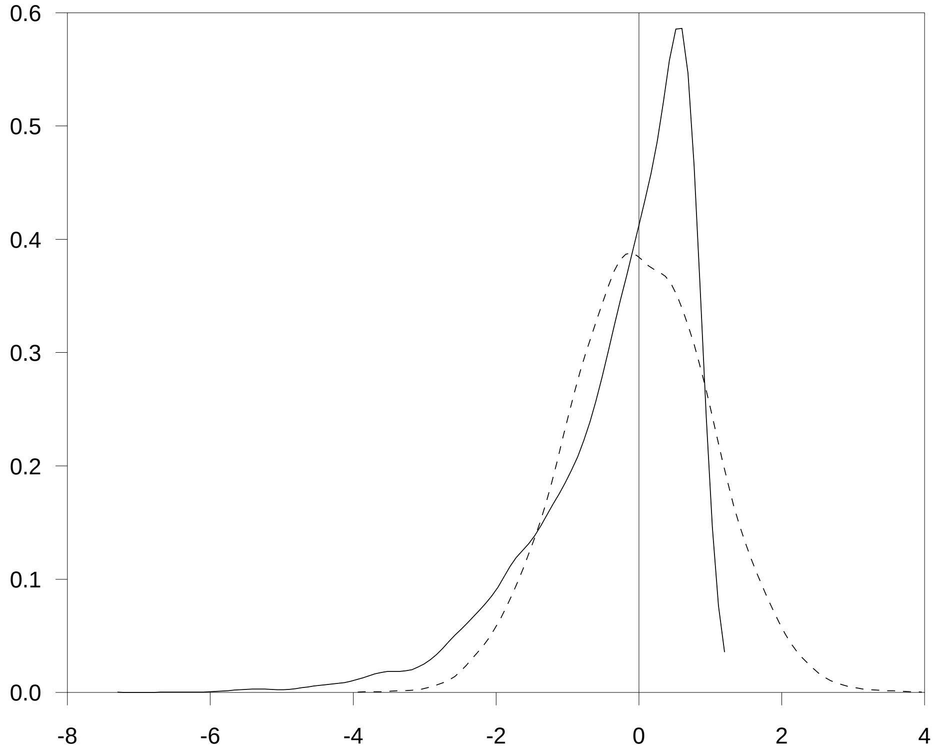


Figure 9.1: Distributions of $t(\hat{\theta})$: Exact Distribution (solid line) and Asymptotic Distribution (dashed line, $N(0,1)$)

be a statistic of interest. Note that in most cases the statistic is written as $T_n = T_n(x_1, \ldots, x_n, \theta)$, where θ is an unknown parameter. For example,

$$\begin{aligned} T_n &= \hat{\theta}, \quad \text{(estimator)} \\ T_n &= \hat{\theta} - \theta, \quad \text{(bias)} \\ T_n &= \frac{(\hat{\theta} - \theta)}{S(\hat{\theta})}, \quad \text{(t-statistic)} \end{aligned}$$

Since that the parameter θ itself is a function of F, i.e., $\theta = \theta(F)$, it is clear that

$$T_n = T_n(x_1, \ldots, x_n, \theta) = T_n(x_1, \ldots, x_n, \theta(F)) = T_n(x_1, \ldots, x_n, F).$$

Let

$$G_n(\tau, F) = P(T_n \leq \tau | F)$$

be the *exact distribution function* of T_n when the data are sampled from the distribution F. Clearly, T_n depends on $\{x_i\}_{i=1}^n$ and θ, so its distribution depends on F and θ. But $\theta = \theta(F)$, so G depends on F through two channels: $\{x_i\}_{i=1}^n$ and θ. Since the distribution function of T_n depends on F, so does each moment of T_n. For example, if $T_n = \bar{X}_n$, then $Var_F(\bar{X}_n) = \frac{\sigma^2}{n}$, where $\sigma^2 = \int (x - \mu)^2 dF(x)$ and $\mu = \int x dF(x)$. Thus, the variance of T_n is a function of F.

Remarks

- Ideally, inference would be based on exact sampling distribution, $G_n(\tau, F)$. This is generally impossible since F is unknown.

- Asymptotic inference is based on approximating $G_n(\tau, F)$ with $G_\infty(\tau, F) = \lim_{n\to\infty} G_n(\tau, F)$. When $G_\infty(\tau, F) = G_\infty(\tau)$ does not depend on F, we say that T_n is *asymptotically pivotal* and use the distribution function $G_\infty(\tau)$ for inferential purposes. For example, the limit distribution of many econometric statistics are $N(0, 1)$ or χ^2, which is independent of F or θ. In most applications, however, asymptotic pivotal statistics are not available. Moreover, even if the asymptotic pivotal statistic is available, the asymptotic approximation may be very poor as shown in the above example.

Efron (1979) proposed a different approximation: the bootstrap. The most attractive feature of the bootstrap method is that it can be used even when T_n is complicated to compute and difficult to analyze. It is not necessary for T_n to have a known asymptotic distribution.

It is proposed that first estimate F by a consistent estimate $\hat{F}_n$, and then plug F_n into $G_n(\tau, F)$ to obtain

$$G_n^*(\tau) = G_n(\tau, \hat{F}_n)$$

as an estimate of $G_n(\tau, F)$. We call $G_n^*(\tau)$ the bootstrap distribution, and the bootstrap inference is based on $G_n^*(\tau)$.

Recall that $F(x) = P(X_i \leq x) = E(1(X_i \leq x))$, where $1(\cdot)$ is the indicator function. Therefore, according to analogy principle, a natural choice of F_n is the empirical distribution function (EDF):

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x).$$

By the WLLN, for any x,

$$\hat{F}_n(x) \xrightarrow{p} F(x),$$

a consistent estimator.

Hence, under some conditions, we have

$$\lim_{n \to \infty} G_n^*(\tau) = G_n(\tau, F)$$

and

$$\lim_{n \to \infty} G_n^*(\tau) = G_\infty(\tau, F)$$

That is, the bootstrap distribution function, $G_n^*(\tau)$ is close to the finite sampling distribution of T_n: $G_n(\tau, F)$ when n is large. Since we know that the asymptotic distribution of T_n is $G_\infty(\tau, F) = \lim_{n \to \infty} G_n(\tau, F)$, the bootstrap distribution function, $G_n^*(\tau)$ is also close to the asymptotic distribution of T_n: $G_\infty(\tau, F)$ when n is large.

Here, we have used a very sloppy notations and descriptions to give you some ideas about the consistency of the bootstrap. For rigorous treatments, see Horowitz (2001). Although some unusual conditions may cause inconsistency, Horowitz (2001) suggests that the bootstrap is consistent in most applications in econometric practice.

Again, use $T_n = \bar{X}_n$ as an example, and suppose that we are interested in $Var_F(T_n)$. The bootstrap idea has two steps:

Step 1: Estimate $Var_F(T_n)$ with $Var_{\hat{F}_n}(T_n)$.

Step 2: Approximate $Var_{\hat{F}_n}(T_n)$ using Monte Carlo simulation.

9.5 Nonparametric Bootstrap

Efron (1979) proposed a Monte Carlo simulation to approximate G_n^*. The procedure is as follows.

Step 1: Draw a bootstrap sample $\{x_i^*\}_{i=1}^n$ from $\{x_i\}_{i=1}^n$ **with replacement**. Note that the bootstrap sample will necessarily have some ties and missions.

Step 2: The bootstrap statistic $T_n^* = T_n(x_1^*, \ldots, x_n^*, F_n)$ is calculated for each bootstrap sample. When the statistic T_n is a function of F, it is typically through dependence on a parameter, θ. Hence, we have the bootstrap statistic $T_n^* = T_n(x_1^*, \ldots, x_n^*, \theta_n)$. Typically, $\theta_n = \hat{\theta}$.

Step 3: Repeat Steps 1 and 2 B times and yield B values of T_{nb}^*: $\{T_{n1}^*, \ldots, T_{nB}^*\}$. Thus, the EDF of T_{nb}^* is

$$\hat{G}_n^*(\tau) = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb}^* \leq \tau).$$

As $B \to \infty$,

$$\hat{G}_n^*(\tau) \xrightarrow{p} G_n^*(\tau).$$

It is desirable for B to be large, for instance, $B = 1000$ or $B = 5000$.

Tips for Nonparametric Bootstrap Here is a practical guide to conduct resampling from $\{x_1, x_2, \ldots, x_n\}$.

1. First, we draw n random numbers v's from the uniform distribution, $U[0,1]$.

2. For each v_i, compute

$$\kappa_i = \begin{cases} round(v_i \times n) & \text{if } v_i \neq 0, \\ 1 & \text{if } v_i = 0. \end{cases}$$

   Where *round* is an operator to round to the **next** integer. Clearly, $\kappa_i \in [1, n]$.

3. Pick up the bootstrap sample x_i^* as the κ_i-th x_i.

For example, suppose $n = 10$ and v_i are

$$0.631, \quad 0.277, \quad 0.745, \quad 0.202, \quad 0.914, \quad 0.136, \quad 0.851, \quad 0.878, \quad 0.120, \quad 0.00$$

Then κ_i will be

$$7, \ 3, \ 8, \ 3, \ 10, \ 2, \ 9, \ 9, \ 2, \ 1$$

Therefore, the bootstrap sample is

$$\{x_7, \ x_3, \ x_8, \ x_3, \ x_{10}, \ x_2, \ x_9, \ x_9, \ x_2, \ x_1\}$$

Clearly, as claimed above, the bootstrap sample will necessarily have some ties (such as x_2, x_3 and x_9) and missions (such as x_4, x_5 and x_6).

Bootstrapping in Different Statistical Softwares In GAUSS, you need to write your own bootstrapping program following the procedure mentioned above. For RATS, an instruction called `BOOT` is available. STATA provides a more sophisticated command: `bootstrap`.

Example 34 (R Code for Non-Parametric Bootstrap).

```
set.seed(567812)
# I.I.D. standard normal random variables
X = rnorm(10,0,1)
X
# Bootstrap resampling
Xstar = sample(X,replace=T)
Xstar
```

9.6 Bootstrap Bias and Standard Error

9.6.1 Bootstrap Estimation of Bias

The bias of $\hat{\theta}$ is

$$\omega_n = E(\hat{\theta} - \theta).$$

Let $T_n(\theta) = \hat{\theta} - \theta$, then bias can be rewritten as

$$\omega_n = E[T_n(\theta)] = \int \tau dG_n(\tau, F).$$

The bootstrap counterpart are

$$\hat{\theta}^* = \hat{\theta}(x_1^*, \ldots, x_n^*),$$

and

$$T_n^* = \hat{\theta}^* - \hat{\theta}.$$

The bootstrap bias is

$$\omega_n^* = \int \tau dG_n^*(\tau),$$

and the simulation estimate of ω_n^* is

$$\begin{aligned}
\hat{\omega}_n^* &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\
&= \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \hat{\theta}) \\
&= \left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right) - \hat{\theta} \\
&= \overline{\hat{\theta}^*} - \hat{\theta}.
\end{aligned}$$

Given $\hat{\theta}$ is biased, the unbiased estimator of θ would be

$$\tilde{\theta} = \hat{\theta} - \omega_n,$$

so that $E(\tilde{\theta}) = E(\hat{\theta}) - E(\hat{\theta}) + \theta = \theta$. The unbiased bootstrap estimator would be

$$\begin{aligned}\tilde{\theta}^* &= \hat{\theta} - \hat{\omega}_n^* \\ &= \hat{\theta} - (\overline{\hat{\theta}^*} - \hat{\theta}) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}.\end{aligned}$$

9.6.2 Bootstrap Estimation of Variance (Standard Error)

Let $T_n = \hat{\theta}$, then variance is

$$\begin{aligned}V_n &= Var(\hat{\theta}) \\ &= Var(T_n) \\ &= E(T_n - E[T_n])^2.\end{aligned}$$

Let $T_n^* = \hat{\theta}^*$, then its variance is

$$V_n^* = E(T_n^* - E(T_n^*))^2$$

The simulation estimate of V_n^* is

$$\hat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2.$$

A bootstrap standard error is $\sqrt{\hat{V}_n^*}$.

Early work on the application of bootstrap methods merely consisted of using the bootstrap distribution to get standard errors. But as commented by Bruce Hansen:

> While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct asymptotic confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspected. It appears superior to calculate bootstrap confidence intervals.

Hence, it seems make little sense to purely replace the asymptotic standard errors with bootstrap standard errors. However, when constructing bootstrap p-values, we still need to compute the bootstrap standard errors first.

9.7 Bootstrap Confidence Interval

9.7.1 Percentile Intervals

Suppose that $G_n(\tau, F)$ is the distribution function of T_n. Let $q_n(\alpha, F)$ be its **quantile function** such that

$$\alpha = G_n(q_n(\alpha, F), F).$$

Let

$$q_n^*(\alpha) = q_n(\alpha, F_n)$$

denote the quantile function of the bootstrap distribution. Given $T_n = \hat{\theta}$ be the estimate of a parameter of interest. In $100 \cdot (1 - \alpha)\%$ CI of sample, $\hat{\theta}$ is covered by the region

$$\left[q_n\left(\frac{\alpha}{2}\right), \; q_n\left(1 - \frac{\alpha}{2}\right) \right].$$

This motivates a confidence interval for θ proposed by Efron

$$CI^* = \left[q_n^*\left(\frac{\alpha}{2}\right), \; q_n^*\left(1 - \frac{\alpha}{2}\right) \right].$$

This is often called the *percentile confidence interval.* The simulation estimate of CI is

$$\widehat{CI}^* = \left[\hat{q}_n^*\left(\frac{\alpha}{2}\right), \; \hat{q}_n^*\left(1 - \frac{\alpha}{2}\right) \right],$$

where $\hat{q}_n^*(\cdot)$ is the sample quantile of the bootstrap statistics $\{T_{n1}^*, \ldots, T_{nB}^*\}$. That is, we simulate $\{T_{n1}^*, \ldots, T_{nB}^*\}$, then sort them in ascending order. Finally, find the $B\alpha$-th T_{nb}^* as the quantile $q_n^*(\alpha)$. The interval $\widehat{CI}$ is a popular bootstrap confidence interval often used in empirical practice.

Remarks on Percentile Intervals

- For instance, with 1000 replications, a 95% interval is obtained by the 25th and 975th T_{nb}^*.
- Advantages
  1. Easy to compute.
  2. Does not require $S(\hat{\theta})$
- Disadvantages:
  1. It may perform poorly when $\hat{\theta}$ does not have symmetric distribution.

9.7.2 Percentile-t Equal-tailed Intervals

Let

$$T_n(\theta) = \frac{\hat{\theta} - \theta}{S(\hat{\theta})}.$$

Since

$$\begin{aligned}
1 - \alpha &= P\left(q_n\left(\frac{\alpha}{2}\right) \leq T_n(\theta_0) \leq q_n\left(1 - \frac{\alpha}{2}\right)\right) \\
&= P\left(q_n\left(\frac{\alpha}{2}\right) \leq \frac{\hat{\theta} - \theta_0}{S(\hat{\theta})} \leq q_n\left(1 - \frac{\alpha}{2}\right)\right) \\
&= P\left(\hat{\theta} - S(\hat{\theta})q_n\left(1 - \frac{\alpha}{2}\right) \leq \theta_0 \leq \hat{\theta} - S(\hat{\theta})q_n\left(\frac{\alpha}{2}\right)\right),
\end{aligned}$$

an exact $100 * (1 - \alpha)$ confidence interval for θ_0 would be

$$\left[\hat{\theta} - S(\hat{\theta})q_n\left(1 - \frac{\alpha}{2}\right), \quad \hat{\theta} - S(\hat{\theta})q_n\left(\frac{\alpha}{2}\right))\right].$$

This motivates a bootstrap analog

$$\widehat{CI}_t^* = \left[\hat{\theta} - S(\hat{\theta})\hat{q}_n^*\left(1 - \frac{\alpha}{2}\right), \quad \hat{\theta} - S(\hat{\theta})\hat{q}_n^*\left(\frac{\alpha}{2}\right))\right],$$

where $\hat{q}_n^*(\cdot)$ is the sample quantile of the bootstrap statistics $\{T_{n1}^*, \ldots, T_{nB}^*\}$, where

$$T_n^* = \frac{\hat{\theta}^* - \hat{\theta}}{S(\hat{\theta}^*)}$$

This is often called a *percentile-t confidence interval.* Note that unless the distribution of the T_n^* happens to be symmetric around the origin, this will not be a symmetric interval.

9.8 Bootstrap P-values (Hypothesis Testing)

9.8.1 One-sided Tests

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ at significance level α. Let

$$T_n = \frac{\hat{\theta} - \theta}{S(\hat{\theta})}$$

be the test statistic of interest. We first simulate the bootstrap distribution of

$$T_n^* = \frac{\hat{\theta}^* - \hat{\theta}}{S(\hat{\theta}^*)},$$

where $S(\hat{\theta}^*)$ is the bootstrap standard error. We then find the bootstrap critical value $q_n^*(1-\alpha)$ such that

$$P(T_n^* > q_n^*(1-\alpha)) = \alpha,$$

and reject H_0 if $T_n(\theta_0) > q_n^*(1-\alpha)$.

On the other hand, we may compute the bootstrap p-value:

$$p^* = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb}^* > T_n(\theta_0)).$$

9.8.2 Two-sided Tests

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at significance level α. Let

$$T_n = \frac{\hat{\theta} - \theta}{S(\hat{\theta})}$$

be the test statistic of interest. Again, simulate the bootstrap distribution of

$$T_n^* = \frac{\hat{\theta}^* - \hat{\theta}}{S(\hat{\theta}^*)}.$$

Sort $|T_{nb}^*|$ and find $100 \cdot (1-\alpha)\%$ quantile, $q_n^*(1-\alpha)$. Reject H_0 if

$$|T_n(\theta_0)| > q_n^*(1-\alpha).$$

The bootstrap p-value is:

$$p^* = \frac{1}{B}\sum_{b=1}^{B} 1(|T_{nb}^*| > |T_n(\theta_0)|).$$

Remarks

- Note that the bootstrap test statistic T_n^* is centered at the estimate $\hat{\theta}$, and the standard error, $S(\hat{\theta}^*)$ is calculated on the bootstrap sample. That is, $T_n^* = (\hat{\theta}^* - \hat{\theta})/S(\hat{\theta}^*)$ but NOT $(\hat{\theta}^* - \theta_0)/S(\hat{\theta}^*)$ or $(\hat{\theta}^* - \hat{\theta})/S(\hat{\theta})$.

  The guideline is proposed by Hall and Wilson (1991) and is often referred to as the **Hall and Wilson rule**. As suggested by Hansen (2006), he states "[w]hen in doubt use $\hat{\theta}$". He also emphasizes that using θ_0 rather than $\hat{\theta}$ is a "typical mistake made by practitioners".

  However, as indicated in Maddala and Kim (1998), the guideline has been violated in econometric practice (particularly in time-series econometrics) BUT with good reasons. We will talk about this later.

- The bootstrap tests are invariant to strictly monotonic transformation of the test statistic. If T_n is a test statistic, and $g(T_n)$ is a strictly monotonic function of it, then a bootstrap based on $g(T_n)$ will yield exactly the same inferences as a bootstrap test based on T_n. The reason for this is simple. The position of $T_n(\theta_0)$ in the sorted list of T^*_{nb} is exactly the same as the position of $g(T_n(\theta_0))$ in the sorted list of $g(T^*_{nb})$.

9.9 Bootstrap Methods for Regression Models

Consider the following regression model:

$$y_t = \beta x_t + \varepsilon_t$$

$$\varepsilon_t \sim^{i.i.d.} (0, \sigma^2).$$

Suppose that we are interested in testing $H_0 : \beta = \beta_0$.

We can of course use nonparametric bootstrap to sample (y, x) pairs from data randomly with replacement. It is fully nonparametric, and works in nearly any context without imposing any condition. However, it may be inefficient in contexts where more is known about F. For instance, the regression model considered above.

Therefore, we turn to another bootstrap method, which is typically called *residual bootstrap*. The procedure is as follows.

- Step 1: Estimate the regression model and obtain estimator, $\hat{\beta}$ and $\hat{\sigma}$. Then get the residuals $\hat{\varepsilon} = \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_T\}$.
- Step 2: Get bootstrap residuals, ε^* from $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_T\}$ by EITHER
  - **nonparametric method:** randomly sample from $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_T\}$ with replacement, OR
  - **parametric method:** Generate bootstrap residuals ε^* from a parametric distribution, such as $\varepsilon^*_t \sim N(0, \hat{\sigma}^2)$.
- Step 3: The bootstrap sample of regressor, x^*_t can be generated by (1) nonparametric bootstrap, (2) parametric bootstrap or simply (3) $x^*_t = x_t$.
- Step 4: Consider two sampling schemes for the generation of the bootstrap samples

$$\mathbb{S}_1 : \; y^*_t = \hat{\beta} x^*_t + \varepsilon^*_t$$

$$\mathbb{S}_2 : \; y^*_t = \beta_0 x^*_t + \varepsilon^*_t$$

Both use ε^* but they differ the way y^*_t is generated.

- Step 5: Consider two t-statistics

$$\mathbb{T}_1 : \; T_n = \frac{\hat{\beta}^* - \hat{\beta}}{S(\hat{\theta}^*)}$$

$$\mathbb{T}_2 : \; T_n = \frac{\hat{\beta}^* - \beta_0}{S(\hat{\theta}^*)}$$

Four different combinations $[\mathbb{S}_1, \mathbb{S}_2] \times [\mathbb{T}_1, \mathbb{T}_2]$ can be applied. Note that if the regressor include lagged dependent variables, for instance,

$$y_t = \beta y_{t-1} + \varepsilon_t,$$

the bootstrap DGP is implemented recursively, so that y_t^* depends on its own lagged values. You may use the unconditional mean, the pre-sample value or drawings from the unconditional distribution of y_t as the starting value. In practice, we may generate $T + R$ bootstrap sample and then discard the first R observation to avoid the effects of initial values. Typically, $R = 50$.

Remarks

1. Clearly, $\mathbb{S}_1 \times \mathbb{T}_1$ is consistent with the Hall and Wilson rule.

2. VanGiersbergen and Kiviet (1993) suggest, on the basis of a Monte Carlo study of an AR(1) model, the use of $\mathbb{S}_2 \times \mathbb{T}_2$ is better than the use of $\mathbb{S}_1 \times \mathbb{T}_1$ in finite sample. However, the limiting distributions of $\mathbb{T}_1$ under $\mathbb{S}_1$ and $\mathbb{T}_2$ under $\mathbb{S}_2$ are identical even with dynamic models. Finally, they suggest that $\mathbb{S}_1 \times \mathbb{T}_2$ or $\mathbb{S}_2 \times \mathbb{T}_1$ should not be used. That is the reason in most applications of Time-series econometrics, we do not follow the Hall and Wilson rule.

3. It is advisable to rescale the residuals so that they have correct variance:

$$\ddot{\varepsilon}_t \equiv \left(\frac{T}{T-k} \right)^{1/2} \hat{\varepsilon}_t.$$

   Then the bootstrap ε^* is resampled from $\ddot{\varepsilon}$.

4. It is clear that a residual bootstrap with nonparametric method in Step 2 is a semi-parametric bootstrap. However, in practice, if nonparametric method is used in Step 2, it is generally called the nonparametric bootstrap. On the other hand, if parametric method is used in Step 2, it is clearly a parametric bootstrap.

5. Finally, in Step 1, we obtain the *unrestricted residuals* $\hat{\varepsilon}$ from unrestricted estimation. However, MacKinnon (2006) has shown that in AR(1) model with high persistency (AR coefficient is equal to 0.9), using *restricted residuals* $\tilde{\varepsilon}$ from restricted estimation that impose the restrictions of the null hypothesis works extremely well in small sample. If sample size is large, it seems to make little difference which residuals we use.

9.10 Some Final Remarks on the Bootstrap

1. In section 9.9, we have shown how to use residual bootstrap to implement bootstrap for dependent data (time-series data). The major other way of bootstrapping dependent processes is to divide the data sequence into blocks, and resample the blocks rather than individual data values. This approach is called *block bootstrap*.

2. In most cases, bootstrap tests work better than the asymptotic tests. However, here are situations in which bootstrap tests perform badly.

   (a) Underlying residuals are serially correlated.

   (b) Underlying residuals are heteroskedastic.

   (c) Simultaneous equation models.

3. For some more surveys of the bootstrap method, see MacKinnon (2002) and MacKinnon (2006).

Chapter 10

Stochastic Process

10.1 Stochastic Process

Definition 55. A stochastic process is a sequence of random variables

$$\{X(t),\ t \in T\},$$

where T is called the index set for the process.

In many applications involving stochastic process, the index t is thought of as time. If the index for the random variables is interpreted as representing time, the stochastic process is called a **time series**.

For a fixed t, the random variable $X(t)$ has its own distribution. The outcome $X(t) = x$ is called the **state** of the stochastic process, and the state is an element of what is termed the **state space**, S. The state space could be either countable or uncountable.

If the distribution is unchanged over time, the time series is said to be **stationary** (a formal definition will be given later).

If the set T is countable ($t = 0, \pm 1, \pm 2, \pm 3, \ldots$), $X(t)$ is called a discrete time series. If the set T is uncountable ($-\infty < t < \infty$), it is called a continuous time series. Generally, we use X_t rather than $X(t)$ to denote a discrete time series so that the notation alone can often help describe a key feature of the process.

Definition 56 (Joint Distribution). Given a stochastic process $\{X_t\}_{t=1}^n$. The joint density is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_n|x_1, x_2 \ldots, x_{n-1}) f(x_{n-1}|x_1, x_2 \ldots, x_{n-2}) \cdots f(x_3|x_1, x_2) f(x_2|x_1) f(x1), \\ &= \prod_{i=1}^{n} f(x_i|\text{past}_i). \end{aligned}$$

Definition 57 (Strictly Stationary Process). A stochastic process $\{X_t\}$ is said to be strictly stationary if the joint distribution of $(X_t, X_{t-1}, \ldots, X_{t-k})$ is independent of t for all k.

For instance, the joint distribution of (X_5, X_2) is the same as that of (X_{11}, X_8). Moreover, if $\{X_t\}$ is strictly stationary, so is $\{f(X_t)\}$, where $f(\cdot)$ is continuous.

Definition 58 (Weakly Stationary Process). A stochastic process $\{X_t\}$ is said to be weakly stationary (or covariance stationary) if

1. $E(X_t^2) < \infty$.
2. $E(X_t) = \mu$ is independent of t.
3. $Cov(X_t, X_{t-k}) = \gamma(k)$ is independent of t for all k.

Where $\gamma(k)$ is called the **autocovariance function**.

Note that

1. $\gamma(0) = Cov(X_t, X_t) = Var(X_t)$.
2. $\gamma(k) = \gamma(-k)$ if $\{X_t\}$ is weakly stationary.

We can present some examples of stochastic processes. A very important class of weakly stationary process is a white noise process, a process with zero mean and no serial correlation.

Definition 59 (White Noise Process). A *weakly stationary* process $\{X_t\}$ is called a white noise process if

$$E(X_t) = 0,$$

$$Cov(X_t, X_{t-k}) = 0 \text{ for } k \neq 0.$$

Moreover, a special case of a white noise process is defined below.

Definition 60 (Independent White Noise Process). If $\{X_t\}$ is a i.i.d. random sequence with

$$E(X_t) = 0,$$

$$Var(X_t) < \infty,$$

then it is called an independent white noise process.

Definition 61 (Random Walk). Let $X_1, X_2, \ldots,$ be i.i.d. random variables with $E(X_i) = 0$ and $E(X_i^2) < \infty$. Define $S_0 = 0$ and $S_t = \sum_{i=1}^{t} X_i$ for $t \geq 1$. The the stochastic process $\{S_t\}$ is called a random walk.

Note that a random walk process can be represented as

$$S_{t+1} = S_t + X_{t+1}.$$

10.2 Martingales

Definition 62 (Martingales). A stochastic process $\{X_t\}$ is called a martingale if

$$E(X_{t+1}|X_t, X_{t-1}, \dots, X_1) = X_t.$$

Here is a theorem that is helpful in time-series forecasting and macroeconomic theory.

Theorem 72. If a stochastic process $\{X_t\}$ is a martingale and let $I_t = \{X_t, X_{t-1}, \dots, X_1\}$, then

$$E(X_{t+k}|I_t) = X_t.$$

Proof. Since $\{X_t\}$ is a martingale,

$$E(X_{t+k}|I_{t+k-1}) = E(X_{t+k}|\underbrace{X_1, X_2, \dots, X_t, X_{t+1}}_{\mathbf{Z}_1}, \underbrace{X_{t+2}, \dots, X_{t+k-1}}_{\mathbf{Z}_2}) = X_{t+k-1}.$$

That is

$$E(X_{t+k}|\mathbf{Z}_1, \mathbf{Z}_2) = X_{t+k-1}.$$

Take a condition expectation (conditional on $\mathbf{Z}_1$) on both sides,

$$E\big[E(X_{t+k}|\mathbf{Z}_1, \mathbf{Z}_2)\big|\mathbf{Z}_1\big] = E\big[X_{t+k-1}\big|\mathbf{Z}_1\big].$$

Since by SCSWR (see Theorem 13), the left hand side of the above equation is

$$E\big[E(X_{t+k}|\mathbf{Z}_1, \mathbf{Z}_2)|\mathbf{Z}_1\big] = E(X_{t+k}|\mathbf{Z}_1),$$

we thus have

$$E(X_{t+k}|\mathbf{Z}_1) = E(X_{t+k-1}|\mathbf{Z}_1).$$

Using the same argument, we can obtain

$$E(X_{t+k-1}|\mathbf{Z}_1) = E(X_{t+k-2}|\mathbf{Z}_1),$$

and so on and so forth. Therefore,

$$E(X_{t+k}|\mathbf{Z}_1) = E(X_{t+1}|\mathbf{Z}_1),$$

or

$$E(X_{t+k}|X_1, X_2, \dots, X_t) = E(X_{t+1}|X_1, X_2, \dots, X_t) = X_t.$$

Where the second equality comes from the fact that $\{X_t\}$ is a martingale.

□

Definition 63 (Martingale Difference Sequences, MDS). A sequence $\{Y_t\}$ is said to be a martingale difference sequence if

$$E(Y_{t+1}|Y_1, Y_2, \ldots, Y_t) = 0.$$

Clearly, if $\{Y_t\}$ is a MDS, then for all $j \geq 1$,

$$E(Y_t|Y_{t-j}) = E[E(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1)|Y_{t-j}] = E[0|Y_{t-j}] = 0.$$

Hence,

$$E(Y_t) = E[E(Y_t|Y_{t-j})] = 0.$$

Moreover,

$$\begin{aligned} Cov(Y_t, Y_{t-k}) &= E(Y_t Y_{t-k}), \\ &= E[E(Y_t Y_{t-k}|Y_{t-k})], \\ &= E[Y_{t-k} E(Y_t|Y_{t-k})], \\ &= E[Y_{t-k} \cdot 0] = 0. \end{aligned}$$

You should try to show the following two properties:

1. If Y_t is a martingale, then ΔY_t is a MDS.
2. If X_t is a MDS, then

$$Y_t = \sum_{i=1}^{t} X_i = X_1 + X_2 + \cdots + X_t$$

is a martingale.

10.3 Markov Process

Definition 64 (Markov Chain). Let X_t be a random variable that can assume only an integer value with state space, $S = \{1, 2, \ldots, N\}$. A (time-homogeneous) Markov chain is a discrete time and discrete state stochastic process $\{X_t,\ t = 0, 1, 2, \ldots\}$ which satisfies the condition of one-step Markov dependence. Namely,

$$P(X_t = j|X_{t-1} = i, X_{t-2} = k, \ldots, X_0 = m) = P(X_t = j|X_{t-1} = i),$$

for all $t \geq 1$ and all $\{j, i, k, \ldots, m\} \in S$. Moreover, we require that

$$P(X_t = j|X_{t-1} = i) = P(X_1 = j|X_0 = i)$$

for all t and all states $i, j \in S$.

That is, a Markov chain moves to a future state with probabilities depending only on the current state. Information on states prior to the current state do not alter the probabilities. And the condition that $P(X_t = j|X_{t-1} = i) = P(X_1 = j|X_0 = i)$ is called a time-homogeneous condition. We will omit the term "time-homogeneous" without confusion that all of the Markov chains considered in this lecture note are time-homogeneous.

Moreover, the conditional probability $P(X_t = j|X_{t-1} = i)$ is called the transition probability. and typically denoted as p_{ij}. Note that

$$P_{i1} + p_{i2} + \cdots + p_{iN} = 1.$$

The transition probabilities p_{ij} are often arranged in an $(N \times N)$ square matrix $\mathbf{P}$ known as the transition matrix of the Markov chain:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \cdots & \cdots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}$$

Note that in the transition matrix, the sum of each row must equal one and all entries are non-negative. Moreover, a state j is absorbing if $p_{jj} = 1$. Finally, for a Markov chain, the joint pdf of $\{X_1, X_2, \ldots, X_n\}$ can be simplified to

$$f(x_1, x_2, \ldots, x_n) = f(x_n|x_{n-1})f(x_{n-1}|x_{n-2})\cdots f(x_3|x_2)f(x_2|x_1)f(x_1).$$

Theorem 73 (**Chapman-Kolmogorov I**). Let $\{X_t,\ t = 0, 1, 2, \ldots\}$ be a Markov chain with state space S and define $p_{ij}^{(m.n)} = P(X_n = j|X_m = i)$ for $n > m$ and $i, j \in S$. Then

$$p_{ij}^{(m,n)} = \sum_{k \in S} p_{ik}^{(m,r)} p_{kj}^{(r,n)},$$

for $m < r < n$.

Proof. By law of total probability,

$$\begin{aligned} p_{i,j}^{(m,n)} &= P(X_n = j|X_m = i), \\ &= \sum_{k \in S} P(X_n = j, X_r = k|X_m = i). \end{aligned}$$

But

$$\begin{aligned}
&\sum_{k\in S} P(X_n = j, X_r = k | X_m = i) \\
&\quad = \sum_{k\in S} \frac{P(X_n = j, X_r = k, X_m = i)}{P(X_m = i)} \\
&\quad = \sum_{k\in S} \frac{P(X_n = j, X_r = k, X_m = i)}{P(X_r = k, X_m = i)} \frac{P(X_r = k, X_m = i)}{P(X_m = i)} \\
&\quad = \sum_{k\in S} P(X_n = j | X_r = k, X_m = i) P(X_r = k | X_m = i) \\
&\quad = \sum_{k\in S} P(X_n = j | X_r = k) P(X_r = k | X_m = i) \\
&\quad = \sum_{k\in S} p_{kj}^{(r,n)} p_{ik}^{(m,r)} \\
&\quad = \sum_{k\in S} p_{ik}^{(m,r)} p_{kj}^{(r,n)}
\end{aligned}$$

□

We now define the n-step transition probability as

$$p_{ij}^{(n)} = p_{ij}^{(t,t+n)} = p(X_{t+n} = j | X_t = i),$$

and the matrix with $p_{ij}^{(n)}$ as $\mathbf{P}^{(n)}$. Then the Chapman-Kolmogorov Theorem I can be restated as the following theorem.

Theorem 74 (Chapman-Kolmogorov II). Let $\{X_t,\ t = 0, 1, 2, \ldots\}$ be a Markov chain with state space S and define $p_{ij}^{(n)} = P(X_{t+n} = j | X_t = i)$ for $i, j \in S$. Then

$$p_{ij}^{(a+b)} = \sum_{k\in S} p_{ik}^{(a)} p_{kj}^{(b)}.$$

Proof. Simply set $m = 0$, $r = a$, and $n = a + b$. □

According to Chapman-Kolmogorov II, it is nothing more than the equation for matrix multiplication. That is, Chapman-Kolmogorov II implies that

$$\mathbf{P}^{(a+b)} = \mathbf{P}^{(a)} \mathbf{P}^{(b)}.$$

Theorem 75. Given the n-step transition matrix $\mathbf{P}^{(n)}$, we have

$$\mathbf{P}^{(n)} = \mathbf{P}^{n}.$$

Proof. By definition, $\mathbf{P}^{(1)} = \mathbf{P}$. Using Chapman-Kolmogorov II, we have

$$\mathbf{P}^{(2)} = \mathbf{P}^{(1)}\mathbf{P}^{(1)} = \mathbf{P}\mathbf{P} = \mathbf{P}^2,$$

and

$$\mathbf{P}^{(3)} = \mathbf{P}^{(2)}\mathbf{P}^{(1)} = \mathbf{P}^2\mathbf{P} = \mathbf{P}^3.$$

Continuing this way, it can be shown that

$$\mathbf{P}^{(n)} = \mathbf{P}^n.$$

□

Let $\mu_n = [\mu_n(1), \mu_n(2), \ldots, \mu_n(N)]$ be a row vector where

$$\mu_n(i) \equiv P(X_n = i)$$

is the marginal probability that the chain is in state i at time n. Then

$$\begin{aligned}\mu_n(i) &= P(X_n = i), \\ &= \sum_i P(X_n = i, X_0 = i), \\ &= \sum_i P(X_n = i|X_0 = i)P(X_0 = i), \\ &= \sum_i p_{ij}^{(n)} P(X_0 = i), \\ &= \sum_i \mu_0(i) p_{ij}^{(n)}.\end{aligned}$$

That is,

$$\mu_n = \mu_0 \mathbf{P}^n.$$

10.4 Continuous-Time Stochastic Process

Definition 65 (Winer Process). $W(r) : r \in [0,1] \mapsto \mathbb{R}$, a continuous-time stochastic process is called a Winer process or standard Brownian Motion process if the process has the following properties

1. $W(0) = 0$.
2. $W(r) \sim N(0, r)$.
3. For $0 < r_1 < r_2 < \cdots < r_k < 1$,

$$\begin{bmatrix} W(r_1) - W(0) \\ W(r_2) - W(r_1) \\ \vdots \\ W(r_k) - W(r_{k-1}) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 - r_1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & r_k - r_{k-1} \end{bmatrix}\right),$$

that is, $W(r)$ has independent increments.

10.5 Asymptotic Theory for Stochastic Process

Definition 66 (Ergodicity). A stationary time series is ergodic if

$$\gamma(k) \longrightarrow 0 \text{ as } k \longrightarrow \infty$$

Theorem 76 (Ergodic Theorem). If X_t is strictly stationary and ergodic, and $E(X_t) < \infty$, then as $T \longrightarrow \infty$,

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{p} E(X_t),$$

$$\hat{\gamma}(k) \xrightarrow{p} \gamma(k)$$

Theorem 77 (MDS-CLT). If ε_t is a strictly stationary and ergodic MDS and $E(\varepsilon_t \varepsilon_t') = \Omega < \infty$, then as $T \longrightarrow \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \varepsilon_t \xrightarrow{d} N(0, \Omega)$$

Theorem 78 (**Functional CLT).** Define a partial sum process

$$S_T(r) = \sum_{t=1}^{[Tr]} u_t,$$

where $u_t \sim^{i.i.d.} (0, \sigma^2)$, and $[Tr]$ denotes the largest integer that is less than or equal to Tr, $r \in [0, 1]$.

Then for any r,

$$\frac{1}{\sigma\sqrt{T}} S_T(r) \xrightarrow{d} W(r).$$

Note that when $r = 1$, we have the conventional CLT

$$\begin{aligned}\frac{1}{\sigma\sqrt{T}} S_T(1) &= \frac{1}{\sigma\sqrt{T}} \sum_{t=1}^{T} u_t, \\ &= \frac{\frac{\sum_t u_t}{T}}{\sqrt{\frac{\sigma^2}{T}}} \xrightarrow{d} W(1) = N(0, 1).\end{aligned}$$

Chapter 11

Mathematical Appendix

11.1 Gaussian Integral

Theorem 79 (Gaussian Integral). The following integral

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi},$$

is called the Gaussian integral.

Proof. Let

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx$$

Hence,

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy \end{aligned}$$

Let $y = r\cos\theta$ and $x = r\sin\theta$, then since $x^2 + y^2 = r^2$, and

$$\mathbf{J} = \left|\begin{array}{cc} \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \\ \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \end{array}\right| = \left|\begin{array}{cc} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{array}\right| = r,$$

we have

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r dr d\theta = 2\pi \int_0^{\infty} e^{-r^2} r dr \\ &= 2\pi \int_{-\infty}^{0} \frac{1}{2} e^s ds \quad (\text{let } s = -r^2) \\ &= \pi \int_{-\infty}^{0} e^s ds \\ &= \pi \left[e^s\right]_{-\infty}^{0} = \pi \end{aligned}$$

That is,

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

□

11.2 Gamma Function

Definition 67 (Gamma Function)**.** For every $\alpha > 0$, the Gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

The Gamma function has the following properties.

Theorem 80.

1. $\Gamma(1) = 1$
2. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
3. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 1$
4. $\Gamma(n) = (n-1)!$ for every positive integer n
5. $\int_0^{\infty} x^{\alpha-1} e^{-\xi x} dx = \left(\frac{1}{\xi}\right)^{\alpha} \Gamma(\alpha)$ for $\alpha > 0$ and $\xi > 0$

Proof.

1. By definition,

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = -e^{-x}\Big]_0^{\infty} = 1$$

2. By Gaussian Integral in Theorem 79.

3.

$$\Gamma(\alpha + 1) = \int_0^{\infty} x^{\alpha} e^{-x} dx$$

Let $u = x^{\alpha}$, $v = -e^{-x}$, then $dv = e^{-x}dx$. By integral by parts,

$$\begin{aligned}\Gamma(\alpha+1) &= \int_{0}^{\infty} udv \\ &= uv]_{0}^{\infty} - \int_{0}^{\infty} vdu \\ &= x^{\alpha}(-e^{-x})]_{0}^{\infty} - \int_{0}^{\infty} -e^{-x}\alpha x^{\alpha-1}dx \\ &= -x^{\alpha}e^{-x}]_{0}^{\infty} + \alpha\int_{0}^{\infty} x^{\alpha-1}e^{-x}dx \\ &= 0 + \alpha\Gamma(\alpha) \\ &= \alpha\Gamma(\alpha).\end{aligned}$$

Where $-x^{\alpha}e^{-x}]_{0}^{\infty} = 0$ comes from:

$$\begin{aligned}\lim_{x\to\infty}\left[\frac{x^{\alpha}}{e^{x}}\right] &= \lim_{x\to\infty}\left[\frac{e^{\alpha\log x}}{e^{x}}\right] \\ &= \lim_{x\to\infty}\left[e^{\alpha\log x - x}\right] \\ &= \lim_{x\to\infty}\left[e^{x\left[\alpha\frac{\log x}{x}-1\right]}\right]\end{aligned}$$

According to L'Hôpital's Rule,

$$\lim_{x\to\infty}\frac{\log x}{x} = \lim_{x\to\infty}\frac{\frac{1}{x}}{1} = 0$$

Therefore,

$$\lim_{x\to\infty}\left[\alpha\frac{\log x}{x} - 1\right] = -1$$

and

$$\lim_{x\to\infty}\left[e^{x\left[\alpha\frac{\log x}{x}-1\right]}\right] = 0$$

4. Accordingly,

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) \\ &= (n-1)(n-2)\cdots 1\Gamma(1) \\ &= (n-1)!\end{aligned}$$

5. For $\xi > 0$, let $y = \xi x$, then $dy = \xi dx$, and

$$\begin{aligned}\int_{0}^{\infty} x^{\alpha-1}e^{-\xi x}dx &= \int_{0}^{\infty}\left(\frac{y}{\xi}\right)^{\alpha-1} e^{-y}\left(\frac{1}{\xi}\right)dy \\ &= \xi^{-\alpha}\int_{0}^{\infty} y^{\alpha-1}e^{-y}dy \\ &= \left(\frac{1}{\xi}\right)^{\alpha}\Gamma(\alpha)\end{aligned}$$

□

Chapter 12

Answers to Exercises

1 1. $\left(\bigcap_{k=1}^{n} A_k\right)^c = \bigcup_{k=1}^{n} A_k^c$.

(a) First we would like to prove that

$$\left(\bigcap_{k=1}^{n} A_k\right)^c \subseteq \left(\bigcup_{k=1}^{n} A_k^c\right).$$

Suppose not,

$$\exists x \in \left(\bigcap_{k=1}^{n} A_k\right)^c \text{ but } \quad x \notin \left(\bigcup_{k=1}^{n} A_k^c\right)$$

$$\begin{aligned} \text{Since} \quad & x \notin \left(\bigcup_{k=1}^{n} A_k^c\right) \\ & \Rightarrow x \notin A_1^c \quad \text{and} \quad x \notin A_2^c \quad \text{and} \quad \cdots \quad \text{and} \quad x \notin A_n^c \\ & \Rightarrow x \in A_1 \quad \text{and} \quad x \in A_2 \quad \text{and} \quad \cdots \quad \text{and} \quad x \in A_n \\ & \Rightarrow x \in \bigcap_{k=1}^{n} A_k \\ & \Rightarrow x \notin \left(\bigcap_{k=1}^{n} A_k\right)^c \text{ contradiction} \end{aligned}$$

(b) We then prove that

$$\left(\bigcap_{k=1}^{n} A_k\right)^c \supseteq \left(\bigcup_{k=1}^{n} A_k^c\right).$$

Suppose not,

$$\exists x \in \left(\bigcup_{k=1}^{n} A_k^c\right) \quad \text{but} \quad x \notin \left(\bigcap_{k=1}^{n} A_k^c\right)$$

$$\begin{aligned}
\text{Since}\quad & x \notin \left(\bigcap_{k=1}^{n} A_k^c\right)\\
& \Rightarrow x \notin A_1^c \quad \text{or} \quad x \notin A_2^c \quad \text{or} \quad \cdots \quad \text{or} \quad x \notin A_n^c\\
& \Rightarrow x \in A_1 \quad \text{or} \quad x \in A_2 \quad \text{or} \quad \cdots \quad \text{or} \quad x \in A_n\\
& \Rightarrow x \in \bigcup_{k=1}^{n} A_k\\
& \Rightarrow x \notin \left(\bigcup_{k=1}^{n} A_k\right)^c \text{contradiction}
\end{aligned}$$

From (a) and (b), we have $(\bigcap_{k=1}^{n} A_k)^c = \bigcup_{k=1}^{n} A_k^c$.

2. $(\bigcup_{k=1}^{n} A_k)^c = \bigcap_{k=1}^{n} A_k^c$.

   (a) First we prove that

$$\left(\bigcup_{k=1}^{n} A_k\right)^c \subseteq \bigcap_{k=1}^{n} A_k^c.$$

   Suppose not,

$$\exists x \in \left(\bigcup_{k=1}^{n} A_k\right)^c \quad \text{but} \quad x \notin \bigcap_{k=1}^{n} A_k^c$$

$$\begin{aligned}
\text{Since}\quad & x \notin \bigcap_{k=1}^{n} A_k^c\\
& \Rightarrow x \notin A_1^c \quad \text{or} \quad x \notin A_2^c \quad \text{or} \quad \cdots \quad \text{or} \quad x \notin A_n^c\\
& \Rightarrow x \in A_1 \quad \text{or} \quad x \in A_2 \quad \text{or} \quad \cdots \quad \text{or} \quad x \in A_n\\
& \Rightarrow x \in \bigcup_{k=1}^{n} A_k\\
& \Rightarrow x \notin \left(\bigcup_{k=1}^{n} A_k\right)^c \text{contradiction}
\end{aligned}$$

   (b) We then prove that

$$\left(\bigcup_{k=1}^{n} A_k\right)^c \supseteq \bigcap_{k=1}^{n} A_k^c.$$

   Suppose not,

$$\exists x \in \bigcap_{k=1}^{n} A_k^c \quad \text{but} \quad x \notin \left(\bigcup_{k=1}^{n} A_k\right)^c$$

$$
\begin{aligned}
\text{Since}\quad & x \notin \left(\bigcup_{k=1}^{n} A_k\right)^c \\
& \Rightarrow x \in \bigcup_{k=1}^{n} A_k \\
& \Rightarrow x \in A_1 \quad \text{or} \quad x \in A_2 \quad \text{or} \quad \cdots \quad \text{or} \quad x \in A_n \\
& \Rightarrow x \notin A_1^c \quad \text{or} \quad x \notin A_2^c \quad \text{or} \quad \cdots \quad \text{or} \quad x \notin A_n^c \\
& \Rightarrow x \notin \bigcap_{k=1}^{n} A_k^c \quad \text{contradiction}
\end{aligned}
$$

From (a) and (b), we have $(\bigcup_{k=1}^{n} A_k)^c = \bigcap_{k=1}^{n} A_k^c$.

See Pages 4–5 in Roussas (2002) for an alternative proof.

2

1. Check if $F(x)$ is indeed a CDF.
   (a) $F(x) \geq 0, \forall x \in \mathbb{R}$.
   (b) $F(\infty) = pI_{\{\infty \geq 0\}} + (1-p)\Phi(\infty) = p + (1-p) = 1$, and $F(-\infty) = 0$.
   (c) $F(x)$ is increasing.
   (d) $F(x)$ is right continuous.
2. Plot $F(x)$. See Figure 12.1.
3. Find out the pdf $f(x)$.

$$
\begin{aligned}
f(x) &= \frac{d}{dx}F(x), \\
&= p\frac{d}{dx}I_{\{x\geq 0\}} + (1-p)\frac{d}{dx}\Phi(x), \\
&= p\delta(x) + (1-p)\phi(x),
\end{aligned}
$$

where $\delta(x) = 0$ if $x \neq 0$; $\delta(x) = \infty$ if $x = 0$.

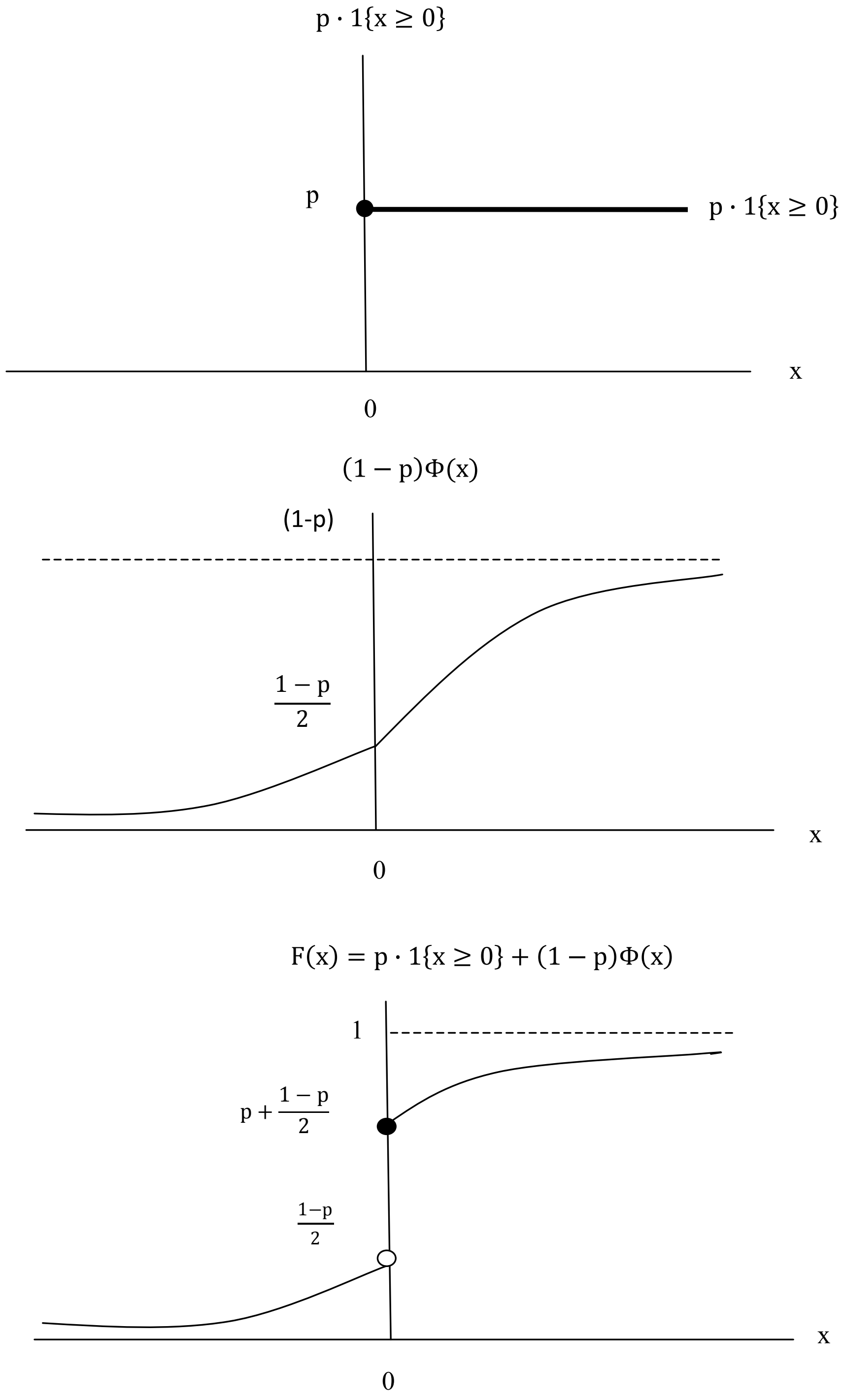
3

$$
\begin{aligned}
P(X \leq x | X \geq a) &= \frac{P(X \leq x, X \geq a)}{P(X \geq a)}, \\
&= \frac{P(a \leq X \leq x)}{P(X \geq a)}, \\
&= \begin{cases} 0 & \text{if } x < a, \\ \frac{F(x)-F(a)}{1-F(a)} & \text{if } x \geq a. \end{cases}
\end{aligned}
$$

Therefore,

$$
g_{X|X\geq a}(x) = \frac{d}{dx}P(X \leq x | X \geq a) = \begin{cases} 0 & \text{if } x < a, \\ \frac{f(x)}{1-F(a)} & \text{if } x \geq a. \end{cases} .
$$

Figure 12.1: Mixed Distribution


$p \cdot 1\{x \geq 0\}$
p
$p \cdot 1\{x \geq 0\}$
x
0
$(1-p)\Phi(x)$
(1-p)
$\frac{1-p}{2}$
x
0
$F(x) = p \cdot 1\{x \geq 0\} + (1-p)\Phi(x)$
1
$p + \frac{1-p}{2}$
$\frac{1-p}{2}$
x
0