Review Notes for EC770

Little Tiger

4th Dec 07

Contents

Preface

This is note for probability Theory based on classroom notes.

Introduction

This is intended to be a complete review notes including all statistics courses in Ph.D exams: Statistics, Econometrics, Time Series and Cross section Analysis. For now, only statistics is included.

Part I

EC770 Statistics

Chapter 1

Probability Theory

1.1 Lecture 1

Sample Space Ω: the collection of all possible outocme of an experiment
Event E: A collection of outcome
Experiment: a process that could be repeated

Set Theory

Terminology: empty set $\varnothing$, finite and infinte sets, disjoint sets.
Operation of Set: union $\cup$, intersection $\cap$, complement ^{C}, difference $-$, subset $\subseteq$, element of a set $\in$.

Property of Set Operation

1. Commutative: $A \cup B = B \cup A$, $A \cap B = B \cap A$

2. Associative: $A \cup (B \cup C) = (A \cup B) \cup C$, $(A \cap B) \cap C = A \cap (B \cap C)$

3. Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cup (B \cap C) = (A \cup B) \cap (A \cap C)$

4. DeMorgan's Law: $(A \cap B)^C = A^C \cup B^C$, $(A \cup B)^C = A^C \cap B^C$

5. Reflexive Law: $A \cup A = A$, $A \cap A = A$

Collection of sets

1. Partition: Sets A_i $(i = 1, 2, \ldots, n)$ where $A_i \cap A_j = \varnothing$ $(i \neq j)$ are paritition of S if $S = \bigcup_{i=1}^{n} A_i$. Note that n can be infinite.

2. Field: closed under finite union and closed under complementation. Field $\mathcal{A}$ is a collection of sets $A_1, \ldots, A_n$ satisfying the following properties:

   (a) If $A_1, A_2, \dots, A_n \in \mathcal{A}$, then $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$.
   (b) If $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.

3. $\sigma-$field: closed under countable unition and closed under complementation. Field $\mathcal{A}$ is a collection of sets $A_1, A_2, \dots$ satisfying the following properties:

   (a) If $A_1, A_2, \dots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.
   (b) If $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.

   If $\mathcal{A}_1, \mathcal{A}_2$ are σ-fields, $\mathcal{A}_1 \cap \mathcal{A}_2$ is also σ-field. In general, $\mathcal{A}_1 \cup \mathcal{A}_2$ is not necessary a σ-field. However, there exists a unique smallest σ-field that contains all elements of $\mathcal{A}_1$ and $\mathcal{A}_2$.

1.2 Lecture 2

Definition of Probability

(Kolmogorov Axiom of probability) Given a sample space Ω and an associated σ-field that contains subsets of Ω, a probability defined as P, is a function (of sets) from $\mathcal{A}$ to $[0, 1]$ satisfying the following properties:

1. $\forall A \in \mathcal{A}$, $P(A) \geq 0$

2. $P(\Omega) = 1$

3. Additive Property: For disjoint events $A_1, A_2, \dots \in \mathcal{A}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Therefore, probability space is defined as $(\Omega, \mathcal{A}, P)$

Property of Probability

1. If $\cup_{i=1}^{n} A_i = \Omega$ and $A_1, \dots A_n$ is partition of Ω, then $\sum_{i=1}^{n} P(A_i) = 1$.

   Proof: Direct application of additive property. $P(\Omega) = P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) = 1$

2. $P\left(A^C\right) = 1 - P(A)$

   Proof: special case of property 1 when $A_1 = A$ and $A_2 = A^C$.

3. $P(\varnothing) = 0$

   Proof: special case of property 2 when $A_1 = \Omega$ and $A_2 = \varnothing$

4. If $A \subset B$, then $P(A) \leq P(B)$.

   Proof: From non-negativity of probability.

$$\begin{aligned} B &= B \cap \Omega = B \cap \left(A \cup A^C\right) \\ &= (B \cap A) \cup \left(B \cap A^C\right) \\ &= A \cup \left(B \cap A^C\right) \end{aligned}$$

   so we have

$$\begin{aligned} P(B) &= P\left(A \cup \left(B \cap A^C\right)\right) \\ &= P(A) + P\left(B \cap A^C\right) \\ &\geq P(A) \end{aligned}$$

5. $0 \leq P(A) \leq 1$

   Proof: Monotonicty of probability. Non-negativity is from $P(A) \geq 0$ and bounded by 1 is by second and third axiom.

6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   Proof:

$$\begin{aligned} A \cup B &= A \cup (\Omega \cap B) = A \cup \left[\left(A \cup A^C\right) \cap B\right] \\ &= A \cup \left[(A \cap B) \cup \left(A^C \cap B\right)\right] \\ &= \left[A \cup (A \cap B)\right] \cup \left[A \cup \left(A^C \cap B\right)\right] \\ &= A \cup \left[A \cup \left(A^C \cap B\right)\right] = A \cup \left(A^C \cap B\right) \\ &= (A \cap \Omega) \cup \left(A^C \cap B\right) = \left[A \cap \left(B \cup B^C\right)\right] \cup \left(A^C \cap B\right) \\ &= (A \cap B) \cup \left(A \cap B^C\right) \cup \left(A^C \cap B\right) \end{aligned}$$

   so that

$$P(A \cup B) = P(A \cap B) + P\left(A \cap B^C\right) + P\left(A^C \cap B\right)$$

   However, we know that

$$\begin{aligned} P(A) &= P(A \cap B) + P\left(A \cap B^C\right) \\ P(B) &= P(A \cap B) + P\left(A^C \cap B\right) \end{aligned}$$

   Hence,

$$\begin{aligned} P(A \cup B) &= P(A \cap B) + [P(A) - P(A \cap B)] + [P(B) - P(A \cap B)] \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Inequalities of Probability

1. Bonferroni's Inequlity:

$$P(A \cap B) \geq P(A) + P(B) - 1$$

   Proof: $P(A \cup B) = P(A) + P(B) - P(A \cap B) \geq P(A) + P(B) - 1$ since $0 \leq P(A \cap B) \leq 1$

2. Bode's inequality:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

   Proof: Omitted.

3. Extenion of Property 6:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - [P(A \cap B) + P(A \cap C) + P(B \cap C)] \\ &\quad + P(A \cap B \cap C) \end{aligned}$$

   or more generally,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \cdots + (-1)^{q+1} \sum_{i_1<\ldots<i_q} P\left(\bigcap_{i=1}^{q} A_{i_j}\right) + \cdots$$

Probability for limit of sequence of events

1. Let A_i be an increasing sequence of events ($A_i \subseteq A_{i+1}$). Define $A = \lim_{i\to\infty} A_i = \bigcup_{i=1}^{\infty} A_i$. Then $P(A) = \lim_{i\to\infty} P(A_i)$.

   Proof: Define $A_x = (-\infty, x]$ which is an increasing sequence set in x as $A_y \subseteq A_z$ if $y \leq z$. Hence, by property 4,

$$P(X \in A_y) \leq P(X \in A_z)$$

   Define $D_k = A_k - A_{k+1}$, $A_0 = \varnothing$ and $D_1 = A_1$, then D_k is sequence of disjoint sets. By construction, we have

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} D_i; \quad \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} D_i$$

   so that

$$\begin{aligned} P\left(\lim_{i\to\infty} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} D_i\right) = \sum_{i=1}^{\infty} P(D_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(D_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_k - A_{k+1}) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} [P(A_k) - P(A_{k+1})] = \lim_{n\to\infty} P(A_n) \end{aligned}$$

2. Let A_i be a decreasing sequence of events ($A_i \supseteq A_{i+1}$). Define $A = \lim_{i\to\infty} A_i = \bigcap_{i=1}^{\infty} A_i$. Then $P(A) = \lim_{i\to\infty} P(A_i)$.

   Proof: Similar.

1.3 Lecture 3

Conditional probability

Definition: If $P(A) > 0$, then the conditional probability $P(B \mid A)$ is defined as the proportion of total probability $P(A)$ that is represented by the probability of $P(A \cap B)$. In formula,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

Multiplicative Rule

From the definition of conditional probability, we have

$$P(A \cap B) = P(B \mid A)\, P(A)$$

or generally, for finite k,

$$P\left(\bigcap_{i=1}^{k} A_i\right) = P(A_1)\, P(A_2 \mid A_1) \cdots P\left(A_k \mid \bigcap_{i=1}^{k-1} A_i\right)$$

Independent Event

Event A and B are independent if

$$P(A \cap B) = P(A)\, P(B)$$

or if $P(A) > 0$ and $P(B) > 0$, then A and B are independent if

$$P(A \mid B) = P(A) \text{ and } P(B \mid A) = P(B)$$

For k events, $A_1, \ldots, A_k$ are independent iff for all $j = 2, \ldots, k$ and $(i_1, \ldots, i_j) \subseteq (1, \ldots, k)$

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P\left(A_{i_j}\right)$$

Conditional Independence

A, B, C are events in probability space with $P(A) > 0$, then B and C are conditionally independent given A iff

$$P(B \cap C \mid A) = P(B \mid A)\, P(C \mid A)$$

Law of Total Probability

Consider A_i be the partition of Ω, that is, $\sum_{i=1}^{k} P(A_i) = 1$, $A_i \cap A_j = \varnothing$, $\bigcup_{i=1}^{k} A_i = \Omega$. For any event $B \in \Omega$, we have

$$\begin{aligned} P(B) &= P\left(\bigcup_{i=1}^{k} (B \cup A_i)\right) = \sum_{i=1}^{k} P(B \cup A_i) \\ &= \sum_{i=1}^{k} P(A_i)\, P(B \mid A_i) \end{aligned}$$

Bayes Theorem

Let event $B_1, \ldots, B_k$ be partition of Ω, $P(B_i) > 0$ and event A in the same space with $P(A) > 0$, then

$$P(B_j \mid A) = \frac{P(B_j) P(A \mid B_j)}{\sum_{i=1}^{k} P(B_i) P(A \mid B_i)}$$

1.4 Lecture 4

Random Variables

A random variable X is a function $X(\omega) : \Omega \to \mathbb{R}$ such that $\forall B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{F}$ where $\mathcal{F}$ is σ-field in original sample space and $\mathcal{B}$ is the Borel σ-field which is the smallest σ-field that contains all open intervals of $\mathbb{R}$.
Random variable tranforms the probability space from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B}, P_X)$

Distributive Function of a random variable

Also known as cumulative distribution function, or c.d.f.. In symbol, it is

$$F(x) = P_X(X \leq x) = P_X(X \in (-\infty, x]) \quad \forall x \in \mathbb{R}$$

Properties of c.d.f.

1. $0 \leq F(x) \leq 1$

   Proof: $0 \leq P(x) \leq 1$

2. $F(x)$ is non-decreasing in x.

   Proof: $A_x = (-\infty, x]$ is an increasing sequence set in x, so $A_x \subseteq A_y$ if $x \leq y$. Hence,
   $$P(X \in A_x) \leq P(X \in A_y)$$

3. Define $F(-\infty) = \lim_{n\to\infty} F(-n)$. $F(-\infty) = 0$

   Proof: Taking $A_n = (-\infty, n]$,
   $$\begin{aligned} F(-\infty) &= \lim_{n\to-\infty} P(X \in A_n) = P\left(X \in \lim_{n\to-\infty} A_n\right) \\ &= P(X \in A_{-\infty}) = P(X \in \varnothing) = 0 \end{aligned}$$

4. Define $F(\infty) = \lim_{n\to\infty} F(n)$. $F(\infty) = 1$

   Proof: Similar.

5. $F(x)$ is right continuous, that is
   $$\lim_{\varepsilon \searrow 0} F(x + \varepsilon) = F(x)$$

Proof: let $A_n = \left(-\infty, y + \frac{1}{n}\right]$ which converge to $A_\infty = (-\infty, y]$. Then

$$\begin{aligned}\lim_{\varepsilon\to 0} F(x+\varepsilon) &= \lim_{n\to\infty} P(X \in A_n) = P\left(X \in \lim_{n\to\infty} A_n\right) \\ &= P(X \in (-\infty, y]) = F(x)\end{aligned}$$

1.5 Lecture 5

Continuous Random Variable

If $F(x)$ is differentiable, we consider the derivative of distribution function $F(x)$ and define it as the probability density function:

$$f(x) = \frac{dF(x)}{dx}$$

which implies

$$F(x) = \int_{-\infty}^{x} f(u)\, du$$

with following properties:

1. $f(x) \geq 0$

   Proof: $F(x)$ is non-negative for all x.

2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$

   Proof: $F(\infty) = 1$

3. $P(a \leq x \leq b) = \int_b^a f(x)\, dx$

   Proof: $P(x \leq b) - P(x \leq a) = F(b) - F(a) = \int_{-\infty}^{b} f(u)\, du - \int_{-\infty}^{b} f(u)\, du = \int_b^a f(x)\, dx$

Discrete Random Variable

Random variable X takes value on discrete points $x_1, \ldots, x_n$ with $P(X = x_j) = p_j$. The probability function is $f(x_j) = p_j$ with the following properties:

1. $f(x_j) \geq 0$

2. $\sum_j f(x_j) = 1$

3. $F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i)$

Distribution of Two Random Variables

Joint Distribution Function of X, Y

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

Marginal distribution function of X and Y

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X \leq x, Y < \infty) = F(x, \infty) \\ F_Y(y) &= P(Y \leq y) = P(X < \infty, Y \leq y) = F(\infty, y) \end{aligned}$$

Properties of Joint Distribution Function

1. $0 \leq F_X(x) \leq 1$, $0 \leq F_Y(y) \leq 1$

   Proof: Any probability value is bounded by 0 and 1

2. $F(x,y)$ is non-decreasing. That is, for all $a, b > 0$,

   $$F(x+a, y+b) - F(x+a, y) - F(x, y+b) + F(x,y) \geq 0$$

   Proof: Simple Graphic would justify it.

3. $F(-\infty, y) = 0$ and $F(x, -\infty) = 0$

4. $F(\infty, \infty) = 1$

5. $F(x,y)$ is right continuous, that is,

   $$\lim_{\varepsilon_1 \downarrow 0, \varepsilon_2 \downarrow 0} F(x+\varepsilon_1, y+\varepsilon_2) = \lim_{\varepsilon_1 \downarrow 0} F(x+\varepsilon_1, y) = \lim_{\varepsilon_2 \downarrow 0} F(x, y+\varepsilon_2) = F(x,y)$$

Conditional Independence

Events A and B are independent if for all Borel set $A, B \in \mathcal{B}$,

$$P(X \in A \mid Y \in B) = \frac{P(X \in A, Y \in B)}{P(Y \in B)}$$

or

$$\begin{aligned} P_{X|Y}(A \mid B) &= P_X(A) \quad \forall A, B \in \mathcal{B} \\ P_{X \cap Y}(A \cap B) &= P_X(A) P_Y(B) \quad \forall A, B \in \mathcal{B} \end{aligned}$$

Independence Condition from Marginal Distribution

Given existence of $F(x,y), F_X(x), F_Y(y)$, we have

$$F(x,y) = F_X(x) F_Y(y) \quad \forall x, y$$

if and only if X and Y are independent.

Proof: "Only if": We need to show for all $(a_1, b_1]$ and $(a_2, b_2]$, we have

$$P\left((a_1, b_1] \cap (a_2, b_2]\right) = P_X\left((a_1, b_1]\right) \cdot P_Y\left((a_2, b_2]\right)$$

Now, we have

$$\begin{aligned}
P\left((a_1, b_1] \cap (a_2, b_2]\right) &= P\left(a_1 < x \le b_1, a_2 < y \le b_2\right) \\
&= P\left(X \le b_1, Y \le b_2\right) - P\left(X < a_1, Y \le b_2\right) - P\left(X \le b_1, Y < a_2\right) + P\left(X \le a_1, Y \le a_2\right) \\
&= F\left(b_1, b_2\right) - F\left(a_1, b_2\right) - F\left(b_1, a_2\right) + F\left(a_1, a_2\right) \\
&= F_X\left(b_1\right) F_Y\left(b_2\right) - F_X\left(a_1\right) F_Y\left(b_2\right) - F_X\left(b_1\right) F_Y\left(a_2\right) + F_X\left(a_1\right) F_Y\left(a_2\right) \\
&= F_X\left(b_1\right)\left[F_Y\left(b_2\right) - F_Y\left(a_2\right)\right] - F_X\left(a_1\right)\left[F_Y\left(b_2\right) - F_Y\left(a_2\right)\right] \\
&= \left[F_X\left(b_1\right) - F_X\left(a_1\right)\right]\left[F_Y\left(b_2\right) - F_Y\left(a_2\right)\right] \\
&= P_X\left((a_1, b_1]\right) \cdot P_Y\left((a_2, b_2]\right)
\end{aligned}$$

"If": Since X and Y are independent, taking $A = (-\infty, x], B = (-\infty, y]$, we have

$$P\left((a_1, b_1] \cap (a_2, b_2]\right) = P_X\left((a_1, b_1]\right) \cdot P_Y\left((a_2, b_2]\right)$$

so that

$$F(x, y) = F_X(x) F_Y(y) \quad \forall x, y$$

1.6 Lecture 6

Joint Probability Function

Joint probability function is denoted as $f(x_i, y_j) = P(X = x_i, Y = y_j) = p_{ij}$
Marginal probability function for X: $P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} = p_{i\cdot}$
Marginal probability function for Y: $P(Y = y_i) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} = p_{\cdot j}$
Conditional probability funciton:

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}$$

Joint Probability Density Function

If $F(x)$ is differentiable, then

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

is called joint density function with following properties (state without proof):

1. $f(x, y) \ge 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx dy = 1$
3. $\int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, du dv = F(x, y)$

Marginal Density Fucntion

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy; \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Marginal distribution Fucntion

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du; \quad F_Y(y) = \int_{-\infty}^{y} f_Y(v)\,dv$$

Conditional Distribution Fucntion

$$\begin{aligned}
P(X \le x \mid Y = y) &= F(x \mid y) \\
&= \lim_{\delta \to 0} P(X \le x \mid y - \delta \le Y \le y + \delta) \\
&= \lim_{\delta \to 0} \frac{P(X \le x, y - \delta \le Y \le y + \delta)}{P(y - \delta \le Y \le y + \delta)} \frac{2\delta}{2\delta} \\
&= \lim_{\delta \to 0} \frac{\int_{-\infty}^{x} \int_{y-\delta}^{y+\delta} f(u,v)\,dvdu}{\int_{y-\delta}^{y+\delta} f_Y(v)\,dv} \frac{2\delta}{2\delta} \\
&= \lim_{\delta \to 0} \frac{\int_{-\infty}^{x} \dfrac{F(u, y+\delta) - F(u, y+\delta)}{2\delta} du}{\dfrac{F_Y(y+\delta) - F_Y(y-\delta)}{2\delta}} \\
&= \frac{\int_{-\infty}^{x} f(u,y)\,du}{f_Y(y)}
\end{aligned}$$

Conditional Probability Density Fucntion

$$f(x \mid y) = \frac{\partial F(x,y)}{\partial y} = \frac{f(x,y)}{f(y)}$$

Generalized Multiplicative Rule

$$f(x,y) = f_X(x)\, f_Y(y \mid x) = f_Y(y)\, f_X(x \mid y)$$

Generalized Law of Total Probability

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{-\infty}^{\infty} f_X(x \mid y)\, f_Y(y)\,dy$$

Generalzied Bayes Theorem

$$f(x \mid y) = \frac{f_X(x)\, f_Y(y \mid x)}{\int_{-\infty}^{\infty} f_X(x \mid y)\, f_Y(y)\,dy}$$

Joint Distribution for k random variable

$$F(X_1, \ldots, X_k) = P(X_1 \le x_1, \ldots, X_n \le x_n)$$

For discrete case, probability distribution is completely determined by

$$P(X_1 = x_1, \ldots, X_k = x_k)$$

For continuous case,

$$f(x_1, \ldots, x_k) = \frac{\partial^k F(x_1, \ldots, x_k)}{\partial x_1 \cdots \partial x_k}$$

1.7 Lecture 7

Function of Random Variable

A function of a random variable $g(X)$ is also a random variable as long as the mapping is mearsurable, that is,

$$\forall B \in \mathcal{B}, \quad P(g(X) \in B)$$

is defined if $X^{-1}\left(g^{-1}(B)\right) \in \mathcal{F}$.

Transformation of Discrete random variable

Discrete Case: For $Y = r(X)$,

$$P(Y = y) = P(r(X) = y) = \sum\nolimits_{i: r(x_i) = y} P(X = x_i)$$

Transformation of Continuous random variable

Method I (Inversion method): For $Y = r(X)$,

$$P(Y \le y) = P(r(X) \le y) = \int_{x: r(x) \le y} f(x)\, dx$$

Method II (Jacobian transformation):
Thm: Suppose $P(a < X < b) = 1$, $Y = r(X)$ is a continuous and monotone function for $a < X < b$. When $a < X < b$, $\alpha < Y < \beta$. Then the p.d.f. of Y is given by

$$g(y) = f\left(r^{-1}(y)\right)\left|\frac{dx}{dy}\right|$$

for all $\alpha < y < \beta$ where $f(\cdot)$ is pdf of X. In particular,
if $r(\cdot)$ is an increasing fucntion, $g(y) = f\left(r^{-1}(y)\right)\frac{dx}{dy}$.
if $r(\cdot)$ is a decreasing fucntion, $g(y) = -f\left(r^{-1}(y)\right)\frac{dx}{dy}$.

Proof: If r is increasing, then

$$G(y) = P(Y \leq y) = P(r(X) \leq y) = P\left(X \leq r^{-1}(y)\right) = F\left(r^{-1}(y)\right)$$

which implies

$$g(y) = f\left(r^{-1}(y)\right) \frac{dx}{dy}$$

Similarly, if r is decreasing,

$$G(y) = P(Y \leq y) = P(r(X) \leq y) = P\left(X \geq r^{-1}(y)\right) = 1 - F\left(r^{-1}(y)\right)$$

which implies

$$g(y) = -f\left(r^{-1}(y)\right) \frac{dx}{dy}$$

Transformation of Several random variables

$X_1, \ldots, X_n$ are random variables and the new random variables $Y_1, \ldots, Y_m$ are represented by function

$$\begin{cases} Y_1 = r_1(X_1, \ldots, X_n) \\ \vdots \\ Y_m = r_m(X_1, \ldots, X_n) \end{cases}$$

Discrete Case:

$$P(Y_1 = y_1, \ldots Y_n = y_n) = \sum_{x=(x_1,\ldots,x_n); r_j(x)=y_j} P(X = x_j)$$

Continuous Case:
Method I (Inversion method)

$$\begin{aligned} G(y_1, \ldots, y_m) &= P(Y_1 \leq y_1, \ldots, Y_m \leq y_m) \\ &= \int \cdots \int_{r_j(x_1,\ldots,x_n) \leq y_j; j=1,\ldots,m} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \end{aligned}$$

Method II (Jacobian matrix)
For case of one-to-one mapping ($m = n$)
If $P((x_1, \ldots, x_n) \in S) = 1$ and there is a one-to-one correspondence between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, $Y_j = r_j(X_1, \ldots, X_n)$, then the joint p.d.f. of $(Y_1, \ldots, Y_n)$ can be obtained from

$$g(y_1, \ldots, y_n) = f(x_1(y), \ldots, x_n(y)) \left| \frac{dx}{dy} \right|$$

where

$$\left| \frac{dx}{dy} \right| = \begin{bmatrix} \frac{dx_1}{dy_1} & \cdots & \frac{dx_1}{dy_n} \\ \vdots & \ddots & \vdots \\ \frac{dx_n}{dy_1} & \cdots & \frac{dx_n}{dy_n} \end{bmatrix}$$

is called the Jacobian of Transformation.
Speical Case I: r is linear combination of $X_1, \ldots, X_n$.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = Y = AX = A \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

and A is non-singular, then we have

$$X = A^{-1}Y; \quad \left| \frac{dx}{dy} \right| = \left| A^{-1} \right|$$

so that

$$g(y) = \frac{1}{|A|} f\left(A^{-1}y\right)$$

Useful Transformation

$X_1, \ldots, X_n \sim F(\cdot)$, $f(x_i)$ are iid, $Y_n = \max\{X_1, \ldots, X_n\}$ and $Z_n = \min\{X_1, \ldots, X_n\}$.
Now,

$$\begin{aligned} P(Y_n \leq y) &= P(\max\{X_1, \ldots, X_n\} \leq y) \\ &= P(X_1 \leq y, \ldots, X_n \leq y) \\ &= P(X_1 \leq y) \ldots P(X_n \leq y) \\ &= [F(y)]^n \equiv G(y) \end{aligned}$$

and if X_i are continuous,

$$g(y) = n[F(y)]^{n-1} f(y)$$

Now,

$$\begin{aligned} P(Z_n \leq z) &= P(\min\{X_1, \ldots, X_n\} \leq z) \\ &= 1 - P(\min\{X_1, \ldots, X_n\} \geq z) \\ &= 1 - P(X_1 \geq z, \ldots, X_n \geq z) \\ &= 1 - [P(X_1 \geq z) \ldots P(X_n \geq z)] \\ &= 1 - [1 - F(z)]^n \equiv H(y) \end{aligned}$$

and if X_i are continuous,

$$h(z) = \frac{dH(z)}{dz} = n[1 - F(y)]^{n-1} f(y)$$

so

$$\begin{aligned} P(y \leq X_1, \ldots, X_n \leq z) &= F(Y_n \leq y, Z_n \leq z) \\ &= P(Y_n \leq y) - P(Y_n \leq y, Z_n > z) \\ &= P(Y_n \leq y) - P(z < X_1 \leq y, \ldots, z < X_n \leq y) \\ &= [F(y)]^n - [F(y) - F(z)]^n \end{aligned}$$

Transformation by Convolution

X_1 and X_2 are two random variables and let $Y = X_1 + X_2$ and $Z = X_2$ so that

$$\begin{bmatrix} Y \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = AX \quad \text{where } \left|A^{-1}\right| = 1$$

where

$$\begin{aligned} g(y,z) &= f(x_1(\cdot), x_2(\cdot)) \frac{1}{|A|} \\ &= f(y-z, z) \end{aligned}$$

so that

$$g(y) = \int f(y-z, z)\, dz$$

Furthermore, if X_1 and X_2 are independent, then

$$g(y) = \int f_1(y-z) f_2(z)\, dz$$

1.8 Lecture 8

Expectation of Discrete Random Variable

For a discrete random variable X taking values $\{x_j\}_{j=1}^J$, with probabilility function $P(X = x_j) = p_j$, the expectation of X is defined as

$$EX = \sum\nolimits_{j=1}^{J} x_j p_j$$

Rmk: For $EX < \infty$, we need to have $\sum_{j=1}^{J} |x_j| p_j < \infty$.

Expectation of Continuous Random Variable

The expectation of continuous random variable X with pdf $f(x)$ is

$$EX = \int_{-\infty}^{\infty} x f(x)\, dx$$

Rmk 1: EX exists if $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$.
Rmk 2: Since $dF(X) = f(x)\, dx$, we have $EX = \int_{-\infty}^{\infty} x dF(x)$.

Properties of Expectation

Rmk: For the following proof, we only show the continuous case as discrete case is similar.

1. If $P(X=c)=1$, then $EX=c$.

   Proof: Immediate from definition.

2. If EX exists, then $E(cX)=cX$.

   Proof: As constant can be taken out from summation/integral.

$$\int_{-\infty}^{\infty} cf(x)\,dx = c\int_{-\infty}^{\infty} f(x)\,dx = c$$

3. If $Y=aX+b$ and EX exists, $EY=aEX+b$.

   Proof: EY exists because

$$|Y| \le |a|\,|X| + |b| \Rightarrow \int_{-\infty}^{\infty} |y|\,dF(y) \le |a| \int_{-\infty}^{\infty} |x|\,dF(x) + |b| < \infty$$

4. If $Y=X_1+X_2$, EX_1 and EX_2 exist, then $EY=EX_1+EX_2$.

   Proof: EY exists because

$$\begin{aligned}
& \iint |x_1+x_2|\, f(x_1,x_2)\,dx_1dx_2 \\
\le & \iint (|x_1|+|x_2|)\, f(x_1,x_2)\,dx_1dx_2 \\
= & \iint |x_1|\, f(x_1,x_2)\,dx_1dx_2 + \iint |x_2|\, f(x_1,x_2)\,dx_1dx_2 \\
= & \int |x_1|\, f(x_1)\,dx_1 + \int |x_2|\, f(x_2)\,dx_2 < \infty
\end{aligned}$$

   where the result is just additive property of integration.

5. Linear property: If EX_i exists, $E(a_1X_1+\cdots+a_kX_k+b) = a_1EX_1 + \cdots a_kEX_k + b$

   Proof: General case of property 4.

6. If $P(X\ge a)=1$, a is a constant, then $EX\ge a$.

   Proof:

$$\begin{aligned}
EX &= \int x dF(x) = \int_{x<a} x dF(x) + \int_{x\ge a} x dF(x) \\
&= \int_{x\ge a} x dF(x) \ge \int_{x\ge a} a dF(x) = aP(X\ge a) = a
\end{aligned}$$

7. If $P(X\ge a)=1$ and $EX=a$, then $P(X>a)=0$ and $P(X=a)=1$.

   Proof:

$$EX = \int x dF(x) = \int_{x<a} x dF(x) + \int_{x\ge a} x dF(x)$$

which implies

$$a = \lim_{\varepsilon \to 0} \int_a^{a+\varepsilon} x dF(x) + \int_{a+\varepsilon}^{\infty} x dF(x)$$

and this can only be true if $P(X > a) = 0$ and $P(X = a) = 1$.

1.9 Lecture 9

Expectation of function of random variable

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) \, dF(x)$$

Moments

Special Case I: $g(X) = X^r$, we call EX^r the r-th raw moment of X.
Thm: If EX^r exists, then for any $0 < s < r$, $EX^s < \infty$.
Proof:

$$\begin{aligned} \int |x|^s \, dF(x) &= \int_{|x|<1} |x|^s \, dF(x) + \int_{|x|\geq 1} |x|^s \, dF(x) \\ &\leq \int_{|x|<1} dF(x) + \int_{|x|\geq 1} |x|^s \, dF(x) \\ &\leq \int_{|x|<1} dF(x) + \int_{|x|\geq 1} |x|^r \, dF(x) < \infty \end{aligned}$$

Rmk 1: If k^{th} moment does not exist, $(k+1)$th moment will also not exist.
Rmk 2: When $r = 1$, we call EX as mean of distribution.
Special Case II: $g(X) = (X - EX)^r$, we call $E(X - EX)^r$ the r-th central moment of X.
Rmk 1: EX^r exists then $E(X - EX)^r$ also exists.
Rmk 2: $E(X - EX)^r = \binom{r}{0} EX^r - \binom{r}{1} EX^{r-1} EX + \binom{r}{2} EX^{r-2} EX^r + \cdots + (-1)^r \binom{r}{r} EX^r$.
Rmk 3: When $r = 2$, we call $E(X - EX)^2$ the variance of distribution.

Properties of Variance

1. $Y = aX + b$, then $VarY = a^2 VarX$

   Proof: From defintiion

2. If $VarX = 0$, then there exists c such that $P(X = c) = 1$

   Proof: $Var(X) = 0$ implies $E(X - EX)^2 = 0$. Then this means for all X, $X = EX$, so that all X would always be equal to one constant, say c.

Skewness and Kurtosis

Skewness: Measure the symmetry of the distribution:

$$\beta_1 = \frac{E\left(X - EX\right)^3}{\sigma^3}$$

Kurtosis: Measure the tail of distribution:

$$\beta_2 = \frac{E\left(X - EX\right)^4}{\sigma^4}$$

Median and Mode

Median m is defined as $P\left(X \leq m\right) = P\left(X \geq m\right) = 0.5$.
For continuous variable,

$$m = F^{-1}\left(0.5\right)$$

where F^{-1} is inverse function of distribution function F.
Mode μ_0 is defined as $f\left(\mu_0\right) = \sup_X f\left(X\right)$.

Quartile, Interquartile range and Range

If $F\left(x\right)$ is strictly monotone, $F^{-1}\left(\tau\right)$ is called τ-th quartile of X where $\tau \in [0,1]$
We call $Q\left(\tau\right) = F^{-1}\left(\tau\right)$ as quartile function
In general, the τ-th quartile of a distribution is defined as

$$Q_x\left(\tau\right) = \inf\left\{x : F\left(x\right) \geq \tau\right\}$$

when $\tau = 0.5$, $Q_x\left(\tau\right)$ is the median.
when $\tau = 0.25, 0.5, 0.75$, $Q_x\left(\tau\right)$ are quartiles.
when $\tau = 0.2, 0.4, 0.6, 0.8, Q_x\left(\tau\right)$ are quintiles.
when $\tau = 0.1, 0.2, \ldots, 0.9$, $Q_x\left(\tau\right)$ are deciles.
Interqartile range IQR is defined as $IQR = Q_x\left(0.75\right) - Q_x\left(0.25\right)$
Range is defined as $range = \sup x - \inf x = Q_x\left(1\right) - Q_x\left(0\right)$.

Bivariate Moments

1. $EXY = \iint xyf\left(x,y\right)dxdy = \iint xydF\left(x,y\right)$
2. $Eg\left(X,Y\right) = \int g\left(x,y\right)dF\left(x,y\right)$
3. $E\left(X^rY^s\right)$ is called the product of raw moment of order r and s.
4. $E\left[\left(X - EX\right)^r\left(Y - EY\right)^s\right]$ is called the product of central moment of order r and s.
5. Particularly, when $r = 1$, $s = 1$, we have

$$\begin{aligned} Cov\left(X,Y\right) &= E\left[\left(X - EX\right)\left(Y - EY\right)\right] \\ &= EXY - EXEY \end{aligned}$$

6. Furthermore, if X and Y are independent, then $Cov(X, Y) = 0$.

   Proof: $Cov(X, Y) = EXY - EXEY = EXEY - EXEY = 0$.

7. If $cov(X, Y) = 0$, then X and Y are uncorrelated.

Correlation

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{VarX}\sqrt{VarY}}$$

Thm: Correlation coefficient is bounded by 1, $|corr(X, Y)| \le 1$.
Proof: By Cauchy-Scharwz inequality,

$$[E(AB)]^2 \le EA^2 EB^2$$

Taking $A = X - EX$ and $Y = Y - EY$, we have

$$E[(X - EX)(Y - EY)]^2 \le E(X - EX)^2 E(Y - EY)^2$$

or

$$[cov(X, Y)]^2 \le VarX \cdot VarY$$

so that

$$\frac{[cov(X, Y)]^2}{VarX \cdot VarY} \le 1$$

which implies

$$|corr(X, Y)| \le 1$$

1.10 Lecture 10

Generating Function

Given a random variable X, generating function $G_X(t)$ is defined as

$$G_X(t) = Eg(x, t) = \int g(x, t)\, dF(x)$$

Probability Generating Function

$$\begin{aligned} g(x, t) &= t^x \\ P_X(t) &= \int t^x dF(x) = Et^x \end{aligned}$$

Moment Generating Function

$$g(x,t) = e^{tx}$$

$$M_X(t) = Ee^{tx} = \int e^{tx} dF(x)$$

For small value of t,

$$e^{tx} = 1 + tx + \cdots + \frac{(tx)^j}{j!}$$

so that

$$M_X(t) = \int \sum_{j=0}^{\infty} \frac{(tx)^j}{j!} dF(x) = \sum_{j=0}^{\infty} \frac{t^j}{j!} \int x^j dF(x)$$

which implies

$$\left. \frac{\partial^j M_x(t)}{\partial t^j} \right|_{t=0} = EX^j$$

Characteristic Function

$$g(x,t) = e^{itx}$$

$$\phi_X(t) = \int e^{itx} dF(x) = Ee^{itx}$$

Rmk 1: Characteristic Function is always well-defined.

$$\int e^{itx} dF(x) < \infty$$

since

$$e^{itx} = \cos tx + i \sin tx$$

so that

$$\left| e^{itx} \right| = 1.$$

Rmk 2: If X is vector of random variable, the c.f. would be, for example, say $X = (Y, Z)$

$$\phi_X(t,s) = Ee^{i(tY+sZ)} = E\left(e^{itY} e^{isZ}\right)$$

Properties of Characteristic Function

1. There is one to one correspondence between c.f. and p.d.f..

2. If the r-th moment exists, then

$$EX^r = \frac{1}{i^r} \left. \frac{\partial \phi_X(t)}{\partial t^r} \right|_{t=0}.$$

Proof:

$$\begin{aligned}
\frac{\partial \phi_X(t)}{\partial t^r} &= i^r \int x^r e^{itx} dF(x) \\
&= i^r \int x^r dF(x) \text{ as } t \to \infty \\
&= i^r EX
\end{aligned}$$

3. Relationship between m.g.f. and c.f.:

$$\phi_X(t) = M_X(it)$$

Proof: From definition, $M_X(t) = Ee^{tX}$ and $\phi_X(t) = Ee^{itX}$

4. Linear function of random variable:

$$\phi_{a+bX} t = e^{ita} \phi_X(bt)$$

Proof: From definition, $\phi_{a+bX}(t) = Ee^{it(a+bX)} = e^{ita} Ee^{i(tb)X} = e^{ita} \phi_X(bt)$

5. If X and Y are independent random variables, we have

$$\phi_{X+Y}(t, s) = \phi_X(t) \phi_Y(s)$$

6. If $\left. \frac{\partial \phi_X(t)}{\partial t^r} \right|_{t=0}$ exists, then

$$\begin{cases} E\left|X^r\right| < \infty & \text{if } r \text{ is even} \\ E\left|X^{r-1}\right| < \infty & \text{if } r \text{ is odd} \end{cases}$$

7. If $|X^r| < \infty$, then

$$\phi_X(t) = \sum_{j=0}^{k} \frac{EX^j}{j!} (it)^j + o\left(t^k\right)$$

where $o\left(t^k\right)$ is defined as

$$\lim_{t \to \infty} \frac{o\left(t^k\right)}{t^k} = 0.$$

8. The density function $f(x)$ can be obtained by inverse transformation

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) dt.$$

1.11 Lecture 11

Normal Distribution

The density function is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}$$

for all $x \in \mathbb{R}$, $-\infty < \mu < \infty$,$\sigma > 0$ with the mean and variance equal to

$$EX = \mu; \quad VarX = E(x-\mu) = \sigma^2.$$

The rth central moment is

$$\begin{aligned} E(X-\mu)^r &= \begin{cases} 0 & \text{if } r \text{ is odd} \\ (2k-1)(2k-3)\cdots 3 \cdot 1 \cdot \sigma^{2k} & \text{if } r \text{ is even and } r = 2k \end{cases} \\ &= c_r \sigma^r \text{ where } c_r = \begin{cases} 0 & \text{if } r \text{ is odd} \\ (2k-1)(2k-3)\cdots 3 \cdot 1 & \text{if } r \text{ is even and } r = 2k \end{cases} \end{aligned}$$

The mgf and cf are

$$M_X(t) = \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}; \quad \phi_X(t) = \exp\left\{iut - \frac{\sigma^2 t^2}{2}\right\}$$

Standard Normal Distribution

When $\mu = 0$ and $\sigma^2 = 1$, then we call the distribution to be standard normal and the p.d.f. and c.d.f. are denoted by $\phi(x)$ and $\Phi(x)$ respectively.

Property of Normal distribution

1. If $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$

   Proof: $\phi_Y(t) = \phi_{aX+b}(t) = e^{itb}\phi_X(at) = e^{itb}\left[e^{i\mu(at) - \sigma^2(at)^2/2}\right] = e^{i(a\mu+b)t - (a\sigma)^2 t^2/2}$

2. If $X_1, \ldots, X_k \sim N(\mu_i, \sigma_i^2)$ are i.i.d., then

   $$\sum_{i=1}^k X_i \sim N\left(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \sigma_i^2\right).$$

   Proof: Let $Y = \sum x_i$ for $i = 1, \ldots, k$ of $\phi_{X_i}(t) = \exp\left\{iu_i t - \frac{\sigma_i^2 t^2}{2}\right\}$. Since x_i are independent,

   $$\begin{aligned} \phi_Y(t) &= \prod_{i=1}^k \phi_{X_i}(t) = \prod_{i=1}^k \exp\left\{iu_i t - \frac{\sigma_i^2 t^2}{2}\right\} \\ &= \exp\left\{it\sum_{i=1}^k u_i - \frac{t^2 \sum_{i=1}^k \sigma_i^2}{2}\right\}. \end{aligned}$$

3. If $Y = a + \sum_{i=1}^{k} b_i x_i$, then $Y \sim N\left(a + \sum_{i=1}^{k} b_i \mu_i, \sum_{i=1}^{k} b_i^2 \sigma_i^2\right)$

   Proof: Combination of 1 and 2 proof.

Multivariate normal distribution

X_i $(i = 1, \ldots, p)$ are $N\left(\mu_i, \sigma_i^2\right)$ with covariance matrix $\Sigma_{p \times p}$ where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{p1} & \cdots & \cdots & \sigma_{pp} \end{bmatrix}$$

such that $\sigma_{ii} = \sigma_i^2$ and $\sigma_{ij} = cov\left(X_i, X_j\right)$.
Let $X = \left(X_1, X_2, \ldots X_p\right)^T$ is a p-dimensional normal distribution with mean $\mu = \left(\mu_1, \mu_2, \ldots, \mu_p\right)^T$ and covariance matrix Σ. The joint density of X is

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right\} \quad \text{for } x \in \mathbb{R}^p$$

and we denote this by $X \sim N(\mu, \Sigma)$.
Prop: If $X \sim N(\mu, \Sigma)$, then $AX \sim N\left(A\mu, A\Sigma A^T\right)$
Proof: Omitted.

Distribution dervied from Normal

Lognormal Distribution

X is lognormal if $Y = \log X \sim N\left(\mu, \sigma^2\right)$
c.f. is $\phi_X(t) = \exp\left\{i\mu t - \sigma^2 t^2/2\right\}$
m.g.f. $M_X(t) = EX^t = \exp\left\{\mu t + \sigma^2 t^2/2\right\}$
$EX = \exp\left(\mu + \sigma^2/2\right)$ and $VarX = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$

χ^2 Distribution

If X_i $(i = 1, \ldots, p)$ are independent $N(0, 1)$, then $Y = \sum_{i=1}^{p} X_i^2$ is called a random variable of χ^2 distribution with p degree of freedom, which is denoted as $Y \sim \chi_p^2$.
The following are properties of chi-square distribution:

1. If $Z_i \sim N\left(u_i, \sigma_i^2\right)$ are independent, then

$$\sum_{i=1}^{p}\left(\frac{Z_i - \mu_i}{\sigma_i}\right)^2 \sim \chi_p^2$$

2. $EX = p$ and $VarX = 2p$. Other higher moments can be found be using the formula:

$$E\left(X-\mu\right)^r = c_r\sigma^r \text{ where } c_r = \begin{cases} 0 & \text{if } r \text{ is odd} \\ (2k-1)(2k-3)\cdots 3\cdot 1 & \text{if } r \text{ is even and } r = 2k \end{cases}$$

3. If $Y_1 \sim \chi^2(p)$ and $Y_2 \sim \chi^2(q)$, Y_1 and Y_2 are independent, then $Y_1 + Y_2 \sim \chi^2(p+q)$.

Non-central χ^2- distribution

If X_i are independent $N(\mu_i, 1)$ and $\mu_i \neq 0$, then

$$Y = \sum_{i=1}^{p} X_i^2 \sim \chi_p^2\left(\sum_{i=1}^{p}\mu_i^2\right)$$

where $\sum_{i=1}^{p}\mu_i^2$ is called non-central parameter.
Properties: For $Y \sim \chi_p^2(\lambda)$, $EX = p + \lambda$ and $VarX = 2p + 4\lambda$
Proof:

$$\begin{aligned} Y &= \sum X_i^2 = \sum\left(X_i - \mu_i + \mu_i\right)^2 \\ &= \sum\left[\left(X_i-\mu_i\right)^2 + \mu_i^2 + 2\left(X_i-\mu_i\right)\mu_i\right] \\ &= \sum\left(X_i-\mu_i\right)^2 + \sum\mu_i^2 \end{aligned}$$

Student's t distribution

If $X \sim N(0,1)$, $Y \sim \chi_p^2$ and X, Y are independent, then

$$t = \frac{X}{\sqrt{Y/p}}$$

has t distribution with degree of freedom p.
Rmk1: $f(x) = c_p\left(1+\dfrac{x^2}{p}\right)^{-\frac{p+1}{2}}$ where c is a constant depend on p
Rmk2: $t_p \to N(0,1)$ as $p \to \infty$.
Rmk3: For $X \sim t_p$ having moments up to order $(p-1)$,

1. EX^r does not exist when $r \geq p$.

2. For $r < p$,

$$EX^r = \frac{p^{r/2}\Gamma\left(\frac{r+1}{2}\right)\Gamma\left(\frac{p-r}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{p}{2}\right)}$$

where

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx.$$

3. $EX = 0$ for $p > 1$ and $VarX = \frac{p}{p-2}$ for $p > 2$.

F-distribution

If $Y_1 \sim \chi^2_m$, $Y_2 \sim \chi^2_n$ are independnet, then

$$F_{m,n} = \frac{Y_1/m}{Y_2/n}$$

Rmk1: For $r < n/2$, EX^r exists.
Rmk2: For $n > 2$, $EX = \dfrac{n}{n-2}$; For $n > 4$, $VarX = \dfrac{2n^2\left(m+n-2\right)}{m\left(n-2\right)^2\left(n-4\right)}$; For $n \to \infty$, $EX = 1$
Rmk3: $t_q^2 = \dfrac{\chi_1^2}{\chi_q^2/q} \sim F\left(1, q\right)$; This can be seen by $Var\left(t_q\right) = \frac{q}{q-2} \to 1$ as $q \to \infty$ and $E\left(t_q\right) = 0$.

Other Distribution

Gamma Distribution

$$\begin{aligned} f\left(x\right) &= \frac{1}{\Gamma\left(t\right)}\lambda^t x^{t-1} e^{-\lambda x}; x \geq 0, t > 0 \\ EX &= \frac{t}{\lambda}; \quad VarX = \frac{t}{\lambda^2} \end{aligned}$$

Bernoulli Distribution

$$X = \begin{cases} 0 & \text{with probability } p \\ 1 & \text{with probability } 1-p \end{cases}$$

so that

$$p\left(x\right) = p^x \left(1-p\right)^{1-x}$$

Binomial Distribution

If $Y = X_1 + X_2 + \cdots + X_n$ where X_i are i.i.d. Bernoulli distribution, then $Y \sim B\left(n, p\right)$

Possion Distribution

X takes the value from $0, 1, 2, \cdots$. The probability function is

$$P\left(X = x\right) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Indicator Fucntion

For event $A \in \Omega$, I_A is defined as

$$I_A = \begin{cases} 0 & \omega \notin A \\ 1 & \omega \in A \end{cases}$$

Rmk: This is speical case of Bernoulli.

1.12 Lecture 12

Markov Inequality

Let $g(\cdot)$ be a non-negative function of a random variable X and $E[g(X)]$ exists, then for any $c > 0$, we have

$$P[g(X) \geq c] \leq \frac{E[g(X)]}{c}$$

Proof:

$$\begin{aligned} E[g(X)] &= \int g(x)\,dF(x) = \int_{g(x)\geq c} g(x)\,dF(x) + \int_{g(x)<c} g(x)\,dF(x) \\ &\geq \int_{g(x)\geq c} g(x)\,dF(x) \geq \int_{g(x)\geq c} c\,dF(x) = cP[g(X) \geq c] \end{aligned}$$

If X is non-negative, that is $P(X \geq 0) = 1$. Then, let $g(X) = X$, we have

$$P(X \geq c) \leq \frac{EX}{c}.$$

In general, consider $g(x) = |x|$, then

$$P(|X| \geq c) \leq \frac{E|X|}{c}$$

Chebyshev's Inequality

Consider $g(x) = (x-\mu)^2$ where $\mu = EX$, then we have

$$P\left[(x-\mu)^2 \geq c\right] \leq \frac{E\left[(x-\mu)^2\right]}{c} = \frac{VarX}{c}$$

Furthermore, taking $c = \varepsilon^2\sigma^2$ then,

$$P\left[(x-\mu)^2 \geq \varepsilon^2\sigma^2\right] \leq \frac{\sigma^2}{\varepsilon^2\sigma^2} = \frac{1}{\varepsilon^2}$$

which implies

$$P[|x-\mu| \geq \varepsilon\sigma] \leq \frac{1}{\varepsilon^2}$$

Generalized Markov Inequality

Let $g(\cdot)$ be a non-negative function on $\mathbb{R}$ and $g(\cdot)$ is non-decreasing on the range of a random variable z. Then

$$P(z \geq a) \leq \frac{E[g(z)]}{g(a)}$$

Proof:

$$P(z \geq a) = P(g(z) \geq g(a)) \leq \frac{E[g(z)]}{g(a)}$$

where the last step is application of Markov inequality.

Berkstein Inequality

Let $g(t) = e^{bt}$ for $b > 0$, then

$$P(z \geq a) \leq e^{-ab} E e^{bz} \leq \inf_{b \geq 0} e^{-ab} E e^{bz}$$

Corollory 1: If $X \sim Binomial(n, p)$, then $P(|x - np| \geq n\varepsilon) \leq 2e^{-n\varepsilon^2/2}$
Corollory 2: If u_i are independent r.v. and $P(u_i \leq b) = 1$, $v = \sum_{i=1}^{n} E\left(u_i^2\right)$, then for all $a > 0$, we have

$$P\left[\sum_{i=1}^{n} \left(u_i^2 - v\right) \geq a\right] \leq \exp\left\{-\frac{a^2}{2(v + ba/3)}\right\}$$

Jensen's Inequality

If g is convex on $S \subseteq \mathbb{R}$ and S is convex and closed, $P(X \in S) = 1$, $E[g(x)] < \infty$ and $EX < \infty$, then

$$E[g(X)] \geq g[E(X)].$$

Covariance inequality

Cauchy-Schwarz Inequality

$$(EXY)^2 \leq EX^2 EY^2$$

Holder's Inequality

If $p, q > 1$ and $\dfrac{1}{p} + \dfrac{1}{q} = 1$, then

$$E|XY| \leq (E|X^p|)^{1/p} (E|X^q|)^{1/q}$$

Notation: $\|X\|_p = (E|X^p|)^{1/p}$ called the L_p norm so that

$$\|XY\| \leq \|X\|_p \|Y\|_q$$

Minokowski Inequality

For $p > 1$,

$$E(|X + Y|^p)^{1/p} \leq (E|X^p|)^{1/p} + (E|X^q|)^{1/q}$$

or

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$$

Lyapunov's Inequality

For any $p \geq q > 0$,

$$\|X\|_p \geq \|Y\|_q.$$

Point-wise Convergence

We say f_n converge to f point-wisely if for all x, $f_n(x) \to f(x)$.
Rmk: Not enough to ensure asymptotic property. Need stronger condition. We need to define distance d between f_n and f to be less than a very small number. If d is norm, it is norm convergence.

Norm

Let V be a collection of functions. $\|\cdot\|$ is a norm of V if

1. $\|f\| \geq 0$ for all $f \in V$
2. $\|f\| = 0$ iff f is a zero function
3. $\forall a \in \mathbb{R}, \forall f \in V, \|af\| = |a| \|f\|$
4. $\forall f, g \in V, \|f+g\| \leq \|f\| + \|g\|$

Rmk: If V satisfies these four properties, we say V is endowed with norm.

Norm Convergence

If $\|f_n - f\| \to 0$ as $n \to 0$, we say $f_n(\cdot)$ converges to f with respect to norm $\|\cdot\|$.
Rmk: Difference definition of norm leads to different definition of convergence.

Convergence in Probability (Weak Convergence)

$\{Z_n\}$ is a sequence of random variable. Then Z_n converges to a in probability if

$$\forall \varepsilon > 0, P(\{|Z_n - a| < \varepsilon\}) \to 1 \text{ as } n \to \infty$$

or

$$\lim_{n\to\infty} P(\{|Z_n - a| < \varepsilon\}) = 1$$

and we denote this as c.

Almost Sure Convergence (Strong Convergence)

Z_n converges to a with probability 1 if

$$P\left(\left\{\lim_{n\to\infty} Z_n = a\right\}\right) = 1$$

and we deonte it as $Z_n \overset{a.s.}{\to} a$.
Rmk: Almost sure event A means $P(A) = 1$.

Convergence in r-th Mean

If $E\left|X_n\right|^r < \infty$ for all n and $E\left|Z_n - a\right|^r \to 0$ as $n \to \infty$, then Z_n converges to a in r-th mean.
Rmk: When $r = 2$, we have convergence in mean square error.

Convergence in Distribution (Weakest Convergence)

$\{Z_n\}$ is a sequence of random variable with c.d.f. $\{F_n\}$ and Z is a random variable with c.d.f. F. If,

$$\forall x, \quad \lim_{n\to\infty} F_n\left(x\right) = F\left(x\right)$$

then sequence $\{Z_n\}$ converges in distribution to Z and we denote is as $Z_n \xrightarrow{D} Z$.

Interrationship between convergences

1. If $Z_n \xrightarrow{P} a$, then we have $Z_n \xrightarrow{D} Z$.
2. If $Z_n \xrightarrow{a.s.} a$, then we have $Z_n \xrightarrow{P} a$.
3. If Z_n converges in r-th mean, we have $Z_n \xrightarrow{P} a$.
4. If Z_n converges in k-th mean and for all $r \le k$, we have Z_n converges in r-th mean.

Special relationship between convergences

1. If $Z_n \xrightarrow{D} Z$ where Z distribution with function which admits a constant value with probability 1, then $Z_n \xrightarrow{P} a$.
2. If $Z_n \xrightarrow{P} Z$ and $P\left(|Z_n| \le k\right) = 1$ for all n, then Z_n converges in r-th mean.
3. If $\sum_{n=1}^{\infty} P\left(|Z_n - Z| > \varepsilon\right) < \infty$, for all $\varepsilon > 0$, then $Z_n \xrightarrow{a.s.} Z$.

Law of Large number and Central Limit Theorem

$\{X_i\}$ are i.i.d. random variables and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

1. If $EX_i = \mu$ then $\bar{X} \xrightarrow{D} \mu$.
2. (WLLN) If $EX_i = \mu$ and $VarX_i = \sigma^2 < \infty$, then $\bar{X} \xrightarrow{P} \mu$.

   Proof: First note that $E\bar{X}_n = \mu$ and $Var\bar{X}_n = \sigma^2/n$. By Chebyshev's Inequality, we have
   $$P\left(|\bar{x} - \mu| \ge \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2 n}$$

so that

$$1 - P\left(|\bar{x} - \mu| \le \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2 n}$$

or

$$P\left(|\bar{x} - \mu| \le \varepsilon\right) \ge 1 - \frac{\sigma^2}{\varepsilon^2 n}$$

Taking $n \to \infty$, we have

$$\lim_{n \to \infty} P\left(|\bar{x} - \mu| \le \varepsilon\right) \ge 1.$$

However, probability is bounded by 1, so we have

$$\lim_{n \to \infty} P\left(|\bar{x} - \mu| \le \varepsilon\right) = 1.$$

3. (SLLN) If $EX_i = \mu$, then $\bar{X} \overset{a.s.}{\to} \mu$
4. (CLT) If $EX_i = \mu$ and $VarX_i = \sigma^2 < \infty$, then

$$\frac{\sqrt{n}\left(\bar{x} - \mu\right)}{\sigma} \to N\left(0, 1\right).$$

Rmk:

$$\frac{\sqrt{n}\left(\bar{x} - \mu\right)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)$$

Chapter 2

Statistical Inference

2.1 Lecture 14

Estimation

For data set $\{X_i\}_{i=1}^n$, we wish to estimate θ, the unknown parameter of a parametric model P_θ.

Point Estimator

Given a sample $\{X_i\}_{i=1}^n$ and consider a pararmeter θ. Let T be a function of the sample and $\hat{\theta} = T(X_1, \ldots, X_n)$ is called a point estimator.

Interval Estimator

Given $\{X_i\}_{i=1}^n$, let T_1 and T_2 be functions of sample. Let $\hat{\theta}_L = T_1(X_1, \ldots, X_n)$ and $\hat{\theta}_U = T_2(X_1, \ldots, X_n)$. If

$$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\right) = 1 - \alpha$$

then $\left[\hat{\theta}_L, \hat{\theta}_U\right]$ is an interval estimator with confidence probability $1 - \alpha$.

Unbiasedness

Define $bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$.

An unbiased estimator $\hat{\theta}$ means $bias\left(\hat{\theta}\right) = 0$ or $E\left(\hat{\theta}\right) = \theta$.

A biased estimator $\hat{\theta}$ means $bias\left(\hat{\theta}\right) \neq 0$ or $E\left(\hat{\theta}\right) \neq \theta$.

Variance of Estimator

$$Var\left(\hat{\theta}\right) = E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2$$

Mean-square Error (MSE)

Combined measure of unbiasedness and variance of estimator.

$$\begin{aligned}
MSE\left(\hat{\theta}\right) &= E\left(\hat{\theta} - \theta\right)^2 \\
&= E\left[\hat{\theta} - E\left(\hat{\theta}\right) + E\left(\hat{\theta}\right) - \theta\right]^2 \\
&= E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2 + 2E\left\{\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]\left[E\left(\hat{\theta}\right) - \theta\right]\right\} + E\left[\hat{\theta} - \theta\right]^2 \\
&= E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2 + 2\left[E\left(\hat{\theta}\right) - \theta\right] E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right] + E\left[\hat{\theta} - \theta\right]^2 \\
&= Var\left(\hat{\theta}\right) + bias^2\left(\hat{\theta}\right)
\end{aligned}$$

Consistency

If $\hat{\theta} \to \theta$ as $n \to \infty$, then $\hat{\theta}$ is a consistent estimator. In particular,
if $\hat{\theta} \xrightarrow{P} \theta$, $\hat{\theta}$ is a weak consistent estimator and
if $\hat{\theta} \xrightarrow{a.s.} \theta$, $\hat{\theta}$ is a strong consistent estimator.
Rmk: $\hat{\theta} \xrightarrow{P} \theta$ means $\forall \varepsilon > 0$, $P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \to 0$ as $n \to \infty$. By Markov inequality,

$$P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) \le \frac{E\left|\hat{\theta} - \theta\right|^2}{\varepsilon^2}.$$

Since ε is fixed, $E\left|\hat{\theta} - \theta\right|^2 = MSE\left(\hat{\theta}\right) \to 0$ as $n \to \infty$ implies consistency and so it suffices to show $bias^2\left(\hat{\theta}\right) \to 0$ and $Var\left(\hat{\theta}\right) \to 0$ as $n \to \infty$ for consistency of $\hat{\theta}$.

Sufficiency

T is a sufficient statistic if the conditional distribution of $(X_1, \ldots, X_n)$ given T is indpendent of θ.
Rmk: Informally, it means T summarizes all information (from the sample) relevant to θ.

Factorization Theorem

Let $(X_1, \ldots, X_n)$ be random sample from population whose distribution is dependent on θ. $T = T\left(X_1, \ldots, X_n\right)$ is a sufficient statistic for θ if and only if the joint density of X and θ can be written as

$$f\left(X_1, \ldots, X_n; \theta\right) = h\left(T, \theta\right) H\left(X_1, \ldots, X_n\right)$$

2.2 Lecture 15

Point Estimation of μ

Natural candidate: sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Unbiasedness:

$$\begin{aligned} E\left(\bar{x}\right) &= E\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E\left(x_i\right) \\ &= \frac{1}{n} n\mu = \mu \end{aligned}$$

Hence, $\bar{x}$ is unbiased estimator.
Consistency:

$$\begin{aligned} Var\left(\bar{x}\right) &= Var\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2} Var \sum_{i=1}^{n} x_i \\ &= \frac{1}{n^2} \sum_{i=1}^{n} Var\left(x_i\right)] \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

therefore,

$$\begin{aligned} MSE\left(\bar{x}\right) &= bias^2\left(\bar{x}\right) + Var\left(\bar{x}\right) \\ &= 0 + \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty \end{aligned}$$

Hence, as $\bar{x} \to \mu$, $\bar{x}$ is consistent estimator.

Point Estimation of σ^2

Natural candidate: sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Unbiasedness:

$$\begin{aligned}
E\left(s^2\right) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\
&= \frac{1}{n}E\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right] \\
&= \frac{1}{n}\left(nEx_i^2\right) - E\bar{x}^2 = Ex^2 - E\bar{x}^2
\end{aligned}$$

Now consider

$$\begin{aligned}
E\bar{x}^2 &= E\left[\frac{\sum_{i=1}^{n}x_i}{n}\right]^2 \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}x_i\right]^2 \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}x_i^2 + \sum_{i\neq j}^{n}x_ix_j\right] \\
&= \frac{1}{n^2}\left[nEx^2 + n(n-1)ExEx\right] \\
&= \frac{1}{n}Ex^2 + \frac{n-1}{n}(Ex)^2
\end{aligned}$$

so that

$$\begin{aligned}
E\left(s^2\right) &= Ex^2 - E\bar{x}^2 \\
&= Ex^2 - \left[\frac{1}{n}Ex^2 + \frac{n-1}{n}(Ex)^2\right] \\
&= \frac{n-1}{n}\left[Ex^2 - (Ex)^2\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Hence, s^2 is biased estimator. The bias would converge to zero as n goes to infinity,

$$\begin{aligned} bias\left(s^2\right) &= E\left(s^2\right) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= -\frac{1}{n}\sigma^2 \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

so that s^2 is asympotically unbiased.
Unbiased Version of sample variance would be

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$$

Its unbiasedness could be proved easily as

$$\begin{aligned} E\left(\hat{\sigma}^2\right) &= E\left(\frac{n}{n-1}s^2\right) \\ &= \frac{n}{n-1}E\left(s^2\right) \\ &= \frac{n}{n-1}\frac{n-1}{n}\sigma^2 \\ &= \sigma^2 \end{aligned}$$

Rmk: The reason for $n-1$ but not n can be explained by degree of freedom or using the mathematical rank concept.

Distribution of $\bar{x}$

Since $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we have

$$\phi_{\hat{\mu}}(t) = \phi_{\frac{1}{n}\Sigma x_i}(t)$$

so that if we know the distribution of x_i, then we know the distribution of sample mean.
Even if we don't know the sample mean, by central limit theorm, we have

$$\frac{\sqrt{n}(\mu - \hat{\mu})}{\sigma} \xrightarrow{D} N(0,1).$$

Distribution of s^2

For simplicity, we assume $X_i \overset{i.i.d.}{\sim} N\left(\mu, \sigma^2\right)$.
Then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ would have distribution $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Notice that $\dfrac{X_i-\mu}{\sigma} \sim N(0,1)$ so that $\dfrac{(X_i-\mu)^2}{\sigma^2} \sim \chi_1^2$. We have to show $\dfrac{ns^2}{\sigma^2} = \sum_{i=1}^{n} \dfrac{(X_i-\mu)^2}{\sigma^2} \sim \chi_{n-1}^2$

Consider orthogonal transformation of

$$L = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \cdots & \frac{1}{\sqrt{n(n-1)}} \end{bmatrix}$$

such that $L(X-\mu) = Y$, $Y = [Y_1 \ldots Y_n]^T$, $X-\mu = [X_1-\mu \ldots X_n-\mu]^T$ and $Y^TY = [X-\mu]^T[X-\mu]$ implies

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i-\mu)^2$$

Jacobian matrix is

$$|J| = \left|\frac{dX}{dY}\right| = 1$$

which implies

$$\begin{aligned} Y_1 &= \sqrt{n}\left(\bar{X}-\mu\right) \\ \sum_{i=2}^{n} Y_i^2 &= \sum_{i=1}^{n} \left(X_i-\bar{X}\right)^2 \end{aligned}$$

So that

$$\begin{aligned} \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} &= \sum_{i=1}^{n} Y_i^2 \\ &= Y_1 + \sum_{i=2}^{n} Y_i^2 \\ &= \sqrt{n}\left(\bar{X}-\mu\right) + \sum_{i=1}^{n} \left(X_i-\bar{X}\right)^2 \\ &= \chi_1^2 + \chi_{n-1}^2 \end{aligned}$$

Therefore, we have the result that

$$\frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} \frac{\left(X_i-\bar{X}\right)^2}{\sigma^2} \sim \chi_{n-1}^2$$

and s^2 is independent of $\bar{X}$ so that

$$\frac{\hat{\mu}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

However, we usually have no information on σ^2 so we have to replace σ^2 by $\hat{\sigma}^2$. The distribution would then be

$$\begin{aligned}
\frac{\hat{\mu}-\mu}{\sqrt{\hat{\sigma}^2/n}} &= \frac{(\hat{\mu}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{\hat{\sigma}^2/n}/\sqrt{\sigma^2/n}} \\
&= \frac{(\hat{\mu}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{\hat{\sigma}^2/\sigma^2}} \\
&= \frac{(\hat{\mu}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}}} \\
&= \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/n-1}} \sim t_{n-1}
\end{aligned}$$

2.3 Lecture 16

Method of Moment (MOM)

Random sample $X_1,\ldots,X_n$ and parametric model P_θ where $\theta=(\theta_1,\ldots,\theta_k)$ is k-dimensional parameter.
Population moment

$$EX_i^r = \int x^r dF_\theta(x) = \mu_r(\theta)$$

and sample moment

$$\hat{\mu}_r(\theta) = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

Equate first k raw moments: For $r=1,\ldots,k$

$$\hat{\mu}_r(\theta) = \mu_r(\theta)$$

Generalized Method of Moment (GMM)

Instead of equating population moment to sample moment, we impose more general moment conditions:

$$E[h(X_i,\theta)] = 0$$

where $h_r = x_i^r - \mu_r(\theta)$ is the speical case of MOM.

We have k parameter as $\theta = (\theta_1, \ldots, \theta_k)$ and r restriction as $h(X_i, \theta) = [h_1(X_i, \theta), \ldots, h_r(X_i, \theta)]^T$. Therefore,
if $k = r$, then the system is exactly identified;
if $k > r$, then the system is under identified;
if $k < r$, then the system is over identified.
Rmk: under-identified case is not considered usually in economics
When $k = r$, set

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

When $k < r$, we would try to minimize

$$\left\| \frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta) \right\|$$

with a weighting martrix W, so that, we have to

$$\min_{\theta} \left[\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta)\right]^T W \left[\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta)\right]$$

Properties of GMM Estimator

Exact identication

Consistency: Yes, if it satisfies idendification condition,

$$\hat{\theta} \to \theta_0, \theta \notin B_\varepsilon(\theta_0) \text{ s.t. } \frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta) = 0$$

Distribution of $\hat{\theta}$: Assuming consistency, using Taylor's expansion,

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \theta)}{\partial \theta}\left(\hat{\theta} - \theta\right) + O_p\left(\left\|\hat{\theta} - \theta\right\|^2\right) = 0$$

As $\theta \to \theta$, then $O_p\left(\left\|\hat{\theta} - \theta\right\|^2\right) \to 0$ and we have

$$\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \theta)}{\partial \theta}\left(\hat{\theta} - \theta\right) = 0$$

so that

$$\left(\hat{\theta} - \theta\right) \simeq \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \theta)}{\partial \theta}\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} h(x_i, \theta)\right]$$

If X_i are i.i.d., then by Law of large number,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(x_i, \theta)}{\partial \theta} \xrightarrow{P} E\left[\frac{\partial h(x_i, \theta)}{\partial \theta}\right] \equiv D$$

and

$$\frac{1}{n}\sum_{i=1}^{n} h\left(x_i,\theta\right) \xrightarrow{P} E\left[h\left(x_i,\theta\right)\right] = 0$$

However, given $E\left[h\left(x_i,\theta\right)\right] = 0$, then X_i are mean zero i.i.d., then by central limit theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} h\left(x_i,\theta\right) \xrightarrow{D} N\left(0,\Sigma\right)$$

Hence,

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \xrightarrow{D} D^{-1}N\left(0,\Sigma\right) = N\left(0,D^{-1}\Sigma D^{-1}\right)$$

Over-indentification

Similar derivation with similar result.

Maximum likelihood estimation (MLE)

$\left(X_1,\ldots,X_n\right)$ are i.i.d. with density function $f\left(x_i,\theta\right)$ and joint density function

$$f\left(x;\theta\right) = \prod_{i=1}^{n} f\left(x;\theta\right).$$

Consider $f\left(x;\theta\right)$ as function of θ given $\left(X_1,\ldots,X_n\right)$ is called likelihood function.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L\left(\theta\right)$$

where $L\left(\theta\right) = f\left(x;\theta\right)$
Log-tranformation:

$$\ell\left(x_i,\theta\right) \equiv \log L\left(\theta\right) = \sum_{i=1}^{n}\log f\left(x_i,\theta\right)$$

with F.O.C.

$$\frac{\partial \log L\left(\theta\right)}{\partial\theta} = 0 \Rightarrow \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\log f\left(x_i,\theta\right)\right] = 0$$

Rmk 1: GMM consider $\frac{1}{n}\sum_{i=1}^{n} h\left(x_i,\theta\right) = 0$
Rmk 2: score function is defined $s\left(x,\theta\right) = \dfrac{\partial\ell\left(x,\theta\right)}{\partial\theta}$
Rmk 3: MLE depends on whole distribution and GMM just need moment restriction
Using GMM method, we know

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \to N\left(0,I^{-1}\left(\theta\right)\right)$$

where

$$I\left(\theta\right) = E\left[\frac{\partial\ell\left(x_i,\theta\right)}{\partial\theta}\frac{\partial\ell\left(x_i,\theta\right)}{\partial\theta}^T\right]$$

which is called information matrix.

Properties of MLE

MLE is efficient which implies smallest variance and covariance matrix. For any other consistent estimator, say $\tilde{\theta}$, if

$$\sqrt{n}\left(\hat{\theta}-\theta\right) \to N\left(0, V\right)$$

then $V \geq I^{-1}\left(\theta\right)$. Therefore, $V - I^{-1}\left(\theta\right)$ is semi-positive definite matrix.
If $\hat{\theta}$ is M.L.E., then $\hat{\beta} = g\left(\hat{\theta}\right)$ is also M.L.E.

MLE for normal i.i.d random sample

Likelihood function would be

$$\ell\left(\mu, \sigma^2\right) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{\sum\left(x_i-\mu\right)^2}{2\sigma^2}$$

with F.O.C.

$$\begin{aligned}
\frac{\partial \ell\left(\mu,\sigma^2\right)}{\partial \mu} &= \frac{2\sum\left(x_i-\mu\right)}{\sigma^2} = 0 \\
\frac{\partial \ell\left(\mu,\sigma^2\right)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{\sum\left(x_i-\mu\right)^2}{2\sigma^4} = 0
\end{aligned}$$

so that

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

2.4 Lecture 17

Confidence Interval

Let $\hat{\theta}$ be point estimate. Confidence interval $\left[\hat{\theta}_L, \hat{\theta}_U\right] \ni \theta$ with confidence $1-\alpha$ if

$$P\left(\theta \in \left[\hat{\theta}_L, \hat{\theta}_U\right]\right) = 1-\alpha$$

Assuming unbiasedness, we have $E\left(\hat{\theta}\right) = \theta$. Procedure to find confidence interval:

1. Find the distribution of estimator $P\left(\hat{\theta} < x\right) = F\left(x\right)$
2. Find x_L and x_U such that $F\left(x_L\right) = \alpha/2$ and $F\left(x_U\right) = \alpha/2$
3. Then we to find $P\left(x_L \leq \hat{\theta} \leq x_U\right) = 1-\alpha$ or

$$P\left(\theta - x_U \leq \theta - \hat{\theta} \leq \theta - x_L\right) = 1-\alpha.$$

Confidence interval for normal random sample

C.I. for $\hat{\mu}_{MLE}$

Known σ^2: by i.i.d. property,

$$\hat{\mu}_{MLE} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

so that

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Hence,

$$P\left(\Phi(\alpha/2) \le \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \le \Phi^{-1}(1 - \alpha/2)\right) = 1 - \alpha$$

or

$$P\left(\hat{\mu} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \le \mu \le \hat{\mu} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

We have C.I. to be

$$\left[\hat{\mu} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}, \hat{\mu} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right]$$

Rmk: C.I. narrows down if σ falls or n grows.
Unknown σ^2: by i.i.d. property

$$t_{\hat{\mu}} = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

with similar result by replace Φ^{-1} to d.f. of t distribution.

C.I. for $\hat{\sigma}^2$

Unbiased point estimate would be

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Recall the distribution of $\hat{\sigma}^2$ would be

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

so that, taking c as inverse of d.f. of χ^2_{n-1}

$$P\left(c(\alpha/2) \le \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \le c\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

or

$$P\left(\frac{(n-1)\hat{\sigma}^2}{c(\alpha/2)} \le \sigma^2 \le \frac{(n-1)\hat{\sigma}^2}{c(1 - \alpha/2)}\right) = 1 - \alpha$$

C.I. for GMM

By C.L.T., we have

$$\sqrt{n}\left(\hat{\theta}-\theta\right)\xrightarrow{D}N\left(0,D^{-1}\Sigma D^{-1}\right)$$

with standardization,

$$\sqrt{n}\left(D^{-1}\Sigma D^{-1}\right)^{-1/2}\left(\hat{\theta}-\theta\right)\xrightarrow{D}N\left(0,I_k\right)$$

so that

$$\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}\xrightarrow{D}N\left(0,1\right)$$

Rmk: In practice, we usually have aympotic C.I. rather than actual C.I.

2.5 Lecture 18

Hypothesis Testing

To formulate a decision rule δ such that given the data $\{X_1,\ldots,X_n\}$, it is possible to infer whether a given hypothesis is supported.

Simple Hypothesis

A hypothesis is simple if together with basic assumption, it specifies the distribution completely.
Suppose for a parametric model P_θ where $\theta\in\Theta$, if a hypothesis is $\theta=\theta_0$ so that the distribution is compeletely known, then such a hypothesis is simple.

Composite Hypothesis

Otherwise, it is called a composite hypothesis. With such a hypothesis, we could not know the distribution completely.

Null hypothesis

Denoted as H_0 and it is hypothesis to be tested.

Alternative hypothesis

Denoted as H_1 which is usually complement of H_0 with respect to sample space Θ.

Testing Statistics

Suppose we have a null hypothesis $H_0 : \theta \in \Theta_0$ and a testing statistics $T_n = T(X_1, \ldots, T_n) \in \mathcal{A}$.
Construction of a testing procedure is to set the following rules:

1. If $T_n \in \mathcal{A}_0 \subseteq \mathcal{A}$, then reject H_0.
2. Otherwise, if $T_n \notin \mathcal{A}_0$, then accept H_0.

Hence, $\mathcal{A}$ is called critical region or rejection region and $\mathcal{A} - \mathcal{A}_0$ is called acceptance region.

Decision error

	Accept H_0	Reject H_0
H_0 is true	correct decision	Type I error
H_0 is not true	Type II error	correct decision

Rmk 1: Cannot simultaneously reduce both types of errors.
Rmk 2: Denote α =significance level $= P(\text{Type I error}) = P(\text{Rejection } H_0 \mid H_0)$.
Rmk 3: Denote $\beta = P(\text{Type II error}) = P(\text{Accept } H_0 \mid H_1)$.
Rmk 3: Denote $1-\beta$ =power of test $= 1-P(\text{Type II error}) = 1-P(\text{Accept } H_0 \mid H_1)$.

Choosing testing procedure

1. Specifies α, which controls type I error.
2. Minimize type II error among testing statistics

$$\begin{aligned} P(\text{Type II error}) &= P(\text{Accept } H_0 \mid H_1) \\ &= 1 - P(\text{Reject } H_0 \mid H_1) \end{aligned}$$

so that min $P(\text{Type II error})$ is equivalent to max $P(\text{Reject } H_0 \mid H_1)$ or to maximize power of the test.

Relationship to confidence Interval for normal random sample

Known σ^2:
Given null hypothesis $H_0 : \mu = \mu_0$, we have

$$\frac{\mu - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

so that with significance level $1-\alpha$, we have

$$P\left(\left|\frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right| \leq c\right) = 1 - \alpha$$

so that

$$P\left(\left|\frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right|>c\right)=\alpha$$

which is exactly the definition of type I error if the testing procedure is to reject H_0 if

$$\left|\frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right|>\Phi(\alpha)$$

Therefore, the rejection region would be

$$\left\{\left|\frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right|>\Phi(\alpha)\right\}$$

Rmk: Confidence level : $1-$ significance level
Knonwn σ^2:
Using $\hat{\sigma}$ to replace unknown population counterpart:

$$\frac{\mu-\mu_0}{\hat{\sigma}/\sqrt{n}}\sim t_{n-1}$$

so that the rejection region would be

$$\left\{\left|\frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right|>c_\alpha\right\}$$

where c_α is the inverse of distribution function of t distribution.

2.6 Lecture 19

Power function

Power function of testing procedure δ is defined as

$$\pi(\theta\mid\delta)=P(\text{rejecting } H_0\mid\theta)$$

For simple hypotheses $H_0:\theta=\theta_0$ and $H_1:\theta=\theta_1$, given level of significance α,

$$\pi(\theta_0\mid\delta)=\alpha$$

and the power of test would be

$$\pi(\theta_1\mid\delta)=1-\beta$$

Rmk: For two different testing procedures δ_1 and δ_2, one comparison method is to compare the power of $\pi(\theta_1\mid\delta_1)$ and $\pi(\theta_1\mid\delta_2)$.

Neyman-Pearson Lemma

Given a sample $\{x_1, \ldots, x_n\}$ and likelihood function $L(\theta)$, consider a simple hypothesis $H_0 : \theta = \theta_0$ against simple alternative $H_1 : \theta = \theta_1$. The following test is the most powerful test:

$$\text{Reject } H_0 \text{ if } \frac{L(\theta_0)}{L(\theta_1)} < c$$

Rmk 1: Since rejection region $\mathcal{A}_0$ is $\left\{\frac{L(\theta_0)}{L(\theta_1)} < c\right\}$, the value of c could be found when α is given because

$$P\left(\left\{\frac{L(\theta_0)}{L(\theta_1)} < c\right\} \middle| H_0\right) = \alpha$$

Rmk 2: If f is continuous, we have

$$\begin{aligned} \alpha &= P(\mathcal{A}_0 \mid H_0) = \int_{\mathcal{A}_0} f(x, \theta_0)\, dx; \\ 1 - \beta &= P(\mathcal{A}_0 \mid H_1) = \int_{\mathcal{A}_0} f(x, \theta_1)\, dx \end{aligned}$$

Rmk 3: The meaning of most powerful test means that if we have another testing procedure with rejection region $\mathcal{B}_0$, we would have (i) $P(\mathcal{B}_0 \mid H_0) = \alpha$ and (ii) $P(\mathcal{B}_0 \mid H_1) \le P(\mathcal{A}_0 \mid H_1)$.

Significance level of composite null hypothesis

Suppose $H_0 : \theta \in \Theta_0$ and $H_1 : \theta = \theta_1$, if for all $\theta \in \Theta_0$, such that

$$P_\theta(\text{Type I error} \mid H_0) = \int_{\mathcal{A}_0} f(x, \theta)\, dx = \pi(\theta) < \alpha$$

then, it has significance level of α.

Power of composite alternative hypothesis

Suppose $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \Theta_1$, if for all $\theta \in \Theta_1$, such that

$$P_\theta(\mathcal{A}_0 \mid \theta) = \int_{\mathcal{A}_0} f(x, \theta)\, dx = \pi(\theta) = 1 - \beta$$

then, it has power of $1 - \beta$.

Uniformly most powerful Test (UMP Test)

Suppose $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta$, testing procedures δ_i, test δ^* is UMP test if for all level of α such that for the

1. same level of significance α,

$$\pi(\theta \mid \delta_i) \leq \alpha, \text{ for all } \theta \in \Theta_0$$

2. δ^* is highest power among δ_i,

$$\pi(\theta \mid \delta^*) \geq \pi(\theta \mid \delta_i), \text{ for all } \theta \in \Theta_1$$

Monotone Likelihood Ratio

Let $f_n(x \mid \theta)$ be joint p.d.f. of $\{x_1, \ldots, x_n\}$ and consider a statistic $T = T(x_1, \ldots, x_n)$ if for two values θ_1 and θ_2 in Θ, say $\theta_1 < \theta_2$, the likelihood ratio

$$\frac{f_n(x \mid \theta_2)}{f_n(x \mid \theta_1)}$$

depends on x only through $T(x)$, and this ratio is an increasing fucntion of $T(x)$ over the range of all possible values of $T(x)$, then $f_n(x \mid \theta)$ has a monotone likelihood ratio in statistics T.

Testing with monotone likelihood ratio

For $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$:
Suppose $f_n(x \mid \theta)$ has a monotone likelihood ratio in $T(x)$. Let c and α be constant such that $P(T \geq c \mid \theta = \theta_0) = \alpha$. Then the test "rejects $H_0 : \theta \leq \theta_0$ if $T \geq c$" is a UMP test at the level of significance α.
For $H_0 : \theta \geq \theta_0$ and $H_1 : \theta < \theta_0$:
Suppose $f_n(x \mid \theta)$ has a monotone likelihood ratio in $T(x)$. Let c and α be constant such that $P(T \leq c \mid \theta = \theta_0) = \alpha$. Then the test "rejects $H_0 : \theta \geq \theta_0$ if $T \leq c$" is a UMP test at the level of significance α.

Appendix A

The Appendix

Now the Appendix is empty.

Afterword

The back matter often includes one or more of an index, an afterword, acknowledgements, a bibliography, a colophon, or any other similar item. In the back matter, chapters do not produce a chapter number, but they are entered in the table of contents. If you are not using anything in the back matter, you can delete the back matter TeX field and everything that follows it.

Acknowledgement

Under construction

Bibliography

In writing the notes, I have consulted the classnotes by Prof. Xiao and textbook "Probability and Statistics" by DeGroot.