Mathematical Statistics

Dr.Kamel Rekab

Lecture Notes

Contents

Lecture 1

Set Theory

Definition 1.1: *The set* **S, (***sometimes also* **Ω),** *which is the set of all possible events of a random experiment is called the* ***sample space*** *for the experiment*

If the experiment consists of flipping a coin, then the sample space of this experiment is:

$$\mathrm{S} = \{\mathrm{H}, \mathrm{T}\}$$

If a fair coin is flipped twice, the sample space is:

$$\mathrm{S} = \{(H,H),(T,H),(H,T),(T,T)\}$$

Definition 1.2: *An* ***event (E)****, is any subset of sample space*

Let E be an event of "observing Head at least once" out of two flips of a coin, then:

$$E = \{(T,H),(H,T),(T,T)\}$$

For n events $E_1, E_2, .., E_n$,

Union:	$F_n = \bigcup_{i=1}^{n} E_i \ \Leftrightarrow if\ e \in F_n, then\ \exists\ i = 1,2, \dots n\ s.t. e \in E_i$	Example: $n = 2$: $E_1 \cup E_2 = E_2 \cup E_1$
Intersection:	$F_n = \bigcap_{i=1}^{n} E_i \Leftrightarrow if\ e \in F_n, then\ \forall\ i = 1,2, \dots n\ s.t. e \in E_i$	Example: $n = 2$: $E_1 \cup E_2 = E_2 \cup E_1$ (Also denoted as $E_1E_2 = E_2E_1$)
Complement:	$F = \bar{E}\ (or\ E^C, E') \Leftrightarrow \forall e, s.t. e \in F, e \notin E$	Example: $\bar{E} = S - E$
The largest event:	S, "the universe" (the true event, always occurs, 100%)	
The null event:	$\emptyset$, "impossible" (empty set, 0% chance of occurring)	

Theorem 1.3 (Distributive Law): *For any n events $E_1, E_2, .., E_n$,* defined on a sample space S:

a. $$F \cap \left[\bigcup_{i=1}^{n} E_i\right] = \bigcup_{i=1}^{n} (F \cap E_i) \tag{1.1}$$

b. $$F \cup \left[\bigcap_{i=1}^{n} E_i\right] = \bigcap_{i=1}^{n} (F \cup E_i) \tag{1.2}$$

Proof of (1.1) by Example:

If $S = \{1,2,3,4,5,6,7,8,9,10\}; F = \{1,2,7\}; E_1 = \{7,8,9\}; E_2 = \{2,7,9,10\}$, want to show that:

$$F \cap [E_1 \cup E_2] = [F \cap E_1] \cup [F \cap E_2]$$

$LHS: F \cap [E_1 \cup E_2] = \{1,2,7\} \cap \{2,7,8,9,10\} = \{2,7\}$

$RHS: [F \cap E_1] \cup [F \cap E_2] = \{7\} \cup \{2,7\} = \{2,7\}$

$\Longrightarrow LHS = RHS$

Proof of (1.2) by Definition:

Let $A = F \cup [\bigcap_{i=1}^{n} E_i]$ and $B = \bigcap_{i=1}^{n}(F \cup E_i)$, in order to prove (1.2), it suffices to show that both $A \subset B$ and $B \subset A$ hold. Let e be any element in A, i.e. $e \in F \cup [\bigcap_{i=1}^{n} E_i]$, by definition of union and intersection:

$$e \in F \cup \left[\bigcap_{i=1}^{n} E_i\right] \Longrightarrow e \in F \; or \; e \in \bigcap_{i=1}^{n} E_i$$

Case I: If $e \in F$, then $e \in F \cup E_i, \forall i = 1, \dots, n$, therefore $e \in \bigcap_{i=1}^{n}(F \cup E_i)$

Case II: If $e \in \bigcap_{i=1}^{n} E_i$, then $e \in E_i, \forall i = 1, \dots, n$, so that $e \in F \cup E_i, \forall \; i = 1, \dots, n$, i.e. $e \in \bigcap_{i=1}^{n}(F \cup E_i)$

Thus, the above two cases confirm that $e \in F \cup [\bigcap_{i=1}^{n} E_i] \Longrightarrow e \in \bigcap_{i=1}^{n}(F \cup E_i) \Longrightarrow A \subset B$

Similarly, $B \subset A$ and the final conclusion of (1.2) can be reached. ■

Theorem 1.4 (Demorgan's Law): *For any n events* $E_1, E_2, \dots, E_n$, defined on a sample space S:

a.
$$\left(\bigcup_{i=1}^{n} E_i\right)^{C} = \bigcap_{i=1}^{n}\left(E_i{}^{C}\right) \tag{1.3}$$

b.
$$\left(\bigcap_{i=1}^{n} E_i\right)^{C} = \bigcup_{i=1}^{n}\left(E_i{}^{C}\right) \tag{1.4}$$

Theorem 1.5 (Mutually Exclusive): *Two events E_1 and E_2 are disjoint* ***mutually exclusive*** *(or* ***disjoint)*** *if they do not overlap, or:*

$$E_1 \cap E_2 = \emptyset$$

For example, if we throw a fair die twice and let:

$$E = \{Sum\ of\ the\ up\ faces\ is\ 10\}$$
$$F = \{Product\ of\ the\ up\ faces\ is\ 30\}$$

What is the event that either E or F happends?

$S = \{(1,1), (1,2), (1,3), \dots (6,6)\}(36\ outcomes); E = \{(4,6), (5,5), (6,4)\}; F = \{(5,6), (6,5)\}$

Clearly, E and F don't overlap, or they are disjoint sets. Therefore, the union of E and F is simply by joining their outcomes together, i.e. $E \cup F = \{(4,6), (5,5), (6,4), (5,6), (6,5)\}$.

Lecture 2

Probability Theory

Definition 2.1: *Let **S** be the sample space of a random experiment. The **probability** is a **function** that maps an event **E** to $\boldsymbol{P(E)}$ that satisfies the following axioms:*

(i) $P(E) \geq 0$, *for all* $E \subset S$

(ii) $P(S) = 1$

(iii) *If* $E_1, E_2, \ldots$,*are pairwise disjoint, i.e.* $E_i \cap E_j = \emptyset$, *for* $i \neq j$,*then:*

$$P\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} P(E_i) \tag{2.1}$$

$\boldsymbol{P(E)}$ *is also interpreted as the chance that event* E *occurs*

Theorem 2.2: *For any events* E *and* F, *the following are true:*

a. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$ (2.2*a*)

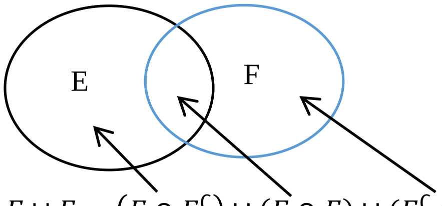
b. $P(E^{\mathrm{C}}) = 1 - P(E)$ (2.2*b*)

c. $P(\emptyset) = 0$ (2.2*c*)

d. $If\ E \subset F, then\ P(E) \leq P(F)$ (2.2*d*)

Proof:

a. For any two events E and F,



$E \cup F = (E \cap F^{\mathrm{C}}) \cup (E \cap F) \cup (E^{\mathrm{C}} \cap F)$, $E = (E \cap F) \cup (E^{\mathrm{C}} \cap F)$, $F = (E \cap F^{\mathrm{C}}) \cup (E \cap F)$

And obviously, $(E \cap F^{\mathrm{C}})$, $(E \cap F)$, $(E^{\mathrm{C}} \cap F)$ are pairwise disjoint

By the *third axiom*, we should have the following equalities:

$$P(E \cup F) = P(E \cap F^{\mathrm{C}}) + P(E \cap F) + P(E^{\mathrm{C}} \cap F) \tag{2.2a.1}$$

$$P(E) = P(E \cap F) + P(E \cap F^{\mathrm{C}})$$

$$\Longrightarrow P(E \cap F^{\mathrm{C}}) = P(E) - P(E \cap F) \tag{2.2a.2}$$

$$P(F) = P(E \cap F) + P(E^{\mathrm{C}} \cap F)$$

$$\Longrightarrow P(E^{\mathrm{C}} \cap F) = P(F) - P(E \cap F) \tag{2.2a.3}$$

Combining (2.2*a*.1), (2.2*a*.2), (2.2*a*.3), gives:

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

b. For the sample space S and any event E, it is easy to see that $E^{\mathrm{C}} \cap E = \emptyset, E^{\mathrm{C}} \cup E = S$. Therefore

$$P\left(E^{\mathrm{C}} \cup E\right) = P\left(E^{\mathrm{C}}\right) + P(E) \tag{2.2b.1}$$

$$P\left(E^{\mathrm{C}} \cup E\right) = P(S) = 1 \tag{2.2b.2}$$

(2.2*b*.1) and (2.2*b*.2) together will give (2.2*b*).

c. Immediately implied from (2.2*b*.1), (2.2*b*.2) and equalities (2.2*a*), (2.2*b*) we have:

$$\begin{aligned} P(\emptyset) \;=\; P\left(E^{\mathrm{C}} \cap E\right) \;&=\; P\left(E^{\mathrm{C}} \cup E\right) - P\left(E^{\mathrm{C}}\right) - P(E) \\ &=\; P(S) - P\left(E^{\mathrm{C}}\right) - P(E) \\ &=\; 1 - [1 - P(E)] - P(E) = 0 \end{aligned}$$

d. Let G be a non-empty event, such that $E \cap G = \emptyset, E \cup G = F$, either by applying the *third axiom* or equality (2.2*a*), we have:

$$P(F) - P(E) = P(G) \tag{2.2d.1}$$

The first axiom gives that $P(G) \geq 0$ and together with (2.2*d*.1), we have:

$$P(F) - P(E) \geq 0$$

which immediately gives (2.2*d*). ■

Theorem 2.3 (Inclusion-Exclusion Identity):

a. *For any three events A, B, C,*

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{2.3a}$$

b. *For a finite sequence of events $E_1, E_2 \ldots E_n$,*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i \leq j}^{n} P\left(E_i \cap E_j\right) + \sum_{i \leq j \leq k}^{n} P\left(E_i \cap E_j \cap E_k\right) - \cdots (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right) \tag{2.3b}$$

Proof:

a. Let A, B, C be any events (not necessarily disjoint),

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

b. Let $E_1, E_2 \ldots E_n$ be a sequence any events (not necessarily disjoint). By math induction,

First, for the base case $n = 1$, it is clearly true that: $P(E) = P(E)$

Assume it is true for the case of $n - 1$, namely,

$$P\left(\bigcup_{i=1}^{n-1} E_i\right) = \sum_{i=1}^{n-1} P(E_i) - \sum_{i \leq j}^{n-1} P\left(E_i \cap E_j\right) + \sum_{i \leq j \leq k}^{n-1} P\left(E_i \cap E_j \cap E_k\right) - \cdots + (-1)^{n} P\left(\bigcap_{i=1}^{n-1} E_i\right) \tag{2.3b.1}$$

Then for the case of n, by (2.2a),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left[\left(\bigcup_{i=1}^{n-1} E_i\right) \cup E_n\right] = P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n) - P\left[\left(\bigcup_{i=1}^{n-1} E_i\right) \cap E_n\right]$$

(by 1.1)
$$= P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n) - P\left[\bigcup_{i=1}^{n-1} (E_i \cap E_n)\right]$$

$$= P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n) - \left[\sum_{i=1}^{n-1} P(E_i \cap E_n) - \sum_{i\le j}^{n-1} P\left((E_i \cap E_n) \cap E_j\right) + \cdots + (-1)^n P\left(\bigcap_{i=1}^{n-1} E_i \cap E_n\right)\right]$$

(by 2.3b.1)
$$= \left[\sum_{i=1}^{n-1} P(E_i) - \sum_{i\le j}^{n-1} P(E_i \cap E_j) + \sum_{i\le j\le k}^{n-1} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^n P\left(\bigcap_{i=1}^{n-1} E_i\right)\right] + P(E_n)$$

$$- \left[\sum_{i=1}^{n-1} P(E_i \cap E_n) - \sum_{i\le j}^{n-1} P\left((E_i \cap E_n) \cap E_j\right) + \cdots + (-1)^n P\left(\bigcap_{i=1}^{n-1} E_i\right) + (-1)^n P\left(\bigcap_{i=1}^{n} E_i\right)\right]$$

$$= \left[\sum_{i=1}^{n-1} P(E_i) + P(E_n)\right] - \left[\sum_{i\le j}^{n-1} P(E_i \cap E_j) + \sum_{i=1}^{n-1} P(E_i \cap E_n)\right] + \cdots + (-1)\cdot(-1)^{n-2} P\left(\bigcap_{i=1}^{n} E_i\right)$$

$$= \sum_{i=1}^{n} P(E_i) - \sum_{i\le j}^{n} P(E_i \cap E_j) + \sum_{i\le j\le k}^{n} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n-1} P(\bigcap_{i=1}^{n} E_i) \qquad ■$$

Theorem 2.4 (Boole's Inequality): *Let $E_1, E_2 \ldots E_n$ be sequence of any events,*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \le \sum_{i=1}^{n} P(E_i) \qquad (2.4)$$

Proof: By math induction,

First, for the base case $n = 1, P(E) = P(E)$, and of course $P(E) \le P(E)$

Assume it is true for the case of $n - 1$, namely,

$$P\left(\bigcup_{i=1}^{n-1} E_i\right) \le \sum_{i=1}^{n-1} P(E_i) \qquad (2.4.1)$$

Then for the case of n,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left[\left(\bigcup_{i=1}^{n-1} E_i\right) \cup E_n\right] = P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n) - P\left[\left(\bigcup_{i=1}^{n-1} E_i\right) \cap E_n\right] \le P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n)$$

(by 2.4.1)
$$\le \sum_{i=1}^{n-1} P(E_i) + P(E_n) = \sum_{i=1}^{n} P(E_i) \qquad ■$$

Theorem 2.5 (Bonferroni Inequality):

a. *For any two events E, F,*

$$P(E \cap F) \geq P(E) + P(F) - 1 \qquad (2.5a)$$

b. *For a finite sequence of events $E_1, E_2 \dots E_n$,*

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - (n-1) \qquad (2.5b)$$

Proof:

a. For any two events E, F, which are not necessarily disjoint, it is apparent that $E \cup F \subset S$ and by (2.2*d*) and the *second axiom*:

$$P(E \cup F) \leq P(S) = 1 \qquad (2.5a.1)$$

And by (2.2*a*) in Theorem 2.2,

$$P(E) + P(F) - P(E \cap F) = P(E \cup F) \qquad (2.5a.2)$$

Combining (2.5*a*.1) and (2.5*a*.2), gives

$$P(E) + P(F) - P(E \cap F) \leq 1$$

which gives (2.5*a*) after some rearranging.

b. Let $E_1, E_2 \dots E_n$ be a sequence any events (not necessarily disjoint). Apply *Boole's Inequality* to the complements of all the events, then,

$$P\left(\bigcup_{i=1}^{n} {E_i}^C\right) \leq \sum_{i=1}^{n} P\left({E_i}^C\right) = \sum_{i=1}^{n} [1 - P(E_i)] \qquad (2.5b.1)$$

Apply Demorgan's law on left-hand side of (2.5*b*.1),

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} {E_i}^C\right) &= P\left[\left(\bigcap_{i=1}^{n} E_i\right)^C\right] \\ &= 1 - P\left(\bigcap_{i=1}^{n} E_i\right) \qquad (2.5b.2) \end{aligned}$$

Replace the left-hand side of (2.5*b*.1) with (2.5*b*.2) and get:

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq 1 - \sum_{i=1}^{n} [1 - P(E_i)] \qquad (2.5b.3)$$

The right-hand side of inequality (2.5*b*.3) can be further simplified to be $1 - n + \sum_{i=1}^{n} P(E_i)$, which will finally give (2.5*b*). ■

Theorem 2.6: *Let $E_1, E_2 \dots E_n$ form a* ***partition*** *of S. Let F be any event, then,*

$$P(F) = \sum_{i=1}^{n} P(F \cap E_i) \qquad (2.6)$$

Proof: Since $E_1, E_2 \dots E_n$ form a **partition**, we have that i). $E_i \cap E_j = \emptyset$ for all $i \neq j$, and ii). $\bigcup_{i=1}^{n} E_i = S$. Hence,

$$F = F \cap S = F \cap \left(\bigcup_{i=1}^{n} E_i \right) = \bigcup_{i=1}^{n} (F \cap E_i)$$

where the last equality comes from Distributive law. We then have,

$$P(F) = P\left(\bigcup_{i=1}^{n} (F \cap E_i) \right)$$

Since $E_1, \dots E_n$ are disjoint, so are $F \cap E_i$, for $i = 1, \dots n$, and from the *third axiom* of probability we have,

$$P\left(\bigcup_{i=1}^{n} (F \cap E_i) \right) = \sum_{i=1}^{n} P(F \cap E_i)$$

establishing (2.6). ■

Theorem 2.7 *For a sequence of events* $\{E_n\}_{n=1,2,\dots}$ *such that* $E_1 \subset E_2 \subset \cdots \subset E_n \uparrow E$, *then*

$$\lim_{n\to\infty} P(E_n) = P(E)$$

Proof: Since $E_1 \subset E_2 \subset \cdots \subset E_n \subset \cdots$, define $A_1 = E_1, A_2 = E_1^C \cap E_2, \dots, A_n = E_{n-1}^C \cap E_n, \dots$, so that $\{A_n\}$ are disjoint sets, by the *third axiom* of probability, we have,

$$\lim_{n\to\infty} P(E_n) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i \right) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{\infty} P(A_i) = P\left(\bigcup_{i=1}^{\infty} A_i \right) = P(E) \qquad ■$$

Lecture 3

Counting

Theorem 3.1 (General Principle of Counting*):* *If an experiment has* n_1 *possible outcomes, and each outcome has* n_2 *possible outcomes. Then the total number of outcomes is* $n_1 \cdot n_2$

Definition 3.2 (Permutation): *number of permutations for* k *objects is*

$$k! = k \times (k-1) \times \dots \times 2 \times 1 \tag{3.1}$$

where $0! = 1$ *by convention*

Definition 3.3 (Combination): *To choose a set of* k *objects out of a total of* n *objects when order is irrelevant, then the number of such sets (combinations) is*

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \tag{3.2}$$

which reads as "n choose k". And it always holds that:

$$\binom{n}{k} = \binom{n}{n-k} \tag{3.3}$$

Example 3.3.1: Prove that $(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}$ is true by math induction

Proof: When $n = 1$, it is obvious that $(a+b)^1 = \binom{1}{0}a^0b^1 + \binom{1}{1}a^1b^0 = a + b$

Assume for the case of $n-1$, the equality holds as below,

$$(a+b)^{n-1} = \sum_{k=0}^{n-1}\binom{n-1}{k}a^k b^{n-1-k}$$

Based on this assumption, for the case of n,

$$\begin{aligned}
(a+b)^n &= \sum_{k=0}^{n-1}\binom{n-1}{k}a^k b^{n-1-k}\cdot(a+b) = \sum_{k=0}^{n-1}\binom{n-1}{k}\left[a^{k+1}b^{n-1-k} + a^k b^{n-k}\right] \\
&= \sum_{j=0}^{n-1}\binom{n-1}{j}a^{j+1}b^{n-1-j} + \sum_{k=0}^{n-1}\binom{n-1}{k}a^k b^{n-k} \\
&= \sum_{j=0}^{n-1}\binom{n-1}{(j+1)-1}a^{j+1}b^{n-(j+1)} + \sum_{k=0}^{n-1}\binom{n-1}{k}a^k b^{n-k} \\
&= \sum_{k=1}^{n}\binom{n-1}{k-1}a^k b^{n-k} + \sum_{k=0}^{n-1}\binom{n-1}{k}a^k b^{n-k} \\
&= \sum_{k=1}^{n-1}\binom{n-1}{k-1}a^k b^{n-k} + \sum_{k=1}^{n-1}\binom{n-1}{k}a^k b^{n-k} + \binom{n-1}{n-1}a^n b^{n-n} + \binom{n-1}{0}a^0 b^{n-0} \\
&= \left[\sum_{k=1}^{n-1}\binom{n-1}{k-1} + \sum_{k=1}^{n-1}\binom{n-1}{k}\right]a^k b^{n-k} + 1\cdot a^n b^0 + 1\cdot a^0 b^n \\
\textit{(by 3.3)}\quad &= \sum_{k=1}^{n-1}\binom{n}{k}a^k b^{n-k} + \binom{n}{n}a^n b^0 + \binom{n}{0}a^0 b^n \\
&= \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}
\end{aligned}$$

By the principle of math induction, we can conclude that the stated equality is true. ■

<u>**Equally Likely Outcomes**</u>

Definition 3.4: If the sample space S is a finite set and all the outcomes in S are equally likely, then, for any event E,

$$P(E) = \frac{\#\ of\ elements\ in\ E}{\#\ of\ elements\ in\ S} \qquad (3.4)$$

Example 3.4.1: A fair die will be rolled 3 times, what's the probability that the sum is no less than 17?

Solution: Let $E = \{sum\ of\ 3\ dice\ is\ no\ less\ than\ 17\} = \{(5,6,6), (6,5,6), (6,6,5), (6,6,6)\}$

# of all outcomes of rolling a die 3 times $= 6 \times 6 \times 6 = 216$

# of ways of having 3 dice sum up to be no less than 17 = # of elements in E = 4

Thus, the probability of the event asked is:

$$P(E) = \frac{4}{216} = \frac{1}{54} \qquad \blacksquare$$

Example 3.4.2: There are 50 consecutive numbers, namely 1,2,3…..,50: 6 of the numbers are 'success' and the rest are 'failure'. Randomly choose 6 numbers from 1 to 50 without putting them back, what is the probability that "all 6 numbers are successes" ? What is the probability that "4 out of the 6 numbers are successes"?

Solution: Let $E = \{all\ 6\ numbers\ are\ sucesses\}$ and $F = \{4\ out\ of\ 6\ are\ sucesses\}$

# of all outcomes of selecting 6 cards without replacement = $\binom{50}{6} = 15{,}890{,}700$

# of ways of getting 6 successes = 1 (there is only one possible combination of getting 6 cards all success)

# of ways of getting 4 successes (and two others should be failures) = $\binom{6}{4}\binom{44}{2} = 14{,}190$

Thus,

$$P(E) = \frac{1}{15{,}890{,}700}; \qquad P(F) = \frac{14{,}190}{15{,}890{,}700} = \frac{473}{529{,}690} \qquad \blacksquare$$

Conditional Probability and Independence

Definition 3.5: *If E and F are events in S, and $P(F) > 0$, then the conditional probability of E given F, written as $P(A|B)$, is*

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{3.5}$$

Theorem 3.6:

a. *If E and F are events in S,*

$$P(E \cap F) = P(F) \cdot P(E|F) \tag{3.6a}$$

b. *If $E_1, E_2, \dots, E_n$ are events in S,*

$$P\left(\bigcap_{i=1}^{n} E_i\right) = P(E_1 | \cap_{i=2}^{n} E_i) P(E_2 | \cap_{i=3}^{n} E_i) \dots P(E_n) \tag{3.6b}$$

Proof: (3.6*a*) can be easily derived from the definition of conditional probability (3.5). In order to show (3.6*b*), math induction will be used.

For the base case $n = 1$, $P(E) = P(E)$

Assume the equality (3.6*b*) is true for $n - 1$ events, i.e.

$$P\left(\bigcap_{i=1}^{n-1} E_i\right) = P(E_1)P(E_2|E_1)P(E_3|E_2 \cap E_1) \dots P\left(E_{n-1}\middle| \cap_{i=1}^{n-2} E_i\right)$$

Then, for the case of n events,

$$
\begin{aligned}
P\left(\bigcap_{i=1}^{n} E_i\right) &= P\left(\bigcap_{i=1}^{n-1} E_i \cap E_n\right) = P\left(E_n \middle| \cap_{i=1}^{n-1} E_i\right) P\left(\bigcap_{i=1}^{n-1} E_i\right) \\
&= P\left(E_n \middle| \cap_{i=1}^{n-1} E_i\right) P(E_1) P(E_2|E_1) P(E_3|E_2 \cap E_1) \ldots P\left(E_{n-1} \middle| \cap_{i=1}^{n-2} E_i\right) \\
&= P(E_1) P(E_2|E_1) P(E_3|E_2 \cap E_1) \ldots P\left(E_{n-1} \middle| \cap_{i=1}^{n-2} E_i\right) P\left(E_n \middle| \cap_{i=1}^{n-1} E_i\right) \qquad \blacksquare
\end{aligned}
$$

Theorem 3.7 (Bayes Theorem): *Let $E_1, E_2, \ldots \ldots E_n$ form a partition of S, and let F be any event. Then, for each $i = 1,2, \ldots, n$*

$$P(E_i|F) = \frac{P(E_i F)}{P(F)} = \frac{P(F|E_i) \cdot P(E_i)}{\sum_{i=1}^{n} P(F|E_i)\, P(E_i)} \tag{3.7}$$

Definition 3.8: *Two events E, F are independent, if*

$$P(E|F) = P(E) \tag{3.8.1}$$

$$\text{OR} \quad P(E \cap F) = P(E)P(F) \tag{3.8.2}$$

*Remark: Two events that are independent cannot be mutually exclusive.

Definition 3.9: *A collection of events $E_1, \ldots, E_n$ are mutually independent if for any sub-collection $E_{i1}, \ldots, E_{ik}$, we always have:*

$$P\left(\bigcap_{j=1}^{n} E_{ij}\right) = \prod_{j=1}^{n} P\left(E_{ij}\right) \tag{3.9}$$

Theorem 3.10: *Suppose that E and F are independent, then*

a. E^C *and* F *are independent*
b. E *and* F^C *are independent*
c. E^C *and* F^C *are independent*

Proof: By Bayes' Theorem

$$P(F) = P(F|E)P(E) + P\left(F\middle|E^{\mathrm{C}}\right)P\left(E^{\mathrm{C}}\right)$$

Since E and F are independence, by (3.8.1), we can then have

$$P(F) = P(F)P(E) + P\left(F\middle|E^{\mathrm{C}}\right)P\left(E^{\mathrm{C}}\right)$$

After some rearranging by putting only $P\left(F\middle|E^{\mathrm{C}}\right)P\left(E^{\mathrm{C}}\right)$ on one side and the rest on the other, we get:

$$P\left(F\middle|E^{\mathrm{C}}\right)P\left(E^{\mathrm{C}}\right) = P(F)P\left(E^{\mathrm{C}}\right)$$

where the left-hand side is nothing but $P(F \cap E^C)$, concluding the proof.

The second and third statements can also be proved in a similar fashion ■

Lecture 4

Random Variable

Definition 4.1 (Discrete Random Variable): *A **discrete random variable** is a random variable, which takes value in the countable set with positive probability*

Example 4.1.1: If an unfair coin is flipped twice, which yielding a 1/3 chance of getting a head and 2/3 chance of getting a tail. Let the random variable X denote the times of getting a head and find the probability for each possible value of X.

Solution: There are three possible values for X to take: 0,1,2, with their corresponding probabilities:

$$P(X=0)=P(\{T,T\})=\frac{2}{3}\cdot\frac{2}{3}=\frac{4}{9}$$

$$P(X=1)=P(H,T)+P(T,H)=\frac{1}{3}\cdot\frac{2}{3}\times 2=\frac{4}{9}$$

$$P(X=2)=P(H,H)=\frac{1}{3}\cdot\frac{1}{3}=\frac{1}{9}$$

■

Definition 4.2 (Probability Mass Function): *The **probability mass function (pmf)** of the discrete random variable X, is any function $P(X=x)$, or $p(x)$, satisfies:*

(i) $p(x)\geq 0, \forall x$

(ii) $\sum_{x} P(x)=1$

Example 4.2.1: $p(x)=\binom{n}{x}(0.3)^x(0.7)^{n-x}, x=0,1,2,\dots,n$, show that this $p(x)$ is a pmf

Solution: In order to show this, we just need to check the two conditions of pmf:

Condition (i) check: Since $\binom{n}{x}\geq 0, (0.3)^x>0, (0.7)^{n-x}>0\ \forall n,x$, then $p(x)\geq 0$

Condition (ii) check: $\sum_{x=0}^{n}\binom{n}{x}(0.3)^x(0.7)^{n-x}=(0.3+0.7)^n=1^n=1$

By satisfying both conditions, we can conclude that this $p(x)$ is a pmf ■

Definition 4.3: *The cumulative distribution function (cdf) of a **discrete** random variable, denoted by $F(t)$ is defined by*

$$F(t)=P(X\leq t)=\sum_{x\leq t} P(x)$$

Example 4.3.1: Construct and graph the cdf $F(t)$ based on the given probability distribution table.

Solution:

X	*0*	*1*	*2*
$p(x)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

$$F(t) = \begin{cases} 0, & t < 0 \\ \frac{4}{9}, & 0 \le t < 1 \\ \frac{4}{9} + \frac{4}{9} = \frac{8}{9}, & 1 \le t < 2 \\ 1, & t \ge 2 \end{cases}$$

The cdf is a step function as specified

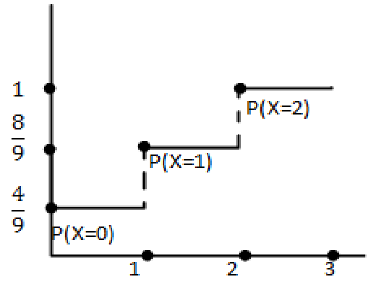


Figure 4.3.1 Illustration of the cdf

Definition 4.4 (Continuous Random Variable): *A* ***continuous random variable*** *is a variable that takes a variable continuously vary in an interval or many intervals (uncountable)*

Definition 4.5 (Probability Density Function): *The* ***probability density function (pdf)*** *of a continuous random variable X, is any function $f_X(x)$, sometimes written $f(x)$, satisfies:*

(i) $f(x) \ge 0, \forall x$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

Example 4.5.1: Given $f(x) = \begin{cases} 2e^{-2x}, & x \ge 0 \\ 0, & O.W. \end{cases}$, show that $f(x)$ is a pdf.

Solution: In order to show this, we just need to check the two conditions of pdf:

Condition (i) check: Since $e^{-2x} \ge 0, \forall x$, then $f(x) \ge 0$

Condition (ii) check: $\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{0} 0dx + \int_{0}^{+\infty} 2e^{-2x}\,dx = 2\left[-\frac{e^{-2x}}{2}\right]\Big|_0^{\infty} = -[0-1] = 1$

By satisfying both conditions, we can conclude that this $f(x)$ is a pdf ■

Definition 4.6: *The cumulative distribution function (cdf) of a* ***continuous*** *random variable, denoted by $F(t)$ is defined by:*

$$F(t) = P(X \le t) = \int_{-\infty}^{t} f(x)dx$$

Example 4.6.1: $f(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & O.W. \end{cases}$. Find $F(t)$ and graph it

Solution:

$$F(t) = P(X \le t) = \begin{cases} \int_{-\infty}^{t} 0dx = 0\,, & t < 0 \\ \int_{-\infty}^{0} 0dx + \int_{0}^{t} 1dx = t, 0 \le t < 1 \\ \int_{0}^{1} 1\,dx + \int_{1}^{t} 0dx = 1, & t \ge 1 \end{cases}$$

The cdf is a continuous function as specified

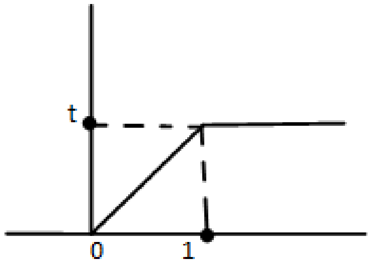


Figure 4.6.1 Illustration of the cdf

Example 4.6.2: Find $F(t)$ for the pdf specified in Example 4.5.1

Solution:

$$F(t) = P(X \le t) = \begin{cases} \int_{-\infty}^{t} 0dx = 0, & t < 0 \\ \int_{-\infty}^{0} 0dx + \int_{0}^{t} 2e^{-2x}dx = 2\left[-\frac{e^{-2x}}{2}\right]\Big|_{0}^{t} = 1 - e^{-2t}, & t \ge 0 \end{cases}$$

***Remark:** The relation between a pdf and its corresponding cdf is:

(i) $$F(t) = \int_{-\infty}^{t} f(x)dx$$

(ii) $$f(x) = F'(x) = \frac{d}{dx}F(x) \tag{4.1}$$

Theorem 4.7: *A function $F(t)$ is a cdf if and only if the following three conditions hold:*

(i) $\lim_{t\to-\infty} F(t) = 0$ *and* $\lim_{t\to\infty} F(t) = 1$

(ii) $F(t)$ *is non-decreasing*

(iii) $F(t)$ *is right continuous; that is, for every* x_0, $\lim_{t\downarrow x_0} F(t) = F(x_0)$

This theorem also holds for the discrete case where is proof will be very similar only by changing integral with sum.

Proof: *(i)* can be easily shown by applying the definition formula of cdf, that is

$$\lim_{t\to-\infty} F(t) = \lim_{t\to-\infty} \int_{-\infty}^{t} f(x)dx = \int_{-\infty}^{-\infty} f(x)dx = 0$$

and also by property of a pdf, we have

$$\lim_{t\to\infty} F(t) = \lim_{t\to\infty} \int_{-\infty}^{t} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

In order to establish *(ii)*, let $t_1 < t_2 < \cdots < t_n < t_{n+1} < \cdots < t$, then,

$$F(t_n) \ = \int_{-\infty}^{t} f(x)dx = \int_{-\infty}^{\infty} f(x)I_{\{x\le t_n\}}dx$$

Defne $g(x,t_n) = f(x)I_{\{x\le t_n\}}$, since $I_{\{x\le t_n\}} \le I_{\{x\le t_{n+1}\}}$, and $0 \le f(x) \le 1$, then $\forall x$

$$g(x,t_1) \le \cdots \le g(x,t_n) \le g(x,t_{n+1}) \le \cdots \le g(x,t)$$

Hence,

$$\int_{-\infty}^{\infty} g(x,t_n)dx \le \int_{-\infty}^{\infty} g(x,t_{n+1})dx\,, for\ n \ge 1$$

Or equivelently,

$$F(t_n) = \int_{-\infty}^{\infty} f(x)I_{\{x\le t_n\}}dx \le \int_{-\infty}^{\infty} f(x)I_{\{x\le t_{n+1}\}}dx = F(t_{n+1})$$

In order to show *(iii)*, let $t > t_2 > \cdots > t_n > t_{n+1} > \cdots > x_0$, then, $\forall x$

$$g(x,t) \ge \cdots \ge g(x,t_n) \ge g(x,t_{n+1}) \ge \cdots \ge g(x,x_0)$$

where $g(x,t_n) = f(x)I_{\{x\le t_n\}}$ and $g(x,x_0)$ is bounded and integrable. Thus, by ***Monotone Convergence Theorem***[1], we have the following limit holds for each x_0,

$$\begin{aligned}\lim_{t\downarrow x_0} F(t) = \lim_{n\to\infty} F(t_n) \ &= \lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)I_{\{x\le t_n\}}dx = \int_{-\infty}^{\infty} \lim_{n\to\infty} f(x)I_{\{x\le t_n\}}dx \\ &= \int_{-\infty}^{\infty} f(x)I_{\{x\le x_0\}}dx = \int_{-\infty}^{x_0} f(x)dx = F(x_0)\end{aligned}$$

■

Lecture 5

Distributions of Functions of a R.V. and Expectation

Theorem 5.1: *Let X has pdf $f_X(x)$ and $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ definted as $\mathcal{X} = \{x: f_X(x) > 0\}$ and $\mathcal{Y} = \{y: y = g(x), x \in \mathcal{X}\}$. Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by:*

$$f_Y(y) = \begin{cases} f_X\big(g^{-1}(y)\big)\left|\dfrac{d}{dy}g^{-1}(y)\right|, & y \in \mathcal{Y} \\ 0, & O.W. \end{cases} \tag{5.1}$$

[1] Monotone Convergence Theorem states that : If $\{f_n\}$ is a sequence of measurable functions, with $0 \le f_n \le f_{n+1}$ for every n, then $\int \lim_{n\to\infty} f_n\, d\mu = \lim_{n\to\infty} \int f_n d\mu$

Proof: For $g(X)$ is monotone, there are only two possible ways for $g(X)$ to behave, as long as $y \in \mathcal{Y}$

(i) g' exists and $g' > 0$ (g is an increasing function)
(ii) g' exists and $g' < 0$ (g is a decreasing function)

For case (i),

Step i).
$$F_Y(y) = P(Y \le y) = P[g(X) \le y] = P\{g^{-1}[g(X)] \le g^{-1}(y)\}$$
$$= P[X \le g^{-1}(y)] = F_X\left(g^{-1}(y)\right)$$

Step ii).
$$f_Y(y) = \frac{d}{dy}[F_Y(y)] = f_X\left(g^{-1}(y)\right) \cdot \frac{d}{dy}[g^{-1}(y)]$$
$$= f_X\left(g^{-1}(y)\right) \cdot \left|\frac{d}{dy}[g^{-1}(y)]\right| \left(\text{for } \frac{d}{dy}[g^{-1}(y)] > 0\right)$$

For case (ii),

Step i).
$$F_Y(y) = P(Y \le y) = P[g(X) \le y] = P\{g^{-1}[g(X)] \ge g^{-1}(y)\}$$
$$= 1 - P[X \le g^{-1}(y)] = 1 - F_X\left(g^{-1}(y)\right)$$

Step ii).
$$f_Y(y) = \frac{d}{dy}[1 - F_Y(y)] = f_X\left(g^{-1}(y)\right) \cdot \left(-\frac{d}{dy}[g^{-1}(y)]\right)$$
$$= f_X\left(g^{-1}(y)\right) \cdot \left|\frac{d}{dy}[g^{-1}(y)]\right| \left(\text{for } \frac{d}{dy}[g^{-1}(y)] < 0\right)$$

After checking $f_Y(y)$ does represent a valid density function, we can conclude that if $g(X)$ is monotone,

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right|, y \in \mathcal{Y}$$ ■

Example 5.1.1: X has pdf $f_X(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & O.W. \end{cases}$. Let $Y = g(X) = \log\left(\frac{1}{X}\right)$, find $f_Y(y)$.

Solution:

Step i). Derive the cdf of Y

$$F_Y(y) = P(Y \le y) = P(\log\left(\frac{1}{X}\right) \le y) = P\left(\frac{1}{X} \le e^y\right) = P(X \ge e^{-y})$$
$$= 1 - P(X \le e^{-y}) = 1 - F_X(e^{-y})$$

Step ii). Apply equality(4.1) and derive the pdf of Y

$$f_Y(y) = F_Y'(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}[1 - F_X(e^{-y})]$$
$$= -f_X(e^{-y}) \cdot e^{-y} \cdot (-1) = e^{-y} f_X(e^{-y})$$

Step iii). Relate to the specific pdf of X

$$f_X(e^{-y}) = \begin{cases} 1, & e^{-y} \le 1 \\ 0, & O.W. \end{cases} = \begin{cases} 1, & y \ge 0 \\ 0, & y < 0 \end{cases}$$

Thus, $$f_Y(y) = \begin{cases} e^{-y}, & y \ge 0 \\ 0, & y < 0 \end{cases} \quad (5.1.1)$$

Step iv). Check if the resulting pdf $f_Y(y)$ is a valid density function by verifying if the two conditions in Definition 4.5 are satisfied. (The proof of it is similar to that in Example 4.5.1)

The other way to show it is simply by following Theorem 5.1, for $\log\left(\frac{1}{X}\right)$ is a monotone function and the result will be the same as (5.1.1). ■

Example 5.2: When $g(X)$ is not a one-to-one function, we can still find the pdf for $Y = g(X)$ for some cases. For example, let $f_X(x)$ be pdf of X and $Y = X^2$, find $f_Y(y)$.

Solution:

Step i).
$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^2 \le y) = P\left(|X| \le \sqrt{y}\right), y \ge 0 \\ &= P\left(-\sqrt{y} \le X \le \sqrt{y}\right) = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right) \end{aligned}$$

Step ii).
$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\left[F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right)\right] = f_X\left(\sqrt{y}\right) \cdot \frac{1}{2\sqrt{y}} - f_X\left(-\sqrt{y}\right) \cdot \left(-\frac{1}{2\sqrt{y}}\right) \\ &= \frac{1}{2\sqrt{y}}\left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right] \end{aligned}$$

After checking the validity of $f_Y(y)$ being a density function, we can finally have:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}\left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right], & y \ge 0 \\ 0, & y < 0 \end{cases} \tag{5.2}$$

Example 5.2.1: X has pdf $f_X(x) = \begin{cases} 30x^2(1-x)^2, & 0 < x < 1 \\ 0, & O.W. \end{cases}$. Let $Y = g(X) = X^2$, find $f_Y(y)$.

Solution:

Step i). & ii). Derive the pdf of Y.

Since $g(X)$ is the same as in Example 5.2, we can just borrow the result from (5.2)

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}}\left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right], & y \ge 0 \\ 0, & y < 0 \end{cases}$$

Step iii). Relate to the specific pdf of X

$$f_X\left(\pm\sqrt{y}\right) = \begin{cases} 30y\left(1 \mp \sqrt{y}\right)^2, & 0 < \sqrt{y} < 1 \\ 0, & O.W. \end{cases} = \begin{cases} 30y\left(1 \mp \sqrt{y}\right)^2, & 0 < y < 1 \\ 0, & O.W. \end{cases}$$

Thus,
$$f_Y(y) = \begin{cases} 15\sqrt{y}\left[\left(1-\sqrt{y}\right)^2 + \left(1+\sqrt{y}\right)^2\right], & 0 < y < 1 \\ 0, & O.W. \end{cases} \tag{5.2.1}$$

Step iv). Check if the resulting pdf $f_Y(y)$ is a valid density function by verifying if the two conditions in Definition 4.5 are satisfied. (The proof of it is similar to that in Example 4.5.1). If so, (5.2.1) gives what we are looking for. ■

Expectation

Definition 5.3: *The* ***expectation****, or the* ***expected value****, or the* ***mean*** *of a random variable X, denotes by $E(X)$ (or μ or μ_X), is defined to be*

$$E(X) = \begin{cases} \sum_{x} xP(X = x)\,, & if\ X\ is\ discrete \\ \int_{R} xf_X(x)dx\,, & if\ X\ is\ continuous \end{cases} \tag{5.3}$$

Example 5.3.1: Suppose X has a $Binomial(n, p)$ distribution, that is, it has pmf given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0,1,2 \dots \dots, n, 0 < p < 1$$

Then, prove that $E(X) = np$

Proof: By definition (5.3),

$$E(X) = \sum_{k} kP(X = k) = \sum_{x=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} k \cdot \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

$$= n \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k} = n \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k}$$

Let $j = k - 1$ and make a change of variable,

$$E(X) = n \cdot \sum_{j=0}^{n} \binom{n-1}{j} p^{j+1} (1-p)^{n-1-j} = np \cdot \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} = np \cdot 1 = np \qquad \blacksquare$$

Think about the experiment of flipping a fair coin, the number of getting heads follows a binomial distribution with n to be the total number of flips and p to be $1/2$. If the coin is flipped for 10 times, then the expected value of number of heads can be calculated as $10 \times 1/2 = 5$.

Example 5.3.2: Suppose X has a $Geometric(p)$ distribution, that is, it has pmf given by

$$P(X = x) = pq^{x-1}, \quad x = 1,2 \dots \dots, \infty, 0 < p < 1, q = 1 - p$$

Then, prove that $E(X) = \dfrac{1}{p}$

Proof: By definition (5.3),

$$E(X) = \sum_{x=1}^{\infty} xpq^{x-1} = p \sum_{x=1}^{\infty} xq^{x-1} = p \sum_{x=1}^{\infty} \frac{d}{dq}(q^x) = p \frac{d}{dq}\left(\sum_{x=1}^{\infty} q^x\right) = p \lim_{n\to\infty} \frac{d}{dq}\left(\sum_{x=1}^{n} q^x\right)$$

$$= p \lim_{n\to\infty} \frac{d}{dq}\left(\frac{1-q^n}{1-q} q\right) = \frac{d}{dq}\left(\frac{q}{1-q}\right) = p\left[\frac{(1-q)+q}{(1-q)^2}\right]$$

$$= p\left[\frac{1}{(1-q)^2}\right] = p \cdot \frac{1}{p^2} = \frac{1}{p} \qquad \blacksquare$$

Lecture 6

Recall from last lecture that the expectation of a continuous random variable is defined as

$$E(X) = \int_R x f_X(x) dx \tag{6.1}$$

Example 6.1.1: Suppose X has a $Uniform(0,1)$ distribution, that is, it has pdf given by

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & O.W. \end{cases}$$

Find the expectation, $E(X)$

Solution: By definition (5.3),

$$E(X) \quad = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x\,dx = \frac{x^2}{2}\bigg|_0^1 = \frac{1}{2} \qquad \blacksquare$$

Example 6.1.2: Suppose X has an $Exponential(1)$ distribution, that is, it has pdf given by

$$f_X(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & O.W. \end{cases}$$

Find the expectation, $E(X)$

Solution: By definition (5.3) and integration by parts,

$$E(X) \quad = \int_0^{\infty} x \cdot e^{-x} dx = -e^{-x}x\Big|_0^{\infty} - \left(-\int_0^{\infty} e^{-x}\right) = 0 + [-e^{-x}]\Big|_0^{\infty} = 1 \qquad \blacksquare$$

Example 6.1.3: Suppose X has a $Normal(0,1)$ distribution, that is, it has pdf given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Find the expectation, $E(X)$

Solution: By definition (5.3) and integration by parts,

$$E(X) \quad = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}}\left[-e^{-\frac{x^2}{2}}\right]\Big|_{-\infty}^{\infty} = 0$$

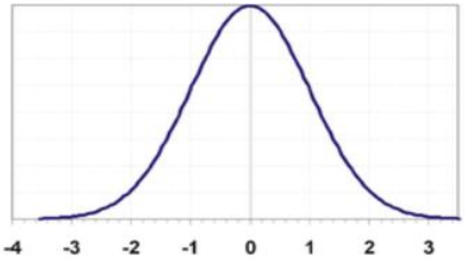


Figure 6.3 pdf of $Normal(0,1)$

Or it can be observed that the pdf is an even function (symmetric about $x = 0$), as shown in figure 6.3. Therefore, by property of even function, $g(-x) = g(x)$, we have:

$$E(X) \;= \int_{-\infty}^{\infty} x g(x) dx = \int_{-\infty}^{0} (-x) g(-x) dx + \int_0^{\infty} x g(x) dx$$

$$= \int_{-\infty}^{\infty} [-x + x] g(x)\,dx = 0$$

$\blacksquare$

Theorem 6.2: *Let X, Y, Z be random variables whose expectations exist and let a, b, c be constants. Then,*

a. $E(X+Y) = E(X) + E(Y)$ (6.2*a*)

b. $E(X-Y) = E(X) - E(Y)$ (6.2*b*)

c. $E(aX) = aE(X)$ (6.2*c*)

d. $E(aX+bY) = aE(X) + bE(Y)$ (6.2*d*)

e. $If\ Y \le X \le Z, then\ E(Y) \le E(X) \le E(Z)$ (6.2*e*)

Proof of this theorem will be given after conditional probability is introduced.

Example 6.2.1: Show that $E(X^2) \ge E^2(X)$

Proof: Let $E(X) = c$, and clearly $(X-c)^2 \ge 0$, by (6.2*e*), we have

$$E(X-c)^2 \ge 0$$

where the left-hand side can be expanded to be $E(X^2 - 2cX + c^2)$.Then, the linearity property(6.2*a*) and the condition $E(X) = c$ immediately give

$$E(X^2) - c^2 \ge 0$$

which is equivalent to $E(X^2) \ge c^2$, namely $E(X^2) \ge E^2(X)$ and the equality holds if and only $X = c$ ■

Expectation of function of a R.V. and Variance

Theorem 6.3 (Law of the Unconscious Statistician): *The* ***expectation****, or* ***expected value*** *of a random variable $g(X)$, denoted by $E[g(X)]$, is defined to be*

$$E(X) = \begin{cases} \sum_x g(x)P(X=x)\,, & if\ X\ is\ discrete \\ \int_R g(x)f_X(x)dx\,, & if\ X\ is\ continuous \end{cases} \tag{6.3}$$

Example 6.3.1: Suppose X has a $Uniform(1)$ distribution, that is, it has pdf given by

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & O.W. \end{cases}$$

Find the expectation of X^2, $E(X^2)$

Solution:

Approach I: By making change of variable. Let $Y = X^2$, Using similar technique for solving Example 5.2.1, we can get:

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] = \begin{cases} \frac{1}{2\sqrt{y}}, & 0 < y < 1 \\ 0, & O.W. \end{cases}$$

Then by definition (6.1),

$$E(X^2) = E(Y) = \int_0^1 y \cdot \frac{1}{2\sqrt{y}} = \int_0^1 \frac{\sqrt{y}}{2} = \frac{1}{2}\cdot\frac{y^{\frac{3}{2}}}{3/2}\Big|_0^1 = \frac{1}{3}$$

Approach II: By applying Theorem 6.3,

$$E(X^2) \quad = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

The second approach turns out to be less tedious then the first one. ■

Theorem 6.4 (Jensen Inequality): *If $g(X)$ is a convex function of X, then*

$$E\big(g(X)\big) \geq g[E(X)] \tag{6.4}$$

Proof:

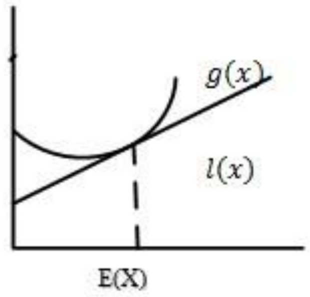


Figure 6.4

Let $l(x)$ be a tangent line to $g(x)$ at the point $(E(X), g[E(X)] = l[g(X)])$

Write $l(x) = a + bx$ for some constants a, b

As shown in Figure 6.4, by the convexity of $g(x)$, we always have:

$$g(x) \geq l(x) = a + bx$$

And since expectations preserve inequalities (6.2e):

$$\begin{aligned} E[g(X)] &\geq E(a + bX) \\ \textit{(by 6.2a)} &= a + bE(X) \\ &= l[E(X)] \\ &= g[E(X)] \end{aligned}$$

Or it can be proved by using Taylor expansion and expand the convex function $g(x)$ around $x = \mu$ for each value of X, that is

$$g(x) \;= g(\mu) + \frac{g'(\mu)(x-\mu)}{1!} + \frac{g''(x^*)(x-x^*)^2}{2!}$$

where $x^* \in [\min(x,\mu), \max(x,\mu)]$. Then, we take expectation on both sides and have,

$$E[g(X)] \;= g(\mu) + E(X-\mu)\frac{g'(\mu)}{1!} + \frac{E[g''(x^*)(x-x^*)^2]}{2!}$$

By convexity, we know that $g''(x) \geq 0, \forall x$, then the second power term has to be nonnegative, therefore,

$$E[g(X)] \;\geq g(\mu) + E(X-\mu)\frac{g'(\mu)}{1!} = g(\mu) = g[E(X)]$$

which is by simply following the identity that $E(X) = \mu$. ■

Example 6.4.1: Revisit Example 6.2.1, but use Jensen Inequality to show the inequality $E(X^2) \geq E^2(X)$.

Let $g(x) = x^2$, so $g'(x) = 2x$ and $g''(x) = 2 > 0$, indicating that $g(x)$ is a convex function. Thus, we can apply Jensen Inequality on $g(x)$ and validating the inequality $E(X^2) \geq E^2(X)$ to be true.

Definition 6.5: *The **variance** of a random variable X, denoted by $Var(X)$ (or σ^2, or σ_X^2), is defined as*

$$Var(X) \; = E[(X-\mu)^2] \; \left(or\; E\left[\left(X - E(X)\right)^2\right]\right) \tag{6.5}$$

*and the **standard deviation** of X, denoted by $SD(X)$ (or σ),is defined as $\sqrt{Var(X)}$.*

Theorem 6.6: *The **variance** of a random variable X can also be calculated by*

$$Var(X) \; = E(X^2) - \mu^2 \; \left(or\; E(X^2) - E^2(X)\right) \tag{6.6}$$

Proof: Expand the right-hand side of (6.5), which gives

$$\begin{aligned} Var(X) \; &= E[(X-\mu)^2] = E(X^2 - 2\mu X + \mu^2) \\ \textit{(by 6.2a)} \; &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

■

Example 6.6.1: Suppose X has a $Uniform(0,1)$ distribution, that is, it has pdf given by

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & O.W. \end{cases}$$

Find the variance, $Var(X)$

Solution: We have previously calculated that $E(X) = 1/2$ in Example 6.1.1 and $E(X^2) = 1/3$ in Example 6.3.1. Then, applying (6.6), we have

$$Var(X) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

■

Example 6.6.2: Suppose X has an $Exponential(1)$ distribution, that is, it has pdf given by

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & O.W. \end{cases}$$

Find the variance, $Var(X)$

Solution: We have previously calculated that $E(X) = 1$ in Example 6.1.2. To get $Var(X)$, we need to calculate $E(X^2)$ first, by (6.3), we have

$$E(X^2) \; = \int_0^\infty x^2 \cdot e^{-x}\, dx = -x^2 e^{-x} \Big|_0^\infty + 2\int_0^\infty xe^{-x} dx = 2$$

Using (6.6), we have

$$Var(X) = 2 - 1 = 1$$

■

Example 6.6.2: Suppose X has a $Poisson(\lambda)$ distribution, that is, it has pdf given by

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0,1,2\ldots\ldots$$

Find the variance, $Var(X)$

Solution: To get $Var(X)$, we need to calculate $E(X)$ and $E(X^2)$ first, by (5.3) and (6.3) we have

$$E(X) = \sum_{x=0}^{\infty} x\cdot\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\cdot\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\cdot\lambda\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}\lambda\cdot e^{\lambda} = \lambda$$

$$E(X^2) = \sum_{x=0}^{\infty} x^2\cdot\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\cdot\lambda\sum_{x=1}^{\infty} x\cdot\frac{\lambda^{x-1}}{(x-1)!} = e^{-\lambda}\cdot\lambda\sum_{y=0}^{\infty}(y+1)\frac{\lambda^y}{y!}$$

$$= e^{-\lambda}\cdot\lambda\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!} + e^{-\lambda}\cdot\lambda\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = e^{-\lambda}\cdot\lambda^2\sum_{z=0}^{\infty}\frac{\lambda^z}{z!} + \lambda = \lambda^2 + \lambda$$

Using (6.6), we have

$$Var(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

■

Example 6.6.3: Suppose X has a $Binomial(n, p)$ distribution, that is, it has pdf given by

$$p(x) = \binom{n}{k} p^k(1-p)^{n-k}, x = 1,2,\ldots n$$

Find the variance, $Var(X)$

Solution: We have shown that $E(X) = np$ in Example 5.3.1. Then we calculate (X^2) , by (6.3) we have

$$E(X^2) = \sum_{x=0}^{n} x^2\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=0}^{n} x^2\frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x}$$

$$= \sum_{x=0}^{n} nx\frac{(n-1)!}{(x-1)!\,(n-x)!}p^x(1-p)^{n-x} = np\sum_{x=1}^{n} x\binom{n-1}{x-1}p^{x-1}(1-p)^{n-x}$$

$$= np\sum_{y=0}^{n-1}(y+1)\binom{n-1}{y}p^y(1-p)^{n-1-y} \quad (y = x-1)$$

$$= \underbrace{np\sum_{y=0}^{n-1} y\binom{n-1}{y}p^y(1-p)^{n-1-y}}_{E(Y)=(n-1)p} + np\underbrace{\sum_{y=0}^{n-1}\binom{n-1}{y}p^y(1-p)^{n-1-y}}_{1}$$

*Since $\binom{n-1}{y}p^y(1-p)^{n-1-y}$ can be viewed as the distribution of $Binomial(n-1, p)$

Therefore, $E(X^2) = np\cdot(n-1)p + np\cdot 1 = n(n-1)p^2 + np = n^2p^2 - np^2 + np$

Using (6.6), we have

$$Var(X) = n^2p^2 - np^2 + np - (np)^2 = np - np^2 = np(1-p)$$

■

Theorem 6.7: *If X is a random variable with finite variance, then for any constants a and b,*

a. $Var(a) = 0$

b. $Var(aX + b) = a^2 Var(X)$

Proof: Since (*a*) is a direct application of (*b*), we will just prove (*b*). From the definition, we have

$$\begin{aligned} Var(aX+b) &= E[(aX+b) - E(aX+b)]^2 = E(aX - a\mu)^2 \\ &= E[(aX)^2 - 2(a\mu)(aX) + (a\mu)^2] = E[(aX)^2] - 2(a\mu)E(aX) + (a\mu)^2 \\ \text{(by 6.6)} \quad &= E[(aX)^2] - (a\mu)^2 = a^2[E(X^2) - \mu^2] = aVar(X) \end{aligned}$$ ■

Example 6.7.1: Suppose X is a random variable with expectation μ and variance σ^2. Let $Z = \frac{X-\mu}{\sigma}$, specify $E(Z)$ and $Var(X)$ in terms of μ and σ^2.

Proof: From Theorem 6.2 and Theorem 6.7, we have

$$\begin{aligned} E(X) &= E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = 0 \\ Var(X) &= Var\left(\frac{X-\mu}{\sigma}\right) = \frac{Var(X)}{\sigma^2} = 1 \end{aligned}$$ ■

Lecture 7

Theorem 7.1 (Markov's Inequality): *Let X be a random variable with finite expected value of $|X|^p$, then for any $a > 0$,*

$$P\{|X| > a\} \leq \frac{E[|X|^p]}{a^p} \tag{7.1}$$

Proof: By definition of expectation, for $a > 0$ and $p \geq 1$,

$$\begin{aligned} E[|X|^p] &= \int_R |X|^p f_X(x)dx = \int_{\{|X|\leq a\}} |X|^p f_X(x)dx + \int_{\{|X|>a\}} |X|^p f_X(x)dx \\ &\geq \int_{\{|X|>a\}} |X|^p f_X(x)dx \left(since \int_{\{|X|\leq a\}} |X|^p f_X(x)dx \geq 0 \right) \\ &\geq \int_{\{|X|>a\}} |a|^p f_X(x)dx = a^p \int_{\{|X|>a\}} f_X(x)dx = a^p \cdot P\{|X| > a\} \end{aligned}$$

The inequality (7.1) can be immediately obtained after some rearranging. ■

Example 7.1.1: Let $a > 0$ and $p = 2$, then for a random variable X, by following (7.1),

$$P\{|X| > a\} \leq \frac{E[X^2]}{a^2} \tag{7.1.1}$$

Theorem 7.2 (Chebychev's Inequality): *If X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $k > 0$,*

$$P\{|X-\mu| \geq k\sigma\} \leq \frac{1}{k^2} \tag{7.2}$$

Proof: By Markov's Inequality and let $a = k\sigma$, and the random variable to be $X - \mu$, then,

$$P\{|X-\mu| \geq k\sigma\} \geq \frac{E|X-\mu|^2}{k^2\sigma^2}$$

where $E|X-\mu|^2 = E(X-\mu)^2 = \sigma^2$ by definition of variance. Therefore,

$$P\{|X-\mu| \geq k\sigma\} \geq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \qquad \blacksquare$$

Moment and Moment Generating Function

Definition 7.3: *For each integer n, the n*^th^ ***moment** of X, is defined to be $E(X^n)$. And the n*^th^ ***central moment** of X is defined to be $E[(X-\mu)^n]$*

Some important moments include: the **first moment,** or the **mean,** $E(X)$; **the second central moment,** or the **variance**, $E[(X-\mu)^2]$; the **third central moment,** or the **skewness,** $E[(X-\mu)^3]$; the **fourth moment**, or the **kurtosis**, $E(X^4)$

Definition 7.4: *Let X be a random variable with pdf $f_X(x)$, or pmf $p(x)$. The moment generating function (mgf), denoted by $M_X(t)$, is defined as,*

$$M_X(t) = E(e^{tX}) = \begin{cases} M_X(t) = \int_R e^{tx} \cdot f_X(x)dx, & X \sim C.R.V. \\ M_X(t) = \sum_x e^{tx}p(x), & X \sim D.R.V. \end{cases} \tag{7.3}$$

Example 7.4.1: Suppose X has a $Poisson(\lambda)$ distribution, find the mgf $M_X(t)$

Solution: Based on the pmf given in Example 6.6.2, we can get,

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty} \frac{e^{tx-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{e^{tx}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(e^t\lambda)^x}{x!} = e^{-\lambda} \cdot e^{(e^t\lambda)} = e^{\lambda(e^t-1)} \end{aligned} \qquad \blacksquare$$

Example 7.4.2: Suppose X has a $Exponential(\lambda)$ distribution, that is, it has pdf given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & O.W.' \end{cases} \qquad \lambda > 0$$

Find the mgf $M_X(t)$

Solution: Based on the pdf given, we can get,

$$
\begin{aligned}
M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx}\cdot\lambda e^{-\lambda x}dx = \int_0^\infty \lambda e^{(t-\lambda)x}dx \\
&= \frac{\lambda}{t-\lambda}e^{(t-\lambda)x}\Big|_0^\infty = \frac{\lambda}{t-\lambda}(0-1) = \frac{\lambda}{\lambda-t}\,, \quad for\; t<\lambda
\end{aligned}
\qquad ■
$$

Example 7.4.3: Suppose X has a $Normal(0{,}1)$ distribution, find the mgf $M_X(t)$

Solution: Based on the pdf given in Example 6.1.3, we can get,

$$
\begin{aligned}
M_X(t) &= E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{tx-\frac{x^2}{2}}dx \\
&= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2-2tx}{2}}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2-2tx+t^2-t^2}{2}}dx \\
&= e^{\frac{1}{2}t^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-t)^2}{2}}dx \qquad \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-t)^2}{2}}\; is\; the\; pdf\; of\; N(t,1)\right) \\
&= e^{\frac{1}{2}t^2}
\end{aligned}
\qquad ■
$$

Theorem 7.5: If X has mgf $M_X(t)$, then $\forall\; n \in N$

$$M_X^{(n)}(0) = E(X^n) \tag{7.4}$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n}M_X(t)|_{t=0} \tag{7.5}$$

Proof: Assume the mgf is differentiable and we can differentiate under the integral sign, then

$$
\begin{aligned}
M_X'(t) &= \frac{d}{dt}E(e^{tX}) = \frac{d}{dt}\left[\int e^{tx}f_X(x)dx\right] = \int \frac{d}{dt}[e^{tx}f_X(x)]dx \\
&= \int f_X(x)\left[\frac{d}{dt}(e^{tx})\right]dx = \int xe^{tx}f_X(x)dx \\
&= E(X^1e^{tX}) && (7.5.1)
\end{aligned}
$$

where if take $t=0$, $M_X'(0) = E(X)$

Use math induction on (7.5.1), if $M_X^{(n-1)}(t) = E\left(X^{(n-1)}e^{tX}\right)$ is true, then

$$
\begin{aligned}
M_X^{(n)}(t) &= \frac{d}{dt}M_X^{(n-1)}(t) = \frac{d}{dt}E\left(X^{(n-1)}e^{tX}\right) = \frac{d}{dt}\left[\int x^{n-1}e^{tx}f_X(x)dx\right] \\
&= \int x^{n-1}f_X(x)\left[\frac{d}{dt}(e^{tx})\right]dx = \int x^n e^{tx}f_X(x)dx \\
&= E(X^n e^{tX}) && (7.5.2)
\end{aligned}
$$

Thus, let $t=0$, (7.5.2) will finally result in (7.4). ■

Another way to see the relation between mgf and moments is by applying *Taylor expansion* on $E(e^{tX})$. Recall that Taylor expansion of a real function $f(x)$ around a real number a is

$$f(a)+\frac{f'(a)}{1}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f^{(3)}(a)}{3!}(x-a)^3+\frac{f^{(4)}(a)}{4!}(x-a)^4+\cdots$$

Let $f(X)=e^{tX}, a=0$, and by linearity of expectation, we can have

$$M_X(t)=E(e^{tX})=1+tE(X)+\frac{t^2}{2!}E(X^2)+\frac{t^3}{3!}E(X^3)+\frac{t^4}{4!}E(X^4)+\cdots$$

or can be compactly written as

$$M_X(t)=\sum_{k=0}^{\infty}\frac{t^k}{k!}E\left(X^k\right) \tag{7.6}$$

which shows that the mgf can approximately represents a linear sum of all the moments. ■

Example 7.5.1: To illustrate Theorem 7.5, consider a specific case where X has a distribution of $Binomial(n,p)$. Find $M_X(t)$ and $E(X)$.

Solution: Based on the pmf given in Example 5.3.1, we can get,

$$\begin{aligned} M_X(t) &= E(e^{tX})=\sum_{x=0}^{n}e^{tx}\binom{n}{x}p^xq^{n-x}=\sum_{x=0}^{n}\binom{n}{x}(e^tp)^xq^{n-x} \\ &= (e^tp+q)^n \end{aligned}$$

By Theorem 7.5, the expectation can be calculated by making use of the mgf,

$$\begin{aligned} E(X)=M_X'(0) &= \frac{d}{dt}M_X(t)|_{t=0}=\frac{d}{dt}(e^tp+q)^n|_{t=0}=n(e^tp+q)^{n-1}\cdot pe^t|_{t=0} \\ &= n(e^0p+q)^{n-1}\cdot pe^0=np \end{aligned}$$ ■

Theorem 7.6 (Uniqueness of MGF): *Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist. If the moment generating functions also exist and $M_X(t)=M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u)=F_Y(u)$ for all u. Or simply saying that, the MGF is a one-to-one related to distribution.*

Proof of this theorem is beyond the scope of this class, which involves the theorem of *convergence of characteristic function*. The main application of this theorem is to determine what distribution a random variable has. For example, if the mgf of X exist and has the form $\left(\frac{1}{3}e^t+\frac{2}{3}\right)^5$, then it can be concluded that X should follow a distribution of $Binomial\left(5,\frac{1}{3}\right)$.

Theorem 7.7: *Let X be a random variable with mgf $M_X(t)$, then for constants a,b,*

$$M_{aX+b}(t)=M_X(at)\cdot e^{bt} \tag{7.7}$$

Proof: From the definition of mgf, we have

$$\begin{aligned} M_{aX+b}(t) &= E\left[e^{t(aX+b)}\right]=E\left[e^{(at)X+bt}\right] \\ &= E\left[e^{(at)X}\cdot e^{bt}\right]=M_X(at)\cdot e^{bt} \end{aligned}$$ ■

Example 7.7.1: Suppose Z has a $Normal(0,1)$ distribution. For constants μ and σ, define another random variable X such that $X = \sigma Z + \mu$, find the mgf for X

Solution: By Theorem 7.7, we have

$$M_X(t) = M_{\sigma Z+\mu}(t) = M_Z(\sigma t) \cdot e^{\mu t}$$

where $M_Z(\sigma t) = e^{\frac{1}{2}(\sigma t)^2}$ by Example 7.4.3. Therefore,

$$M_X(t) = e^{\frac{1}{2}(\sigma t)^2} \cdot e^{\mu t} = e^{\mu t+\frac{1}{2}\sigma^2 t^2} \qquad \blacksquare$$

Lecture 8

<u>Discrete Distributions</u>

Binomial Distribution

Definition 8.1: *A **Bernoulli trial** is a random experiment that:*

- *(i) with only two possible outcomes, either a Success (S), or a Failure(F)*
- *(ii) the probability of Success (p) remains the same for each trial*

*Let Y be result of a Bernoulli trial, then Y has a **Bernoulli distribution** which is:*

$$Y = \begin{cases} 1, & with\, probability\; p \\ 0, & with\, probability\; 1-p \end{cases}, \qquad 0 \le p \le 1$$

If n such identical independent Bernoulli trials are performed and let a random variable X counts the number of total successes in n trials, then X can be modeled by a $\boldsymbol{Binomial(n,p)}$ distribution, often expressed as $\boldsymbol{X \sim Bin(n,p)}$, which has a pmf given by:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0,1,2,\dots\dots,n \tag{8.1}$$

For one particular sequence of n independent Bernoulli trials with exactly x successes and $n-x$ failures, such as:

$$\underbrace{S, S, \dots S}_{x}, \underbrace{F, F, \dots F}_{n-x}$$

should have probability $p^x(1-p)^{n-x}$ followed by property of independence, where p is the probability of a success. (8.1) is therefore obtained since there are $\binom{n}{k}$ such sequences constituting the event $\{X = x\}$.

As derived in Example 5.3.1 and 6.6.3, *the mean and variance of the $Binomial(n,p)$ distribution* are:

$$E(X) = np, \quad Var(X) = np(1-p)$$

And the *mgf of the $Binomial(n,p)$ distribution*, as shown in Example 7.5.1, is:

$$M_X(t) = (e^t p + q)^n$$

Example 8.1.1: Choose n chips one at a time **with replacement** from the urn as shown in the figure below, which consisting of a red chips, b blue chips and c green chips. Let X denote the number of **red chips**, specify $P(X = x), E(X), Var(X)$ in terms of a, b, c

Figure 8.1.1

Proof: Each chip can only yield two possible outcomes: either red or non-red, with the probability of getting a red remains the same to be $\frac{a}{a+b+c}$ (the chip is picked with replacement), indicating X can be modelled by a $Binomial\left(n, \frac{a}{a+b+c}\right)$ distribution. Hence, by (8.1), we have the pmf,

$$P(X = x) = \binom{n}{x}\left(\frac{a}{a+b+c}\right)^{x}\left(\frac{b+c}{a+b+c}\right)^{n-x}, \qquad x = 0,1,2,\dots n$$

and as defined, the expectation and variance are

$$E(X) = n\frac{a}{a+b+c}, \qquad Var(X) = n\frac{a(b+c)}{(a+b+c)^2}$$ ■

Poisson Distribution

Definition 8.2: *A random variable, X, that counts the number of events occurring in a fixed interval of time and /or space if these events occur with a known average rate and independently of the time since the last event, can be modeled by a $\boldsymbol{Poisson(\lambda)}$ distribution, often expressed as $\boldsymbol{X \sim \mathcal{P}(\lambda)}$, which has a pmf given by:*

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0,1,2 \dots \dots \tag{8.2}$$

where λ is called ***intensity parameter***

As derived in Example 6.6.2, t*he mean and variance of the $Poisson(\lambda)$ distribution* are:

$$E(X) = \lambda, \quad Var(X) = \lambda$$

And the *mgf of the $Binomial(n, p)$ distribution*, as shown in Example 7.4.1, is:

$$M_X(t) = e^{\lambda(e^t - 1)}$$

Theorem 8.3: *If $X \sim Binomial(n, p)$ and $Y \sim Poisson(\lambda)$, with $\lambda = np$, then for large n and small np,*

$$P(X = x) \approx P(Y = y) \tag{8.3}$$

Proof: The pmf of Binomial distribution is given by,

$$P(X = x) = \binom{n}{x} p^{x}(1-p)^{n-x}$$

which can be rewritten as,

$$P(X = x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^{x}\left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^{x}}{x!}\cdot\frac{n!}{(n-x)!\,n^{x}}\cdot\frac{\left(1-\frac{\lambda}{n}\right)^{n}}{\left(1-\frac{\lambda}{n}\right)^{x}} \tag{8.3.1}$$

Then, take limit on both sides of (8.3.1),

$$\begin{aligned}\lim_{n\to\infty} P(X = x) &= \lim_{n\to\infty}\frac{\lambda^{x}}{x!}\cdot\frac{n!}{(n-x)!\,n^{x}}\cdot\frac{\left(1-\frac{\lambda}{n}\right)^{n}}{\left(1-\frac{\lambda}{n}\right)^{x}} \\ &= \frac{\lambda^{x}}{x!}\cdot\lim_{n\to\infty}\frac{n!}{(n-x)!\,n^{x}}\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}\end{aligned}$$

The first two limits are easily seen,

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}=1$$

$$\lim_{n\to\infty}\frac{n!}{(n-x)!\,n^x}=\lim_{n\to\infty}\frac{n(n-1)\dots(n-x+1)}{n^x}=\frac{n^x}{n^x}=1$$

We can use the special limit, $\lim_{\varepsilon\to 0}\log(1+\varepsilon)=\varepsilon$, to calculate the last limit. Since λ is relatively small with respect to n, then

$$\lim_{n\to\infty}\log\left(1+\left(-\frac{\lambda}{n}\right)\right)=-\frac{\lambda}{n}$$

Hence, $$\lim_{n\to\infty}n\cdot\log\left(1+\left(-\frac{\lambda}{n}\right)\right)=-\lambda$$

Also, by continuity of exponential function,

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=\lim_{n\to\infty}\exp\left[n\cdot\log\left(1+\left(-\frac{\lambda}{n}\right)\right)\right]=e^{-\lambda}$$

Therefore, $$\lim_{n\to\infty}P(X=x)=\frac{\lambda^x}{x!}\cdot 1\cdot 1\cdot\ e^{-\lambda}=\frac{e^{-\lambda}\lambda^x}{x!}=\frac{e^{-\lambda}\lambda^y}{y!}=P(Y=y)$$ ■

Geometric Distribution

Definition 8.4: *Among n Bernoulli trials with success probability p, let a random variable, X, to be the number of trials until the first success occurs, then such X can be modeled by a $\boldsymbol{Geometric(p)}$ distribution, often expressed as $\boldsymbol{X\sim\mathcal{G}(p)}$, which has a pmf given by:*

$$P(x)=(1-p)^{x-1}p,\qquad x=1,2\dots\dots \tag{8.4}$$

The derivation of (8.4) is similar to Binomial distribution. The event $\{X=x\}$ can only occur when the first $x-1$ times are all failures while the last one yields a success, namely,

$$\underbrace{F,F,\dots F}_{x-1},\underbrace{S}_{1}$$

which gives the probability in (8.4) straightforwardly.

As derived in Example 5.3.2 and can be derived by a similar manner, t*he mean and variance of the $Geometric(p)$ distribution* are:

$$E(X)=\frac{1}{p},\quad Var(X)=\frac{1-p}{p^2}$$

And the *mgf of the $Geometric(p)$ distribution*, is:

$$M_X(t)=\frac{pe^t}{1-qe^t}$$

The derivation of mgf is left after next section.

Negative Binomial Distribution

Definition 8.5: *In a sequence of n independent Bernoulli trials with success probability p, let the random variable, X, denote the trial at which the rth success occurs, where r is a fixed integer. Then, X has a $Negative\ Binomial(r, p)$ distribution, expressed as $X \sim \mathcal{NB}(r, p)$, with pmf given by,*

$$P(x) = \binom{x-1}{r-1} p^r q^{x-r}, \qquad x = r, r+1, \dots \dots \tag{8.5}$$

Negative Binomial distribution is a general case of Geometric distribution, or $\mathcal{NB}(1, p)$ is equivalent to $\mathcal{G}(p)$. The coefficient $\binom{x-1}{r-1}$ computes the number of events that there are exactly $r-1$ successes in the first $x-1$ trials.

By similar approach as in Example 5.3.2, t*he mean and variance of the $\mathcal{NB}(r, p)$ distribution* are:

$$E(X) = \frac{r}{p}, \quad Var(X) = r\frac{1-p}{p^2}$$

And the *mgf of the $\mathcal{NB}(r, p)$ distribution*, is:

$$M_X(t) = \left(\frac{pe^t}{1-qe^t}\right)^r$$

Proof: By the definition of mgf, we have,

$$\begin{aligned}
M_X(t) &= E[e^{tX}] = \sum_{x=r}^{\infty} e^{tx}\binom{x-1}{r-1} p^r (1-p)^{x-r} \\
&= \sum_{y=0}^{\infty} e^{t(r+y)}\binom{r+y-1}{r-1} p^r (1-p)^y && \Big| (y = x - r) \\
&= \sum_{y=0}^{\infty} e^{t(r+y)} \frac{(r+y-1)!}{(r-1)!\,y!} p^r (1-p)^y = \frac{e^{tr}p^r}{(r-1)!}\sum_{y=0}^{\infty}\frac{(r+y-1)!}{y!} e^{ty}(1-p)^y \\
&= \frac{e^{tr}p^r}{(r-1)!}\sum_{y=0}^{\infty}\frac{(r+y-1)!}{y!}[e^t(1-p)]^y \\
&= \frac{e^{tr}p^r}{(r-1)!}\sum_{y=0}^{\infty}\frac{(r+y-1)!}{y!} z^y \ (*) && \Big| \big(z = e^t(1-p)\big)
\end{aligned}$$

It is true and can be easily induced that:

$$\frac{d^k}{dx^k}x^a = a(a-1)\dots(a-k+1)x^{a-k} = \frac{a!}{(a-k)!}x^{a-k} \tag{8.5.1}$$

$$\frac{d^{r-1}}{dz^{r-1}}\frac{1}{1-z} = \frac{d}{dz}(1-z)^{-(r-1)} = (r-1)!\,(1-z)^{-r} \tag{8.5.2}$$

Let $a = r + y - 1,\ k = r - 1,\ x = z,\ a - k = y$, and make changes of variables on $(*)$, we have

$$\begin{aligned} M_X(t) &= \frac{e^{tr}p^r}{(r-1)!}\sum_{y=0}^{\infty}\frac{(r+y-1)!}{y!}z^y = \frac{e^{tr}p^r}{(r-1)!}\sum_{y=0}^{\infty}\frac{d^{r-1}}{dz^{r-1}}z^{r+y-1} \\ &= \frac{e^{tr}p^r}{(r-1)!}\frac{d^{r-1}}{dz^{r-1}}\sum_{y=0}^{\infty}z^{r+y-1} = \frac{e^{tr}p^r}{(r-1)!}\frac{d^{r-1}}{dz^{r-1}}\left(\frac{1}{1-z}\right) \end{aligned}$$

Therefore, by (8.5.2), for $|z| < 1$, or equivalently $|t| < -\ln(1-p)$,

$$\begin{aligned} M_X(t) &= \frac{e^{tr}p^r}{(r-1)!}\cdot(r-1)!\,(1-z)^{-r} = \left(\frac{e^t p}{1-z}\right)^r \\ &= \left(\frac{e^t p}{1-e^t(1-p)}\right)^r = \left(\frac{pe^t}{1-qe^t}\right)^r \end{aligned}$$

■

Hypergeometric Distribution

Definition 8.6: *In contrast to the Binomial distribution, the* $\boldsymbol{Hypergeometric(N,k,n)}$ *distribution, often denoted by* $X \sim \mathcal{H}(N,k,n)$*, describes the probability of* X *successes in* n *draws (which are identical independent Bernoulli trials)* ***without replacement*** *from a* ***finite population*** *of size* N *containing exactly* k *successes. Such random variable* X *has a pmf given by:*

$$P(X=x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \qquad x = 0,1,2\ldots,n; N,n,k > 0 \tag{8.6}$$

The *mean* of the Hypergeometric distribution is given by:

$$E(X) = \sum_{x=0}^{n} x\cdot\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n}\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}\cdot k = k\frac{\binom{N-1}{n-1}}{\binom{N}{n}}\sum_{x=0}^{n}\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N-1}{n-1}} = k\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{kn}{N}$$

The *variance* of the Hypergeometric distribution can be found by first computing the second moment:

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{n} x^2\cdot\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n}\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}}\cdot k = k\sum_{x=0}^{n}(x-1+1)\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}} \\ &= k\sum_{x=0}^{n}(x-1)\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}} + k\sum_{x=0}^{n}\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}} \\ &= k(k-1)\frac{\binom{N-2}{n-2}}{\binom{N}{n}}\sum_{x=0}^{n}\frac{\binom{k-2}{x-2}\binom{N-k}{n-x}}{\binom{N-2}{n-2}} + k\frac{\binom{N-1}{n-1}}{\binom{N}{n}}\sum_{x=0}^{n}\frac{\binom{k-1}{x-1}\binom{N-k}{n-x}}{\binom{N}{n}} \\ &= k(k-1)\frac{n(n-1)}{N(N-1)} + k\frac{n}{N} \end{aligned}$$

$$Var(X) = k(k-1)\frac{n(n-1)}{N(N-1)} + k\frac{n}{N} - \left(\frac{kn}{N}\right)^2 = \frac{kn}{N}\left(\frac{(N-k)(N-n)}{N(N-1)}\right)$$

In fact, Example 3.4.2 (pick 6 cards without replacement from a total of 50 cards among which only 6 of them give "success"), was solved by following a $\mathcal{H}(50,6,6)$ distribution, where the probability of "no successes among 6 picks" was calculated by evaluating pmf (8.6) with $N = 50, k = 6, n = 6, x = 0$; and the probability of "4 out of 6 cards are successes" was calculated by evaluating (8.6) with $N = 50, k = 6$, $n = 6, x = 4$.

Multinomial Distribution

Definition 8.7: *Among n independent trials, each trial results in exactly one of some fixed finite number k possible outcomes, with probability $p_1, \dots, p_k$, such that $\sum_{i=1}^{k} p_i = 1$. Then if the variables X_i indicate the number of times outcome number i is observed over the n trials, the vector $\boldsymbol{X} = (X_1, X_2 \dots, X_k)$ follows a $\boldsymbol{Multinomial(n, p_1, \dots, p_k)}$ distribution, with the pmf given by,*

$$P(\boldsymbol{X} = \boldsymbol{x}) = P(X_1 = x_1, \dots, X_k = x_k) = \begin{cases} \dfrac{n!}{x_1! \cdot \dots \cdot x_k!} p_1^{x_1} \cdot \dots \cdot p_k^{x_k}, & if \displaystyle\sum_{i=1}^{k} x_i = n \\ 0, & O.W. \end{cases} \tag{8.7}$$

For each outcome i, X_i can be seen as a Binomial random variable, with the *mean* and *variance* to be:

$$E(X_i) = np_i, \quad Var(X_i) = np_i(1 - p_i)$$

Example 8.7.1: Recall Example 8.1.1 of choosing n chips one at a time **with replacement** from the urn as shown in the figure below, which consisting of a red chips, b blue chips and c green chips. Instead of counting balls of just one color, let $\boldsymbol{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3})$, where $\boldsymbol{X_1}$ denote the number of **red chips**, $\boldsymbol{X_2}$ blue chips, and $\boldsymbol{X_3}$ green chips, specify the probability such that $P\{\boldsymbol{X} = (a_1, b_1, c_c)\}$

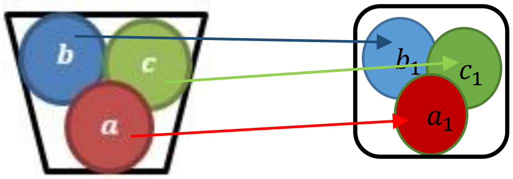


Figure 8.1.1

Solution: Simply applying (8.7), we have

$$P\{\boldsymbol{X} = (a_1, b_1, c_c)\} = \begin{cases} \dfrac{n!}{a_1! b_1! c_1!} \left(\dfrac{a}{a+b+c}\right)^{a_1} \left(\dfrac{b}{a+b+c}\right)^{b_1} \left(\dfrac{c}{a+b+c}\right)^{c_1}, & if\ a_1 + b_1 + c_1 = n \\ 0, & O.W. \end{cases}$$

Lecture 9

<u>Common Continuous Distribution</u>

Uniform Distribution

Definition 9.1: *The $\boldsymbol{Uniform(a, b)}$ distribution, or $\mathcal{U}(a, b)$, is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by:*

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & O.W. \end{cases} \tag{9.1}$$

It is easy to see $f_X(x) \ge 0$, and $\int f_X(x)dx = \int_a^b \frac{1}{b-a} dx = 1$. We also have its *mean* and *variance* to be:

$$E(X) = \frac{b-a}{2}, \quad Var(X) = \frac{(b-a)^2}{12}$$

The mean can be visualized as the center of the interval as in Figure 9.1 below.

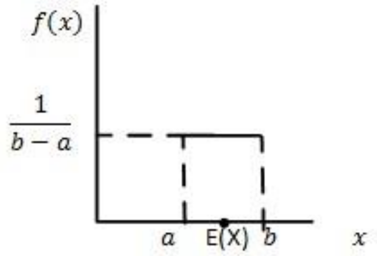


Figure 9.1

And the *moment generating function(mgf)* of the Uniform distribution exists and is found to be:

$$M_X(t) = \frac{e^{tb} - e^{ta}}{(b-a)t}$$

Exponential Distribution

Definition 9.2: *The* $\boldsymbol{Exponential(\lambda)}$ *distribution, or* $Exp(\lambda)$, *is the probability distribution that describes the waiting time between randomly occurring events with rate of change* λ,*and its pdf is given by:*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & O.W.' \end{cases} \quad \lambda > 0 \tag{9.2}$$

The *mean* and *variance* of the $Exp(\lambda)$ are,

$$E(X) = \frac{1}{\lambda}, \quad Var(X) = \frac{1}{\lambda^2}$$

In Example 7.4.2, it has been shown that the *mgf* of the Exponential distribution is,

$$M_X(t) = \frac{\lambda}{\lambda - t}, \qquad t < \lambda$$

Example 9.2.1: If $X \sim \mathcal{U}(0,1)$, i.e. X has a *standard uniform distribution*, $Y = \log\left(\frac{1}{X}\right)$, find the pdf of Y

Solution: Applied by the similar technique used in Example 5.1.1, we have the cdf of the new random variable Y to be:

$$F_Y(y) = P(Y \leq y) = P\left(\log\frac{1}{X} \leq y\right) = P(X \geq e^{-y}) = 1 - F_X(e^{-y})$$

Then differentiate the last equality, we obtain the pdf,

$$f_Y(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & O.W. \end{cases}$$ ■

Example 9.2.2: If $X \sim \mathcal{U}(0,1)$, and $Y = aX + b, a > 0$, derive the pdf of Y

Solution: Similarly as in last example, we first find the cdf of Y,

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X(\frac{y-b}{a})$$

The pdf can be obtained after differentiating,

$$f_Y(y) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right) = \begin{cases} \frac{1}{a}, & b < y < a + b \\ 0, & O.W. \end{cases}$$

which, by definition, is just the pdf of a $\mathcal{U}(b, a+b)$ distribution. In other words, the linear transformation of a uniform random variable is still a uniform random variable ■

Example 9.2.3: If $X \sim Exp(1)$, and $Y = \lambda X$ with λ to be a constant, derive the pdf of Y

Solution: First find the cdf of Y,

$$F_Y(y) = P(Y \leq y) = P(\lambda X \leq y) = P\left(X \leq \frac{y}{\lambda}\right) = F_X\left(\frac{y}{\lambda}\right)$$

After differentiation, we have the pdf,

$$f_Y(y) = \frac{1}{\lambda} f_X\left(\frac{y}{\lambda}\right) = \frac{1}{\lambda} e^{-\frac{y}{\lambda}} = \begin{cases} \frac{1}{\lambda} e^{-\frac{y}{\lambda}}, y \geq 0 \\ 0, \quad\quad O.W. \end{cases}$$

which, by definition, is the pdf of $Exp\left(\frac{1}{\lambda}\right)$ distribution. In other words, the linear transformation of an exponential random variable is still an exponential random variable ■

Gamma Distribution

Definition 9.3: *The $\boldsymbol{Gamma(a, p)}$ distribution, or $\gamma(a, p)$, is the probability distribution that describes the waiting time until the $\boldsymbol{a}$th random event occurs with a rate of change p,and its pdf is given by:*

$$f_X(x) = p^a \frac{e^{-px} x^{a-1}}{\Gamma(a)}, \qquad x > 0 \tag{9.3}$$

where $\boldsymbol{\Gamma(a) = \int_0^\infty e^{-y} y^{a-1} dy}$, is defined as the Gamma function.

Theorem 9.3.1: *There are some useful properties or special cases about Gamma function, which are:*

a. $\Gamma(a+1) = a\Gamma(a)$ (9.3.1*a*)

b. *If n is an integer, then $\Gamma(n+1) = n!$* (9.3.1*b*)

c. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ (9.3.1*c*)

Example 9.3.2: Prove that $\gamma(1, p)$ is equivalent to $Exp(p)$

Proof: Let $X \sim \gamma(1, p)$ and $Y \sim Exp(p)$, so by definition, we have

$$f_X(x) = \begin{cases} p^1 \frac{e^{-px} x^{1-1}}{\Gamma(1)} = pe^{-px}, x \geq 0 \\ 0, \quad\quad O.W. \end{cases} \text{ and } f_Y(y) = \begin{cases} pe^{-py}, y \geq 0 \\ 0, \quad\quad O.W. \end{cases}$$

Thus, for every $x = y$, we always have $f_X(x) = f_y(y)$, namely $X \equiv Y$. ■

A generalization of this example is that: *if $X_1, \dots X_n$ are identical independent random variable of $Exp(p)$, then, $\sum_{i=1}^n X_i \sim \gamma(n, p)$* (called *Erlang* distribution, which is a special case of Gamma distribution with n to be an integer). Proof of this generalization is left after the introduction of *convergence of distribution*.

The *mean* and *variance* of the $Gamma(a, p)$ are,

$$E(X) = \frac{a}{p}, \quad Var(X) = \frac{a}{p^2}$$

And the *mgf* of the $Gamma(a, p)$ distribution is,

$$M_X(t) = \left(\frac{p}{p-t}\right)^a, \qquad t < p$$

Proof: From definition of mgf, we have,

$$\begin{aligned} M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} \cdot p^a \frac{e^{-px} x^{a-1}}{\Gamma(a)} dx = \int_0^\infty p^a \frac{e^{(t-p)x} x^{a-1}}{\Gamma(a)} dx \\ &= \frac{p^a}{(p-t)^a} \int_0^\infty (p-t)^a \frac{e^{-(p-t)x} x^{a-1}}{\Gamma(a)} dx, \qquad t - p < 0 \\ &= \frac{p^a}{(p-t)^a} \int_0^\infty f_{\gamma(a,p-t)}(x) dx = \frac{p^a}{(p-t)^a} = \left(\frac{p}{p-t}\right)^a, \qquad t < p \end{aligned}$$

■

Beta Distribution

Definition 9.4: *The $\boldsymbol{Beta(a,b)}$ distribution, or $\beta(a,b)$, is a family of continuous probability distribution defined on the interval $[0,1]$ parametrized by two positive shape parameters, denoted by a and b, which has the pdf given by,*

$$f_X(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \qquad 0 \le x \le 1 \tag{9.4}$$

where $\boldsymbol{B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1} dy}$ is defined as the Beta function, which is related to Gamma function through the following identity:

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

The *mean* of the $Beta(a,b)$ distribution is given by,

$$\begin{aligned} E(X) &= \int_0^1 x \cdot \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} dx = \frac{B(a+1,b)}{B(a,b)} \int_0^1 \frac{x^{(a+1)-1}(1-x)^{b-1}}{B(a+1,b)} dx \\ &= \frac{B(a+1,b)}{B(a,b)} = \frac{\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}} = \frac{a}{a+b} \end{aligned}$$

The *variance* of the $Beta(a,b)$ distribution can be found by first calculating the second moment,

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \cdot \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)} dx = \frac{B(a+2,b)}{B(a,b)} \int_0^1 \frac{x^{(a+2)-1}(1-x)^{b-1}}{\mathrm{B}(a+2,b)} dx \\ &= \frac{B(a+2,b)}{B(a,b)} = \frac{\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}}{\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}} = \frac{a(a+1)}{(a+b)(a+b+1)} \\ Var(X) &= E(X^2) - E^2(X) = \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 = \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

Example 9.4.1: Prove that $\beta(1,1)$ is identical to $\mathcal{U}(0,1)$

Proof: Let $X \sim \beta(1,1)$ and $Y \sim \mathcal{U}(0,1)$, so by definitions, we have

$$f_X(x) = \begin{cases} \frac{x^{1-1}(1-x)^{1-1}}{B(1,1)} = \frac{1}{B(1,1)} = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} = 1, & 0 \le x \le 1 \\ 0, & O.W. \end{cases} \text{ and } f_Y(y) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & O.W. \end{cases}$$

Thus, for every $x = y$, we always have $f_X(x) = f_y(y)$, namely $X \equiv Y$. ■

Lecture 10

Cauchy Distribution

Definition 10.1: *The $Cauchy(\theta)$ distribution is a symmetric, bell-shaped distribution on $(-\infty, \infty)$ with pdf*

$$f_X(x) = \frac{1}{\pi[1 + (x - \theta)^2]} \tag{10.1}$$

For all θ, (10,1) defines a proper pdf. Since,

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} \frac{1}{\pi(1 + (x - \theta)^2)} dx = \frac{1}{\pi}\left[arctg(x - \theta)\Big|_{-\infty}^{\infty}\right] = \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1$$

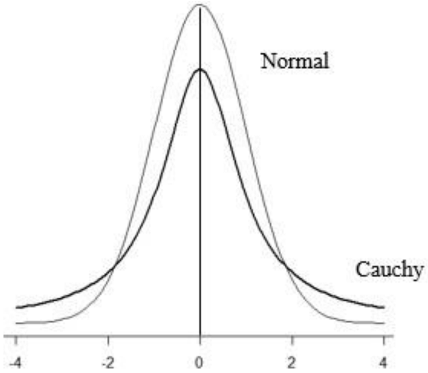


Figure 10.1

As depicted in Figure 10.1, Cauchy distribution looks very similar to normal distribution. However, there is a big difference, indeed. Besides that Cauchy has a thicker tail, the *mean* of the Cauchy distribution **does not exist**, which can be shown as follows (θ is taken to be 0 for simplicity):

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1 + x^2)} \propto \int_{-\infty}^{\infty} \frac{1}{x} dx = \log(x)\Big|_{-\infty}^{\infty} = \infty$$

Normal Distribution

Definition 10.2: *The simplest case of a $Normal(\mu, \sigma^2)$ distribution, (usually denoted by $N(\mu, \sigma^2)$) is known as the* ***standard normal distribution*** *where $\mu = 0, \sigma = 1$, which is described by this pdf:*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{10.2.1}$$

And a general normal distribution, with mean $\mu \in R$ and variance $\sigma^2 \ge 0$, is defined by the pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{10.2.2}$$

More specifically, if $Z \sim N(0,1)$*, then* $X = \sigma Z + \mu$ *will have a* $N(\mu, \sigma^2)$ *distribution. Conversely, if* $X \sim N(\mu, \sigma^2)$*, then* $Z = \frac{X-\mu}{\sigma}$ *will have a standard normal distribution.*

The constant factor $\frac{1}{\sqrt{2\pi}}$ ensures that the total area under the curve $f_Z(z)$ in (10.2.1) is equal to one, which can be proved by first showing the identity $\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$. Let $I = \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy$, then

$$I^2 \ = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \iint_{-\infty}^{\infty} e^{-\frac{z^2+y^2}{2}} dz dy$$

Applying polar coordinates transformation on y, z, we define

$$y = r sin\theta \quad \text{and} \quad z = r cos\theta$$

Then, $y^2 + z^2 = r^2$ and $dydz = rdrd\theta$. With proper limits of integration for r, θ, we have,

$$I^2 \ = \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} r dr d\theta = 2\pi \int_0^{\infty} e^{-\frac{r^2}{2}} \cdot r dr = 2\pi \left(-e^{-\frac{r^2}{2}} \right) \Big|_0^{\infty} = 2\pi$$

So that $I = \sqrt{2\pi}$ and the identity is established and further gives that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$.

To show that $f_X(x)$ in (10.2.2) is also a valid pdf, we start from the relation in definition 10.2, $X = \sigma Z + \mu$, where $Z \sim N(0,1)$, then,

$$z = \frac{x-\mu}{\sigma} \quad \text{and} \quad dz = \frac{1}{\sigma} dx$$

By making a change of variable, we have

$$\int_{-\infty}^{\infty} f_X(x) dx \ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1$$

The *mean* of the $N(0,1)$ distribution is given by,

$$\begin{aligned} E(Z) \ &= \int_{-\infty}^{\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} z dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} d\left(\frac{z^2}{2}\right) \\ &= -\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} = 0 \end{aligned}$$

The *variance* of the $N(0,1)$ distribution can be found by the definition,

$$\begin{aligned} Var(Z) \ &= E(Z - E(Z))^2 = E(Z^2) = \int_{-\infty}^{\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \cdot e^{-\frac{z^2}{2}} z dz \\ &= -z \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} - \left(-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz \right) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 1 \end{aligned}$$

Therefore, the *mean* and *variance* of a general normal distribution $N(\mu, \sigma^2)$ can be easily derived by applying Theorem 6.2, and Theorem 6.7 on the relation $X = \sigma Z + \mu$,

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu$$

$$Var(X) = Var(\sigma Z + \mu) = \sigma^2 Var(Z) = \sigma^2$$

As given by Example 7.7.1, the mgf of a $N(\mu, \sigma^2)$ distribution is,

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Example 10.2.1: If $Z \sim N(0,1)$, and $Y = aZ + b$, show that $X \sim N(b, a^2)$

Solution: By Theorem 7.7, we have the mgf of Y to be,

$$M_Y(t) \ = M_{aZ+b}(t) = M_Z(at) \cdot e^{bt} = e^{\frac{1}{2}(at)^2} \cdot e^{bt} = e^{bt + \frac{1}{2}a^2 t^2}$$

Then, the uniqueness of mgf gives that $Y \sim N(b, a^2)$

Example 10.2.2: If $X \sim N(\mu, \sigma^2)$, and $Z = \frac{X-\mu}{\sigma}$, show that $Z \sim N(0,1)$

Solution: By Theorem 7.7, we have the mgf of Y to be,

$$M_Y(t) \ = M_{\frac{X-\mu}{\sigma}}(t) = M_X\left(\frac{t}{\sigma}\right) \cdot e^{-\frac{\mu t}{\sigma}} = e^{\mu\left(\frac{t}{\sigma}\right) + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma}\right)^2} \cdot e^{-\frac{\mu t}{\sigma}} = e^{\frac{1}{2}t^2}$$

Then, the uniqueness of mgf gives that $Z \sim N(0,1)$

Lognormal Distribution

Definition 10.3: *If X is a r.v. whose logarithm is normally distributed (that is, $\log X \sim N(\mu, \sigma^2)$), then X has a $lognormal(\mu, \sigma^2)$ distribution, with pdf given by:*

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \qquad x > 0 \tag{10.3}$$

The pdf can be obtained by straightforward transformation of the normal pdf using Theorem 5.1, yielding

$$\begin{aligned} f_X(x) \ &= \frac{d}{dx} F_X(x) = \frac{d}{dx} F_Y(\log x) = \frac{1}{x} f_Y(\log x) = \frac{1}{x} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} \\ &= \begin{cases} \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0, & O.W. \end{cases} \end{aligned}$$

The *mean* and *variance* of the lognormal r.v. X can be found by relating to the mgf of its log transformation, $\log X$, which by definition has a $N(\mu, \sigma^2)$ distribution,

$$E(X^t) = E\left(e^{t \log X}\right) = M_{\log X}(t)$$

As specified in last section, $M_{\log X}(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$, with $t = 1$, and $t = 2$, we have,

$$E(X) = e^{\mu \cdot 1 + \frac{1}{2}\sigma^2 1^2} = e^{\mu + \frac{1}{2}\sigma^2}; \quad E(X^2) = e^{\mu \cdot 2 + \frac{1}{2}\sigma^2 2^2} = e^{2\mu + 2\sigma^2}$$

$$Var(X) = E(X^2) - E^2(X) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$$

Lecture 11

Convergence in Distribution

Definition 11.1: *A sequence of random variables, $X_1, X_2, \ldots$, converges in distribution to a random variable X, written as $X_n \overset{\mathcal{D}}{\Rightarrow} X$, if and only if, for every continuity point t*

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \tag{11.1}$$

Theorem 11.2: *Suppose $\{X_i\}_{i=1,2,\ldots}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that*

$$\lim_{i\to\infty} M_{X_i}(t) = M_X(t), \quad \textit{for all } t \textit{ in a neighborhood of 0}$$

and $M_X(t)$ is an mgf. Then there exists a unique cdf $F_X(x)$ whose moments are determined by $M_X(t)$ and , for all continuity points x, we have,

$$\lim_{i\to\infty} F_{X_i}(t) = F_X(t)$$

Theorem 11.3: *If $X \sim Bin(n,p)$, then $E(X) = np$ and $Var(X) = np(1-p)$, and as n being large enough so that there are enough values of X to make an approximation by a continuous distribution which by rule of thumb is: $np \geq 5$ and $n(1-p) \geq 5$. In this case, X can be approximated by $N(np, np(1-p))$*

Proof: The proof mainly relies on Central Limit Theorem (proof of CLT will be given later, here we just use its result), which states that: *for a sequence of identical independent random variables $\{X_i\}_{i=1,2,\ldots,n}$ with finite mean μ and variance σ^2 and if n is large enough, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has a $N\left(\mu, \frac{\sigma^2}{n}\right)$ distribution.* Definition 8.1 gives that a $Bin(n,p)$ is equivalent to sum of n identical independent $Bernoulli(p)$, then $E(Y_i) = p, Var(Y_i) = p(1-p)$. Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, where $Y_i \sim Bernoulli(p)$, we have that:

$$X = n \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i = n\bar{Y} \;\; and \;\; M_{\bar{Y}}(t) = e^{pt + \frac{p(1-p)}{2n}t^2}$$

Since $\bar{Y} \sim N\left(p, \frac{p(1-p)}{n}\right)$, which is a straightforward derivation by CLT

Therefore, from Theorem 7.7 we can have the mgf of X calculated to be:

$$M_X(t) = M_{n\bar{Y}}(t) = M_{\bar{Y}}(nt) = e^{p\cdot nt + \frac{p(1-p)}{2n}(nt)^2} = e^{npt + \frac{1}{2}np(1-p)t^2}$$

Observing that the last expression is the mgf of a $N(np, np(1-p))$ distribution, hence, uniqueness of mgf finally yields that, $X \sim N(np, np(1-p))$. ■

One-Parameter Exponential Family

Definition 11.4 (Indicator Function): *The **indicator function** of a subset A is a function that only takes values 0 or 1, which is defined as*

$$I_A(x) = \begin{cases} 1, x \in A \\ 0, x \notin A \end{cases} \tag{11.4}$$

whose expected value is just the probability of A,

$$E[I_A(x)] = 1 \cdot P(X \in A) + 0 \cdot P(X \notin A) = P(A)$$

Definition 11.5: *A pdf or pmf belongs to a **one-parameter exponential family**, if it can be expressed as:*

$$f_{\theta}(x) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A(x) \tag{11.5}$$

with θ denotes the unknown parameter, and support A doesn't depend on θ. The notation $f_{\theta}(x)$ is used to stress that the pdf or pmf is determined by parameter θ

Example 11.5.1: Suppose $X \sim \mathcal{U}(0, \theta)$, does pdf of X, $f_X(x)$, belong to one-parameter exponential family?
Solution: As defined, the pdf of X is,

$$f_\theta(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & O.W. \end{cases}$$

which can be expressed as,

$$f_\theta(x) = \exp(-\log\theta) \cdot I_A(x),\ where\ A = (0, \theta)$$

Since A depends on θ, $f_\theta(x)$ does not belong to one-parameter exponential family.

Example 11.5.2: Prove if the pdf or pmf of the specified r.v. X belongs to one-parameter exponential family.

1). $X \sim \mathcal{B}in\ (n, \theta)$

Solution: As defined, the pmf of X is $f_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, x = 0{,}1 \ldots n$, which can be expressed as

$$\begin{aligned} f_\theta(x) &= \exp\left\{\log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)\right\} \cdot I_{\{0,1,2\ldots,n\}}(x) \\ &= \exp\left\{x\left(\log\frac{\theta}{1-\theta}\right) + \log\binom{n}{x} + n\log(1-\theta)\right\} \cdot I_{\{0,1,2\ldots,n\}}(x) \end{aligned}$$

with $c(\theta) = \log\frac{\theta}{1-\theta}$, $T(x) = x$, $S(x) = \log\binom{n}{x}$, $d(\theta) = n\log(1-\theta)$, $A = \{0,1,\ldots,n\}$, the specified pmf of X has the same structure as (11.4) and A does not depend on θ. Therefore, pmf of $\mathcal{B}in\ (n, \theta)$ belongs to one-parameter exponential family.

2). $X \sim \mathcal{P}\ (\theta)$

Solution: As defined, the pmf of X is $f_\theta(x) = e^{-\theta}\frac{\theta^x}{x!}, x = 0{,}1{,}2 \ldots$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\{-\theta + x\log\theta - \log(x!)\} \cdot I_{\{0,1,2\ldots,\}}(x) \\ &= \exp\{x\log\theta - \log(x!) - \theta\} \cdot I_{\{0,1,2\ldots,\}}(x) \end{aligned}$$

with $c(\theta) = \log\theta, T(x) = x, S(x) = -\log(x!), d(\theta) = -\theta, A = \{0,1,2,\ldots\ldots\}$, the specified pmf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pmf of $\mathcal{P}\ (\theta)$ belongs to one-parameter exponential family.

3). $X \sim \mathcal{G}\ (\theta)$

Solution: As defined, the pmf of X is $f_\theta(x) = (1-\theta)^{x-1}\theta, x = 1{,}2, \ldots\ldots$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\{(x-1)\log(1-\theta) + \log\theta\} \cdot I_{\{1,2\ldots\}}(x) \\ &= \exp\left\{x\log(1-\theta) + \log\frac{\theta}{1-\theta}\right\} \cdot I_{\{1,2\ldots\}}(x) \end{aligned}$$

with $c(\theta) = \log(1-\theta), T(x) = x, S(x) = 0, d(\theta) = \log\frac{\theta}{1-\theta}, A = \{0,1,2, \dots \dots\}$, the specified pmf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pmf of $\mathcal{G}(\theta)$ belongs to one-parameter exponential family.

4). $X \sim \mathcal{NB}(r, \theta)$

Solution: As defined, the pmf of X is $f_\theta(x) = \binom{x-1}{r-1}\theta^r(1-\theta)^{x-r}, x = r, r+1, \dots \dots$which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\left\{\log\binom{x-1}{r-1} + r\log\theta + (x-r)\log(1-\theta)\right\} \cdot I_{\{r,r+1\dots\}}(x) \\ &= \exp\left\{x\log(1-\theta) + \log\binom{x-1}{r-1} + r\log\frac{\theta}{1-\theta}\right\} \cdot I_{\{r,r+1\dots\}}(x) \end{aligned}$$

with $c(\theta) = \log(1-\theta),\ T(x) = x,\ S(x) = \log\binom{x-1}{r-1},\ d(\theta) = r\log\frac{\theta}{1-\theta},\ A = \{r, r+1, \dots \dots\}$, the specified pmf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pmf of $\mathcal{NB}(r, \theta)$ belongs to one-parameter exponential family.

5). $X \sim Exp(\theta)$

Solution: As defined, the pdf of X is $f_\theta(x) = \begin{cases} \theta e^{-\theta x}, & x \ge 0 \\ 0, & O.W. \end{cases}$, which can be expressed as,

$$f_\theta(x) = \exp\{\log\theta - x\theta\} \cdot I_{\{x\ge0\}}(x) = \exp\{-x\theta + \log\theta\} \cdot I_{\{x\ge0\}}(x)$$

with $c(\theta) = \theta, T(x) = -x, S(x) = 0, d(\theta) = \log\theta, A = [0, \infty)$, the specified pdf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pdf of $Exp(\theta)$ belongs to one-parameter exponential family

6). $X \sim N(\theta, 1)$

Solution: As defined, the pdf of X is $f_\theta(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}}, -\infty < x < \infty$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\left\{-\frac{1}{2}\log(2\pi) - \frac{x^2 - 2x\theta + \theta^2}{2}\right\} \cdot I_R(x) \\ &= \exp\left\{x\theta - \frac{1}{2}x^2 - \frac{1}{2}(\log(2\pi) + \theta^2)\right\} \cdot I_R(x) \end{aligned}$$

with $c(\theta) = \theta, T(x) = x, S(x) = -\frac{1}{2}x^2, d(\theta) = -\frac{1}{2}(\log 2\pi + \theta^2), A = (-\infty, \infty)$,the specified pdf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pdf of $N(\theta, 1)$ belongs to one-parameter exponential family

7). $X \sim N(0, \theta)$

Solution: As defined, the pdf of X is $f_\theta(x) = \frac{1}{\sqrt{2\pi\theta}}e^{-\frac{x^2}{2\theta}}, -\infty < x < \infty$, which can be expressed as,

$$f_\theta(x) = \exp\left\{-\frac{1}{2}\log(2\pi\theta) - \frac{1}{2\theta}x^2\right\} \cdot I_R(x)$$

$$= \exp\left\{-\frac{1}{2\theta}x^2 - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta\right\} \cdot I_R(x)$$

with $c(\theta) = \frac{1}{2\theta},\ T(x) = -x^2,\ S(x) = -\frac{1}{2}\log(2\pi),\ d(\theta) = -\frac{1}{2}\log(\theta),\ A = (-\infty, \infty)$,the specified pdf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pdf of $N(\theta, 1)$ belongs to one-parameter exponential family

8). $X \sim \gamma(a, \theta)$

Solution: As defined, the pdf of X is $f_\theta(x) = \begin{cases} \frac{\theta^a e^{-\theta x} x^{a-1}}{\Gamma(a)}, x > 0 \\ 0, \quad\quad O.W. \end{cases}$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\{a\log\theta - x\theta + (a-1)\log x - \log\Gamma(a)\} \cdot I_{\{x \geq 0\}}(x) \\ &= \exp\{-x\theta + [(a-1)\log x - \log\Gamma(a)] + a\log\theta\} \cdot I_{\{x \geq 0\}}(x) \end{aligned}$$

with $c(\theta) = \theta,\ T(x) = -x,\ S(x) = (a-1)\log x - \log\Gamma(a),\ d(\theta) = a\log\theta,\ A = (0, \infty)$, the specified pdf of X has the same structure as (11.4), where A does not depend on θ. Therefore, pdf of $\gamma(a, \theta)$ belongs to one-parameter exponential family. ■

Definition 11.5: *The natural (one-parameter) exponential family is an exponential family in which the distribution function can be reparametrized as*

$$f_\eta(x) = \exp\{\eta x - \psi(\eta)\} \cdot \lambda(x) \tag{11.5}$$

where η is defined as the natural parameter, which is a function of the actual parameter θ

Take $Bernoulli(\theta)$ distribution as an example, its pdf $f_\theta(x) = \theta^x(1-\theta)^{1-x}, x = 0,1$ can be written as

$$f_\theta(x) = \exp\left[x\log\frac{\theta}{1-\theta} + \log(1-\theta)\right] \cdot I_{\{0,1\}}(x)$$

Let $\eta = \log\frac{\theta}{1-\theta}$, then $\theta = \frac{e^\eta}{1+e^\eta}$, make a change of variable from θ to η, we have

$$f_\eta(x) = \exp[x\eta - \log(1 + e^\eta)] \cdot I_{\{0,1\}}(x) \tag{11.5.1}$$

which has the same structure as (11.5) if let $\psi(\eta) = \log(1 + e^\eta)$. Therefore, $f_\eta(x)$ belongs to natural one-parameter exponential family.

Theorem 11.6: *If a random variable X has pdf from the natural one-parameter exponential family defined in (11.5), it has the following properties:*

a. $$E(X) = \psi'(\eta) \tag{11.6a}$$

b. $$Var(X) = \psi''(\eta) \tag{11.6b}$$

Proof: As defined, X has pdf $f_\eta(x) = \exp\{\eta x - \psi(\eta)\}\lambda(x)$. Allow interchanging the order of integration and differentiation and from definition of expectation and variance, we have,

$$\begin{aligned} E(X) &= \int x \cdot \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx = \int \big(x - \psi'(\eta) + \psi'(\eta)\big) \cdot \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \\ &= \int \big(x - \psi'(\eta)\big) \cdot \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx + \int \psi'(\eta) \cdot \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \end{aligned}$$

$$
\begin{aligned}
&= \int \frac{d}{d\eta} \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx + \psi'(\eta)\cdot \int \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \\
&= \frac{d}{d\eta}\int \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx + \psi'(\eta)\int \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \\
&= \frac{d}{d\eta} 1 + \psi'(\eta)\cdot 1 = 0 + \psi'(\eta) = \psi'(\eta)
\end{aligned}
$$

$$
\begin{aligned}
Var(X) &= \int \left(x - \psi'(\eta)\right)^2 \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \\
&= \int \left[\frac{d^2}{d\eta^2} \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx + \psi''(\eta)\exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx\right] \\
&= \frac{d^2}{d\eta^2}\int \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx + \psi''(\eta)\int \exp\{\eta x - \psi(\eta)\}\lambda(x)\,dx \\
&= \frac{d^2}{d\eta^2} 1 + \psi''(\eta)\cdot 1 = 0 + \psi''(\eta) = \psi''(\eta) \qquad \blacksquare
\end{aligned}
$$

Example 11.6.1: Justify if the following pdf or pmf of the specified random variable X belongs to natural one-parameter exponential family and find $E(X)$, $Var(X)$ using Theorem 11.6

1). $X \sim \mathcal{P}(\theta)$

Solution: It has been shown that the pmf of X can be written as,

$$f_\theta(x) \ = \exp\{x\log\theta - \log(x!) - \theta\}\cdot I_{\{0,1,2\dots,\}}(x) = \exp[x\log\theta - \theta]\,\lambda(x)$$

where $\lambda(x)$ contains all the factors which don't involve the parameter θ.

Let $\eta = \log\theta$, then $\theta = e^\eta$ and $\psi(\eta) = \theta = e^\eta$, thus the natural form of the pmf of X is,

$$f_\eta(x) \ = \exp\{\eta x - e^\eta\}\,\lambda(x)$$

Therefore,

$$E(X) = \psi'(\eta) = \frac{d}{d\eta}e^\eta = e^\eta = \theta$$

$$Var(X) = \psi''(\eta) = \frac{d}{d\eta}e^\eta = e^\eta = \theta$$

2). $X{\sim}Exp(\theta)$

Solution: It has been shown that the pdf of X can be written as,

$$f_\theta(x) \ = \exp\{-x\theta + \log\theta\}\cdot I_{\{x\ge 0\}}(x) = \exp[-x\theta - (-\log\theta)]\,\lambda(x)$$

where $\lambda(x)$ contains all the factors which don't involve the parameter θ.

Let $\eta = -\theta$, then $\theta = -\eta$ and $\psi(\eta) = -\log\theta = -\log(-\eta)$, thus the natural form of the pdf of X is,

$$f_\eta(x) \ = \exp\{\eta x - (-\log(-\eta))\}\,\lambda(x)$$

Therefore,

$$E(X) = \psi'(\eta) = \frac{d}{d\eta}(-\log(-\eta)) = -\frac{1}{\eta} = \frac{1}{\theta}$$

$$Var(X) = \psi''(\eta) = \frac{d}{d\eta}\left(-\frac{1}{\eta}\right) = \frac{1}{\eta^2} = \frac{1}{\theta^2}$$

3). $X \sim N(\theta, 1)$

Solution: It has been shown that the pdf of X can be written as,

$$f_\theta(x) \ = \exp\left\{x\theta - \frac{1}{2}x^2 - \frac{1}{2}(\log(2\pi) + \theta^2)\right\} \cdot I_R(x) = \exp\left[x\theta - \left(\frac{1}{2}\theta^2\right)\right] \lambda(x)$$

where $\lambda(x)$ contains all the factors which don't involve the parameter θ.

Let $\eta = \theta$, and $\psi(\eta) = \frac{1}{2}\theta^2 = \frac{1}{2}\eta^2$, thus the natural form of the pdf of X is,

$$f_\eta(x) \ = \exp\left\{\eta x - \frac{1}{2}\eta^2\right\} \lambda(x)$$

Therefore,

$$E(X) = \psi'(\eta) = \frac{d}{d\eta}\left(\frac{1}{2}\eta^2\right) = \eta = \theta$$

$$Var(X) = \psi''(\eta) = \frac{d}{d\eta}(\eta) = 1$$

4). $X \sim \gamma(a, \theta)$

Solution: It has been shown that the pdf of X can be written as,

$$\begin{aligned} f_\theta(x) \ &= \exp\{-x\theta + [(a-1)\log x - \log\Gamma(a)] + a\log\theta\} \cdot I_{\{x \geq 0\}}(x) \\ &= \exp[-x\theta - (-a\log\theta)]\, \lambda(x) \end{aligned}$$

where $\lambda(x)$ contains all the factors which don't involve the parameter θ.

Let $\eta = -\theta$, then $\theta = -\eta$, and $\psi(\eta) = -a\log\theta = -a\log(-\eta)$, thus the natural form of the pdf is,

$$f_\eta(x) \ = \exp\{\eta x - [-a\log(-\eta)]\}\, \lambda(x)$$

Therefore,

$$E(X) = \psi'(\eta) = \frac{d}{d\eta}[-a\log(-\eta)] = -\frac{a}{\eta} = \frac{a}{\theta}$$

$$Var(X) = \psi''(\eta) = \frac{d}{d\eta}\left(-\frac{a}{\eta}\right) = \frac{a}{\eta^2} = \frac{a}{\theta^2}$$ ■

Lecture 12

***k*-Parameter Exponential Family**

Definition 12.1: *A pdf or pmf belongs to a k-parameter exponential family, if it can be expressed as:*

$$f_{\boldsymbol{\theta}}(x) = \exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})T_i(x) + S(x) + d(\boldsymbol{\theta})\right\} \cdot I_A(x) \qquad (12.1)$$

with $\boldsymbol{\theta}$ denotes the vector of unknown parameters, and support A doesn't depend on $\boldsymbol{\theta}$

Example 12.1.1: Prove if the pdf or pmf of the specified r.v. X belongs to 2-parameter exponential family.

1). $X \sim N(\mu, \sigma^2)$, both μ, σ^2 are unknown, or $\boldsymbol{\theta} = (\mu, \sigma^2)$

Solution: As defined, the pdf of X is $f_{\boldsymbol{\theta}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, which can be expressed as

$$\begin{aligned} f_\theta(x) &= \exp\left\{-\frac{1}{2}\left[\frac{x^2}{\sigma^2} - \frac{2\mu x}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right] + \log\left[\frac{1}{\sigma\sqrt{2\pi}}\right]\right\} \cdot I_R(x) \\ &= \exp\left\{\frac{\mu x}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \log\left[\frac{1}{\sigma}\right] + \log\left[\frac{1}{\sqrt{2\pi}}\right]\right\} \cdot I_R(x) \end{aligned}$$

with $T_1(x) = x, c_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}$; $T_2(x) = x^2$, $c_2(\boldsymbol{\theta}) = -\frac{1}{\sigma^2}$; $d(\boldsymbol{\theta}) = -\frac{\mu^2}{2\sigma^2} + \log\left(\frac{1}{\sigma}\right), S(x) = \log\left(\frac{1}{\sqrt{2\pi}}\right)$ the specified pdf of X has the same structure as (12.1) and the support R does not depend on $\boldsymbol{\theta}$. Therefore, pdf of $N(\mu, \sigma^2)$ belongs to 2-parameter exponential family.

2). $X \sim \gamma(a, p)$, both a, p are unknown, or $\boldsymbol{\theta} = (a, p)$

Solution: As defined, the pdf of X is $f_\theta(x) = \frac{p^a e^{-px} x^{a-1}}{\Gamma(a)}, x > 0$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\{a \log p - px + (a-1)\log x - \log\Gamma(a)\} \cdot I_{\{x>0\}}(x) \\ &= \exp\{a \log x - px + a \log p - \log\Gamma(a) - \log x\} \cdot I_{\{x>0\}}(x) \end{aligned}$$

with $T_1(x) = \log x, c_1(\boldsymbol{\theta}) = a; T_2(x) = x, c_2(\boldsymbol{\theta}) = -p; d(\boldsymbol{\theta}) = a \log p - \log\Gamma(a), S(x) = -\log x$, the specified pdf of X has the same structure as (12.1), where $(0, \infty)$ does not depend on $\boldsymbol{\theta}$. Therefore, pdf of $\gamma(a, p)$ belongs to 2-parameter exponential family.

3). $X \sim \beta(a, b)$, both a, b are unknown, or $\boldsymbol{\theta} = (a, b)$

Solution: As defined, the pdf of X is $f_{\boldsymbol{\theta}}(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, 0 \le x \le 1$, which can be expressed as,

$$\begin{aligned} f_\theta(x) &= \exp\{(a-1)\log x + (b-1)\log(1-x) - \log B(a,b)\} \cdot I_{\{0\le x\le 1\}}(x) \\ &= \exp\{a \log x + b \log(1-x) - \log B(a,b) + \log x(1-x)\} \cdot I_{\{0\le x\le 1\}}(x) \end{aligned}$$

with $T_1(x) = \log x, c_1(\boldsymbol{\theta}) = a; T_2(x) = \log(1-x), c_2(\boldsymbol{\theta}) = b; d(\boldsymbol{\theta}) = -\log B(a,b), S(x) = \log x(1-x)$, the specified pdf of X has the same structure as (12.1) and the support $[0,1]$ does not depend on $\boldsymbol{\theta}$. Therefore, pdf of $\beta(a, b)$ belongs to 2-parameter exponential family. ■

Location and Scale Family

Theorem 12.2: *Let $f(x)$ be any pdf and let μ and $\sigma > 0$ be any given constant. Then the function*

$$g(x, \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \tag{12.2}$$

is also a pdf.

Proof: We just need to check that $g(x, \mu, \sigma)$ is nonnegative and integrates to 1. Since $f(x)$ is a pdf, $f(x) \ge 0$ for all values of x. So together with $\sigma > 0$, we have $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \ge 0$. Next, we note that

$$\int_{-\infty}^{\infty} g(x, \mu, \sigma) dx = \int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx = \int_{-\infty}^{\infty} f\left(\frac{x-\mu}{\sigma}\right) d\left(\frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\infty} f(y) dy = 1$$

by making a substitution $y = \frac{x-\mu}{\sigma}$ first and then the fact $f(x)$ is a valid pdf. ■

Definition 12.3: *For some $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $f(x-\mu)$, indexed by the parameter μ, is called the* ***location family*** *with standard pdf $f(x)$; the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the* ***scale family*** *with standard pdf $f(x)$; the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ,σ), is called the* ***location-scale family*** *with standard pdf $f(x)$. μ is called the* ***location*** *parameter and σ is called the* ***scale*** *parameter.*

Example 12.3.1: Let $f_X(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & O.W. \end{cases}$, the location family can be formed by replacing x with $x - a$,

$$f_X(x-a) = \begin{cases} e^{-(x-a)}, & x \geq a \\ 0, & O.W. \end{cases}$$

where a is the location parameter

Theorem 12.4: *Let $f(\cdot)$ be any pdf and let μ be any real number, and σ be any positive real number. Then, X is a random variable with pdf $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$*

Proof: To prove the sufficiency of condition, we derive the pdf by differentiating from its cdf,

$$\begin{aligned} F_X(x) &= P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ f_X(x) &= \frac{\partial}{\partial x}F_X(x) = \frac{1}{\sigma}f_Z\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

And to prove necessity of the condition, construct the random variable $Z = \frac{X-\mu}{\sigma}$, then besides immediately getting $X = \sigma Z + \mu$, we also have,

$$f_Z(z) = \sigma f_X(\mu + \sigma z) = \sigma \cdot \frac{1}{\sigma}f\left(\frac{(\mu+\sigma z)-\mu}{\sigma}\right) = f(z)$$ ■

Example 12.5.1: If $X \sim \mathcal{U}(0,a)$, and let $\frac{X-\frac{a}{2}}{a/\sqrt{12}}$, find the density function of Y

Solution: By similar technique use for the proof above, we have,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P\left(\frac{X-\frac{a}{2}}{a/\sqrt{12}} \leq y\right) = P\left(X \leq y\frac{a}{\sqrt{12}} + \frac{a}{2}\right) \\ &= F_X\left(y\frac{a}{\sqrt{12}} + \frac{a}{2}\right) \end{aligned}$$

Where the domain also needs to be altered by: $0 \leq y\frac{a}{\sqrt{12}} + \frac{a}{2} \leq a \Rightarrow -\sqrt{3} \leq y \leq \sqrt{3}$

$$\begin{aligned} f_X(x) &= \frac{a}{\sqrt{12}} f_X\left(y\frac{a}{\sqrt{12}}+\frac{a}{2}\right) = \frac{a}{\sqrt{12}} \cdot \frac{1}{a} = \frac{1}{\sqrt{12}} \\ &= \begin{cases} \frac{1}{\sqrt{12}}, & -\sqrt{3} \le y \le \sqrt{3} \\ 0, & O.W. \end{cases} \end{aligned}$$

■

Theorem 12.5: *Let Z be a random variable with pdf $f(z)$. Suppose $E(Z)$ and $Var(Z)$ exists. If X is a random variable with pdf $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, then*

$$E(X) = \sigma E(Z) + \mu \quad \textit{and} \quad Var(X) = \sigma^2 Var(Z)$$

in particular, if $E(Z) = 0, Var(Z) = 1$, then $E(X) = \mu, Var(X) = \sigma^2$

Proof: By Theorem 12.4, if X is a random variable with pdf $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, there must exist a random variable Z' with pdf $f(z)$ and $X = \sigma Z' + \mu$. Then, properties of expectation and variance give that,

$$E(X) = E(\sigma Z' + \mu) = \sigma E(Z') + \mu = \sigma E(Z) + \mu$$

$$Var(X) = Var\left(\sigma Z^{'} + \mu\right) = \sigma^2 Var(Z') = \sigma^2 Var(Z)$$

■

Lecture 13

Inequalities and Identities

Theorem 13.1: *Let X be a random variable and let $g(x)$ be a nonnegative function. Then for any $r > 0$,*

$$P[g(X) \ge r] \le \frac{E[g(X)]}{r} \tag{13.1}$$

Proof: The proof is similar to that of Theorem 7.1,

$$\begin{aligned} E[g(X)] &= \int_R g(X) f_X(x)dx = \int_{\{g(X)<r\}} g(X) f_X(x)dx + \int_{\{g(X)\ge r\}} g(X) f_X(x)dx \\ &\ge \int_{\{g(X)\ge r\}} g(X) f_X(x)dx \left(since \int_{\{g(X)<r\}} g(X) f_X(x)dx \ge 0 \right) \\ &\ge \int_{\{g(X)\ge r\}} r f_X(x)dx = r \int_{\{g(X)\ge r\}} f_X(x)dx = rP\{g(X) \ge r\} \end{aligned}$$

The (13.1) can be immediately obtained after some rearranging. ■

Recall the Markov's inequality from lecture 7, for a random variable X and any $a > 0$, we have

$$P\{|X| \ge a\} \le \frac{E(X^2)}{a^2}$$

which can also be proved by using Theorem 13.1 and let $g(X) = X^2$, then,

$$P\{|X| \ge a\} = P\{X^2 \ge a^2\} \le \frac{E(X^2)}{a^2}$$

Example 13.1.1: If $X \sim N(0,1)$, find the upper bound for the probability $P\{|X| > 2\}$

Solution: By Markov's inequality, we have,

$$P\{|X| \geq 2\} \leq \frac{E(X^2)}{2^2} = \frac{1}{4} = 0.25$$

Another way to obtain the upper bound is by the distribution of X,

$$P\{|X| \geq 2\} \ = 2P\{X > 2\} = 2\int_2^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \leq \frac{2}{\sqrt{2\pi}}\int_2^\infty \frac{x}{2} e^{-\frac{x^2}{2}}\,dx = \frac{1}{\sqrt{2\pi}}\int_2^\infty xe^{-\frac{x^2}{2}}\,dx$$

$$= \frac{1}{\sqrt{2\pi}}\left(-e^{-\frac{x^2}{2}}\right)\Big|_2^\infty = \frac{e^{-\frac{2^2}{2}}}{\sqrt{2\pi}} \approx 0.054$$

If the distribution of X is known, then the second approach gives a better bound for the probability asked. However, if information about distribution is not provided, you can always apply Markov's Inequality for a rough estimation.

Theorem 13.2 (Stein's Lemma): *Let $X \sim N(\mu, \sigma^2)$ and let g be a differentiable function satisfying $E[g'(x)] < \infty$, then,*

$$E[g(X)(X-\mu)] = \sigma^2\, E[g'(X)] \qquad (13.2)$$

Proof: By definition of expectation of a function and apply integration by part on the left-hand side,

$$E[g(X)(X-\mu)] \ = \int_{-\infty}^{\infty} g(x)(x-\mu)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} g(y+\mu)e^{-\frac{y^2}{2\sigma^2}} \cdot ydy$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\left[g(y+\mu)\left(-\sigma^2 e^{-\frac{y^2}{2\sigma^2}}\right)\Bigg|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g'^{(y+\mu)}\left(-\sigma^2 e^{-\frac{y^2}{2\sigma^2}}\right)dy\right] \qquad (13.2.1)$$

$f(x)$ is maximized at $x = \mu$, or in other words: $\exists C_0$, such that $\forall x$, $f(x) \leq C_0$. Following the given condition that $E[g'(x)] < \infty$, we can have $g(x) < \infty$ verified after establishing:

$$E[g'(X)] \ = \int g'(x) f(x)dx \geq C_0 \int g'(x)\,dx = C_0 g(x) < \infty$$

which indicates the first term in the parenthesis of (13.2.1) is equal to 0,

$$g(y+\mu)\left(-\sigma^2 e^{-\frac{y^2}{2\sigma^2}}\right)\Big|_{-\infty}^{\infty} = g(x)\cdot 0 = 0$$

Therefore, we finally have,

$$E[g(X)(X-\mu)] \ = \frac{1}{\sigma\sqrt{2\pi}}\left[\int_{-\infty}^{\infty} g'(y+\mu)\left(\sigma^2 e^{-\frac{y^2}{2\sigma^2}}\right)dy\right] = \sigma^2 \int_{-\infty}^{\infty} g'(y+\mu)\frac{1}{\sigma\sqrt{2\pi}}\left(e^{-\frac{y^2}{2\sigma^2}}\right)dy$$

$$= \sigma^2 \int_{-\infty}^{\infty} g'(x)\frac{1}{\sigma\sqrt{2\pi}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dy = \sigma^2 E[g'(X)] \qquad \blacksquare$$

Example 13.2.1: If $X \sim N(0,1)$, find the third and fourth moment of X using Stein's lemma

Solution: For standard normal, the Stein's lemma can be simplified as: $E[g(X)\cdot X] = E[g'(X)]$

$$E(X^3) = E(X^2 \cdot X) = E[(X^2)'] = E(2X) = 2E(X) = 0$$

$$E(X^4) = E(X^3 \cdot X) = E[(X^3)'] = E(3X^2) = 3E(X^2) = 3$$

■

Lecture 14

Multiple Random Variable (Discrete)

Definition 14.1: *Let (X,Y) be a discrete bivariate random vector. Then the function $p(x,y)$ that maps from $\mathbb{R}^2$ to $\mathbb{R}$, defined by $p(x,y) = P(X = x, Y = y)$, is called the **joint probability mass function** or **joint pmf** of (X,Y), which should satisfy:*

i) $$p(x,y) \geq 0$$

ii) $$\sum_x \sum_y p(x,y) = 1$$

To stress the pmf is about (X,Y) rather than any other vectors, the pmf is also denoted as $p_{X,Y}(x,y)$. Furthermore, the probability of (X,Y) defined in bounded domains can be calculated by,

$$P(a \leq X \leq b, c \leq Y \leq d) = \sum_{x\in[a,b]} \sum_{y\in[c,d]} p_{X,Y}(x,y) \qquad (14.1a)$$

*And $F_{X,Y}(x,y)$, denotes the **joint cdf** of the discrete random pair (X,Y), is defined by*

$$F_{X,Y}(a,b) = \sum_{x\leq a} \sum_{y\leq b} p(x,y) \qquad (14.1b)$$

Example 14.1.1: Suppose there are two coins, green and yellow, which have the following specified chances of getting a head(0) and a tail(1).

$$p(x) = \begin{cases} \frac{1}{3}, X = 0 \\ \frac{2}{3}, X = 1 \end{cases} ; \quad p(y) = \begin{cases} \frac{1}{2}, Y = 0 \\ \frac{1}{2}, Y = 1 \end{cases}$$

Let X be the outcome of green coin and Y be the outcome of yellow coin. If both coins are flipped once and assume independence, write down the joint probability of X and Y.

Solution: Probabilities are obtained by taking products of the corresponding probabilities because of independence. The joint probability distribution is illustrated by the following table:

Joint Probability Distribution (X,Y)

	$Y = 0$	$Y = 1$
$X = 0$	$p(0,0) = 1/6$	$p(0,1) = 1/6$
$X = 1$	$p(1,0) = 1/3$	$p(1,1) = 1/3$

Definition 14.2: *Let (X, Y) be a discrete bivariate random vector with joint pmf $p(x, y)$. Then the* ***marginal pmfs*** *of X and Y are defined by $p_X(x) = P(X = x)$, $p_Y(y) = P(Y = y)$, given by,*

$$p_X(x) = \sum_y p(x, y) \quad and \quad p_Y(y) = \sum_x p(x, y) \tag{14.2}$$

Let $F_i = \{Y_i = y_i\}$, $E = \{X = x\}$ and if the sequence of events $F_1, F_2, \ldots F_n$ make a partition of the sample space S, the marginal pmf defined by (14.2.1) can be established by Distributive Law,

$$P(X = x) \ = P(E \cap S) = P\left(E \cap \left(\bigcup_{i=1}^{n} F_i\right)\right) = P\left(\bigcup_{i=1}^{n} (E \cap F_i)\right) = \sum_{i=1}^{n} P(E \cap F_i)$$

$$= \sum_{i=1}^{n} P(X = x, Y_i = y_i) = \sum_y p(x, y)$$

Similar reasoning can be used to verify (14.2.2).

Example 14.3.1: The random pair (X, Y) has the distribution

		$X = x$		
$Y = y$		1	2	3
	2	1/12	1/6	1/12
	3	1/6	0	1/6
	4	0	1/3	0

Find the marginal pmf $p_X(x)$ and $p_Y(y)$

Solution: By definition, we have 3 marginal pmfs for each random variable corresponding to 3 values of the other one,

$$p_Y(2) = \frac{1}{12} + \frac{1}{6} + \frac{1}{12} = \frac{1}{3}; \qquad p_Y(3) = \frac{1}{6} + 0 + \frac{1}{6} = \frac{1}{3}; \qquad p_Y(4) = 0 + \frac{1}{3} + 0 = \frac{1}{3};$$

$$p_X(1) = \frac{1}{12} + \frac{1}{6} + 0 = \frac{1}{4}; \qquad p_X(2) = \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2}; \qquad p_X(3) = \frac{1}{12} + \frac{1}{6} + 0 = \frac{1}{4}$$

Definition 14.3: *Let (X, Y) be a discrete bivariate random vector with joint pmf $p(x, y)$ and marginal pmfs $p_X(x)$ and $p_Y(y)$. For any x such that $p_X(x) > 0$, the* ***conditional pmf*** *of Y given that $X = x$ is the function of y denoted by $p(y|x)$ and defined by*

$$p(y|x) = \frac{p(x, y)}{p_X(x)} \tag{14.3.1}$$

For any y such that $p_Y(y) > 0$, the ***conditional pmf*** *of X given that $Y = y$ is the function of x denoted by $p(x|y)$ and defined by*

$$p(x|y) = \frac{p(x, y)}{p_Y(y)} \tag{14.3.2}$$

Definition 14.4: *Let (X, Y) be a discrete bivariate random vector with joint pmf $p(x, y)$ and marginal pmfs $p_X(x)$ and $p_Y(y)$.Then X and Y are called* ***independent random variables*** *if any one of the following equalities holds for every $x \in \mathbb{R}, y \in \mathbb{R}$:*

a. $$p(x|y) = p(x) \tag{14.4a}$$

b. $$p(y|x) = p(y) \tag{14.4b}$$

c. $$p(x, y) = p_X(x)p_Y(y) \tag{14.4c}$$

Lemma 14.4.1: *Let (X, Y) be a discrete bivariate random vector with joint pmf $p(x, y)$. Then X and Y are* ***independent*** *if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathbb{R}, y \in \mathbb{R}$,*

$$p(x, y) = g(x)h(y) \tag{14.4.1}$$

Lemma 14.4.2: *Let (X, Y) be a discrete bivariate random vector with joint cdf $F_{X,Y}(a, b)$. Then X and Y are* ***independent*** *if and only if there exist functions $g(a)$ and $h(b)$ such that, for every $a \in \mathbb{R}, b \in \mathbb{R}$,*

$$F_{X,Y}(a, b) = g(a)h(b) \tag{14.4.2}$$

The continuous-random-variable version of these two lemmas will be introduced later, where the proofs will be given in details.

Example 14.5.1: If X and Y are independent random variable, evaluate the following limits.

1). $\lim_{a,b\to+\infty} F_{X,Y}(a, b)$

Solution: $\lim_{a,b\to+\infty} F_{X,Y}(a, b) = \lim_{a,b\to+\infty} F_X(a)F_Y(b) = \lim_{a\to\infty} F_X(a) \lim_{b\to\infty} F_Y(b) = 1$

2). $\lim_{a\to+\infty} F_{X,Y}(a, b)$

Solution: $\lim_{a\to+\infty} F_{X,Y}(a, b) = \lim_{a\to+\infty} F_X(a)F_Y(b) == \lim_{a\to\infty} F_X(a) \cdot F_Y(b) = F_Y(b)$

3). $\lim_{b\to+\infty} F_{X,Y}(a, b)$

Solution: $\lim_{b\to+\infty} F_{X,Y}(a, b) = \lim_{b\to+\infty} F_X(a)F_Y(b) = \lim_{b\to\infty} F_Y(b) \cdot F_X(a) = F_X(a)$

4). $\lim_{a\to-\infty} F_{X,Y}(a, b)$

Solution: $\lim_{a\to-\infty} F_{X,Y}(a, b) = \lim_{a\to-\infty} F_X(a)F_Y(b) = \lim_{a\to-\infty} F_X(a) \cdot F_X(b) = 0$

The four limits above still hold true if the X, Y are *not independent*, which will be shown in Theorem 15.3

Definition 14.6: *Let (X, Y) be a discrete bivariate random vector with joint pmf $p(x, y)$ and $g(x, y)$ be a real-valued function defined for all possible values (x, y). Then $g(X, Y)$ is itself a random variable and its* ***expected value*** *$E[g(X, Y)]$ is given by,*

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y) \tag{14.6}$$

Theorem 14.7: *For any two random variable X and Y, we always have:*

a. $$E(X + Y) = E(X) + E(Y) \tag{14.7a}$$

Particularly, if X and Y are independent, the following identities are also true:

b. $E\big(g(X)h(Y)\big) = E\big(g(X)\big)E\big(h(Y)\big)$ (14.7*b*)

c. $E(XY) = E(X)E(Y)$ (14.7*c*)

d. $M_{X+Y}(t) = M_X(t)M_Y(t)$*, if mgf of* X,Y,*and* $X+Y$ *exist* (14.7*d*)

Proof:

a. Let $g(X,Y) = X+Y$, and by definition (14.6), we have,

$$\begin{aligned} E(X+Y) &= \sum_x \sum_y (x+y)p(x,y) = \sum_x x \sum_y p(x,y) + \sum_y y \sum_x p(x,y) \\ &= \sum_x x p_X(x) + \sum_y y p_Y(y) = E(X) + E(Y) \end{aligned}$$

b. By definition (14.6) and (14.4), we have,

$$\begin{aligned} E\big(g(X)h(Y)\big) &= \sum_x \sum_y g(x)h(y)p(x,y) = \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \\ &= \sum_x g(x)p_X(x) \sum_y h(y)p_Y(y) = E\big(g(X)\big)E\big(h(Y)\big) \end{aligned}$$

c. Let $g(X) = X, h(Y) = Y$, (14.7*c*) is straightforward application of (14.7*b*)

d. By definition of mgf, and followed by (14.7*b*), we have,

$$M_{X+Y}(t) = E\big[e^{(X+Y)t}\big] = E[e^{Xt}e^{Yt}] = E(e^{Xt})E(E^{Yt}) = M_X(t)M_Y(t)$$

***Remark**: (14.7*a*), (14.7*b*) and (14.7*c*) are not sufficient to ensure independence of X and Y, except X,Y are both *normal random variables*.

Example 14.7.1: Let $X_1, X_2, \dots X_k$ be k independent random variables, with specified distribution, determine what is the distribution of their sum $\sum_{i=1}^k X_i$.

1). $\{X_i\}_{i=1,2,\dots,k} \sim \mathcal{B}ernoulli(p)$

Solution: The pmf of each X_i is, $p(x_i) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$, so the mgf of each X_i

$$M_{X_i}(t) = E[e^{X_i t}] = e^t p + e^0(1-p) = e^t + 1 - p$$

Then, by independence and identity (14.7*d*), the mgf of $\sum_{i=1}^n X_i$ is,

$$\begin{aligned} M_{\sum_{i=1}^k X_i}(t) &= M_{X_1}(t) \cdot M_{X_2}(t) \cdot \ \dots \ \cdot M_{X_k}(t) = (e^t + 1 - p) \dots (e^t + 1 - p) \\ &= (e^t + 1 - p)^k \end{aligned}$$

Uniquesness of mgf will finally identify that $\sum_{i=1}^k X_i \sim Bin(k,p)$

2). $\{X_i\}_{i=1,2,\dots,n} \sim \mathcal{B}in(n_i,p)$

Solution: The mgf of each X_i is,

$$M_{X_i}(t) \;=\; (e^t+1-p)^{n_i}$$

Then, by independence and identity (14.7d), the mgf of $\sum_{i=1}^{n} X_i$ is,

$$M_{\sum_{i=1}^{n} X_i}(t) \;= M_{X_1}(t)\cdot M_{X_2}(t)\cdot\;\ldots\;\cdot M_{X_k}(t) = \;(e^t+1-p)^{n_1}\cdot\cdots\cdot(e^t+1-p)^{n_k}$$

$$=\;(e^t+1-p)^{n_1+n_2+\cdots+n_k} =\;(e^t+1-p)^{\sum_{i=1}^{k} n_i}$$

Uniquesness of mgf will finally identify that $\sum_{i=1}^{n} X_i \;\sim Bin\left(\sum_{i=1}^{k} n_i\,,p\right)$

3). $\{X_i\}_{i=1,2,\ldots,n} \sim \mathcal{P}(\lambda_i)$

Solution: The mgf of each X_i is,

$$M_{X_i}(t) \;=\; \exp\left(\lambda_i(e^t-1)\right)$$

Then, by independence and identity (14.7d), the mgf of $\sum_{i=1}^{n} X_i$ is,

$$M_{\sum_{i=1}^{n} X_i}(t) \;= M_{X_1}(t)\cdot M_{X_2}(t)\cdot\;\ldots\;\cdot M_{X_k}(t) = \prod_{i=1}^{n}\exp\left(\lambda_i(e^t-1)\right)$$

$$= \exp[(\lambda_1+\lambda_2+\ldots+\lambda_n)(e^t-1)] = \exp\left[\left(\sum_{i=1}^{n}\lambda_i\right)(e^t-1)\right]$$

Uniquesness of mgf will finally identify that $\sum_{i=1}^{n} X_i \;\sim\mathcal{P}(\sum_{i=1}^{n}\lambda_i)$

4). $\{X_i\}_{i=1,2,\ldots,n} \sim \mathcal{G}(p)$

Solution: The mgf of each X_i is,

$$M_{X_i}(t) \;=\; \frac{e^t p}{1-e^t(1-p)},\; |t| < \ln\frac{1}{q}$$

Then, by independence and identity (14.7d), the mgf of $\sum_{i=1}^{n} X_i$ is,

$$M_{\sum_{i=1}^{n} X_i}(t) \;= M_{X_1}(t)\cdot M_{X_2}(t)\cdot\;\ldots\;\cdot M_{X_k}(t) = \prod_{i=1}^{r}\frac{e^t p}{1-e^t(1-p)}$$

$$= \left[\frac{e^t p}{1-e^t(1-p)}\right]^r$$

Uniquesness of mgf will finally identify that $\sum_{i=1}^{n} X_i \;\sim\mathcal{NB}(r,p)$

Lecture 15

<u>Multiple Random Variable (Continuous)</u>

Definition 15.1: *Let (X,Y) be a continuous bivariate random vector. Then the function $f(x,y)$ that maps from $\mathbb{R}^2$ to $\mathbb{R}$, defined by $f(x,y) = \;f_{X,Y}(x,y)$, is called the **joint probability density function** or **joint pdf** of (X,Y), which, for every $A \in \mathbb{R}^2$, we should satisfy:*

i) $\quad f(x,y) \geq 0$

ii) $\quad \iint\limits_{\mathbb{R}^2} f(x,y) = 1$

Furthermore, the probability of (X,Y) defined in bounded domains can be calculated by,

$$P[(X,Y)\in A]=\iint\limits_{A} f_{X,Y}(x,y)dydx=\iint\limits_{A} f_{X,Y}(x,y)dxdy \qquad (15.1a)$$

In particular, $F_{X,Y}(x,y)$, denotes the ***joint cdf*** *of the continuous random pair (X,Y) , is defined by*

$$F_{X,Y}(s,t)=\int_{-\infty}^{s}\int_{-\infty}^{t} f_{X,Y}(x,y)\,dydx=\int_{-\infty}^{t}\int_{-\infty}^{s} f_{X,Y}(x,y)\,dxdy \qquad (15.1b)$$

From the bivariate Fundamental Theorem of Calculus, this implies that,

$$f_{X,Y}(x,y)=\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y) \qquad (15.1c)$$

Example 15.1.1: Suppose we want make the function $f_{X,Y}(x,y)=\begin{cases} ce^{-2x-y}, x\geq 0, y\geq 0 \\ 0, \qquad\qquad O.W. \end{cases}$ a joint pdf, where c is a constant, determine the value of c

Proof: To make $f_{X,Y}(x,y)$ a joint pdf, we need to have it satisfied the two conditions in definition 15.1. From condition i)., we must have,

$$c>0$$

From condition ii). , we must have,

$$\int_{0}^{\infty}\int_{0}^{\infty} ce^{-2x-y}dxdy=1$$

By calculating of the double integration, we have,

$$\int_{0}^{\infty}\int_{0}^{\infty} ce^{-2x-y}dxdy=c\int_{0}^{\infty} e^{-2x}\,dx\int_{0}^{\infty} e^{-y}\,dyc\cdot\left[\left(-\frac{1}{2}e^{-2x}\right)\Big|_{0}^{\infty}\right]\cdot\left[(-e^{-y})\Big|_{0}^{\infty}\right]=\frac{1}{2}c=1$$

Therefore, $c=2$, which clearly satisfies the first condition, is the answer. ■

Definition 15.2: *Let (X,Y) be a continuous bivariate random vector with joint pdf $f(x,y)$. Then the* ***marginal pdfs*** *of X and Y are defined by $f_X(x)$, $f_Y(y)$, given by,*

$$f_X(x)=\int_{y} f_{X,Y}(x,y)dy \quad and \quad f_Y(y)=\int_{x} f_{X,Y}(x,y)dx \qquad (15.2)$$

Example 15.3: *For any two random variables, X and Y,*

a) $$\lim_{a,b\to+\infty} F_{X,Y}(a,b)=1 \qquad (15.3a)$$

b) $$\lim_{a\to+\infty} F_{X,Y}(a,b)=F_Y(b) \qquad (15.3b)$$

c) $$\lim_{b\to+\infty} F_{X,Y}(a,b)=F_X(a) \qquad (15.3c)$$

d) $$\lim_{a\ or\ b\to-\infty} F_{X,Y}(a,b)=0 \qquad (15.3d)$$

Proof: from (15.1b), the joint cdf can be expanded as:

$$F_{X,Y}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)\,dydx$$

For *b)*, we have,

$$\underset{a\to+\infty}{lim} F_{X,Y}(a,b) = \lim_{a\to+\infty}\int_{-\infty}^{b}\int_{-\infty}^{a} f_{X,Y}(x,y)\,dxdy = \int_{-\infty}^{b}\left[\lim_{a\to+\infty}\int_{-\infty}^{a} f_{X,Y}(x,y)dx\right]dy$$

(by 15.2)

$$= \int_{-\infty}^{b}\left[\int_{x} f_{X,Y}(x,y)dx\right]dy = \int_{-\infty}^{b} f_Y(y)dy = F_Y(b)$$

and *c)* can be shown in the similar way. Then, we want to establish *a)* by following from *b)*,

$$\underset{a,b\to+\infty}{lim} F_{X,Y}(a,b) = \lim_{b\to+\infty}\left[\underset{a\to+\infty}{lim} F_{X,Y}(a,b)\right] = \lim_{b\to+\infty} F_Y(b) = \lim_{b\to+\infty}\int_{-\infty}^{b} f_Y(y)dy = \int_{-\infty}^{+\infty} f_Y(y)dy = 1$$

Similarly, *d)* can also be proved. ■

Definition 15.4: *Let (X,Y) be a continuous bivariate random vector with joint pdf $f(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the* ***conditional pdf*** *of Y given that $X = x$ is the function of y denoted by $f(y|x)$, or more specifically $f_{Y|X}(y|x)$, and defined by*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{15.4.1}$$

For any y such that $f_Y(y) > 0$, the ***conditional pdf*** *of X given that $Y = y$ is the function of x denoted by $f(x|y)$, or more specifically $f_{X|Y}(x|y)$ and defined by*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{15.4.2}$$

Definition 15.5: *Let (X,Y) be a continuous bivariate random vector with joint pdf $p(x,y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$.Then X and Y are called* ***independent random variables*** *if, for every $x \in \mathbb{R}, y \in \mathbb{R}$,*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{15.5}$$

Lemma 15.5.1: *Let (X,Y) be a continuous bivariate random vector with joint pdf $f(x,y)$.Then X and Y are* ***independent*** *if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \mathbb{R}, y \in \mathbb{R}$,*

$$f(x,y) = g(x)h(y) \tag{15.5.1}$$

Proof: Necessity can be easily verified by following the definition (14.4) and defining $(x) = f_X(x)$, $h(y) = f_Y(y)$. To prove sufficiency of this condition, suppose that $f(x,y) = g(x)h(y)$. We have,

$$\int_{-\infty}^{\infty} g(x)dx \int_{-\infty}^{\infty} h(y)dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)dydx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$$

And the definition of marginal pdf gives that,

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

$$f_Y(y) \ = \int_{-\infty}^{\infty} f(x,y)dx = \int_{-\infty}^{\infty} g(x)h(y)dx = h(y)\int_{-\infty}^{\infty} g(x)dx$$

Then for all x, y, such that $g(x) \neq 0, h(y) \neq 0$, we have,

$$\frac{f_X(x)}{g(x)} = \int_{-\infty}^{\infty} h(y)dy \quad and \quad \frac{f_Y(y)}{h(y)} = \int_{-\infty}^{\infty} g(x)dx$$

By multiplying this two equalities, we have,

$$\frac{f_X(x)f_Y(y)}{g(x)h(y)} \ = \int_{-\infty}^{\infty} g(x)dx \int_{-\infty}^{\infty} h(y)dy = 1$$

After some rearranging, we finally get,

$$f_X(x)f_Y(y) \ = g(x)h(y) = f(x,y)$$

which indicating that X and Y are independent by definition (15.5). ■

Lemma 15.5.2: *Let (X, Y) be a continuous bivariate random vector with joint cdf $F_{X,Y}(s,t)$. Then X and Y are* ***independent*** *if and only if there exist functions $g(s)$ and $h(t)$ such that, for every $s \in \mathbb{R}, t \in \mathbb{R}$,*

$$F_{X,Y}(s,t) = g(s)h(t) \tag{15.5.2}$$

Proof: In order to show Necessity of (15.5.2), suppose X and Y are independent, by definition, we have,

$$F_{X,Y}(s,t) \ = \int_{-\infty}^{s}\int_{-\infty}^{t} f(x,y)\,dydx = \int_{-\infty}^{s}\int_{-\infty}^{t} f_X(x)f_Y(y)\,dydx = \int_{-\infty}^{s} f_X(x)dx \int_{-\infty}^{t} f_Y(y)\,dy$$
$$= F_X(s)F_Y(t)$$

Sufficiency can be shown by establishing the equality (15.5) holds, suppose $F_{X,Y}(s,t) = g(s)h(t)$,

$$f_{X,Y}(s,t) \ = \frac{\partial^2}{\partial s \partial t}F_{X,Y}(s,t) = \frac{\partial^2}{\partial s \partial t}g(s)h(t) = \frac{\partial}{\partial s}[g(s)h'(t)] = g'(s)h'(t)$$

where $g'(s)$ is a function involving only s (values that X can take), and $h'(t)$ is a function only involving t (values that Y can take), then Lemma 15.5.1 gives that X and Y are independent. ■

Example 15.6: Points in $\mathbb{R}^2$ are uniformly distributed over a circle, which centered at $(0,0)$ with radius $r > 0$. Let $m(X, Y)$ be any point in the circle and answer the following questions:

i. Find the joint pdf $f_{X,Y}(x,y)$
ii. Find the probability $P\{\sqrt{X^2 + Y^2} < a\}$, where $a < r$
iii. Find the marginal pdf of Y, $f_Y(y)$
iv. Find the conditional pdf given $Y = y$, $f_{X|Y}(x|y)$
v. Show whether X and Y are independent

Solution:

i. Since (X, Y) are uniformly distributed, the pdf $f_{X,Y}(x, y) = c$ where c is a constant and,

$$\iint_{x^2+y^2\leq r^2} c\, dxdy = 1$$

Then the double integration on the left gives that,

$$\iint_{x^2+y^2\leq r^2} c\, dxdyc = c\pi r^2 = 1$$

yielding that $c = \frac{1}{\pi r^2}$. Therefore, the joint pdf is,

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi r^2}, x^2 + y^2 \leq r^2 \\ 0, \qquad O.W. \end{cases} \tag{15.6}$$

ii. By definition (15.1a) and the joint pdf found in (15.5), we have,

$$P\left\{\sqrt{X^2+Y^2} < a\right\} = P\{X^2 + Y^2 < a^2\} = \iint_{x^2+y^2<a^2} f_{X,Y}(x, y)\, dxdy = \iint_{x^2+y^2<a^2} \frac{1}{\pi r^2} dxdy$$

$$= \frac{1}{\pi r^2} \iint_{x^2+y^2<a^2} dxdy = \frac{1}{\pi r^2} \cdot \pi a^2 = \frac{a^2}{r^2}$$

iii. As defined in Definition 15.2, we have,

$$f_Y(y) = \int_x f_{X,Y}(x, y) dx = \int_{-\sqrt{r^2-y^2}}^{\sqrt{r^2-y^2}} \frac{1}{\pi r^2} dx = \begin{cases} \frac{2\sqrt{r^2-y^2}}{\pi r^2}, -r < y < r \\ 0, \qquad O.W. \end{cases}$$

$$= \frac{1}{\pi r^2} \iint_{x^2+y^2<a^2} dxdy = \frac{1}{\pi r^2} \cdot \pi a^2 = \frac{a^2}{r^2}$$

iv. Definition of conditional pdf gives that,

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{\frac{1}{\pi r^2}}{\frac{2\sqrt{r^2-y^2}}{\pi r^2}} = \frac{1}{2\sqrt{r^2-y^2}}$$

v. By comparing *iii).* and *iv).*, we have $f_X(x) \neq f_{X|Y}(x|y)$, so X and Y are not independent. ■

Definition 15.7: *Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and $g(x, y)$ be a real-valued function defined for all possible values (x, y). Then $g(X, Y)$ is itself a random variable and its* ***expected value*** *$E[g(X, Y)]$ is given by,*

$$E[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)f(x,y)dxdy \tag{15.7}$$

All the identities in **Theorem 14.7** are also applicable to continuous random variables.

Lecture 16

Sum of Two Independent Random Variables

Theorem 16.1: *Let X and Y be* ***independent discrete random variables*** *with pmf $p_X(x)$ defined on $x = 0,1,2,\dots,n$, and $p_Y(y)$ defined on $y = 0,1,2,\dots,n$, and the sum $Z = X + Y$ has pmf given by,*

$$P(Z = n) = \sum_{k=0}^{n} P(X = k)P(Y = n - k) \tag{16.1}$$

Proof: The event $\{Z = n\}$ can be decomposed by a sequence of disjoint events $\{X = k, Y = n - k\}$, for $k = 0,1,2,\dots,n$

$$\begin{aligned} P(Z = n) &= P\{X + Y = n\} = P\left(\bigcup_{k=0}^{n}\{X = k, Y = n - k\}\right) \\ &= \sum_{k=0}^{n} P\{X = k, Y = n - k\} \\ &= \sum_{k=0}^{n} P(X = k)P(Y = n - k) \end{aligned}$$

the last equality is given by independence. ■

Example 16.1.1: If $X \sim \mathcal{P}(\lambda_1), Y \sim \mathcal{P}(\lambda_2)$, determine the pdf of $X + Y$

Solution: From Theorem 16.1 and pdf of Poisson distribution, we have,

$$\begin{aligned} P\{X + Y = n\} &= \sum_{k=0}^{n} P(X = k)P(Y = n - k) = \sum_{k=0}^{n} \frac{e^{-\lambda_1}{\lambda_1}^k}{k!} \cdot \frac{e^{-\lambda_2}{\lambda_2}^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{{\lambda_1}^k}{k!} \cdot \frac{{\lambda_2}^{n-k}}{(n-k)!} = \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \binom{n}{k} {\lambda_1}^k {\lambda_2}^{n-k} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} (\lambda_1 + \lambda_2)^n \end{aligned}$$

which is the pdf of a $\mathcal{P}(\lambda_1 + \lambda_2)$ distribution. ■

Example 16.1.2: If $X \sim Bin(m_1, p), Y \sim Bin(m_2, p)$, determine the pdf of $X + Y$

Solution: From Theorem 16.1 and pdf of Binomial distribution, we have,

$$\begin{aligned}P\{X+Y=m\} &= \sum_{k=0}^{m} P(X=k)P(Y=m-k) = \sum_{k=0}^{m} \binom{m_1}{k} p^k q^{m_1-k} \binom{m_2}{m-k} p^{m-k} q^{m_2-(m-k)} \\ &= p^m q^{(m_1+m_2-m)} \sum_{k=0}^{m} \binom{m_1}{k}\binom{m_2}{m-k} \\ &= p^m q^{(m_1+m_2-m)} \binom{m_1+m_2}{m}\end{aligned}$$

which is the pdf of a $Bin(m_1+m_2,p)$ distribution. ■

Example 16.3: If $X\sim\mathcal{P}(\lambda), Y\sim\mathcal{P}(\lambda)$ are independent, find the conditional pmf $p_{X|X+Y}(x|x+y)$.

Solution: From definition of conditional probability, we have,

$$\begin{aligned}P(X=m|X+Y=n) &= \frac{P(X=m|X+Y=n)}{P(X+Y=n)} = \frac{P(X=m)P(Y=n-m)}{P(X+Y=n)} \\ &= \frac{\frac{e^{-\lambda}\lambda^m}{m!}\cdot\frac{e^{-\lambda}\lambda^{n-m}}{(n-m)!}}{\frac{e^{-2\lambda}(2\lambda)^n}{n!}} = \binom{n}{m}\left(\frac{1}{2}\right)^n = \binom{n}{m}\left(\frac{1}{2}\right)^m\left(\frac{1}{2}\right)^{n-m}\end{aligned}$$

which is by definition the pdf of a $Bin\left(n,\frac{1}{2}\right)$ distribution ■

Theorem 16.2: *Let X and Y be* ***independent continuous random variables*** *with pdf $f_X(x)$ and $f_Y(y)$ defined on $\mathbb{R}^2$, and the sum $Z=X+Y$ has pdf given by,*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \tag{16.2}$$

Proof: By Fundamental Calculus Theorem, the pdf can be obtained by differentiating the cdf, which is,

$$\begin{aligned}F_Z(z) &= P(Z\le z) = P(X+Y\le z) = \iint_{x+y\le z} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\left[f_Y(y)\int_{-\infty}^{z-y} f_X(x)dx\right]dy = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy\end{aligned}$$

where the last equality is followed by independence. Furthermore, the pdf is given by,

$$\begin{aligned}f_Z(z) &= \frac{d}{dz}\left[\int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy\right] = \int_{-\infty}^{\infty}\frac{\partial}{\partial z}[F_X(z-y)f_Y(y)dy] \\ &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy\end{aligned}$$ ■

Example 16.2.1: If $X\sim\mathcal{U}(0,1), Y\sim\mathcal{U}(0,1)$, let $Z=X+Y$, determine the pdf of Z

Solution: For any given $y \in [0,1]$, the construction $z = x + y$ implies $z \in [y, y+1]$, since $x \in [0,1]$. And thus $y \in [z-1, z]$. Then, from Theorem 16.2, we have,

$$f_Z(z) \ = \int_{(\mathbf{0},\mathbf{1})\cap(\boldsymbol{z}-\mathbf{1},\boldsymbol{z})} f_X(z-y)f_Y(y)dy = \begin{cases} 0, & z < 0 \\ \int_0^z 1dy = z, & 0 \le z < 1 \\ \int_{z-1}^1 1dy = 2 - z, & 1 \le z < 2 \\ 0, & z \ge 2 \end{cases}$$

■

Bivariate Transformation

Theorem 16.3: *Let (X, Y) be a bivariate random vector with joint pdf $f_{X,Y}(x, y)$, and (U, V), a new bivariate random vector defined by $U = \varphi_1(X, Y), V = \varphi_2(X, Y)$, where φ_1 and φ_2 are one-to-one transformations maps from $\mathcal{A} = \{(x, y): f_{X,Y}(x, y) > 0\}$ onto $\mathcal{B} = \{(u, v): u = \varphi_1(x, y), v = \varphi_2(x, y),$ for some $(x, y) \in \mathcal{A}\}$, then the joint pdf of (U, V) can be expressed in terms of $f_{X,Y}(x, y)$ by,*

$$f_{U,V}(u, v) = f_{X,Y}\left(\varphi_1{}^{-1}(u, v), \varphi_2{}^{-1}(u, v)\right) \cdot |J| \qquad (16.3)$$

*where J, called the **Jacobian of the transformation,** is the determinant of a matrix of partial derivatives of $x(u, v) = \varphi_1{}^{-1}(u, v)$ and $y(u, v) = \varphi_2{}^{-1}(u, v)$, defined by,*

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left|\frac{\partial x}{\partial u} \cdot \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \cdot \frac{\partial y}{\partial u}\right|$$

In particular, if X and Y are independent, (16.3) can be written as,

$$f_{U,V}(u, v) = f_X\left(\varphi_1{}^{-1}(u, v)\right) \cdot f_Y\left(\varphi_2{}^{-1}(u, v)\right)|J|$$

Example 16.3.1: $U = X + Y, V = Y$, find $f_{U,V}(u, v)$

Solution: The equations $u = x + y, v = y$ can be uniquely solved by,

$$x = u - v \qquad and \qquad y = u$$

Then, the Jacobian is given by,

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

Therefore, by Theorem 16.3, the joint pdf of (U, V) is,

$$f_{U,V}(u, v) = f_X(u - v)f_Y(u)$$

■

Example 16.3.2: Suppose $X \sim Exp(\lambda), Y \sim Exp(\lambda)$, and let $U = X + Y$ and $V = \frac{X}{X+Y}$, answer the following questions:

i. Find $f_{U,V}(u, v)$
ii. Find $f_V(v)$
iii. Find $f_U(u)$
iv. Show that U and V are independent

Solution: The pdf X and Y are $f_X(x) = \lambda e^{-\lambda x}$ and $f_Y(y) = \lambda e^{-\lambda x}$, defined on:

$$\mathcal{A} = \{(x, y): f_{X,Y}(x, y) > 0\} = \{(x, y): x \geq 0, y \geq 0\}$$

i. So the set of possible values for V and U are $0 \leq v \leq 1$, $u \geq 0$, considering the transformation function $U = X + Y$ and $V = \frac{X}{X+Y}$. Thus, this transformation maps from $\mathcal{A}$ onto $\mathcal{B} = \{(u, v): u \geq 0, 0 \leq v \leq 1\}$. For any $(u, v) \in \mathcal{B}$, the equation set $u = x + y$ and $v = \frac{x}{x+y}$ has a unique solution set, such that:

$$\begin{cases} x = uv \\ y = u(1 - v) \end{cases}$$

from which the Jacobian can be derived to be:

$$J = \begin{vmatrix} v & u \\ 1 - v & -u \end{vmatrix} = u$$

From Theorem 16.2, we have the pdf of (U, V) to be,

$$\begin{aligned} f_{U,V}(u, v) &= f_X(u - v) f_Y(u) \cdot u = \lambda e^{-\lambda uv} \cdot \lambda e^{-\lambda u(1-v)} \cdot u \\ &= \begin{cases} \lambda^2 u e^{-\lambda u}, & u \geq 0, 0 \leq v \leq 1 \\ 0, & O.W. \end{cases} \end{aligned}$$

ii. Integrate the joint pdf with respect to u, we have the marginal pdf of V,

$$\begin{aligned} f_V(v) &= \int_0^\infty \lambda^2 u e^{-\lambda u} du = \lambda \int_0^\infty u \lambda e^{-\lambda u} du = \lambda E(X) = 1 \\ &= \begin{cases} 1, & 0 \leq v \leq 1 \\ 0, & O.W. \end{cases} \end{aligned}$$

which is the pdf of a $\mathcal{U}(0,1)$ distribution.

iii. Integrate the joint pdf with respect to v, we have the marginal pdf of U,

$$\begin{aligned} f_U(u) &= \int_0^1 \lambda^2 u e^{-\lambda u} dv = \lambda^2 u e^{-\lambda u} = \begin{cases} \lambda^2 u e^{-\lambda u}, & u > 0 \\ 0, & O.W. \end{cases} \\ &= \begin{cases} \dfrac{\lambda^2 u^{2-1} e^{-\lambda u}}{\Gamma(2)}, & u > 0 \\ 0, & O.W. \end{cases} \end{aligned}$$

which is the pdf of a $\gamma(2, \lambda)$ distribution.

iv. Compare $f_{U,V}(u, v)$ and $f_U(u) \cdot f_V(v)$ by using the results above, we have that any $(u, v) \in \mathcal{B}$

$$f_{U,V}(u, v) = \lambda^2 u e^{-\lambda u} = \lambda^2 u e^{-\lambda u} \cdot 1 = f_U(u) \cdot f_V(v)$$

Therefore, $U = X + Y$ and $V = \frac{X}{X+Y}$ are independent by (15.4). ■

Lecture 17

Covariance and Correlation

Definition 17.1: *The **covariance** of X and Y, denoted by $Cov(X, Y)$ or σ_{XY}, is defined by,*

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\} \tag{17.1a}$$

which is also computed by the following formula,

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{17.1b}$$

In particular, covariance between a random variable and itself is the variance,

$$Cov(X,X) = Var(X) \tag{17.1c}$$

Theorem 17.2: *Let X_1, X_2 and Y be three random variables, and a_1, a_2 be two constants, then the first argument is linear,*

$$Cov(a_1X_1 + a_2X_2, Y) = a_1Cov(X_1, Y) + a_2Cov(X_2, Y) \tag{17.2a}$$

By symmetry, the second argument is also linear,

$$Cov(Y, a_1X_1 + a_2X_2) = a_1Cov(Y, X_1) + a_2Cov(Y, X_2) \tag{17.2b}$$

*Linearity in both the first and second argument is called **Bilinearity**.*

Example 17.2.1: Express $Var(\alpha X + \beta Y)$ in terms of $Var(X), Var(Y), Cov(X,Y)$

Solution: By (17.1*c*) and the property of Bilinearity, we have,

$$\begin{aligned} Var(\alpha X + \beta Y) &= Cov(\alpha X + \beta Y, \alpha X + \beta Y) = Cov(\alpha X, \alpha X + \beta Y) + Cov(\beta Y, \alpha X + \beta Y) \\ &= Cov(\alpha X, \alpha X) + Cov(\beta Y, \alpha X) + Cov(\alpha X, \beta Y) + Cov(\beta Y, \beta Y) \\ &= Var(\alpha X) + Var(\beta Y) + 2Cov(\alpha X, \beta Y) \\ &= \alpha^2 Var(X) + \beta^2 Var(Y) + 2\alpha\beta Cov(X,Y) \end{aligned}$$

■

Theorem 17.3: *If X and Y are independent random variables, then,*

$$Cov(X,Y) = 0 \tag{17.3}$$

But this statement is not reversible, i.e., $Cov(X,Y) = 0$ doesn't imply independence.

Definition 17.4: *The **correlation** of X and Y, denoted by $r_{X,Y}$, is defined by,*

$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{17.4}$$

The value $r_{X,Y}$ is also called correlation coefficient.

Theorem 17.5: *For any random variables X and Y,*

a. $-1 \le r_{x,y} \le 1$ (17.5)

b. $|r_{x,y}| = 1$ *if and only if there exist number $\alpha \neq 0$ and β such that $Y = \alpha X + \beta$. If $r_{x,y} = 1$, then $\alpha > 0$; and if $r_{x,y} = -1$, then $\alpha < 0$*

Proof: Consider a function $h(t)$ defined by,

$$h(t) = E[(X - \mu_X)t + (Y - \mu_Y)]^2$$

which can be expanded as,

$$\begin{aligned} h(t) &= E[(X-\mu_X)^2t^2 + 2(X-\mu_X)(Y-\mu_Y)t + (Y-\mu_Y)^2] \\ &= t^2{\sigma_X}^2 + 2tr_{X,Y}\sigma_X\sigma_Y + {\sigma_Y}^2 \end{aligned}$$

In order to guarantee this quadratic function $h(t)$ is greater or equal to 0, for ${\sigma_X}^2 > 0$, we need,

$$\Delta(h) \;= \left(2r_{X,Y}\sigma_X\sigma_Y\right)^2 - 4{\sigma_X}^2{\sigma_Y}^2 \leq 0$$

which is equivalent to ${r_{X,Y}}^2 \leq 1$ after some rearranging. ■

A better version of this proof by using Cauchy-Schwarz inequality will be introduced later.

Inequalities

Lemma 17.6: *For $a > 0, b > 0, p > 0, q > 0$, if $\frac{1}{p} + \frac{1}{q} = 1$, then,*

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab \tag{17.6}$$

Proof: Define the function $g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$ and differentiating it, we have,

$$\frac{d}{da}g(a) = a^{p-1} - b = 0 \quad \textit{and} \quad \frac{d^2}{da^2}g(a) = (p-1)a^{p-1} > 0$$

giving that $g(a)$ is minimized at $a^{p-1} = b$, with the minimal value

$$\min g(a) = \frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - a^p = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = 0$$

Therefore, for all $a > 0$, $g(a) \geq 0$, i.e., $\frac{1}{p}a^p + \frac{1}{q}b^q - ab \geq 0$, and the inequality is established. ■

Theorem 17.7 (Holder's Inequality): *Let X and Y be two random variables, suppose that for $p > 0$, $q > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$, we should have,*

$$|E(XY)| \leq E|XY| \leq (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}} \tag{17.7}$$

provided the expectations exist.

Proof: It is natural that $-|XY| \leq XY \leq |XY|$, so that $|E(XY)| \leq E|XY|$. Then, in order to show the second inequality, we let

$$a = \frac{|X|}{(E|X|^p)^{\frac{1}{p}}} \quad \textit{and} \quad b = \frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}$$

By Lemma 17.5, we have,

$$\frac{1}{p}\left[\frac{|X|}{(E|X|^p)^{\frac{1}{p}}}\right]^p + \frac{1}{q}\left[\frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}\right]^{\boldsymbol{q}} \geq \left[\frac{|X|}{(E|X|^p)^{\frac{1}{p}}}\right]\left[\frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}\right]$$

By taking expectation on both sides, which preserves the direction of inequality,

$$E\left\{\frac{1}{p}\left[\frac{|X|}{(E|X|^p)^{\frac{1}{p}}}\right]^p\right\} + E\left\{\frac{1}{q}\left[\frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}\right]^q\right\} \geq E\left\{\left[\frac{|X|}{(E|X|^p)^{\frac{1}{p}}}\right]\left[\frac{|Y|}{(E|Y|^q)^{\frac{1}{q}}}\right]\right\}$$

where the left-hand side can be simplified to be equal to $\frac{1}{p} + \frac{1}{q} = 1$. Hence,

$$1 \geq \frac{E|XY|}{(E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}}$$

which is an equivalent expression of $E|XY| \leq (E|X|^p)^{\frac{1}{p}}(E|Y|^q)^{\frac{1}{q}}$. ■

Theorem 17.8 (Cauchy-Schwarz Inequality): *For two random variables X,Y, we have,*

$$|E(XY)| \leq E|XY| \leq [E(X^2)]^{\frac{1}{2}}[E(Y^2)]^{\frac{1}{2}} \tag{17.8}$$

provided the expectations exist.

Proof: Cauchy-Schwarz inequality is a special case of Holder's inequality when $\frac{1}{p} = \frac{1}{q} = \frac{1}{2}$

Example 17.8.1: Prove Theorem 17.4: $-1 \leq r_{X,Y} \leq 1$, by using Cauchy-Schwarz inequality.

Proof: As mentioned before, Cauchy-Schwarz inequality provides an easier proof for Theorem 17.4. Replace X with $X - \mu_X$, and Y with $Y - \mu_Y$ in (17.7), we have

$$|E[(X-\mu_X)(Y-\mu_Y)]| \leq E|(X-\mu_X)(Y-\mu_Y)| \leq \left(E(X-\mu_X)^{\mathbf{2}}\right)^{\frac{1}{2}}\left(E(Y-\mu_Y)^{\mathbf{2}}\right)^{\frac{1}{2}}$$

Square both the left-most and right-most expressions and by definition of covariance, we have,

$$Cov^2(X,Y) \leq \sigma_X{}^2\sigma_Y{}^2$$

which is equivalent to $r_{X,Y}^2 \leq 1$, and thus $-1 \leq r_{X,Y} \leq 1$. ■

Sum of n Independent Random Variables

Definition 17.9: *A sequence of random variables $X_1, X_2, \dots, X_n$ are called **identical independent distributed (i.i.d.) random variables** if they are independent and have the same distribution.*

Theorem 17.10: *For a sequence of i.i.d. random variables $X_1, X_2, \dots, X_n$ with common pdf $f_X(x)$ and common mgf $M_X(t)$, they have joint pdf to be:*

$$f_{X_1,X_2,\dots X_n}(x_1, x_2, \dots x_n) = \left(f_X(x)\right)^n$$

Let $S = \sum_{i=1}^n X_i$, whose mgf can be calculated by:

$$M_S(t) = [M_X(t)]^n$$

Proof: The proofs are simple, for the joint pdf, by independence, we have,

$$f_{X_1,X_2,\dots X_n}(x_1, x_2, \dots x_n) = \prod_{i=1}^n f_{X_i}(x_i) = \left(f_X(x)\right)^n$$

where the last equality comes from that their distribution are the same as $f_X(x)$.

For $S = \sum_{i=1}^n X_i$ and by Theorem 14.7, we have,

$$M_S(t) = M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = [M_X(t)]^n$$

and the last equality is given since they all have the same mgf $M_X(t)$. ■

Example 14.7.1 has shed some lights on the distribution of sum of some i.i.d. random varibles, such as: the sum of n i.i.d. $Bernoulli(p)$ has a $Bin(n,p)$ distribution; sum of r i.i.d. $\mathcal{G}(p)$ has a $\mathcal{NB}(r,p)$ distribution. More such examples will be given next.

Example 17.10.1: Suppose $\{X_i\}_{i=1,2,\ldots,n}$ is a sequence of n i.i.d. random variables with the specified pdf, and let $S=\sum_{i=1}^{n} X_i$. Determine the distribution of S, the sum.

1). $\{X_i\}$ are i.i.d. $Exp(\lambda)$

Solution: As defined, the mgf of $Exp(\lambda)$ has been calculated to be $\boldsymbol{\frac{\lambda}{\lambda-t}}$**.** By Theorem 17.10, we have,

$$M_S(t) \;= M_{\sum_{i=1}^{n} X_i}(t) = [M_X(t)]^n = \left(\frac{\lambda}{\lambda-t}\right)^n$$

Therefore, uniqueness of mgf gives that S has a $\gamma(n,\lambda)$ distribution.

2). $\{X_i\}$ are i.i.d. $\gamma(a,p)$

Solution: As defined, the mgf of $\gamma(a,p)$ has been calculated to be $\left(\frac{p}{p-t}\right)^a$**.** By Theorem 17.10, we have,

$$M_S(t) \;= M_{\sum_{i=1}^{n} X_i}(t) = [M_X(t)]^n = \left[\left(\frac{p}{p-t}\right)^a\right]^n = \left(\frac{p}{p-t}\right)^{na}$$

Therefore, uniqueness of mgf gives that S has a $\gamma(na,p)$ distribution.

3). $\{X_i\}$ are i.i.d. $N(\mu,\sigma^2)$

Solution: The mgf of $N(\mu,\sigma^2)$ has been calculated to be $e^{\mu t+\frac{1}{2}\sigma^2 t^2}$**.** By Theorem 17.10, we have,

$$M_S(t) \;= M_{\sum_{i=1}^{n} X_i}(t) = [M_X(t)]^n = \left(e^{\mu t+\frac{1}{2}\sigma^2 t^2}\right)^n = e^{n\mu t+\frac{1}{2}n\sigma^2 t^2}$$

Therefore, uniqueness of mgf gives that S has a $N(n\mu,n\sigma^2)$ distribution. ■

Lecture 18

Conditional Expectation and Conditional Variance

Definition 18.1: *The **conditional expectation** of Y given that $X=x$ is denoted by $E(Y|X=x)$ and is given by:*

$$E(Y|X=x)=\sum_y y\,p_{Y|X}(y|x) \quad \text{and} \quad E(Y|X=x)=\int_{-\infty}^{\infty} yf_{Y|X}(y|x)dy$$

in the discrete and continuous cases, respectively. Then, $E(Y|X)$ is a random variable whose value depends on the value of X

Theorem 18.2: *If X and Y are two random variables, then,*

$$E[E(Y|X)]=E(Y) \tag{18.2}$$

*It is sometimes called the "**Smoothing Theorem**".*

Proof: By definition of conditional probability, the left expression can be expanded by,

$$E[E(Y|X)] \ = \int \left(\int y \frac{f_{X,Y}(x,y)}{f_X(x)} dy \right) f_X(x) dx = \iint y f_{X,Y}(x,y) dx dy$$
$$= \int y \left(\int f_{X,Y}(x,y) dx \right) dy = \int y \, f_Y(y) dy = E(Y)$$

Proof for discrete case is very similar by simply changing the integral to sum. ■

Example 18.2.1: A prison is in jail and attempt to escape from either Door 1, Door 2 or Door 3 with chances of 30%, 20% and 50%, respectively. If he takes Door 1, the expected days taken for his escape will be 5 days, while Door 2 expects a 2-day escape and Door 3, just 1 day. If the prisoner chooses the door randomly, what is the expected days of escape?

Solution: Let Y be the days to escape the prison, and X be the number of door. Since there are only 3 doors he can choose from, the sample space can be decomposed by events $\{X = 1\}$,$\{X = 2\}$,and $\{X = 3\}$, Then, by Theorem 18.2, we have,

$$E(Y) \ = E[E(Y|X)] = E[Y|X = 1]P(X = 1) + E[Y|X = 2]P(X = 2) + E[Y|X = 3]P(X = 3)$$
$$= 5 \cdot 30\% + 2 \cdot 20\% + 1 \cdot 50\% = 2.4$$ ■

Theorem 18.3: *If X and Y are two random variables, then,*

a. $E(XY|X) = XE(Y|X)$ (18.3*a*)

b. $E[(aX + bY)|Z] = aE(X|Z) + bE(Y|Z)$ (18.3*b*)

given that the expectations exist.

Theorem 18.4: *Let $X.Y,Z$ be three random variables, the Cauchy-Schwarz inequality for conditional expectation is defined by,*

$$E[(|XY|)|Z] \leq [E(X^2|Z)]^{\frac{1}{2}} + [E(Y^2|Z)]^{\frac{1}{2}} \tag{18.4}$$

provided that the expectations exist.

Theorem 18.5: *If X and Y are independent random variables, then,*

$$E(Y|X) = E(Y) \qquad and \qquad E(X|Y) = E(X) \tag{18.5}$$

Theorem 18.6 (Wald Lemma): *Let $X_1, \dots, X_N$ be i.i.d. with the $(X_i) = \mu < \infty$, where N is also a random variable and independent of the sequence $\{X_i\}$, then,*

$$E\left(\sum_{i=1}^{N} X_i\right) = \mu E(N) \tag{18.6}$$

Proof: From Theorem 18.2, we can rewrite the left expectation in to a compound expectation, that is,

$$E\left(\sum_{i=1}^{N} X_i\right) \ = E\left[E\left(\sum_{i=1}^{N} X_i \,|N\right)\right]$$

Expanding the right-hand expression and by linearity of expectations, we have,

$$E\left[E\left(\sum_{i=1}^{N} X_i \,|N\right)\right] \ = \sum_n E\left(\sum_{i=1}^{n} X_i\right) p_N(n) = \sum_n n\mu p_N(n) = \mu \sum_n n p_N(n)$$
$$= \mu E(N)$$ ■

Definition 18.7: *The **conditional variance** of Y given that $X = x$ is denoted by $Var(Y|X = x)$ and is given by:*

$$Var(Y|X = x) = E\left[\left(X - E(X|Y)\right)^2\middle|Y\right] \qquad (18.7)$$

Then, $Var(Y|X)$ is a random variable whose value depends on the value of X.

Theorem 18.8: *The conditional variance of Y given that $X = x$ can also be calculated by,*

$$Var(Y|X) = E(Y^2|X) - E^2(Y|X) \qquad (18.8)$$

Proof: Expand the right-hand side of Definition 18.7 and then apply identities of Theorem 18.2, we have,

$$\begin{aligned} Var(Y|X) &= E\left[\left(Y - E(Y|X)\right)^2\middle|X\right] = E\left[\left(Y^2 - 2YE(Y|X) + E^2(Y|X)\right)\middle|X\right] \\ &= E(Y^2|X) - 2E[YE(Y|X)|X] + E[E^2(Y|X)|X] \\ &= E(Y^2|X) - 2E(Y|X)E(Y|X) + E^2(Y|X) \\ &= E(Y^2|X) - E^2(Y|X) \end{aligned}$$

■

Theorem 18.9: *For any two random variables X and Y,*

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \qquad (18.9)$$

Proof: By definition, we have,

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

where the last expression is by adding and abstracting $E(Y|X)$. Expanding the square, we have,

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

The cross-product in this expression is equal to 0, which can be easily seen by applying the Smoothing Theorem and get,

$$\begin{aligned} E\left[\left(Y - E(Y|X)\right)\left(E(Y|X) - E(Y)\right)\right] &= E\left\{E\left[\left(Y - E(Y|X)\right)\left(E(Y|X) - E(Y)\right)|X\right]\right\} \\ &= E\left\{[E(Y|X) - E(Y)]E\left[\left(Y - E(Y|X)\right)|X\right]\right\} \\ &= E\{[E(Y|X) - E(Y)][E(Y|X) - E(Y|X)]\} \\ &= 0 \end{aligned}$$

Thus, we have the equation of $Var(Y)$ with three terms reduced to two terms, such that,

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

Applying the Smoothing Theorem again on both expectations, we have,

$$\begin{aligned} Var(Y) &= E\left\{E\left[\left(Y - E(Y|X)\right)^2\middle|X\right]\right\} + E\left[\left(E(Y|X) - E\left(E(Y|X)\right)\right)^2\right] \\ &= E\left\{E\left[\left(Y^2 - 2YE(Y|X) + E^2(Y|X)\right)\middle|X\right]\right\} + E\left[\left(E(Y|X) - E\left(E(Y|X)\right)\right)^2\right] \\ &= E[E(Y^2|X) - E^2(Y|X)] + E\left\{\left[E(Y|X) - E\left(E(Y|X)\right)\right]^2\right\} \\ &= E[Var(Y|X)] + Var[E(Y|X)] \end{aligned}$$

■

Theorem 18.10: *Let $X_1, \dots, X_N$ be i.i.d. with the $E(X_i) = \mu < \infty$, and $Var(X_i) = \sigma^2 < \infty$, where N is also a random variable and independent of the sequence $\{X_i\}$, then,*

$$Var\left(\sum_{i=1}^{N} X_i\right) = \mu^2 Var(N) + \sigma^2 E(N) \qquad (18.10)$$

Proof: From Theorem 18.9, we have,

$$Var\left(\sum_{i=1}^{N} X_i\right) = Var\left[E\left(\sum_{i=1}^{N} X_i \,|N\right)\right] + E\left[Var\left(\sum_{i=1}^{N} X_i \,|N\right)\right] \qquad (18.10.1)$$

Expanding the first expression on the right, we have,

$$\begin{aligned} Var\left[E\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right] &= E\left[E\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right]^2 - E^2\left[E\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right] = \sum_n E^2\left(\sum_{i=1}^{n} X_i\right) p_N(n) - [\mu E(N)]^2 \\ &= \sum_n (n\mu)^2\, p_N(n) - [\mu E(N)]^2 = \mu^2 \sum_n n^2\, p_N(n) - \mu^2 E^2(N) \\ &= \mu^2 [E(N^2) - E^2(N)] = \mu^2 Var(N) \qquad (18.10.2) \end{aligned}$$

Expanding the second expression, we get,

$$\begin{aligned} E\left[Var\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right] &= E\left[E\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)^2\right] - E\left[E^2\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right] = E\left[E\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)^2 - E^2\left(\textstyle\sum_{i=1}^{N} X_i \,|N\right)\right] \\ &= \sum_n \left[E\left(\sum_{i=1}^{n} X_i\right)^2 - E^2\left(\sum_{i=1}^{n} X_i\right)\right] p_N(n) = \sum_n Var\left(\sum_{i=1}^{n} X_i\right) p_N(n) \\ &= \sum_n n\sigma^2 p_N(n) = \sigma^2 \sum_n n p_N(n) = \sigma^2 E(N) \qquad (18.10.3) \end{aligned}$$

Combining (18.10.1), (18.10.2) and (18.10.3), the identity is obtained. ■

Lecture 19

Properties of a Random Sample

Definition 19.1: *The random variables $X_1, X_2, \dots X_n$ are called a* ***random sample of size n*** *from the population $f_\theta(x)$ if they are i.i.d. with the same pdf or pmf function $f_\theta(x)$ determined by parameter θ.*

The joint pdf or pmf of a random sample is denoted by $f_\theta(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \dots, x_n)$ is a vector of the **observations**. $f_\theta(\boldsymbol{x})$ is also called likelihood in some occasion. Recall from the property of i.i.d. random variables, the joind pdf or pmf of a random sample of size n is calculated by,

$$f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} f_\theta\,(x_i) \qquad (19.1)$$

Definition 19.2: *Let $X_1, X_2, \dots X_n$ be a random sample of size n from population $f_\theta(x)$ where θ is unknown. A* ***statistic*** *is a function of the random sample, $T(X_1, X_2, \dots X_n)$, which is used to make an inference on the unknown parameter θ.* Sample mean and sample variance are two commonly used statistics.

Definition 19.3: *The* ***sample mean****, denoted by $\bar{X}$ or $\bar{X}_n$, is the arithmetic average of the values in a random sample, is defined by,*

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{19.2}$$

Definition 19.4: *The* ***sample variance****, denoted by S^2, is the statistics defined by,*

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{19.3}$$

The sample standard deviation is the statistics defined as the square root of sample variance,

$$S = \sqrt{S^2} \tag{19.4}$$

Theorem 19.5: *Let $X_1, X_2, \dots X_n$ be a random sample of size n from a population with finite mean μ and variance σ^2, then,*

a. $E(\bar{X}) = \mu$ (19.5*a*)

b. $Var(\bar{X}) = \dfrac{\sigma^2}{n}$ (19.5*b*)

c. $E(S^2) = \sigma^2$ (19.5*c*)

Proof: The proofs mainly depend on Theorem 16.2 and Theorem 16.7.

a. By definition of sample mean, we have,

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{\sum_{i=1}^{n} E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

b. Theorem 16.7 then gives that,

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

c. By definition of sample variance, we have,

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \tag{19.5.1}$$

Expanding the square and the summation on the right,

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

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

By taking expectations on both sides,

$$E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] - nE[(\bar{X}-\mu)^2] = \sum_{i=1}^{n}E(X_i-\mu)^2 - nE[(\bar{X}-\mu)^2]$$

$$= \sum_{i=1}^{n}[Var(X_i-\mu) + [(E(X_i)-\mu)]^2\,] - n[Var(\bar{X}-\mu) + [(E(\bar{X})-\mu)]^2\,]$$

$$= \sum_{i=1}^{n}[\sigma^2+0] - n[\frac{\sigma^2}{n}+0] = n\sigma^2 - \sigma^2$$

Referring back to (19.5.1), we have,

$$E(S^2)\ = E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)}\right] = \frac{n\sigma^2-\sigma^2}{n-1} = \sigma^2 \qquad \blacksquare$$

Theorem 19.6: *Let $X_1, X_2, \dots X_n$ be a random sample of size n from a population with mgf $M_X(t)$. Then the mgf of the sample mean is,*

$$M_{\bar{X}}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n \tag{19.6}$$

Proof: By Theorem 7.7 and definition of sample mean, we have

$$M_{\bar{X}}(t)\ = M_{\frac{\sum_{i=1}^{n}X_i}{n}}(t) = M_{\sum_{i=1}^{n}X_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n}M_{X_i}\left(\frac{t}{n}\right) = \left[M_X\left(\frac{t}{n}\right)\right]^n \qquad \blacksquare$$

<u>Sampling from the Normal Distribution</u>

Definition 19.7: *A special case of Gamma distribution $\gamma(n/2\,,1/2)$, whose pdf is give by,*

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} \cdot e^{-\frac{1}{2}x} \cdot x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, \qquad x > 0 \tag{19.7}$$

is called the chi-squared distribution, denoted by $\mathcal{X}_n^2$ with n degree of freedom

The *mean* and *variance* of $\mathcal{X}_n^2$, in accordance with those of $\gamma(n/2\,,1/2)$, are

$$E(X) = n, \quad Var(X) = 2n$$

And the *mgf* of the $Gamma(a, p)$ distribution is,

$$M_X(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}, \qquad t < \frac{1}{2}$$

Lemma 19.8: *The followings are true for ant chi-squared distribution with n degree of freedom, $\mathcal{X}_n^2$,*

a. *Let* $Z \sim N(0,1)$*, then* $Z^2 \sim \mathcal{X}_1^2$

b. *For a sequence of independent standard normal random variable* $Z_i \sim N(0,1)$*, we have* $\sum_{i=1}^{n} Z_i^2 \sim \mathcal{X}_n^2$

c. *If* $X \sim \mathcal{X}_m^2$ *and* $Y \sim \mathcal{X}_n^2$*, and* X,Y *are independent, then* $X + Y \sim \mathcal{X}_{m+n}^2$

Proof: Statement $a.$ can be proved by either using mgf or directly deriving the pdf and the latter will be demonstrated here. Similar as in Example 5.2, we have,

$$F_{Z^2}(t) = P(Z^2 \leq t) = P\left(-\sqrt{t} \leq Z \leq \sqrt{t}\right) = F\left(\sqrt{t}\right) - F\left(-\sqrt{t}\right)$$

$$f_{Z^2}(t) = \frac{1}{2\sqrt{t}}\left[f_Z\left(\sqrt{t}\right) + f_Z\left(-\sqrt{t}\right)\right] = \frac{1}{2\sqrt{t}}\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{t}{2}} + \frac{1}{\sqrt{2\pi}}e^{-\frac{t}{2}}\right] = \frac{1}{\sqrt{2\pi t}}e^{-\frac{t}{2}}$$

$$= \begin{cases} \dfrac{\left(\frac{1}{2}\right)^{\frac{1}{2}} \cdot e^{-\frac{1}{2}t} \cdot t^{\frac{1}{2}-1}}{\Gamma\left(\frac{1}{2}\right)}, t > 0 \\ 0, \qquad\qquad O.W. \end{cases}$$

which is the pdf of a chi-squared distribution with 1 degree of freedom, $\mathcal{X}_1^2$

b and c can be easily seen as a result of Example 17.9.1 (2), that is: the sum of n independent random variables of $\gamma(a,p)$, gives another Gamma distribution, $\gamma(na,p)$.

Theorem 19.9: *Let* $X_1, X_2, \dots X_n$ *be a random sample of size* n *from a* $N(\mu,\sigma^2)$ *distribution, and let* $\bar{X}$ *to be the sample mean defined in (19.3) and* S^2 *to be the sample variance defined in (19.4). Then,*

a. $\bar{X}$ *and* S^2 *are independent*

b. $$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \tag{19.8}$$

c. $$(n-1)\frac{S^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2 \tag{19.9}$$

Proof:

a. Without loss of generality, let $\mu = 0, \sigma = 1$. Since $X_1, \dots X_n$ are i.i.d., so that the joint pdf is,

$$f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} f_\theta\,(x_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} \cdot e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot e^{-\frac{\sum_{i=1}^{n} x_i^2}{2}}$$

Let $y_1 = \bar{x}, y_i = x_i - \bar{x}$, for $i = 2, \dots, n$, so that,

$$x_1 = ny_1 - \sum_{i=2}^{n} x_2 = ny_1 - \sum_{i=2}^{n}(y_i + y_1) = y_1 - \sum_{i=2}^{n} y_i$$

$$x_i = y_i + y_1$$

The partial dereivatives dan be calculated by, for $i = 2, \dots, n$

$$\frac{\partial x_i}{\partial y_i} = 1 \quad and \quad \frac{\partial x_i}{\partial y_1} = 1 \quad and \quad \frac{\partial x_1}{\partial y_i} = -1 \quad and \quad \frac{\partial x_1}{\partial y_1} = 1$$

Then, the Jacobian of this transformation is,

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & & \cdots \frac{\partial x_1}{\partial y_i} \cdots & & \frac{\partial x_1}{\partial y_n} \\ & & 0 & & \\ \vdots & & \vdots & & \vdots \\ \frac{\partial x_i}{\partial y_1} & \cdots 0 & \frac{\partial x_i}{\partial y_i} & 0 \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ & & 0 & & \\ \frac{\partial x_n}{\partial y_1} & & \cdots 0 \cdots & & \frac{\partial x_n}{\partial y_n} \end{vmatrix} = \begin{vmatrix} 1 & & \cdots -1 \cdots & & -1 \\ & & 0 & & \\ \vdots & & \vdots & & \vdots \\ 1 & \cdots 0 & 1 & 0 \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ & & 0 & & \\ 1 & & \cdots 0 \cdots & & 1 \end{vmatrix} = \begin{vmatrix} n & & \cdots 0 \cdots & & 0 \\ & & 0 & & \\ \vdots & & \vdots & & \vdots \\ 1 & \cdots 0 & 1 & 0 \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ & & 0 & & \\ 1 & & \cdots 0 \cdots & & 1 \end{vmatrix} = n$$

The rule of bivaraite tranformation gives that,

$$\begin{aligned} f_{Y_1,Y_2,..Y_n}(y_1,y_2,\dots y_n) &= f_{X_1,X_2,..X_n}(x_1,x_2,\dots x_n)\cdot|J| = \frac{1}{(2\pi)^{\frac{n}{2}}}\cdot e^{-\frac{\sum_{i=1}^n x_i^2}{2}}\cdot n \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}}\cdot e^{-\frac{1}{2}\left(y_1-\sum_{i=2}^n y_i\right)^2} e^{-\frac{1}{2}\sum_{i=2}^n (y_i+y_1)^2} \\ &= \frac{n}{(2\pi)^{\frac{n}{2}}}\cdot e^{-\frac{1}{2}\left[y_1^2-2y_1\sum_{i=2}^n y_i+\left(\sum_{i=2}^n y_i\right)^2+\sum_{i=2}^n y_i^2+2y_1\sum_{i=2}^n y_i+(n-1)y_1^2\right]} \\ &= \left[\left(\frac{n}{2\pi}\right)^{\frac{1}{2}}\ e^{-\frac{1}{2}ny_1^2}\right]\cdot\left[\frac{n^{\frac{1}{2}}}{(2\pi)^{\frac{n-1}{2}}}e^{-\frac{1}{2}\left[\left(\sum_{i=2}^n y_i\right)^2+\sum_{i=2}^n y_i^2\right]}\right] \qquad (19.9.1) \end{aligned}$$

By Lemma 15.4.1, (19.9.1) can be see as $g(y_1)g(y_2,\dots,y_n)$, so that Y_1 is independent of $Y_2,\dots,Y_n$, or equivalently, $\bar{X}$ is independent of $X_i - X_n$, for $i = 2,\dots,n$. On the other hands, by the fact that

$$\sum_{i=2}^{n}(X_i-\bar{X}) = -(X_i-\bar{X}), \qquad thus \quad (X_i-\bar{X})^2 = \left(\sum_{i=2}^{n}(X_i-\bar{X})\right)^2$$

S^2 can be rewritten as,

$$\begin{aligned} S^2 &= \sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1} = \frac{1}{n-1}\left[(X_i-\bar{X})^2+\sum_{i=2}^{n}(X_i-\bar{X})^2\right] \\ &= \frac{1}{n-1}\left[\left(\sum_{i=2}^{n}(X_i-\bar{X})\right)^2+\sum_{i=2}^{n}(X_i-\bar{X})^2\right] \end{aligned}$$

which only depends on $X_i - X_n$, for $i = 2,\dots,n$. Therefore, S^2 is independent of $\bar{X}$.

b. Theorem 19.6 gives that,

$$M_{\bar{X}}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n = e^{\left(\mu\frac{t}{n}+\frac{1}{2}\sigma^2\frac{t^2}{n^2}\right)\cdot n} = e^{\mu t+\frac{1}{2}\left(\frac{\sigma}{\sqrt{n}}\right)^2 t^2}$$

By uniqueness of mgf, we have $\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$

c. By reexpressing $(n-1)\frac{S^2}{\sigma^2}$, we have,

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

Therefore,

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = (n-1)\frac{S^2}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \tag{19.9.2}$$

Let $Z_i = \frac{X_i-\mu}{\sigma}, Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, which both have $N(0,1)$ distribution. Then for n independent Z_i^2, Theorem 19.8 gives that $\sum_{i=1}^{n} Z_i^2 \sim \gamma\left(\frac{n}{2},\frac{1}{2}\right)$ and $Z^2 \sim \gamma\left(\frac{1}{2},\frac{1}{2}\right)$, with mgfs $M_{\sum_{i=1}^{n} Z_i^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$ and $M_{Z^2}(t) = \left(\frac{1}{1-2t}\right)^{\frac{1}{2}}$.

Therefore, by the equal relation (19.9.2) and property of mgf, we have,

$$M_{(n-1)\frac{S^2}{\sigma^2}}(t) = \frac{M_{\sum_{i=1}^{n} Z_i^2}(t)}{M_{Z^2}(t)} = \frac{\left(\frac{1}{1-2t}\right)^{\frac{n}{2}}}{\left(\frac{1}{1-2t}\right)^{\frac{1}{2}}} = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}$$

which unqiuesly determines that $(n-1)\frac{S^2}{\sigma^2}$ has a $\mathcal{X}_{n-1}^2$ distribution . ■

Lecture 20

<u>Student's t and Fisher distributions</u>

Definition 20.1: *Let $Z \sim N(0,1)$, $X \sim {\mathcal{X}_n}^2$ and Z, X are independent. Then then quantity $Z/\left(\sqrt{X/n}\right)$ has Student's t distribution with n degree of freedom. Equivalently, a random variable T has Student's t distribution with n degrees of freedom, denoted by $T \sim t_n$, if it has pdf,*

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \cdot \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < t < \infty \tag{20.1}$$

The t pdf can be derived by first defining,

$$W = \frac{X}{n} \quad and \quad T = \frac{Z}{\sqrt{W}} \tag{20.1.1}$$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty \quad and \quad f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} \cdot e^{-\frac{1}{2}x} \cdot x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)},\ x > 0$$

(20.1.1) can be uniquely solved for,

$$X = Wn \quad and \quad Z = \sqrt{W}T \tag{20.1.2}$$

and the Jacobian of such transfoamtion is,

$$J=\begin{vmatrix}\dfrac{\partial x}{\partial w} & \dfrac{\partial x}{\partial t}\\ \dfrac{\partial z}{\partial w} & \dfrac{\partial z}{\partial t}\end{vmatrix}=\begin{vmatrix}2n & 0\\ \dfrac{1}{2}tw^{-\frac{1}{2}} & w^{\frac{1}{2}}\end{vmatrix}=2nw^{\frac{1}{2}};\ \ w>0$$

Now, make the transformation defined in (20.1.2),

$$\begin{aligned}f_{W,T}(w,t)&=f_{X,Z}\left(wn,\sqrt{w}t\right)\cdot|J|=f_X(wn)\cdot f_Z\left(\sqrt{w}t\right)\cdot 2nw^{\frac{1}{2}}\\&=\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}\cdot e^{-\frac{1}{2}(wn)}\cdot(wn)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(\sqrt{w}t\right)^2}{2}}\cdot 2nw^{\frac{1}{2}};\ \ w>0,\ -\infty<t<\infty\end{aligned}$$

Then , the marginal pdf of T is given by integrating $f_{W,T}(w,t)$ in terms of w, that is,

$$f_T(t)=\int_0^\infty f_{W,T}(w,t)dw=\int_0^\infty\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}\cdot e^{-\frac{1}{2}(wn)}\cdot(wn)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(\sqrt{w}t\right)^2}{2}}\cdot 2nw^{\frac{1}{2}}dw$$

$$=\frac{2n^{\frac{n}{2}}\cdot\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi}}\int_0^\infty e^{-\frac{n+t^2}{2}w}\,w^{\frac{n}{2}-\frac{1}{2}}dw$$

$$=\frac{n^{\frac{n}{2}}\cdot\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi}}\cdot\frac{\Gamma\left(\frac{n+1}{2}\right)}{\left(\frac{n+t^2}{2}\right)^{\frac{n+1}{2}}}\int_0^\infty\frac{\left(\frac{n+t^2}{2}\right)^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)}e^{-\frac{n+t^2}{2}v}\,w^{\frac{n+1}{2}-1}dw$$

$$=\frac{n^{\frac{n}{2}}\cdot\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi}}\cdot\frac{\Gamma\left(\frac{n+1}{2}\right)}{\left(\frac{n+t^2}{2}\right)^{\frac{n+1}{2}}}\cdot\int_0^\infty f_{\gamma\left(\frac{n+1}{2},\frac{n+t^2}{2}\right)}(w)dw=\frac{n^{\frac{n}{2}}\cdot\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi}}\cdot\frac{\Gamma\left(\frac{n+1}{2}\right)}{\left(\frac{n+t^2}{2}\right)^{\frac{n+1}{2}}}$$

$$=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\cdot\left(\frac{1}{n}\right)^{-\frac{n+1}{2}}\cdot\left(n+t^2\right)^{-\frac{n+1}{2}}=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

■

Example 20.1.1: Let $X_1,X_2,\dots,X_n$ be a random sample from a $N(\mu,\sigma^2)$ distribution, show that the quantity $(\bar{X}-\mu)/\left(S/\sqrt{n}\right)$ has a t_{n-1} distribution.

Proof: Let $Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ and $Y=(n-1)\frac{S^2}{\sigma^2}$, from (19.8) and (19.9), we have that $Z\sim N(0,1)$ and $Y\sim\mathcal{X}_{n-1}^2$, and Z, Y are independent. Then, the quantity can be rewritten as,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\cdot\frac{\sigma}{S}=\frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{S^2}{\sigma^2}}}=\frac{Z}{\sqrt{\frac{Y^2}{n-1}}}$$

which, by definition 20.1, has a student's t distribution with degrees of freedom $n-1$. In statistical inference, the quantity $(\bar{X}-\mu)/\left(S/\sqrt{n}\right)$ is called the *T-statistics*. ■

Theorem 20.2: *If a random variable has a student's t distribution,* t_n*, with degrees of freedom* n*, then*

$$a. \qquad E(T) = 0 \tag{20.2a}$$

$$b. \qquad Var(T) = \frac{n}{n-2} \tag{20.2b}$$

Proof: Let $T = Z/\left(\sqrt{X/n}\right)$, where $Z \sim N(0,1)$, $X \sim \mathcal{X}_n{}^2$and Z, X are independent.

a. By independence, we have,

$$E(T) = E\left(\frac{Z}{\sqrt{X/n}}\right) = \sqrt{n}E(Z)E\left(\frac{1}{\sqrt{X}}\right) = \sqrt{n} \cdot 0 \cdot E\left(\frac{1}{\sqrt{X}}\right) = 0, \qquad n > 1$$

b. In order to find the variance, we calculate the second moment first,

$$E(T^2) = E\left(\frac{Z^2}{X/n}\right) = nE\left(\frac{Z^2}{X}\right) = nE(Z^2)E\left(\frac{1}{X}\right) = nE\left(\frac{1}{X}\right) \tag{20.2.1}$$

Where the expectation of reciprocal of X is computed by following Theorom 6.3,

$$E\left(\frac{1}{X}\right) = \int_0^\infty \frac{1}{x} \cdot \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} e^{-\frac{1}{2}x} \cdot x^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)} dx = \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}-1} e^{-\frac{1}{2}x} \cdot x^{\left(\frac{n}{2}-1\right)-1}}{\Gamma\left(\frac{n}{2}-1\right)} \cdot \frac{\Gamma\left(\frac{n}{2}-1\right)\frac{1}{2}}{\Gamma\left(\frac{n}{2}\right)} dx$$

$$= \frac{1}{2} \cdot \frac{\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n}{2}\right)} \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}-1} e^{-\frac{1}{2}x} \cdot x^{\left(\frac{n}{2}-1\right)-1}}{\Gamma\left(\frac{n}{2}-1\right)} dx = \frac{1}{2} \cdot \frac{\Gamma\left(\frac{n}{2}-1\right)}{\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}-1\right)} = \frac{1}{n-2}$$

Referring back to (20.2.1),we can get the second moment as well as the variance, that is,

$$E(T^2) = nE\left(\frac{1}{X}\right) = n \cdot \frac{1}{n-2} = \frac{n}{n-2}$$

$$Var(T) = E(T^2) - E^2(T) = \frac{n}{n-2} - 0 = \frac{n}{n-2} \qquad \blacksquare$$

Theorem 20.3: *A student's t distribution,* t_n*, with* n *degrees of freedom, has the standard normal as its asymptotic distribution. In other words, if* $T \sim t_n$ *and* $Z \sim N(0,1)$*, then* T *converges to* Z *in distribution.*

Proof: The pdf of a t distribution with n degrees of freedom is defined in 20.1 and take the limit of the pdf by letting n go to infinity, we have,

$$\lim_{n\to\infty} f_T(t) = \lim_{n\to\infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

$$= \frac{1}{\sqrt{\pi}} \lim_{n\to\infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\Gamma\left(\frac{n}{2}\right)} \cdot \lim_{n\to\infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-\frac{1}{2}} \cdot \lim_{n\to\infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}} \tag{20.3.1}$$

where for any given value of t, the last limit is equal to 1, the second limit is equal to $e^{-\frac{1}{2}t^2}$, and the first limit, by expanding the gamma functions, we have,

$$\lim_{n\to\infty} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\Gamma\left(\frac{n}{2}\right)} = \lim_{n\to\infty} \frac{\sqrt{2\pi\frac{(n-1)}{2}}\left(\frac{(n-1)/2}{e}\right)^{\frac{(n-1)}{2}}}{\sqrt{n}\sqrt{2\pi\frac{(n-2)}{2}}\left(\frac{(n-2)/2}{e}\right)^{\frac{(n-2)}{2}}} \qquad \text{(Stirling's formula)}$$

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

$$= \lim_{n\to\infty} \sqrt{\frac{(n-1)^2}{n(n-2)}}\left[\left(1+\frac{1}{n-2}\right)^{n-2}\right]^{\frac{1}{2}} \left(\frac{1}{2e}\right)^{\frac{1}{2}}$$

$$= \left(\frac{1}{2e}\right)^{\frac{1}{2}} \cdot \lim_{n\to\infty} \sqrt{\frac{(n-1)^2}{n(n-2)}} \lim_{n\to\infty} \left[\left(1+\frac{1}{n-2}\right)^{n-2}\right]^{\frac{1}{2}}$$

$$= \left(\frac{1}{2e}\right)^{\frac{1}{2}} \cdot 1 \cdot e^{\frac{1}{2}} = \frac{1}{\sqrt{2}}$$

Referring all the calculated limits back to (20.3.1), we have,

$$\lim_{n\to\infty} f_T(t) = \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\sqrt{2}} \cdot e^{-\frac{1}{2}t^2} \cdot 1 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}, -\infty < t < \infty$$

Which is the pdf of $Z \sim N(0,1)$. Therefore, T converges to Z in distribution. ■

Definition 20.4: *Let $X \sim \mathcal{X}_m{}^2$, $Y \sim \mathcal{X}_n{}^2$and X, Y are independent. Then then quantity $F = (X/m)/(Y/n)$ has Fisher distribution with (m, n) degrees of freedom. Equivalently, a random variable Fhas Fisher distribution with (m, n) degrees of freedom, denoted by $F_{m,n}$, if it has pdf,*

$$f_F(f) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \cdot f^{\frac{m}{2}-1} \cdot \left(\frac{m}{n}\right)^{\frac{m}{2}} \cdot \left(1+\frac{mf}{n}\right)^{-\frac{m+n}{2}}, \ f > 0 \qquad (20.4)$$

The F pdf can be derived by first defining,

$$W = \frac{Y}{n} \quad and \quad F = \frac{X/m}{W} \qquad (20.4.1)$$

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{m}{2}} \cdot e^{-\frac{1}{2}x} \cdot x^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)}, \ x > 0 \quad and \quad f_Y(y) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} \cdot e^{-\frac{1}{2}y} \cdot y^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}, \ y > 0$$

(20.4.1) can be uniquely solved for,

$$X = mFW \quad and \quad Y = nW \qquad (20.4.2)$$

and the Jacobian of such transfoamtion is,

$$J = \begin{vmatrix} \frac{\partial x}{\partial w} & \frac{\partial x}{\partial f} \\ \frac{\partial y}{\partial w} & \frac{\partial y}{\partial f} \end{vmatrix} = \begin{vmatrix} mf & mw \\ n & 0 \end{vmatrix} = nmw; \ w > 0, \ f > 0$$

Now, make the transformation defined in (20.4.2),

$$\begin{aligned} f_{W,F}(w,f) &= f_{X,Y}(mwf, nw)\cdot|J| = f_X(mwf)\cdot f_Y(nw)\cdot nmw \\ &= \frac{\left(\frac{1}{2}\right)^{\frac{m}{2}}\cdot e^{-\frac{1}{2}(mwf)}\cdot(mwf)^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}\cdot e^{-\frac{1}{2}(nw)}\cdot(nw)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}\cdot nmw;\ \ w>0,\ f>0 \end{aligned}$$

Then , the marginal pdf of T is given by integrating $f_{W,F}(w,f)$ in terms of w, that is,

$$f_F(f) = \int_0^\infty f_{W,F}(w,f)dw = \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{m}{2}}\cdot e^{-\frac{1}{2}(mwf)}\cdot(mwf)^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)}\cdot\frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}\cdot e^{-\frac{1}{2}(nw)}\cdot(nw)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)}\cdot nmw$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}}\cdot m^{\frac{m}{2}}\cdot n^{\frac{n}{2}}\cdot f^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^\infty e^{-\frac{mf+n}{2}w}\cdot w^{\frac{m+n}{2}-1}\,dw$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}} m^{\frac{m}{2}} n^{\frac{n}{2}} f^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\cdot\frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{mf+n}{2}\right)^{\frac{m+n}{2}}}\int_0^\infty \frac{1}{\Gamma\left(\frac{m+n}{2}\right)}\left(\frac{mf+n}{2}\right)^{\frac{m+n}{2}} e^{-\frac{mf+n}{2}w} w^{\frac{m+n}{2}-1}\,dw$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}}\cdot m^{\frac{m}{2}}\cdot n^{\frac{n}{2}}\cdot f^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\cdot\frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{mf+n}{2}\right)^{\frac{m+n}{2}}}\cdot\int_0^\infty f_{\gamma\left(\frac{mf+n}{2},\frac{m+n}{2}\right)}(w)dw$$

$$= \frac{\left(\frac{1}{2}\right)^{\frac{m+n}{2}} m^{\frac{m}{2}} n^{\frac{n}{2}} f^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\cdot\frac{\Gamma\left(\frac{m+n}{2}\right)}{\left(\frac{mf+n}{2}\right)^{\frac{m+n}{2}}} = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} f^{\frac{m}{2}-1} m^{\frac{m}{2}} n^{\frac{n}{2}}\left(\frac{mf+n}{2}\right)^{-\frac{m+n}{2}} 2^{-\frac{m+n}{2}}$$

$$= \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\cdot f^{\frac{m}{2}-1}\cdot\left(\frac{m}{n}\right)^{\frac{m}{2}}\cdot\left(1+\frac{mf}{n}\right)^{-\frac{m+n}{2}},\qquad f>0$$

■

Example 20.4.1: Let $X_1,\dots,X_n$ be a random sample from a $N(\mu_X, {\sigma_X}^2)$ population, and $Y_1,\dots,Y_n$ be a random sample from an independent $N(\mu_Y, {\sigma_Y}^2)$ population. Show the quantity $({S_X}^2/{\sigma_X}^2)/({S_Y}^2/{\sigma_Y}^2)$ has a $F_{m-1,n-1}$ distribution.

Proof: Let $F_X = (m-1)\frac{{S_X}^2}{{\sigma_X}^2}$ and $F_Y = (n-1)\frac{{S_Y}^2}{{\sigma_Y}^2}$, from (19.9), we have that $F_X \sim \mathcal{X}^2_{m-1}$ and $F_Y \sim \mathcal{X}^2_{n-1}$, and F_X, F_Y are independent. Then, the quantity can be rewritten as,

$$\frac{{S_X}^2/{\sigma_X}^2}{{S_Y}^2/{\sigma_Y}^2} = \frac{F_X/(m-1)}{F_Y/(n-1)}$$

which, by definition 20.4, has an F distribution with $m-1$ and $n-1$ degrees of freedom. In statistical inference, the quantity $({S_X}^2/{\sigma_X}^2)/({S_Y}^2/{\sigma_Y}^2)$ is called the *F-statistics*. ■

Theorem 20.5: The Fisher distribution has the following properties:

$$a. \quad \text{If } X \sim F_{m,n}, \text{ then } \frac{1}{X} \sim F_{n,m} \tag{20.5a}$$

$$b. \quad \text{If } X \sim t_n, \text{ then } X^2 \sim F_{1,n} \tag{20.5b}$$

$$c. \quad \text{If } X \sim F_{p,q}, \text{ then } \frac{(p/q)X}{1+(p/q)X} \sim \beta\left(\frac{p}{2},\frac{q}{2}\right) \tag{20.5c}$$

Proof: Statement *a.* and *b.* can be shown by similar approach**,**

a. Since $X \sim F_{m,n}$, and thus X can be written as $X = (Y_m/m)/(Y_n/n)$, where $Y_m \sim \mathcal{X}_m^2, Y_n \sim \mathcal{X}_n^2$ and Y_m, Y_n are independent. Then,

$$\frac{1}{X} = \frac{1}{(Y_m/m)/(Y_n/n)} = \frac{Y_n/n}{Y_m/m}$$

which, by definition, is an F random variable with the degrees of freedom interchange, i.e., $F_{n,m}$

b. Let $T \sim t_n$, then T can be written as $T = Z/\sqrt{X/n}$, where $Z \sim N(0,1)$, $X \sim \mathcal{X}_n^2$ and X, Z are independent. By squaring, we have,

$$T^2 = \left(Z/\sqrt{X/n}\right)^2 = \frac{Z^2}{X/n} = \frac{Z^2/1}{X/n}$$

Lemma 19.8 gives that $Z^2 \sim \mathcal{X}_1^2$, and thus $T^2 \sim F_{1,n}$ by definition.

c. Given that $X \sim F_{p,q}$, so we have $f_X(x)$ defined by (20.4), which is,

$$f_X(x) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} x^{\frac{p}{2}-1}\left(\frac{p}{q}\right)^{\frac{p}{2}}\left(1+\frac{px}{q}\right)^{-\frac{p+q}{2}}, \ x > 0$$

Then the pdf of $Y = \frac{(p/q)X}{1+(p/q)X}$ can be derived by differentiating from its cdf,

$$\begin{aligned} F_Y(y) &= P\{Y \le y\} = P\left\{\frac{\frac{p}{q}X}{1+\frac{p}{q}X} \le y\right\} = P\left\{\frac{p}{q}X \le y + \frac{p}{q}yX\right\} \\ &= P\left\{X \le \frac{qy}{p(1-y)}\right\} = F_X\left(\frac{qy}{p(1-y)}\right), \qquad 0 < y < 1 \\ f_Y(y) &= f_X\left(\frac{qy}{p(1-y)}\right)\cdot\frac{q}{p}\cdot\frac{1}{(1-y)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\left(\frac{qy}{p(1-y)}\right)^{\frac{p}{2}-1}\left(\frac{p}{q}\right)^{\frac{p}{2}}\left(1+\frac{p}{q}\cdot\frac{qy}{p(1-y)}\right)^{-\frac{p+q}{2}}\cdot\frac{q}{p}\cdot\frac{1}{(1-y)^2} \\ &= \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)}\cdot(y)^{\frac{p}{2}-1}\cdot(1-y)^{\frac{q}{2}-1}, \qquad 0 < y < 1 \end{aligned}$$

which is the pdf of a $\beta\left(\frac{p}{2},\frac{q}{2}\right)$ distribution. ■

Lecture 21

Order Statistics

Theorem 21.1: *Let $X_1, \dots X_n$ be a random sample from a population with pdf $f_X(t)$ and cdf $F_X(t)$ and define $X^{(n)} = \max(X_1, \dots X_n)$, then,*

$$f_{X^{(n)}}(t) = n\big(F_X(t)\big)^{n-1} f_X(t) \tag{21.1}$$

Proof: By definition of cdf and $X^{(n)}$, we have,

$$\begin{aligned} F_{X^{(n)}}(t) &= P\big(X^{(n)} \le t\big) = P[(X_1 \le t) \cap (X_2 \le t) \cap \dots \cap (X_n \le t)] \\ &= P(X_1 \le t)P(X_2 \le t) \dots P(X_n \le t) = \prod_{i=1}^{n} F_{X_i}(t) = \big(F_X(t)\big)^n \end{aligned}$$

By differentiating the last expression, we have,

$$f_{X^{(n)}}(t) = \frac{d}{dt} F_{X^{(n)}}(t) = n\big(F_X(t)\big)^{n-1} f_X(t)$$ ■

Example 21.1.1: Suppose $X_1, \dots X_n$ is a random sample from a $\mathcal{U}(0,1)$ population, find the pdf of $X^{(n)}$ and its expectation, $E\left(X^{(n)}(t)\right)$ and variance, $Var\left(X^{(n)}(t)\right)$.

Solution: cdf and pdf of the population are needed in the expression (21.1). Then by definition of a standard uniform distribution, we have,

$$\begin{aligned} f_X(t) &= \begin{cases} 1, & 0 < t < 1 \\ 0, & O.W. \end{cases} \\ F_X(t) &= \begin{cases} 0, & t < 0 \\ t, & 0 \le t < 1 \\ 1, & 1 \le t \end{cases} \end{aligned} \tag{21.1.1}$$

Therefore,

$$f_{X^{(n)}}(t) = \begin{cases} nt^{n-1}, & 0 < t < 1 \\ 0, & O.W. \end{cases} = \begin{cases} \dfrac{t^{n-1}(1-t)^{1-1}}{B(n,1)}, & 0 < t < 1 \\ 0, & O.W. \end{cases}$$

which can be observed is the pdf of a $\beta(n,1)$ distribution. For a Beta distribution, its expectation and variance have been derived before, which are,

$$E\left(X^{(n)}(t)\right) = \frac{n}{n+1} \qquad and \qquad Var\left(X^{(n)}(t)\right) = \frac{n}{(n+2)(n+1)^2}$$ ■

Theorem 21.2: *Let $X_1, \dots X_n$ be a random sample from a population with pdf $f_X(t)$ and cdf $F_X(t)$ and define $X^{(1)} = \min(X_1, \dots X_n)$, then,*

$$f_{X^{(1)}}(t) = n\big(1 - F_X(t)\big)^{n-1} f_X(t) \tag{21.2}$$

Proof: By definition of cdf and $X^{(1)}$, we have,

$$\begin{aligned} F_{X^{(1)}}(t) &= 1 - P\big[X^{(1)} \geq t\big] = 1 - P[(X_1 \geq t) \cap (X_2 \geq t) \cap \ldots \cap (X_n \geq t)] \\ &= 1 - P(X_1 \geq t)P(X_2 \geq t)\ldots P(X_n \geq t) = 1 - \prod_{i=1}^{n}\big[1 - F_{X_i}(t)\big] \\ &= 1 - [1 - F_X(t)]^n \end{aligned}$$

By differentiating the last expression, we have,

$$f_{X^{(1)}}(t) = \frac{d}{dt}F_{X^{(1)}}(t) = -n\big(1 - F_X(t)\big)^{n-1}\cdot(-1)f_X(t) = n\big(1 - F_X(t)\big)^{n-1} f_X(t)$$ ■

Example 21.2.1: Suppose $X_1, \ldots X_n$ is a random sample from a $\mathcal{U}(0,1)$ population, find the pdf of $X^{(1)}$ and its expectation, $E\left(X^{(1)}(t)\right)$ and variance, $Var\left(X^{(1)}(t)\right)$.

Solution: By the same cdf and pdf of a $\mathcal{U}(0,1)$ distribution specified in (21.1.1) and (21.2), we have

$$f_{X^{(1)}}(t) = \begin{cases} n(1-t)^{n-1}, 0 < t < 1 \\ 0, \qquad\qquad\qquad O.W. \end{cases} = \begin{cases} \dfrac{(1-t)^{n-1}t^{1-1}}{B(1,n)}, & 0 < t < 1 \\ 0, & O.W. \end{cases}$$

which can be observed is the pdf of a $\beta(1,n)$ distribution. For a Beta distribution, its expectation and variance have been derived before, which are,

$$E\left(X^{(n)}(t)\right) = \frac{1}{n+1} \qquad and \qquad Var\left(X^{(n)}(t)\right) = \frac{n}{(n+1)^2(n+2)}$$ ■

Theorem 21.3: *Let $X_1, \ldots X_n$ be a random sample from a population with pdf $f_X(t)$ and cdf $F_X(t)$ and define $X^{(j)}$ to be the j^{th} smallest element, then,*

$$f_{X^{(j)}}(t) = \binom{n}{(j-1), 1, (n-j)} [F_X(t)]^{j-1} f_X(t)[1 - F_X(t)]^{n-j} \tag{21.3}$$

Example 21.3.1: Suppose $X_1, \ldots X_n$ is a random sample from a $\mathcal{U}(0,1)$ population, find the pdf of $X^{(j)}$ and its expectation, $E\left(X^{(j)}(t)\right)$ and variance, $Var\left(X^{(j)}(t)\right)$.

Solution: By the same cdf and pdf of a $\mathcal{U}(0,1)$ distribution specified in (21.1.1) and (21.3), we have

$$f_{X^{(1)}}(t) = \begin{cases} \dfrac{n!}{(j-1)!\,1!\,(n-j)!}t^{j-1}(1-t)^{n-j}, 0 < t < 1 \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad O.W. \end{cases} = \begin{cases} \dfrac{t^{j-1}(1-t)^{n-j}}{B(j, n-j+1)}, & 0 < t < 1 \\ 0, & O.W. \end{cases}$$

which can be observed is the pdf of a $\beta(j, n-j+1)$ distribution. For a Beta distribution, its expectation and variance have been derived before, which are,

$$E\left(X^{(n)}(t)\right) = \frac{j}{n+1} \qquad and \qquad Var\left(X^{(n)}(t)\right) = \frac{j(n-j+1)}{(n+1)^2(n+2)}$$ ■

It has been shown that $\mathcal{U}(0,1)$ is equivalent to $\beta(1,1)$, we can notice that from example 21.1.1,21.2.1,and 21.3.1, that the order statistics of n i.i.d Beta random variables also has a Beta distribution.

Theorem 21.4: *Let $X_1, \dots X_n$ be a random sample from a population with pdf $f_X(t)$ and cdf $F_X(t)$ and define $X^{(i)}, X^{(j)}$ to be the i^{th}, j^{th} smallest elements, then the joint pdf of them is,*

$$f_{X^{(i)},X^{(j)}}(s,t) = \binom{n}{(i-1),1,(j-i-1),1,(n-j)} [F_X(s)]^{i-1} f_X(s)[F_X(t)-F_X(s)]^{j-i-1} f_X(t)[1-F_X(t)]^{n-j} \tag{21.4}$$

Example 21.4.1: Suppose $X_1, \dots X_n$ is a random sample from a $\mathcal{U}(0,1)$ population, find the joint pdf of $X^{(1)}$, the minimum and $X^{(n)}$, the maximum.

Solution: By the same cdf and pdf of a $\mathcal{U}(0,1)$ distribution specified in (21.1.1) and (21.4), we have

$$\begin{aligned} f_{X^{(1)},X^{(n)}}(s,t) &= \binom{n}{(1-1),1,(n-1-1),1,(n-n)} [F_X(s)]^{1-1} f_X(s)[F_X(t)-F_X(s)]^{n-1-1} f_X(t)[1-F_X(t)]^{n-n} \\ &= \binom{n}{n-2} f_X(s)[F_X(t)-F_X(s)]^{n-2} f_X(t) \qquad (21.4.1) \\ &= \begin{cases} \dfrac{n!}{(n-2)!}(t-s)^{n-2}, 0 < s < t < 1 \\ 0, \qquad\qquad O.W. \end{cases} \end{aligned}$$

■

Theorem 21.5: *Let $X_1, \dots X_n$ be a random sample from a population with pdf $f_X(t)$ and cdf $F_X(t)$ with support $\mathcal{A}$, and define the range $R = X^{(n)} - X^{(1)}$ and the midpoint $M = \frac{X^{(n)}+X^{(1)}}{2}$, then,*

$$f_{M,R}(m,r) = n(n-1) f_X\left(m+\frac{r}{2}\right) f_X\left(m-\frac{r}{2}\right)\left[F_X\left(m+\frac{r}{2}\right) - F_X\left(m-\frac{r}{2}\right)\right]^{n-2} \tag{21.5}$$

Proof: We first find the joint pdf of M and R, given that,

$$M = \frac{X^{(1)}+X^{(n)}}{2} \quad and \quad R = X^{(n)} - X^{(1)}$$

which can be uniquely solved for,

$$X^{(1)} = M - \frac{R}{2} \quad and \quad X^{(n)} = M + \frac{R}{2} \tag{21.5.1}$$

with $0 \le r < x^{(n)} - x^{(1)}, x^{(1)} + \frac{r}{2} < m < x^{(n)} - \frac{r}{2}$. And the Jacobian of such transfoamtion is,

$$J = \begin{vmatrix} \dfrac{\partial X^{(1)}}{\partial M} & \dfrac{\partial X^{(1)}}{\partial R} \\ \dfrac{\partial X^{(n)}}{\partial M} & \dfrac{\partial X^{(n)}}{\partial R} \end{vmatrix} = \begin{vmatrix} 1 & -\dfrac{1}{2} \\ 1 & \dfrac{1}{2} \end{vmatrix} = 1$$

Now, make the transformation defined in (20.5.1), and from (21.4.1) we have,

$$\begin{aligned} f_{M,R}(m,r) &= f_{X^{(1)},X^{(n)}}\left(x^{(1)},x^{(n)}\right)|J| = f_{X^{(1)},X^{(n)}}\left(m-\frac{r}{2}, m+\frac{r}{2}\right) \\ &= \binom{n}{n-2} f_X\left(x^{(1)}\right)[F_X\left(x^{(n)}\right) - F_X(x^{(1)})]^{n-2} f_X\left(x^{(n)}\right) \\ &= n(n-1)[F_X\left(x^{(n)}\right) - F_X(x^{(1)})]^{n-2} f_X\left(x^{(1)}\right) f_X\left(x^{(n)}\right) \\ &= n(n-1) f_X\left(m+\frac{r}{2}\right) f_X\left(m-\frac{r}{2}\right)\left[F_X\left(m+\frac{r}{2}\right) - F_X\left(m-\frac{r}{2}\right)\right]^{n-2} \end{aligned}$$

for $\left[m-\frac{r}{2}, m+\frac{r}{2}\right] \in \mathcal{A}$. ■

Furthermore, the marginal pdf of M is given by integrating $f_{M,R}(m,r)$ in terms of r, that is.

$$f_M(m) \; = \int\limits_{\left[m-\frac{r}{2},m+\frac{r}{2}\right]\in\mathcal{A}} n(n-1)f_X\left(m+\frac{r}{2}\right)f_X\left(m-\frac{r}{2}\right)\left[F_X\left(m+\frac{r}{2}\right)-F_X\left(m-\frac{r}{2}\right)\right]^{n-2} dr$$

The marginal pdf of R is given by integrating $f_{M,R}(m,r)$ in terms of m, that is,

$$f_R(r) \; = \int\limits_{\left[m-\frac{r}{2},m+\frac{r}{2}\right]\in\mathcal{A}} n(n-1)f_X\left(m+\frac{r}{2}\right)f_X\left(m-\frac{r}{2}\right)\left[F_X\left(m+\frac{r}{2}\right)-F_X\left(m-\frac{r}{2}\right)\right]^{n-2} dm$$

Example 21.5.1: Suppose $X_1, \dots X_n$ is a random sample from a $\mathcal{U}(0,\theta)$ population, find the joint pdf of M, the midpoint and R, the range.

Solution: cdf and pdf of the population are needed in the expression (21.1). Then by definition of a standard uniform distribution, we have,

$$f_X(t) \; = \begin{cases} \dfrac{1}{\theta}, & 0 < x < \theta \\ 0, & O.W. \end{cases}$$

$$F_X(t) \; = \begin{cases} 0, & t < 0 \\ \dfrac{t}{\theta}, & 0 \le t < \theta \\ 1, & \theta \le t \end{cases}$$

Therefore, by Theorem 21.5

$$f_{M,R}(m,r) \; = \begin{cases} n(n-1)\left(\dfrac{1}{\theta}\right)^2\left(\dfrac{r}{\theta}\right)^{n-2}, & 0 < r < \theta, \dfrac{r}{2} < \; m < \theta - \dfrac{r}{2} \\ 0, & O.W. \end{cases} \tag{21.5.1}$$

where the domain of m, r comes from $\left\{(m,r)\colon 0 < m - \frac{r}{2} < \theta, 0 < m + \frac{r}{2} < \theta\right\}$

Thus, the marginal pdf of M and marginal pdf of R are,

$$f_M(m) \; = \begin{cases} \displaystyle\int_0^{2m} \frac{n(n-1)}{\theta^n} r^{n-2}\, dr \quad = \frac{n}{\theta^n}(2m)^{n-2}, & 0 < m \le \dfrac{\theta}{2}; \\ \displaystyle\int_0^{2(\theta-m)} \frac{n(n-1)}{\theta^n} r^{n-2}\, dr = \frac{n}{\theta^n}(2(\theta-m))^{n-2}, & \dfrac{\theta}{2} < m \le \theta; \\ 0, & \theta > m \end{cases}$$

$$f_R(r) \; = \begin{cases} \displaystyle\int_{\frac{r}{2}}^{\theta-\frac{r}{2}} \frac{n(n-1)}{\theta^n} r^{n-2} dm = \frac{n(n-1)}{\theta^n} r^{n-2}(\theta - r), & \dfrac{r}{2} < r < \theta - \dfrac{r}{2} \\ 0, & O.W. \end{cases}$$

■

Lecture 22

Convergence in Probability

Definition 22.1: $X_1, X_2 \dots X_n$ *is a sequence of random variables that* ***converges in probability*** *to a random variable* X*, expressed as* $X_n \overset{\mathcal{P}}{\rightarrow} X$*, if* $\forall\ \varepsilon > 0$,

$$\lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1 \qquad (22.1a)$$

$$\textit{or equivalently} \quad \lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0 \qquad (22.1b)$$

Theorem 22.2 (Weak Law of Large Number): *Let* $X_1, X_2 \dots X_n$ *be i.i.d. random varibles with finite* $E(X_i) = \mu$*, and finite* $Var(X_i) = \sigma^2$*. Define* $\bar{X}_n = \frac{\sum_{i=1}^{n} X_n}{n}$*, then,*

$$\bar{X}_n \overset{\mathcal{P}}{\rightarrow} \mu \qquad (22.2)$$

Proof: By Chebychev's Inequality, we have,

$$P\{|\overline{X_n} - \mu| > \varepsilon\} \le \frac{E(\overline{X_n} - \mu)^2}{\varepsilon^2} = \frac{Var(\overline{X_n})}{\varepsilon^2} = \frac{\frac{\sigma^2}{n}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

Hence,

$$\lim_{n\to\infty} P\{|\overline{X_n} - \mu| > \varepsilon\} \le \lim \frac{\sigma^2}{n\varepsilon^2} = 0 \qquad \blacksquare$$

Definition 22.3 (Consistency): *Let* X_n *be an estimator of a random sample of size* n *which are all from the same population whose pdf determined by parameter* θ*, then* X_n *is called a* ***consistent*** *estimator of* θ *if* X_n *converges in probability to* θ*.*

Example 22.3.1: *Let* $X_1, X_2 \dots X_n$ *be a random sample from a population specified by pdf,*

$$f_X(x) = \begin{cases} e^{-(x-\theta)}, & x > \theta \\ 0, & O.W. \end{cases}$$

let $X_n^{(1)} = \min\{X_1 \dots X_n\}$*, show that* $X_n^{(1)}$ *is a consistent estimator of* θ*.*

Proof: The cdf can be derived from pdf, which gives,

$$F_X(t) = \int_0^t e^{-(x-\theta)}\, dx = 1 - e^{-(t-\theta)}, t > \theta$$

Follow the definition of consistency, we have,

$$P\left\{\left|X_n^{(1)} - \theta\right| < \varepsilon\right\} = P\left\{\theta - \varepsilon < X_n^{(1)} < \theta + \varepsilon\right\} = \int_{\theta-\varepsilon}^{\theta+\varepsilon} f_{X_n^{(1)}}(t)dt = \int_{\theta-\varepsilon}^{\theta+\varepsilon} n(1 - F_X(t))^{n-1} f_X(t)dt$$

$$= \int_{\theta-\varepsilon}^{\theta+\varepsilon} ne^{-(n-1)(t-\theta)} \cdot e^{-(t-\theta)} dt = \int_{\theta-\varepsilon}^{\theta+\varepsilon} ne^{-n(t-\theta)} dt$$

$$= \int_{\theta-\varepsilon}^{\theta} ne^{-n(t-\theta)}dt + \int_{\theta}^{\theta+\varepsilon} ne^{-n(t-\theta)}dt = 0 + \int_{\theta}^{\theta+\varepsilon} ne^{-n(t-\theta)}dt$$

$$= \int_{\theta}^{\theta+\varepsilon} ne^{-n(t-\theta)}dt = e^{n\theta}\cdot n\cdot\left(-\frac{1}{n}\right)\cdot e^{-nt}\,\Bigg|\begin{matrix}\theta+\varepsilon\\ \theta\end{matrix}$$

$$= -e^{n(\theta-t)}\,\Bigg|\begin{matrix}\theta+\varepsilon\\ \theta\end{matrix} = 1-e^{-n\varepsilon}$$

Thus, $P\left\{\left|X_n^{(1)}-\theta\right|<\varepsilon\right\}$ goes to 1 since $e^{-n\varepsilon}$ approaches 0, as $n\to\infty$. Then, (22.1*a*) gives that $X_n^{(1)}$ converges to θ in probability and therefore $X_n^{(1)}$ is a consistent estimator of θ. ■

Example 22.3.2: Let $X_1,\dots,X_n$ be i.i.d. random varibles with finite $E(X_i)=\mu$, and finite $Var(X_i)=\sigma^2$. Define $S_n^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$, and find a sufficient condition to make S_n^2 a consistent estimator.

Proof: As shown before**,** $E(S_n^2)=\sigma^2$, so $E(S_n^2-\sigma^2)^2=Var(S_n^2)$. Then Chebychev's Inequality gives, for any $\varepsilon>0$,

$$P\{|S_n^2-\sigma^2|>\varepsilon\} \le \frac{E(S_n^2-\sigma^2)^2}{\varepsilon^2}=\frac{Var(S_n^2)}{\varepsilon^2}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that $\lim\limits_{n\to\infty} Var(S_n^2)=0$. ■

Example 22.3.3: Let $X_1,\dots,X_n$ be i.i.d. $N(\mu,\sigma^2)$ random variables. Define $S_n^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$, or sample mean, show that S_n^2 is a consistent estimator of σ^2.

Proof: From last example, we have,

$$P\{|S_n^2-\sigma^2|>\varepsilon\} \le \frac{Var(S_n^2)}{\varepsilon^2}=\frac{Var\left(\frac{\sigma^2}{n-1}\cdot(n-1)\frac{S_n^2}{\sigma^2}\right)}{\varepsilon^2}=\frac{\frac{\sigma^4}{(n-1)^2}Var\left((n-1)\frac{S_n^2}{\sigma^2}\right)}{\varepsilon^2}$$

By Theorem 19.9, $(n-1)\frac{S_n^2}{\sigma^2}\sim\mathcal{X}_{n-1}^2$with variance $2(n-1)$, then continue from the last expression above

$$P\{|S_n^2-\sigma^2|>\varepsilon\} \le \frac{\frac{\sigma^4}{(n-1)^2}\cdot 2(n-1)}{\varepsilon^2}=\frac{\frac{2\sigma^4}{n-1}}{\varepsilon^2}=\frac{2\sigma^4}{(n-1)\varepsilon^2}$$

Hence,

$$\lim_{n\to\infty} P\{|S_n^2-\sigma^2|>\varepsilon\} \le \lim\frac{2\sigma^4}{(n-1)\varepsilon^2}=0$$

Therefore, (22.1*b*) gives that S_n^2 converges in probability to σ^2 and S_n^2 is a consistent estimator of σ^2. ■

Theorem 22.4: *For a sequence of random variable* $\{X_n\}$, *if* $X_n\overset{\mathcal{P}}{\to}X$, *then for any continuous function* g ,

$$g(X_n)\overset{\mathcal{P}}{\to}g(X) \qquad (22.4)$$

Proof: Since g is a continuous function, then for $\forall\epsilon>0,\exists\,\delta>0$, such that $|X_n-X|<\delta$ implies $|g(X_n)-g(X)|<\epsilon$. Then,

$$P\{|g(X_n) - g(X)| < \epsilon\} \geq P\{|X_n - X| < \varepsilon\}$$

Take limits on both sides,

$$\lim_{n\to\infty} P\{|g(X_n) - g(X)| < \epsilon\} \geq \lim_{n\to\infty} P\{|X_n - X| < \varepsilon\} = 1 \qquad (22.4.1)$$

The *first axiom* of probability gives that,

$$\lim_{n\to\infty} P\{|g(X_n) - g(X)| < \epsilon\} \leq 1 \qquad (22.4.2)$$

By Squeeze Theorem, (22.4.1) and (22.4.2) together yields an equality,

$$\lim_{n\to\infty} P\{|g(X_n) - g(X)| < \epsilon\} = 1 \qquad ■$$

Example 22.4.1: Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Define $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, or sample mean, show that S_n is a consistent estimator of σ.

Proof: Define function $g(a) = \sqrt{a}$, which is continuous for $a \geq 0$. So that $g(S_n^2) = \sqrt{S_n^2} = S_n$ and $g(\sigma^2) = \sqrt{\sigma^2} = \sigma$. Furthermore, it has been shown in Example 22.3.3 that S_n^2 converges in probability to σ^2, then Theorem 22.4 gives that $g(S_n^2)$ also converges in probability to $g(\sigma^2)$, namely S_n converges in probability to σ. ■

Almost Sure Convergence

Definition 22.5: *$X_1, X_2 \dots X_n$ is a sequence of random variables that* ***converges almost surely*** *(or* ***almost everywhere,*** *or* ***with probability one****) to a random variable X, expressed as $X_n \xrightarrow{a.s.} X$, if $\forall\ \varepsilon > 0$,*

$$P\left\{\lim_{n\to\infty} |X_n - X| < \varepsilon\right\} = 1 \qquad (22.5)$$

Theorem 22.6 (Strong Law of Large Number): *Let $X_1, X_2 \dots X_n$ be i.i.d. random varibles with finite $E(X_i) = \mu$, and finite $Var(X_i) = \sigma^2$. Define $\bar{X}_n = \frac{\sum_{i=1}^n X_n}{n}$, then,*

$$\bar{X}_n \xrightarrow{a,s.} \mu \qquad (22.6)$$

Theorem 22.7: *For a sequence of random variables $\{X_n\}$ and a random variable X,*

$$If\ X_n \xrightarrow{a.s.} X, \quad then\ \ X_n \xrightarrow{\mathcal{P}} X \qquad (22.7)$$

Theorem 22.8: *For a sequence of random variables $\{X_n\}$ and a constant C,*

$$X_n \xrightarrow{\mathcal{P}} C \quad if\ and\ only\ if \quad X_n \xrightarrow{a.s.} C \qquad (22.8)$$

Theorem 22.9: *For a sequence of random variable $\{X_n\}$, if $X_n \xrightarrow{a.s.} X$, then for any continuous function g ,*

$$g(X_n) \xrightarrow{a.s.} g(X) \qquad (22.9)$$

Convergence in Distribution

Recall from Definition 11.1 and Theorem 11.2, for a sequence of variables $X_1, X_2, \ldots$

a. $\quad$ *If* $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ *at all continuity point* t, *then* $X_n \overset{\mathcal{D}}{\Rightarrow} X$ $\qquad$ (Convergence in distribution)

b. $\quad$ *If* $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ *for all* t *in neighborhood of 0, then* $X_n \overset{\mathcal{D}}{\Rightarrow} X$ $\qquad$ (Convergence of mgf)

where $X_n \overset{\mathcal{D}}{\Rightarrow} X$ is the notation for "X_n converges in distribution to X".

Theorem 22.10: *For a sequence of random variables* $\{X_n\}$ *and a random variable* X,

$$\text{If } X_n \overset{\mathcal{P}}{\to} X, \quad \text{then } \; X_n \overset{\mathcal{D}}{\Rightarrow} X \tag{22.10}$$

Theorem 22.11: *For a sequence of random variables* $\{X_n\}$ *and a constant* C,

$$X_n \overset{\mathcal{D}}{\Rightarrow} C \quad \text{if and only if} \quad X_n \overset{\mathcal{P}}{\to} C \tag{22.11}$$

Proof: The "if" part can be easily shown by Theorem 22.10. From (22.1*a*), it suffices to prove the "only if" part by showing that: for any $\varepsilon > 0$, $\lim_{n\to\infty} P\{|X_n - C| < \varepsilon\} = 1$, where the limit can be expanded as,

$$\lim_{n\to\infty} P\{|X_n - C| < \varepsilon\} = \lim_{n\to\infty} P\{C - \varepsilon < X_n < C + \varepsilon\} = \lim_{n\to\infty} \int_{C-\epsilon}^{C+\varepsilon} f_{X_n}(t)dt$$

$$= \lim_{n\to\infty} \int_{-\infty}^{C+\varepsilon} f_{X_n}(t)dt - \lim_{n\to\infty} \int_{-\infty}^{C-\varepsilon} f_{X_n}(t)dt = \lim_{n\to\infty} F_{X_n}(C+\varepsilon) - \lim_{n\to\infty} F_{X_n}(C-\varepsilon)$$

$$= F_C(C+\varepsilon) - F_C(C-\varepsilon) = \int_{C-\epsilon}^{C+\varepsilon} f_C(t)dt = f_C(C) = 1$$

since ε can be arbitrarily close to 0. ■

Lecture 23

Convergence of Multiple Random Variables

Theorem 23.1: *For a sequence of random variables* $\{X_n\}$ *and a sequence of random variables* $\{Y_n\}$, *if* $X_n \overset{\mathcal{P}}{\to} X, Y_n \overset{\mathcal{P}}{\to} Y$, *then,*

a. $$X_n + Y_n \overset{\mathcal{P}}{\to} X + Y \tag{23.1a}$$

b. $$X_n Y_n \overset{\mathcal{P}}{\to} XY \tag{23.1b}$$

Proof:

a. According to the property of absolute value such that $|a + b| \le |a| + |b|$, it is true that,

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

Therefore, we have these inclusive relations, such that,

$$\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \cap \left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\} \subset \{|X_n - X| + |Y_n - Y| < \varepsilon\} \subset \{|(X_n + Y_n) - (X + Y)| < \varepsilon\}$$

By monotone property of probability, we then have,

$$P\left\{\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \cap \left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\}\right\} \leq P\{|(X_n + Y_n) - (X + Y)| < \varepsilon\} \qquad (23.1.1)$$

Apply Bonferroni's Inequality on the left joint probability, we have,

$$P\left\{\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \cap \left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\}\right\} \geq P\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} + P\left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\} - 1$$

and take limits on both sides,

$$\lim_{n\to\infty} P\left\{\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} \cap \left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\}\right\} \geq \lim_{n\to\infty} P\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} + \lim_{n\to\infty} P\left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\} - 1$$

where the two limits of probability on the right can be easily seen to be:

$$\lim_{n\to\infty} \{P\left\{|X_n - X| < \frac{\varepsilon}{2}\right\} = 1 \qquad and \qquad \lim_{n\to\infty} P\left\{|Y_n - Y| < \frac{\varepsilon}{2}\right\} = 1$$

Since $X_n \overset{\mathcal{P}}{\to} X, Y_n \overset{\mathcal{P}}{\to} Y$, for $\forall\varepsilon$. Referring the values of limit back to (23.1.1), we finally get,

$$\lim_{n\to\infty} P\{|(X_n + Y_n) - (X + Y)| < \varepsilon\} \geq 1 + 1 - 1 = 1$$

and hence, by *first axiom* of probability and Squeeze theorem, we have,

$$\lim_{n\to\infty} P\{|(X_n + Y_n) - (X + Y)| < \varepsilon\} = 1$$

which is just how $X_n + Y_n \overset{\mathcal{P}}{\to} X + Y$ is defined. ■

b. From (23.1a), we have

$$X_n + Y_n \overset{\mathcal{P}}{\to} X + Y$$

Apply the continuous function $g(a) = a^2$ on $X_n, Y_n, X_n + Y_n$ and Theorem 22.4 gives that,

$$(X_n + Y_n)^2 \overset{\mathcal{P}}{\to} (X + Y)^2 \qquad (23.1.2)$$

$$(X_n)^2 \overset{\mathcal{P}}{\to} (X)^2$$

$$(Y_n)^2 \overset{\mathcal{P}}{\to} (Y)^2$$

The last two convergences together implies that,

$$(X_n)^2 + (Y_n)^2 \overset{\mathcal{P}}{\to} (X)^2 + (Y)^2$$

And expanding 23.1.2, we have,

$$(X_n)^2 + 2X_nY_n + (Y_n)^2 \overset{\mathcal{P}}{\to} (X)^2 + 2XY + (Y)^2 \qquad (23.1.3)$$

Apply the continuous function $h(a) = -a$ on $(X_n)^2 + (Y_n)^2$ and by Theorem 22.4, we have

$$-(X_n)^2 - (Y_n)^2 \overset{\mathcal{P}}{\to} -(X)^2 - (Y)^2 \qquad (23.1.4)$$

(23.1.3) and (23.1.4) together gives,

$$2X_nY_n \overset{\mathcal{P}}{\to} 2XY$$

Then, apply another continuous function $f(a) = \frac{1}{2}a$ on $2X_nY_n$ and Theorem 22.4, we have,

$$X_nY_n \overset{\mathcal{P}}{\to} XY$$ ■

In fact, this theorem can be generalized to any *continuous function* h and have $h(X_n, Y_n) \overset{\mathcal{P}}{\to} h(X, Y)$, if $X_n \overset{\mathcal{P}}{\to} X$ and $Y_n \overset{\mathcal{P}}{\to} Y$

Theorem 23.2 (Central Limit Theorem): *Let $X_1, X_2 \dots X_n$ be i.i.d. random variables with finite $E(X_i) = \mu$, and finite $Var(X_i) = \sigma^2$. Define $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$, then,*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\mathcal{D}}{\Rightarrow} Z \sim N(0,1) \tag{23.2}$$

$$OR \quad \bar{X}_n \overset{\mathcal{D}}{\Rightarrow} X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof: Let $Y_i = (X_i - \mu)/\sigma$, so that $Y_1, \dots, Y_n$ are i.i.d. random variables with $\mu = 0$, $\sigma^2 = 1$. Define:

$$Z_n = \frac{\bar{Y}_n}{1/\sqrt{n}} = \sqrt{n} \cdot \bar{Y}_n$$

And notice that, Z_n is also equal to $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Then the mgf of Z_n is given by,

$$M_{Z_n}(t) = E\left[e^{t\sqrt{n}\cdot\bar{Y}_n}\right] = M_{\bar{Y}_n}(t\sqrt{n}) = M_{\frac{\sum_{i=1}^n Y_i}{n}}(t\sqrt{n}) = M_{\sum_{i=1}^n Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Take log on both sides,

$$\log M_{Z_n}(t) = n \log M_Y\left(\frac{t}{\sqrt{n}}\right)$$

Use Taylor expansion on the exponential function $e^{(t/\sqrt{n})Y}$, we have,

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = E\left[e^{\frac{t}{\sqrt{n}}Y}\right] \approx E\left[1 + \frac{t}{\sqrt{n}}Y + \frac{t^2}{2n}Y^2\right] \approx 1 + \frac{t^2}{2n}$$

The identity: $\lim_{\varepsilon \to 0} \log(1 + \varepsilon) = \varepsilon$ gives that, as $n \to \infty$, $t^2/2n \to 0$, and thus,

$$\log M_Y\left(\frac{t}{\sqrt{n}}\right) = n \log\left(1 + \frac{t^2}{2n}\right) \approx \frac{t^2}{2n}$$

Therefore,

$$M_{Z_n}(t) \approx \exp\left(n \cdot \frac{t^2}{2n}\right) = e^{\frac{t^2}{2}}$$

$Z_n \sim N(0,1)$ follows by uniqueness of mgf. ■

Theorem 23.3 (Slutsky's Theorem): *For a sequence of random variables $\{X_n\}$ and a sequence of random variables $\{Y_n\}$, if $X_n \overset{\mathcal{D}}{\Rightarrow} X, Y_n \overset{\mathcal{P}}{\rightarrow} C$, where X is a andom variable and C is a constant, then*

$$X_n \cdot Y_n \overset{\mathcal{D}}{\Rightarrow} X \cdot C \tag{23.3}$$

Example 23.3.1: Let $X_1, \dots, X_n$ be i.i.d. random varibles with finite $E(X_i) = \mu$, and finite $Var(X_i) = \sigma^2$. Define $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ and $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, and find a sufficient condition for:

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \overset{\mathcal{D}}{\Rightarrow} Z \sim N(0,1)$$

Solution: The expression on the left has been shown to be a T-statistics, which can be rewritten as,

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{S_n}$$

where $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to $N(0,1)$. From Slutsky's Theorem, it can be easily seen that we just need:

$$\frac{\sigma}{S_n} \overset{\mathcal{P}}{\rightarrow} 1 \quad \text{or equivalently} \quad S_n \overset{\mathcal{P}}{\rightarrow} \sigma$$

Example 22.3.2 has shown that a sufficient condition for this convergence in probability is:

$$Var(S_n^2) \rightarrow 0, \quad as\ n \rightarrow \infty$$ ■

Lecture 24

Point Estimation and Sufficiency

Definition 24.1 (Point Estimation): *Let $X_1, X_2, \dots X_n$ be a random sample from a population with pdf $f_{\boldsymbol{\theta}}(x)$, where $\boldsymbol{\theta}$ is the unknown. In order to estimate the* ***population parameter*** *(or vector of parameters) $\boldsymbol{\theta}$, a* ***statistic****, $T(X), X = (X_1, \dots X_n)$, is constructed and calculated, which is to serve as the "best guess" of $\boldsymbol{\theta}$. In classical point of view, the parameter $\boldsymbol{\theta}$ is* ***fixed.*** *$T(X)$, a function of the sample data, is called* ***estimator****.*

The notations $f(x|\boldsymbol{\theta})$ (Beysian notation) and $f_{\boldsymbol{\theta}}(x)$ will be interchangeably used in the context.

For a parameter, take μ as an example, there are many candidates of estimators, $X^{(1)}, X^{(n)}, X_1, \bar{X}$ ..., then which ones of them are better than the others? What are the criteria to evaluate that an estimator is a "good" even the "best" estimator?

Definition 24.2 (Sufficiency): *A* ***statistic*** *$T(\boldsymbol{X})$ is sufficient for θ, if $f_{(\boldsymbol{X}|T(\boldsymbol{X}))}(\boldsymbol{x}|T(\boldsymbol{x}))$ is independent of θ. In other words, $T(X)$ contains all information of θ*

Example 24.2.1: Let $X_1, X_2, \dots X_n$ be a random sample from $Bernoulli(\theta)$, whose pmf is given by,

$$f_\theta(x) = \theta^x (1-\theta)^{1-x}, x = 0,1$$

where θ is fixed and $0 < \theta < 1$, which is unknown. Define an estimator $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$, determine if $T(\boldsymbol{X})$ is sufficient.

Solution: By definition of conditional probability, we have,

$$f(\boldsymbol{x}|T(x)) = P[X_1 = x_1, X_2 = x_2, \dots X_n = x_n \mid \textstyle\sum_{i=1}^{n} X_i = t]$$

$$= \frac{P[X_1 = x_1, X_2 = x_2, \dots X_n = x_n, \sum_{i=1}^{n} X_i = t]}{P[\sum_{i=1}^{n} X_i = t]}$$

$$= \frac{P[X_1 = x_1, X_2 = x_2, \dots X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i]}{P[\sum_{i=1}^{n} X_i = t]} \qquad (24.2.1)$$

Example 14.7.1 showed that the sum of n independent $Bernoulli(\theta)$ random variables has a $\mathcal{B}in(n, \theta)$ distribution, therefore, the pmf of $T(\boldsymbol{X})$ is:

$$P[T(\boldsymbol{X}) = t] = P\left[\sum_{i=1}^{n} X_i = t\right] = \binom{n}{t} \theta^t (1-\theta)^{n-t}$$

Referring back to (24.2.1),

$$f(\boldsymbol{x}|T(x)) = \frac{\theta^{x_1}(1-\theta)^{1-x_1} \cdot \dots \cdot \theta^{x_{n-1}}(1-\theta)^{1-x_{n-1}} \cdot \theta^{t-\sum_{i=1}^{n-1} x_i}(1-\theta)^{1-(t-\sum_{i=1}^{n-1} x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\theta^{\sum_{i=1}^{n-1} x_i}(1-\theta)^{n-1-(\sum_{i=1}^{n-1} x_i)} \cdot \theta^{t-\sum_{i=1}^{n-1} x_i}(1-\theta)^{1-(t-\sum_{i=1}^{n-1} x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} = \frac{t!\,(n-t)!}{n!}$$

The last expression clearly is independent of θ, so $T(\boldsymbol{X})$ as defined is sufficient for θ. ■

Definition 24.3 (Likelihood Function): *Given that a random sample is observed such that $\boldsymbol{X} = \boldsymbol{x}$, the function of θ defined by,*

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

is called the likelihood function.

For example, let $x_1, \dots x_n$ be observations of a random sample from $Bernoulli(\theta)$, then the likelihood function for θ is,

$$L(\theta|\boldsymbol{x}) = f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n}[\theta^{x_i}(1-\theta)^{1-x_i}] = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-(\sum_{i=1}^{n} x_i)}$$

Theorem 24.4 (Factorization Theorem): *Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(X)$ is a sufficient statistic for θ, if and only if $\exists g(t|\theta)$ such that, for all sample points $\boldsymbol{x}$ and all parameter points θ,*

$$f(\boldsymbol{x}|\theta) = g(\,T(\boldsymbol{x})|\theta) \cdot h(\boldsymbol{x}) \qquad (24.4)$$

Or written as $f_\theta(\boldsymbol{x}) = g[T(\boldsymbol{x}), \theta] \cdot h(\boldsymbol{x})$

Or equivalently, a statistic $T(X)$ is sufficient for θ, if the likelihood function $L(\theta|\boldsymbol{x})$ depends on $\boldsymbol{\theta}$ only through $T(X)$

Proof: To prove Necessity of (24.4), suppose $T(\boldsymbol{X})$ is sufficient, choose $g(t|\theta) = f_\theta(T(\boldsymbol{X}) = t)$, and $h(\boldsymbol{x}) = f(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x}))$, then

$$f(\boldsymbol{x}|\theta) = f_\theta(\boldsymbol{X} = \boldsymbol{x}) = f_\theta(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x}))$$

$$
\begin{aligned}
&= f_\theta\big(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x})\big) \cdot f_\theta\big(T(\boldsymbol{X}) = T(\boldsymbol{x})\big) \\
&= f\big(\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = T(\boldsymbol{x})\big) \cdot g(\, T(\boldsymbol{x})|\theta) \\
&= h(\boldsymbol{x}) \cdot g(\, T(\boldsymbol{x})|\theta)
\end{aligned}
$$

To prove Sufficiency of (24.4), suppose that $f(\boldsymbol{x}|\theta) = g(\, T(\boldsymbol{x})|\theta) \cdot h(\boldsymbol{x})$, let $q(t|\theta)$ be the pdf of $T(\boldsymbol{x})$,

$$
\begin{aligned}
\frac{f(\boldsymbol{x}|\theta)}{q(t|\theta)} &= \frac{g(\, T(\boldsymbol{x}) = t|\theta) \cdot h(\boldsymbol{x})}{q(t|\theta)} = \frac{g(\, T(\boldsymbol{x}) = t|\theta) \cdot h(\boldsymbol{x})}{q(T(\boldsymbol{x}) = t|\theta)} \\
&= \frac{g(\, T(\boldsymbol{x})|\theta) \cdot h(\boldsymbol{x})}{\int_{\{x:T(x)=t\}} f(\boldsymbol{x}|\theta)} = \frac{g(\, T(\boldsymbol{x}) = t|\theta) \cdot h(\boldsymbol{x})}{\int_{\{x:T(x)=t\}} g(\, t|\theta) \cdot h(\boldsymbol{x})} \\
&= \frac{g(\, T(\boldsymbol{x}) = t|\theta) \cdot h(\boldsymbol{x})}{g(t|\theta) \cdot \int_{\{\boldsymbol{x}|T(\boldsymbol{x}) = t\}} h(\boldsymbol{x})} = \frac{h(\boldsymbol{x})}{\int_{\{\boldsymbol{x}|T(\boldsymbol{x}) = t\}} h(\boldsymbol{x})}
\end{aligned}
$$

where the last expression is independent of θ and thus $T(\boldsymbol{X})$ is sufficient by definition. ■

Example 24.4.1: Let $X_1, X_2, \ldots X_n$ be a random sample from $\mathcal{U}(\theta)$, whose pdf is given by,

$$
f_\theta(x) = \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & O.W. \end{cases}
$$

Find a sufficient statistic for θ.

Solution: The joint pdf of $\boldsymbol{X}$ can be derived by,

$$
f_\theta(\boldsymbol{x}) \ = \begin{cases} \frac{1}{\theta^n}, & 0 < x^{(n)} < \theta \\ 0, & O.W. \end{cases}
$$

which can be rewritten as,

$$
f_\theta(\boldsymbol{x}) \ = \frac{1}{\theta^n} \cdot I_{\{x^{(n)} < \theta\}}(\theta)
$$

By Factorization Theorem, θ is determined only through $X^{(n)}$, so that $X^{(n)}$ is sufficient for θ ■

Example 24.4.2: Let $X_1, X_2, \ldots X_n$ be a random sample from $\mathcal{P}(\theta)$, whose pmf is given by,

$$
f_\theta(x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0,1,2, \ldots
$$

Find a sufficient statistic for θ.

Solution: The joint pdf of $\boldsymbol{X}$ can be derived by,

$$
f_\theta(\boldsymbol{x}) \ = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)}
$$

which can be rewritten as,

$$
f_\theta(\boldsymbol{x}) \ = e^{-n\theta}\theta^{\sum_{i=1}^n x_i} \cdot \frac{1}{\prod_{i=1}^n (x_i!)}
$$

By Factorization Theorem, θ is determined only through $\sum_{i=1}^n X_i$, so that $\sum_{i=1}^n X_i$ is sufficient for θ ■

Theorem 24.5: *A statistic $T(X)$ is sufficient for θ, then,*

a. *T is also sufficient for $\varphi(\theta)$*

b. *$h(T)$ is also sufficient for θ*

This theorem indicates that the sufficient statistic is not unique.

Lecture 25

Example 25.1.1: Let $X_1, X_2, \ldots X_n$ be a random sample from $N(\theta, 1)$, whose pdf is given by,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

Find a sufficient statistic for θ.

Solution: The joint pdf of $\boldsymbol{X}$ can be derived by,

$$f_\theta(\boldsymbol{x}) \;= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$= (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\left[\sum_{i=1}^{n} x_i^2 - 2\theta\sum_{i=1}^{n} x_i + n\theta^2\right]}$$

which can be rewritten as,

$$f_\theta(\boldsymbol{x}) \;= (2\pi)^{-\frac{n}{2}} e^{\frac{1}{2}\left[2\theta\sum_{i=1}^{n} x_i - n\theta^2\right]} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2}$$

By Factorization Theorem, the joint pdf depends on θ through $\sum_{i=1}^{n} X_i$, so $\sum_{i=1}^{n} X_i$ is sufficient for θ ■

By similar approach, it can be shown that $\sum_{i=1}^{n} X_i$ is sufficient for θ in all such distribution as $\beta(1,\theta)$, $\beta(n,\theta)$, $Exp(\theta)$, $\gamma(1,\theta), \mathcal{P}(\theta)$

Example 25.1.2: Let $X_1, X_2, \ldots X_n$ be a random sample from $N(0, \theta)$, whose pdf is given by,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$$

Find a sufficient statistic for θ.

Solution: The joint pdf of $\boldsymbol{X}$ can be derived by,

$$f_\theta(\boldsymbol{x}) \;= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x_i^2}{2\theta}} = (2\pi\theta)^{-\frac{n}{2}} e^{-\frac{1}{2\theta}\left(\sum_{i=1}^{n} x_i^2\right)}$$

which can be rewritten as,

$$f_\theta(\boldsymbol{x}) \;= (2\pi\theta)^{-\frac{n}{2}} e^{-\frac{1}{2\theta}\sum_{i=1}^{n} {x_i}^2} \cdot I_{\{x_i \in \mathbb{R}^2\}}$$

By Factorization Theorem, the joint pdf depends on θ through $\sum_{i=1}^{n} {X_i}^2$, so $\sum_{i=1}^{n} {X_i}^2$ is sufficient for θ ■

Theorem 25.2 (Sufficiency of Exponential Family): *Let $X_1, X_2, \ldots X_n$ be a random sample from a population with pdf $f_{\boldsymbol{\theta}}(x)$ that belongs to a one-parameter exponential family given by,*

$$f_\theta(x) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A(x)$$

Then, $\sum_{i=1}^{n} T(X_i)$ is a sufficient statistic for θ

Proof: The joint pdf of $\boldsymbol{X}$ is,

$$f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} \exp\{c(\theta)T(x_i) + S(x_i) + d(\theta)\} \cdot I_A(x_i)$$

$$= \exp\left\{c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta)\right\} \cdot \sum_{i=1}^{n} S(x_i) \cdot I_A(x_1, \dots x_n)$$

By Factorization Theorem, let $g[\theta, T(\boldsymbol{x})] = \exp\{c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta)\}$, it can be seen that the joint pdf depends on θ only through $\sum_{i=1}^{n} T(x_i)$. Therefore $\sum_{i=1}^{n} T(x_i)$ is sufficient for θ ■

Definition 25.3: *A vector of statistics* $(T_1(X) \dots T_k(X))$ *is* ***jointly sufficient*** *for a vector of k parameters* θ*, if* $f_{(X|T_1(X), \dots T_k(X))}(x|T_1(\boldsymbol{x}), \dots T_k(\boldsymbol{x}))$ *is independent of* $\boldsymbol{\theta}$**.**

Theorem 25.4: *Let* $X_1, X_2, \dots X_n$ *be a random sample from a population with pdf* $f_{\boldsymbol{\theta}}(x)$ *that belongs to a* k*-parameter exponential family given by,*

$$f_{\boldsymbol{\theta}}(x) = \exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})T_i(x) + S(x) + d(\boldsymbol{\theta})\right\} \cdot I_A(x)$$

Then,

$$(T_1(X) \dots T_k(X)) = \left(\sum_{j=1}^{n} T_i(x_j) \dots \sum_{j=1}^{n} T_k(x_j)\right)$$

is a joint sufficient statistic for $\boldsymbol{\theta}$

Example 25.4.1: Let $X_1, X_2, \dots X_n$ be a random sample from $N(\mu, \sigma^2)$, whose pdf is given by,

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

where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. Find a sufficient statistic for $\boldsymbol{\theta}$.

Solution: The pdf of X can be rewritten as,

$$f_\theta(x) = \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

By Theorem 25.4, $T(\boldsymbol{X}) = \left(\sum_{i=1}^{n} {X_i}^2, \sum_{i=1}^{n} X_i\right)$ is jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$ ■

Lecture 26

Completeness

Definition 26.1: *Let* $f(t|\theta)$ *be pdf or pmf for a statistic* $T(\boldsymbol{X})$*, and* $T(\boldsymbol{X})$*is complete for* θ*, if*

$$E[g(T)] = 0$$

implies $g \equiv 0$ *for all* θ

Example 26.1.1: Suppose the statistic $T(\boldsymbol{X})$ has pdf or pmf as specified, show that $T(\boldsymbol{X})$ is complete.

1). $T(\boldsymbol{X}) \sim \mathcal{P}\ (\theta)$

Proof: From the definition of expectation and pdf of Poisson distribution, we have,

$$E[g(T)]\ = \sum_{k=0}^{n} g(k)\frac{e^{-\theta}\theta^k}{k!} = e^{-\theta}\sum_{k=0}^{n}\frac{g(k)}{k!}\theta^k = 0$$

which holds only if

$$\sum_{k=0}^{n}\frac{g(k)}{k!}\theta^k = \sum_{k=0}^{n}\alpha(k)\theta^k = 0\ , where\ \alpha(k) = \frac{g(k)}{k!}, \forall\theta > 0$$

$\alpha(k)\theta^k$ can be viewed as a Power series, since $\theta > 0$, then this equality holds only if, $\alpha(k) \equiv 0$,and thus $g(k) \equiv 0, \forall k \geq 0$. ■

2). $T(\boldsymbol{X}) \sim \mathcal{B}\ (n, \theta)$

Proof: From the definition of expectation and pdf of Binomial distribution, we have,

$$E[g(T)]\ = \sum_{k=0}^{n} g(k)\binom{n}{k}\theta^k(1-\theta)^{n-k} = (1-\theta)^n\sum_{k=0}^{n} g(k)\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k = 0$$

which holds only if

$$\sum_{k=0}^{n} g(k)\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k = \sum_{k=0}^{n}\alpha(k)\beta^k = 0, where\ \alpha(k) = g(k)\binom{n}{k}, \beta = \frac{\theta}{1-\theta}, \forall\theta \in [0,1]$$

$\alpha(k)\beta^k$ can be viewed as a Power series, since $\beta > 0$, then this equality holds only if, $\alpha(k) \equiv 0$,and thus $g(k) \equiv 0, \forall k \geq 0$ ■

Example 26.1.2: Let $X_1, \dots, X_n$ be a random sample from $\mathcal{U}\ (0, \theta)$, show that $T = X^{(n)}$ is a complete statistic

Proof: Example 21.1.1 has given that the pdf of $X^{(n)}$ from n i.i.d. $\mathcal{U}\ (0, \theta)$ random variables,

$$f_T(t) = nF_X(t)^{n-1}\theta^{-1} = n\left(\frac{t}{\theta}\right)^{n-1}\theta^{-1} = nt^{n-1}\theta^{-n}$$

From the definition of expectation and pdf of Uniform distribution, we have,

$$E[g(T)]\ = \int_0^\theta g(t)\cdot nt^{n-1}\theta^{-n}dt = 0$$

And thus, we need to have,

$$0 = \frac{d}{d\theta}E[g(T)]\ = \frac{d}{d\theta}\int_0^\theta g(t)\cdot nt^{n-1}\theta^{-n}dt = (\theta^{-n})g(\theta)\cdot n\theta^{n-1} + \frac{d}{d\theta}(\theta^{-n})\int_0^\theta g(t)\cdot nt^{n-1}dt$$

$$= (\theta^{-n})g(\theta)\cdot n\theta^{n-1} + 0 = g(\theta)\cdot n\theta^{-1}$$

For all $\theta > 0$, the equality holds only if $g(\theta) \equiv 0$. ■

Theorem 26.2: *Let $X_1, X_2, \dots X_n$ be a random sample from a population with pdf $f_{\boldsymbol{\theta}}(x)$ that belongs to a one-parameter exponential family given by,*

$$f_\theta(x) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A(x)$$

Then, $\sum_{i=1}^n T(X_i)$ is a complete statistic for θ if the range of $c(\theta)$ contains an open interval.

Example 26.2.1: Let $X_1, X_2, \dots X_n$ be a random sample from $\mathcal{B}ernoulli(\theta)$, whose pmf is given by,

$$f_\theta(x) = \theta^x(1-\theta)^{1-x}, x = 0,1$$

Find a complete statistic for θ.

Solution: The pdf of X belongs to one-parameter exponential which can be written as,

$$f_\theta(x) \ = \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right\}$$

where $T(x) = x, c(\theta) = \log\left(\frac{\theta}{1-\theta}\right) \in (-\infty, \infty)$, which contains an open interval. Therefore, $\sum_{i=1}^n X_i$ is a complete statistic for θ ■

Theorem 26.3: *Let $X_1, X_2, \dots X_n$ be a random sample from a population with pdf $f_{\boldsymbol{\theta}}(x)$ that belongs to a k-parameter exponential family given by,*

$$f_{\boldsymbol{\theta}}(x) = \exp\left\{\sum_{i=1}^{k} c_i(\boldsymbol{\theta})T_i(x) + S(x) + d(\boldsymbol{\theta})\right\} \cdot I_A(x)$$

Then,

$$\left(T_1(X) \dots T_k(X)\right) = \left(\sum_{j=1}^{n} T_i(x_j) \dots \sum_{j=1}^{n} T_k(x_j)\right)$$

is complete if $\left(c_1(\boldsymbol{\theta}), \dots, c_k(\boldsymbol{\theta})\right)$ contains a non-empty interior.

Example 26.3.1: Let $X_1, X_2, \dots X_n$ be a random sample from $N(\mu, \sigma^2)$, whose pdf is given by,

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

where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. Find a complete statistic for $\boldsymbol{\theta}$.

Solution: The pdf of X belongs to one-parameter exponential which can be written as,

$$f_\theta(x) \ = \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 - \frac{1}{2}\log(2\pi\sigma^2)\right\} = \exp\left\{-\frac{1}{2\sigma^2}[x^2 - 2\mu x + \mu^2] - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

$$= \exp\left\{\left(-\frac{1}{2\sigma^2}\right)x^2 + \left(\frac{\mu}{\sigma^2}\right)x + d_1(\boldsymbol{\theta}) + d_2(\boldsymbol{\theta})\right\},\ d_1(\boldsymbol{\theta}) = -\frac{\mu^2}{2\sigma^2}, d_2(\boldsymbol{\theta}) = \ -\frac{1}{2}\log(2\pi\sigma^2)$$

where $c_1(\boldsymbol{\theta}) = -1/2\sigma^2, T_1(x) = x^2; c_2(\boldsymbol{\theta}) = \mu/\sigma^2, T_2(x) = x$. Since $(-1/2\sigma^2, \mu/\sigma^2)$ is a two dimensional vector, then $\left(\sum_{i=1}^n {X_i}^2, \sum_{i=1}^n X_i\right)$ is complete for (μ, σ^2). ■

Example 26.3.2: Let $X_1, \dots, X_n$ be i.i.d. random variable with common mean θ and common variance θ^2, and define $\boldsymbol{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n {X_i}^2\right)$, justify if $\boldsymbol{T}$ is complete.

Solution: It suffices to show that there exists a function $h \neq 0$ such that $E[h(\boldsymbol{T})] = 0$. Then,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E(X_i) = n\theta$$
$$E\left[\sum_{i=1}^{n} {X_i}^2\right] = \sum_{i=1}^{n} E\left({X_i}^2\right) = nE(X^2) = n[Var(X) + E^2(X)] = 2n\theta^2$$

So that, after some rearranging,

$$E\left[\frac{\sum_{i=1}^{n} {X_i}^2}{2n}\right] = \theta^2 \tag{26.3.2a}$$

Furthermore,

$$E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] = Var\left[\sum_{i=1}^{n} X_i\right] + E^2\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} Var(X_i) + E^2\left[\sum_{i=1}^{n} X_i\right]$$
$$= n\theta^2 + n^2\theta^2 = n(n+1)\theta^2$$

After some rearranging, we have

$$E\left[\frac{\sum_{i=1}^{n} {X_i}^2}{2n}\right] = \theta^2 \tag{26.3.2b}$$

Construct a function $h(T)$,

$$h(\boldsymbol{T}) = \frac{(\sum_{i=1}^{n} X_i)^2}{n(n+1)} - \frac{\sum_{i=1}^{n} {X_i}^2}{2n} \neq 0$$

From (26.3.2*a*) and (26.3.2*b*), we have,

$$E[h(T)] = \theta^2 - \theta^2 = 0, \forall\theta$$

Which is contradictory with the definition of complete statistics, so that $\boldsymbol{T}$ is NOT complete. ■

Example 26.3.3: Let $X_1, X_2, \dots X_n$ be a random sample from $N(\theta, \theta^2)$, whose pdf is given by,

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(x-\theta)^2}{2\theta^2}}$$

where $\boldsymbol{\theta} = (\theta, \theta^2)$ is unknown. Define $\boldsymbol{T} = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} {X_i}^2\right)$ and justify if $\boldsymbol{T}$ is complete.

Solution: The pdf of X belongs to one-parameter exponential which can be written as,

$$f_\theta(x) = \exp\left\{-\frac{1}{2}\left(\frac{x-\theta}{\theta}\right)^2 - \frac{1}{2}\log(2\pi\theta^2)\right\} = \exp\left\{-\frac{1}{2\theta^2}[x^2 - 2\theta x + \theta^2] - \frac{1}{2}\log(2\pi\theta^2)\right\}$$
$$= \exp\left\{\left(-\frac{1}{2\theta^2}\right)x^2 + \left(\frac{1}{\theta}\right)x + d_1(\boldsymbol{\theta}) + d_2(\boldsymbol{\theta})\right\},\ d_1(\boldsymbol{\theta}) = -\frac{1}{2}, d_2(\boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\theta^2)$$

where $c_1(\boldsymbol{\theta}) = -1/2\theta^2, T_1(x) = x^2; c_2(\boldsymbol{\theta}) = 1/\theta, T_2(x) = x$.

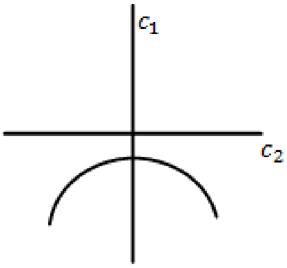


Figure 26.3.3

where $c_1 = -{c_2}^2/2$, as depicted above, does not give a two-dimensional vector; in other words, $(-1/2\theta^2, 1/\theta)$ does not contain an open set in $\mathbb{R}^2$. Therefore, $\left(\sum_{i=1}^{n} {X_i}^2, \sum_{i=1}^{n} X_i\right)$ is NOT complete. ■

Lecture 27

Recall from Definition 24.1 that an estimator is a function $T(\boldsymbol{X})$ of a sample, which can be various. In order to compare among these estimators by evaluating their performance and optimality, a nonnegative function that generally increases as the distance between $T(\boldsymbol{X})$ and the unknown parameter θ increases, will be used. Such function is called loss function and *the less the average loss it has, the better the estimator is.*

Definition 27.2 (Loss Function): *Let $T(\boldsymbol{X})$ be an estimator of a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$, which is from a population with pdf $f_\theta(x)$ where θ is unknown. Then, a* ***loss function****, denoted by $l(T, \theta)$, is a nonnegative, convex function of $T(\boldsymbol{X}) - \theta$. Two commonly used loss functions are:*

Absolute Error Loss (AEL): $$l(T, \theta) = |T - \theta| \tag{27.2.1}$$

and

Squared Error Loss (SEL): $$l(T, \theta) = (T - \theta)^2 \tag{27.2.2}$$

By averaging the SEL, we have the *Mean Squared Error (MSE)*, which can be decomposed into two parts:

$$\begin{aligned} E[(T-\theta)^2] &= E[(T - E(T) + E(T) - \theta)^2] \\ &= E\left[\left(T - E(T)\right)^2\right] + E[(E(T) - \theta)^2] + 2E\left[\left(T - E(T)\right)(E(T) - \theta)\right] \\ &= Var(T) + (E(T) - \theta)^2 + 2(E(T) - \theta)[E(T) - E(T)] \\ &= Var(T) + (E(T) - \theta)^2 \end{aligned} \tag{27.2.3}$$

where the term $(E(T) - \theta)^2$ is called the *bias.*

<u>Unbiased Estimator</u>

Definition 27.3: *An estimator $T(\boldsymbol{X})$ is an unbiased estimator of θ, if its bias is equal to 0, or*

$$E(T(\boldsymbol{X})) = \theta \tag{27.3}$$

Theorem 27.4: *Let T^* and T be two unbiased estimators of θ, then T^* is better then T in terms of squared error loss, if,*

$$Var(T^*) < Var(T), \forall\theta \tag{27.4}$$

This can be easily seen by (27.2.3). Given that the biases of $T^*$and T are both 0, then the estimator with smaller variance yields smaller loss.

Example 27.4.1: Let $X_1, \dots, X_n$ be a random sample from the population $N(\theta, 1)$. Define two estimators $T = \frac{\sum_{i=1}^{n} X_i}{n}, T^* = X_1$, determine which one is better than the other.

Solution: Clearly, both T and T^* are unbiased estimators of θ. Furthermore, we have,

$$Var(T^*) = Var(X_1) = 1$$

$$Var(T) = Var(\bar{X}) = \frac{1}{n}$$

T is better than T^*, since $Var(T) < Var(T^*)$ for $n > 1$ ■

Definition 27.5 (UMVUE): *An estimator T^* is called the* ***uniform minimal variance unbiased estimator (UMVUE)****, if for all θ and any other unbiased estimator T, we have $Var(T^*) \leq Var(T)$.*

However, the UMVUE is not necessarily the best estimator. In other words, there exists a biased estimator, which has smaller MSE than the unbiased estimator.

Example 27.6: Let $X_1, \dots, X_n$ be a random sample from the population (μ, σ^2) where both parameters are unknown. In order to estimate σ^2, define two estimators $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}, S^{*2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$, determine which one is better than the other.

Solution: Since σ^{*2} is not an unbiased estimator (asymptotically unbiased though), instead of comparing their variances, we will compute and compare their MSEs directly. Then,

$$\begin{aligned} E\left[\left(S^{*2} - \sigma^2\right)^2\right] &= Var\left(\sigma^{*2}\right) + \left[E\left(\sigma^{*2}\right) - \sigma^2\right]^2 = \left(\frac{n-1}{n}\right)^2 \cdot Var(\hat{\sigma}^2) + \left[\frac{n-1}{n}\sigma^2 - \sigma^2\right]^2 \\ &= \left(\frac{n-1}{n}\right)^2 \cdot \frac{2\sigma^4}{n-1} + \left[\frac{n-1}{n}\sigma^2 - \sigma^2\right]^2 = \frac{2(n-1)}{n^2}\sigma^4 + \frac{(-1)^2}{n^2}\sigma^4 \\ &= \frac{\sigma^4}{n^2}[2(n-1) + 1] = \frac{2n-1}{n^2}\sigma^4 \end{aligned} \tag{27.6.1}$$

while S^2 has been shown in Theorem 19.9, $(n-1)\frac{S^2}{\sigma^2} \sim \mathcal{X}_{n-1}^2$, so that,

$$E\left((n-1)\frac{S^2}{\sigma^2}\right) = n - 1 \quad \Longrightarrow \quad E(S^2) = \sigma^2$$

$$Var\left((n-1)\frac{S^2}{\sigma^2}\right) = 2(n-1) \ \Longrightarrow \ Var(S^2) = \frac{2}{n-1}\sigma^4$$

Therefore, we have,

$$E[(S^2 - \sigma^2)^2] = Var(S^2) + [E(S^2) - \sigma^2]^2 = Var(S^2) = \frac{2}{n-1}\sigma^4 \tag{27.6.2}$$

Observe that (27.6.1) and (27.6.2) are both multiple of σ^4, the smaller coefficient between $2/(n-1)$ and $(2n-1)/n^2$ will give a smaller MSE. Then, by some simple algebra, it can be seen that for $n \geq 1$

$$2n^2 > 2n^2 - (3n - 1) = (2n - 1)(n - 1)$$

Then, divide $n^2(n-1)$ on both sides, we have,

$$\frac{2}{n-1} > \frac{2n-1}{n^2}$$

Hence, $E\left[\left(S^{*2}-\sigma^2\right)^2\right] < E[(S^2-\sigma^2)^2]$. So, the best estimator is not always unbiased. ■

Theorem 27.7 (Completeness and Unbiasedness): *If an estimator $T(\boldsymbol{X})$ is a complete estimator of θ, then there exists a unique function of $T(\boldsymbol{X})$, $g(T)$, such that $g(T)$ unbiased of θ*

Proof: Suppose T is complete and let $g(T)$ and $h(T)$ be two different functions of T, which are both unbiased of θ. Then, we have,

$$E[g(T)] = \theta \qquad and \qquad E[h(T)] = \theta$$

Hence,

$$E[g(T) - h(T)] = 0$$

Let $K(T) = g(T) - h(T)$, since T is complete, then $E[K(T)] = 0$ can only imply $K(T) \equiv 0$, therefore,

$$g(T) \equiv h(T)$$ ■

Theorem 27.8 (Consistency and Unbiasedness): *If an estimator $T(\boldsymbol{X})$ is asymptotically unbiased of θ and has variance converging to 0 as $n \to \infty$. Then, $T(\boldsymbol{X})$ is a consistent estimator of θ*

Proof: From Chebyshev's Inequality and (27.2.3), we have,

$$\begin{aligned} P\{|T_n - \theta| < \varepsilon\} &\le \frac{E(T_n-\theta)^2}{\varepsilon^2} = \frac{E\left(T_n - E(T_n)\right)^2}{\varepsilon^2} + \frac{E(E(T_n)-\theta)^2}{\varepsilon^2} \\ &= \frac{Var(T_n)}{\varepsilon^2} + \frac{(E(T_n)-\theta)^2}{\varepsilon^2} \end{aligned}$$

Then, take limits on both sides,

$$\begin{aligned} P\{|T_n - \theta| < \varepsilon\} &\le \lim_{n\to\infty} \frac{Var(T_n)}{\varepsilon^2} + \lim_{n\to\infty} \frac{(E(T_n)-\theta)^2}{\varepsilon^2} \\ &= 0 + 0 = 0 \end{aligned}$$

where the two 0s can be easily derived from given conditions. Therefore, T is a consistent estimator of θ by definition. ■

In addition, Unbiasedness is not functional invariance, a property possessed by Sufficiency, which says: if an estimator T is sufficient for θ, then $g(T)$ is sufficient for $g(\theta)$. Namely, if an estimator T is unbiased of θ, then $g(T)$ is not necessarily unbiased of $g(\theta)$.

Lecture 28

Uniformly Minimal Variance Unbiased Estimation (UMVUE)

Theorem 28.1(Blackwell-Rao Theorem): *Suppose that $T(\boldsymbol{X})$ is **sufficient** for θ and $S(\boldsymbol{X})$ is another statistic and $E|S(\boldsymbol{X})| < \infty$ for all θ. Let $T^* = E[S|T]$, which is a function of T and independent of θ. Then:*

a. If $Var(S) < \infty$, for all θ

$$E\left[\left(T^* - q(\theta)\right)^2\right] \le E\left[\left(S - q(\theta)\right)^2\right] \qquad (28.1a)$$

Or we say "T^* is better than S" or "T^* dominates S". And the inequality is strict unless $T^* = S$

b. *If S is an unbiased estimator for $q(\theta)$, then T^* is also unbiased for $q(\theta)$ and*

$$Var(T^*) \leq Var(S), \qquad \forall\theta \in \Omega \tag{28.1b}$$

where Ω is the parameter space

Proof: To show (28.1a), we first expanding the right expectation by adding T^* and then subtracting it,

$$E\left[\left(S-q(\theta)\right)^2\right] = E\left[\left(S-T^*+T^*-q(\theta)\right)^2\right] = E\left[(S-T^*)^2+\left(T^*-q(\theta)\right)^2+2(S-T^*)\left(T^*-q(\theta)\right)\right]$$

$$= E[(S-T^*)^2] + E\left[\left(T^*-q(\theta)\right)^2\right] + 2E\left[(S-T^*)\left(T^*-q(\theta)\right)\right] \tag{28.1.1}$$

To calculate the cross product term, we use the Smooth Theorem and have,

$$\begin{aligned}E\left[(S-T^*)\left(T^*-q(\theta)\right)\right] &= E\left\{E\left[(S-T^*)\left(T^*-q(\theta)\right)\middle|T\right]\right\} = E\left\{\left(T^*-q(\theta)\right)E[(S-T^*)|T]\right\}\\ &= E\left\{\left(E(S|T)-q(\theta)\right)E\left[\left(S-E(S|T)\right)\middle|T\right]\right\}\\ &= E\left\{\left(E(S|T)-q(\theta)\right)E[(S|T)-E(S|T)]\right\}\\ &= E\left\{\left(E(S|T)-q(\theta)\right)[E(S|T)-E(S|T)]\right\} = 0\end{aligned}$$

Referring back to (28.1.1), then

$$E\left[\left(S-q(\theta)\right)^2\right] = E[(S-T^*)^2] + E\left[\left(T^*-q(\theta)\right)^2\right] + 0 = E[(S-T^*)^2] + E\left[\left(T^*-q(\theta)\right)^2\right]$$

Since $E[(S-T^*)^2] \geq 0$ for sure, then,

$$E\left[\left(S-q(\theta)\right)^2\right] \geq E\left[\left(T^*-q(\theta)\right)^2\right]$$

equality holds only when $S = T^*$

(28.1b) is just a special case of (28.1a). Given that S is unbiased for $q(\theta)$, i.e., $E(S) = q(\theta)$, then $E(T^*) = E[E(S|T)] = E(S) = q(\theta)$. Hence,

$$E\left[\left(T^*-q(\theta)\right)^2\right] = Var(T^*)$$

$$E\left[\left(S-q(\theta)\right)^2\right] = Var(S)$$

Then, from (28.1a) we can finally get,

$$Var(T^*) \leq Var(S), \forall\theta \in \Omega \qquad \blacksquare$$

***Remark**: In order to have UMVUE, we can restrict our search among those that are functions of its sufficient estimator

Example 28.1.1: Let $X_1, \dots X_n$ be i.i.d. $\mathcal{P}(\theta)$ with pdf $f_X(x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0,1,2, \dots \dots$ Define an estimator $T^* = E[X_1|\sum_{i=1}^n X_i]$, find T^* and show it is a UMVUE

Solution: We first show that T^* is a UMVUE by following Blackwell-Rao Theorem. From Factorization Theorem, we have that $\sum_{i=1}^n X_i$ is sufficient for θ and $Var(\sum_{i=1}^n X_i) = E(\sum_{i=1}^n X_i) = n\theta < \infty$. Also, X_1 is an unbiased estimator of θ. Hence, $E[X_1|\sum_{i=1}^n X_i] = T^*$ is a UMVUE. Then, in order to specify T^*, we expand the conditional expectation,

$$E[X_1|\textstyle\sum_{i=1}^n X_i] = P[X_1 = x_1|\sum_{i=1}^n X_i = t] = \dfrac{P[X_1 = x_1, \sum_{i=1}^n X_i = t]}{P\left[\sum_{i=1}^n X_i = t\right]}$$

$$= \frac{P[X_1 = x_1, \sum_{i=2}^{n} X_i = t - x_1]}{P[\sum_{i=1}^{n} X_i = t]} \tag{28.1.1}$$

Since the sum of Poisson random variables is still Poisson with parameters adding together. Then, the numerator can be further specified as,

$$P\left[X_1 = x_1, \sum_{i=2}^{n} X_i = t - x_1\right] = P(X_1 = x_1)P\left[\sum_{i=2}^{n} X_i = t - x_1\right] = \frac{e^{-\theta}\theta^{x_1}}{x_1!} \cdot \frac{e^{-(n-1)\theta}[(n-1)\theta]^{(t-x_1)}}{(t-x_1)!}$$

$$= \frac{e^{-n\theta}(n-1)^{(t-x_1)}\theta^t}{x_1!\,(t-x_1)!}$$

and the denominator can further specified as,

$$P\left[\sum_{i=1}^{n} X_i = t\right] = \frac{e^{-n\theta}(n\theta)^t}{t!}$$

Put them back into (28.1.1), we have,

$$E[X_1 | \textstyle\sum_{i=1}^{n} X_i] = \frac{e^{-n\theta}(n-1)^{(t-x_1)}\theta^t}{x_1!\,(t-x_1)!} \cdot \frac{t!}{e^{-n\theta}(n\theta)^t} = \frac{t!}{x_1!\,(t-x_1)!} \cdot \frac{(n-1)^{(t-x_1)}}{n^t}$$

$$= \frac{t!}{x_1!\,(t-x_1)!} \cdot \left(\frac{1}{n}\right)^{x_1} \left(1 - \frac{1}{n}\right)^{t-x_1}$$

which is the pdf of $Bin\left(t, \frac{1}{n}\right)$, whose expectation is the product of the two parameters, therefore,

$$T^* = E[X_1 | \textstyle\sum_{i=1}^{n} X_i] = \frac{1}{n} \cdot t = \frac{t}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

And it can be easily shown that for $n > 1$, $Var(\bar{X}) = \theta/n < \theta = Var(X_1)$, so $\bar{X}$ dominates X_1. ■

Again, if S is unbiased estimator of $q(\theta)$, then $E(S|T)$ of which T is sufficient, does have a smaller variance than S. If an estimator is a UMVUE, it also has to be a function of sufficient statistic.

Lecture 29

Theorem 29.1(Lehmann –Scheffé Theorem): *If T is complete sufficient for θ and $h(T)$ is an unbiased estimator of $q(\theta)$. Then $h(T)$ is the unique UMVUE of $q(\theta)$*

Proof: First show that $h(T)$ is an UMVUE by rewriting $h(T)$ as $E[h(T)|T]$. Given that T is sufficient and $h(T)$ is an unbiased estimator of $q(\theta)$, by Blackwell-Rao Theorem, $h(T) = E[h(T)|T]$ is the UMVUE. Then, from Theorem 27.7, the completeness of T ensures that $h(T)$ is unique

Example 29.1.1: Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{U}\ (0, \theta)$ with unknown θ. Find the UMVUE for θ

Solution: It has been shown from Example 24.4.1 and Example 26.1.2 that $X^{(n)}$, or max$\{X_1, \ldots X_n\}$ is complete sufficient for θ. Then, by Lehmann-Scheffe Theorem, we need to search for such $T = h\left(X^{(n)}\right)$ that $E(T) = \theta$. On the other hands, from the pdf of $X^{(n)}$ given in Example 21.1.1,

$$f_{X^{(n)}}(t) = \begin{cases} nt^{n-1}, & 0 < t < 1 \\ 0, & O.W. \end{cases}$$

we have,

$$E\left(X^{(n)}\right) = \int_0^\theta t \cdot n\frac{t^{n-1}}{\theta^n}dt = \frac{n}{\theta^n}\int_0^\theta t^n dt = \frac{n}{\theta^n}\cdot\frac{t^{n+1}}{n+1}\Big|_0^\theta = \frac{\theta^{n+1}\cdot n}{\theta^n\cdot(n+1)} = \frac{n}{n+1}\theta$$

Then, if we let $T = \frac{n+1}{n}X^{(n)}$,

$$E(T) = E\left[\frac{n+1}{n}X^{(n)}\right] = \frac{n+1}{n}\cdot\frac{n}{n+1}\theta = \theta$$

Hence, $T = \frac{n+1}{n}X^{(n)}$ is the UMVUE for θ. ■

Example 29.1.2: Suppose that $X_1, \dots, X_n$ are i.i.d. $\mathcal{P}(\theta)$ with unknown θ. Find the UMVUE for θ

Solution: From Example 24.4.2 and Example 26.1.1, we know that $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic of θ. It can be seen at once that:

$$E\left(\sum_{i=1}^n X_i\right) = n\theta$$

And we further have,

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{n\theta}{n} = \theta$$

Then, if we let $h(T) = \bar{X}$, we have $E[h(T)] = \theta$ and Lehmann-Scheffe Theorem immediately gives that $h(T) = \bar{X}$ is the unique UMVUE of θ. ■

Given a random sample $X_1, \dots X_n$ from a population with pdf determined by θ, $\bar{X}$ is also the UMVUE of the θ in such distributions as $Bernoulli(\theta), N(\theta, 1)$, which can be justified in the similar way.

Example 29.1.3: Suppose that $X_1, \dots, X_n$ are i.i.d. $Exp(\theta)$ with unknown θ. Let $q(\theta) = 1/\theta$, find the UMVUE for $q(\theta)$

Solution: By sufficiency and completeness for an exponential family, we can show that $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic of θ. It can be seen at once that:

$$E\left(\sum_{i=1}^n X_i\right) = \frac{n}{\theta}$$

And we further have,

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{n}{n\theta} = \frac{1}{\theta}$$

Then, if we let $h(T) = \bar{X}$, we have $E[h(T)] = \theta$ and Lehmann-Scheffe Theorem immediately gives that $h(T) = \bar{X}$ is the unique UMVUE of $1/\theta$. ■

Given a random sample $X_1, \dots X_n$ from a population with pdf determined by θ, $\bar{X}$ is also the UMVUE of the θ in such distributions as $Geo(\theta)$, which can be justified in the similar way.

Example 29.1.4: Suppose that $X_1, \dots, X_n$ are i.i.d. $N(0, \theta^2)$ with unknown θ^2. Find the UMVUE for θ^2

Solution: The pdf of $N(0, \theta^2)$ belongs to a one-parameter exponential family, such that

$$f_{\theta^2}(x) = \frac{1}{\sqrt{2\pi\theta^2}}e^{-\frac{x^2}{\theta^2}} = \exp\left\{-\frac{1}{\theta^2}x^2 - \frac{1}{2}\log(\theta^2) - \frac{1}{2}\log(2\pi)\right\}$$

Let $T(x) = x^2, c(\theta) = -1/\theta^2$ where $c(\theta) \in (-\infty, 0]$. Then, by Theorem 25.2 and Theorem 26.2, we immediately have that $T = \sum_{i=1}^{n} {X_i}^2$ is complete sufficient for θ^2. It can be seen at once that:

$$E\left(\sum_{i=1}^{n} {X_i}^2\right) = n\theta^2$$

And we further have,

$$E\left(\frac{\sum_{i=1}^{n} {X_i}^2}{n}\right) = \frac{n\theta^2}{n} = \theta^2$$

Then, if we let $h(T) = \frac{\sum_{i=1}^{n} {X_i}^2}{n}$, we have $E[h(T)] = \theta^2$ and from Lehmann-Scheffe Theorem, we can conclude $h(T) = \frac{\sum_{i=1}^{n} {X_i}^2}{n}$ is the unique UMVUE of θ^2. ■

Proposition 29.2: *By combining Blackwell-Rao and Lehmann- Scheffé theorems, we have that***:** *let S be an unbiased estimator of $q(\theta)$, and T is complete sufficient for θ. Define $T^* = E(S|T) = h(T)$, namely $E(T^*) = q(\theta)$. Then, $T^* = E[S|T]$ is the unique UMVUE of $q(\theta)$.*

Example 29.2.2: Suppose that $X_1, \dots, X_n$ are i.i.d. (θ) with unknown θ. Let $q(\theta) = e^{-\theta}$, find the UMVUE for $q(\theta)$

Solution: It can be easily seen that $q(\theta) = e^{-\theta} = P_\theta(X_1 = 0) = E\left[I_{\{X_1=0\}}\right]$, in other words, the statistic $I_{\{X_1=0\}}$ is unbiased of $q(\theta)$. Since $T = \sum_{i=1}^{n} X_i$ is complete sufficient for θ, $T^* = E\left[I_{\{X_1=0\}}\middle|T\right]$ is the UMVUE for $q(\theta)$ By Proposition 29.1.2. To calculate $T^* = E\left[I_{\{X_1=0\}}\middle|T\right] = P[X_1 = 0| \sum_{i=1}^{n} X_i = t]$, we first need to specify the conditional pmf $P[X_1 = x_1| \sum_{i=1}^{n} X_i = t]$. From Example 28.1.1, we know that

$$P[X_1 = x_1| \textstyle\sum_{i=1}^{n} X_i = t] \ = \binom{t}{x_1}\left(\frac{1}{n}\right)^{x_1}\left(1 - \frac{1}{n}\right)^{t-x_1}$$

Then, by taking $x_1 = 0$, we have

$$T^* = E\left[I_{\{X_1=0\}}\middle|T\right] = P[X_1 = 0| \textstyle\sum_{i=1}^{n} X_i = t] \ = \binom{t}{0}\left(\frac{1}{n}\right)^{0}\left(1 - \frac{1}{n}\right)^{t-0} = \left(1 - \frac{1}{n}\right)^{t} = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^{n} X_i}$$

Therefore, $T^* = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^{n} X_i}$ is the unique UMVUE for $q(\theta) = e^{-\theta}$. ■

Lecture 30

Fisher Information

Definition 30.1: *Let X be a random variable with pdf or pmf $f_\theta(x)$ and assume the support, defined by the set $\{x: f_\theta(x) > 0\}$, is independent of θ. Then, the* ***fisher information on θ base on a single random variable****, denoted by $I_1(\theta)$, is defined by*

$$I_1(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] \tag{30.1}$$

Example 30.1.1: Suppose $X \sim Bernoulli(\theta)$ with unknown θ, find $I_1(\theta)$

Solution: As defined, the pmf of X is

$$f_\theta(x) = \theta^x(1-\theta)^{(1-x)}, x = 0,1$$

where the support does not depend on θ. From definition of fisher information, we have

$$\begin{aligned}
I_1(\theta) &= E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = E\left[\frac{\partial}{\partial\theta}(x\log\theta + (1-x)\log(1-\theta))\right]^2 \\
&= E\left[\left(\frac{x}{\theta} - \frac{1-x}{1-\theta}\right)^2\right] = E\left[\left(\frac{X-\theta}{\theta(1-\theta)}\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}E[(X-\theta)^2] \\
&= \frac{1}{\theta^2(1-\theta)^2}E\left[\left(X-E(X)\right)^2\right] = \frac{1}{\theta^2(1-\theta)^2}Var(X) \quad \textit{(Since } E(X) = 1\cdot\theta + 0\cdot(1-\theta) = \theta\textit{)} \\
&= \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}
\end{aligned}$$

■

Example 30.1.2: Suppose $X \sim \mathcal{P}(\theta)$ with unknown θ, find $I_1(\theta)$

Solution: As defined, the pmf of X is

$$f_\theta(x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0,1,2,\ldots$$

where the support does not depend on θ. From definition of fisher information, we have

$$\begin{aligned}
I_1(\theta) &= E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = E\left[\frac{\partial}{\partial\theta}(-\theta + x\log\theta - \log x!)\right]^2 \\
&= E\left[\left(-1 + \frac{x}{\theta}\right)^2\right] = E\left[\left(\frac{x-\theta}{\theta}\right)^2\right] = \frac{1}{\theta^2}E[(X-\theta)^2] \\
&= \frac{1}{\theta^2}E\left[\left(X-E(X)\right)^2\right] = \frac{1}{\theta^2}Var(X) \qquad \textit{(Since } E(X) = 1\cdot\theta + 0\cdot(1-\theta) = \theta\textit{)} \\
&= \frac{\theta}{\theta^2} = \frac{1}{\theta}
\end{aligned}$$

■

Example 30.1.3: Suppose $X \sim N(\theta, \sigma^2)$ with unknown θ, find $I_1(\theta)$

Solution: As defined, the pdf of X is

$$f_\theta(x) = (2\pi\sigma^2)^{-\frac{1}{2}}e^{-\frac{(x-\theta)^2}{2\sigma^2}}, x \in (-\infty, \infty)$$

where the support does not depend on θ. From definition of fisher information, we have

$$I_1(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] = E\left[\frac{\partial}{\partial\theta}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2}\right)\right]^2$$

$$= E\left[\left(-\frac{1}{2\sigma^2}\cdot 2(x-\theta)\cdot(-1)\right)^2\right] = E\left[\left(\frac{x-\theta}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4}E[(X-\theta)^2]$$

$$= \frac{1}{\sigma^4}E\left[\left(X-E(X)\right)^2\right] = \frac{1}{\sigma^4}Var(X) \qquad \textit{(Since } E(X) = 1\cdot\theta + 0\cdot(1-\theta) = \theta\textit{)}$$

$$= \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

■

Observe that for all the three examples above, we have this reciprocal relation:

$$I_1(\theta) = \frac{1}{Var(X)}$$

In other words, the higher the value of Fisher Information is, the less variance of the random variable will be, the more precise the parameter is. The “information” that $I(\theta)$ provides can be thought as precision of the parameter.

Theorem 30.2: *There are two other expressions for calculating the Fisher Information, which are equivalent to the definition (30.1):*

a. $$I_1(\theta) = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] \tag{30.2a}$$

b. $$I_1(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] \tag{30.2b}$$

Proof: From the following identity:

$$Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2\right] - E^2\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]$$

We can see that in order to establish (30.2*a*), it suffices to show that the second squared term on the right is zero, or

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = 0 \tag{30.2a.1}$$

By expanding the expectation on the left, we have,

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int_S \frac{\partial}{\partial\theta}\log f_\theta(x)\cdot f_\theta(x)dx = \int_S \frac{f_\theta{}'(x)}{f_\theta(x)}\cdot f_\theta(x)dx = \int_S f_\theta{}'(x)\,dx$$

Since the support does not depend on θ, and $f_\theta(x)$ is differentiable, we can claim that

$$\int_S \frac{\partial}{\partial\theta}f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\int_S f_\theta(x)dx \tag{30.2a.2}$$

This claim will be proved later and right now we just apply this result and get,

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int_S f_\theta{}'(x)\,dx = \int_S \frac{\partial}{\partial\theta}f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\int_S f_\theta(x)\,dx = \frac{\partial}{\partial\theta}1 = 0$$

Then, we show claim (30.2*a*.2) by the definition of derivatives,

$$\int_S \frac{\partial}{\partial\theta} f_\theta(x)\,dx \; = \int_S \lim_{h\to 0}\frac{f(\theta+h,x)-f(\theta,x)}{h}dx = \int_S \lim_{h\to 0}\frac{f(\theta+h,x)}{h}dx - \int_S \lim_{h\to 0}\frac{f(\theta,x)}{h}dx$$

The sample space S, where $x\in S$ and the parameter space Ω, where $\theta\in\Omega$ are independent, therefore the integral and limit are interchangeable, that is for some $h\in\Omega$,

$$\begin{aligned}\int_S \frac{\partial}{\partial\theta} f_\theta(x)\,dx \; &= \int_S \lim_{h\to 0}\frac{f(\theta+h,x)}{h}dx - \int_S \lim_{h\to 0}\frac{f(\theta,x)}{h}dx \\ &= \lim_{h\to 0}\int_S \frac{f(\theta+h,x)}{h}dx - \lim_{h\to 0}\int_S \frac{f(\theta,x)}{h}dx \\ &= \lim_{h\to 0}\frac{\int_S f(\theta+h,x)dx - f(\theta,x)dx}{h} = \frac{\partial}{\partial\theta}\int_S f_\theta(x)\,dx\end{aligned}$$

The quantity in (30.2*b*) can be expressed as:

$$\begin{aligned}E\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] \; &= \int_S \frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\cdot f_\theta(x)dx = \int_S \frac{\partial}{\partial\theta}\left(\frac{f_\theta'(x)}{f_\theta(x)}\right)\cdot f_\theta(x)dx \\ &= \int_S \frac{f_\theta{}''(x)f_\theta(x) - f_\theta'(x)f_\theta'(x)}{f_\theta{}^2(x)}\cdot f_\theta(x)dx \qquad \textit{(Quotient Rule)} \\ &= \int_S \frac{f_\theta{}''(x)f_\theta(x) - f_\theta'(x)f_\theta'(x)}{f_\theta(x)}dx = \int_S f_\theta{}''(x)dx - \int_S \frac{\left(f_\theta'(x)\right)^2}{f_\theta(x)}dx \\ &= \int_S \frac{\partial^2}{\partial\theta^2}f_\theta(x)dx - \int_S \left(\frac{f_\theta'(x)}{f_\theta(x)}\right)^2\cdot f_\theta(x)dx \\ &= \frac{\partial^2}{\partial\theta^2}\int_S f_\theta(x)dx - \int_S \left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2\cdot f_\theta(x)dx \qquad \textit{By claim (30.2a.1)} \\ &= \frac{\partial^2}{\partial\theta^2}1 - E[(\log f_\theta(x))^2] = -E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2\right] = -I_1(\theta) \qquad \blacksquare\end{aligned}$$

Example 30.2.1: Revisit Example 30.1.1, Example 30.1.2, and Example 30.1.3 and use (30.2*a*) to calculate the fisher information

1). $X\sim Bernoulli(\theta)$

$$I_1(\theta) \; = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = Var\left[\frac{X-\theta}{\theta(1-\theta)}\right] = \frac{Var(X)}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

2). $X\sim\mathcal{P}(\theta)$

$$I_1(\theta) \ = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = Var\left[\frac{X-\theta}{\theta}\right] = \frac{Var(X)}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

3). $X \sim \mathcal{N}(\theta, \sigma^2)$

$$I_1(\theta) \ = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = Var\left[\frac{X-\theta}{\sigma^2}\right] = \frac{Var(X)}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

Theorem 30.3: *Let $X_1, X_2, \dots, X_n$ be a random sample from a population with pdf or pmf $f_\theta(x)$ where θ is unknown and independent of the sample, then, the Fisher Information on θ based on this random sample of size n, denoted by $I_n(\theta)$, is defined by,*

$$I_n(\theta) = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X_1, \dots, X_n)\right] = nI_1(\theta) \tag{30.3}$$

Proof: Since the support for $x_1, \dots, x_n$ doesn't dependent on θ, we have,

$$I_n(\theta) \ = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X_1, \dots, X_n)\right] = Var\left[\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f_\theta(X_i)\right] = Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f_\theta(X_i)\right]$$

By interchanging the order of differentiation and summation,

$$I_n(\theta) \ = Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f_\theta(X_i)\right] = Var\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right] = nVar\left[\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right] = nI_1(\theta)$$

where the last equality comes from the property of a random sample. ■

Example 30.3.1: Suppose $X_1, \dots, X_n$ is a random sample from $Bernoulli(\theta)$ with the unknown parameter θ, find the fisher information on θ

Solution: The joint pdf of such random sample has been shown to be,

$$f_\theta(x_1, \dots, x_n) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Then, from (30.2*a*), we have,

$$\begin{aligned} I_n(\theta) \ &= Var\left[\frac{\partial}{\partial\theta}\log f_\theta(X_1, \dots, X_n)\right] = Var\left[\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n} X_i\log\theta + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-\theta)\right)\right] \\ &= Var\left[\frac{\sum_{i=1}^{n} X_i}{\theta} - \frac{n-\sum_{i=1}^{n} X_i}{1-\theta}\right] = Var\left[\frac{\sum_{i=1}^{n} X_i - n\theta}{\theta(1-\theta)}\right] \\ &= \frac{\sum_{i=1}^{n} Var(X_i)}{\theta^2(1-\theta)^2} = \frac{n\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

which by comparing with the result of Example 30.1.1, is clearly equal to $nI_1(\theta)$. ■

Lecture 31

Cramer-Rao Lower Bound

Theorem 31.1: *Let $X_1, \dots X_n$ be a sample with joint pdf $f_\theta(\boldsymbol{x})$ and, $T(\boldsymbol{X})$ be any statistic with $Var\big(T(\boldsymbol{X})\big) < \infty$. Suppose that the two regularity conditions (i), (ii) are satisfied and that $I(\theta) \in (0, \infty)$,* *as well as* $E(T(\boldsymbol{X})) = \psi(\theta)$*. Then,* $\psi(\theta)$ *is differentiable and* $Var(T(\boldsymbol{X}))$ *is bounded by* ***Cramer-Rao Lower Bound (C.R.L.)****:*

$$Var(T(\boldsymbol{X})) \geq \frac{(\psi'(\theta))^2}{I(\theta)} \tag{31.1}$$

where $I(\theta)$ *denotes the Fisher Information of* θ *based on the sample. The two regularity conditions are:*

i) *The support of* $f_\theta(\boldsymbol{x})$ *is independent of* θ

ii) $\frac{d}{d\theta}\int\int\dots\int T(\boldsymbol{x})f_\theta(\boldsymbol{x})d\boldsymbol{x} = \int\int\dots\int T(\boldsymbol{x})\frac{\partial}{\partial\theta}f_\theta(\boldsymbol{x})d\boldsymbol{x}$

Proof: First look at the second regularity condition:

$$LHS = \frac{d}{d\theta}\int\int\dots\int T(\boldsymbol{x})f_\theta(\boldsymbol{x})d\boldsymbol{x} = \frac{d}{d\theta}E[T(\boldsymbol{X})] = \frac{d}{d\theta}\psi(\theta) = \psi'(\theta)$$

$$RHS = \int\int\dots\int T(\boldsymbol{x})\frac{\partial}{\partial\theta}f_\theta(\boldsymbol{x})d\boldsymbol{x} = \int\int\dots\int T(\boldsymbol{x})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\cdot f_\theta(\boldsymbol{x})\,d\boldsymbol{x} = E\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right]$$

Thus, the second regularity condition becomes:

ii)' $\psi'(\theta) = E\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right]$

Then, recall the following facts about covariance and correlation from Lecture 17:

1) $Cov(U,V) = E(UV) - E(U)E(V)$ *(by 17.1b)*

2) $\varphi(U,V) = \dfrac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} \in [0,1]$ *(by 17.4)*

3) $\varphi^2(U,V) \in [0,1]$ *(by 17.5)*

Take $U = T(\boldsymbol{X}), V = \frac{\partial}{\partial\theta}f_\theta(\boldsymbol{X}),$ and from 2) and 3) we have,

$$0 \leq \frac{Cov^2\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right)}{Var(T(\boldsymbol{X}))\cdot Var\left(\frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right)} \leq 1$$

And by multiplying the denominator on both sides, we get,

$$Cov^2\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right) \leq Var(T(\boldsymbol{X}))\cdot Var\left(\frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right) = Var(T(\boldsymbol{X}))\cdot I(\theta) \tag{31.2}$$

where the second variance is just the fisher information written in the form of (30.2*a*). The left-hand side of (31.2) can be further expanded by 1) as follows:

$$\begin{aligned}Cov^2\left(T(\boldsymbol{X}), \frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right) &= \left\{E\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right] - E[T(\boldsymbol{X})]E\left[\frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right]\right\}^2 \\ &= \left\{E\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}logf_\theta(\boldsymbol{X})\right] - E[T(\boldsymbol{X})]\cdot 0\right\}^2 \quad \textit{(by (30.2a.1))}\end{aligned}$$

$$= E^2\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}log f_\theta(\boldsymbol{X})\right] = \big(\psi'(\theta)\big)^2 \qquad \textit{(by regularity condition ii)')}$$

From the results above, (31.3) becomes:

$$(\psi'(\theta))^2 \leq Var\big(T(\boldsymbol{X})\big)\cdot I(\theta)$$

which is equivalent to what we want to establish in (31.1), that is:

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{(\psi'(\theta))^2}{I(\theta)} \qquad \blacksquare$$

Corollary 31.2: *A special case of Theorem 31.1 with $\psi(\theta) = \theta$, for any statistic, $T(\boldsymbol{X})$ that is unbiased of θ and with finite variance,*

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{1}{I(\theta)} \qquad (31.2)$$

The proof is a straight forward application of Theorem 31.1. Since $\psi(\theta) = \theta$, then $\psi'(\theta) = 1$.

Corollary 31.3: *Let $X_1, \dots, X_n$ be a random sample (i.i.d.) from a population with pdf $f_\theta(x)$, while the regularity conditions of Theorem 31.1 are satisfied, then for any statistic $T(\boldsymbol{X})$, s.t. $E\big(T(\boldsymbol{X})\big) = \psi(\theta)$,*

$$Var\big(T(\boldsymbol{X})\big) \geq \frac{(\psi'(\theta))^2}{nI_1(\theta)} \qquad (31.3)$$

This corollary can be easily shown by applying Theorem 30.3: the Fisher Information based on a random sample of size n is equal to $nI_1(\theta)$.

Proposition 31.4: *Let $X_1, \dots X_n$ be a sample with joint pdf $f_\theta(\boldsymbol{x})$ and, $T(\boldsymbol{X})$ by any statistic such that $E\big(T(\boldsymbol{X})\big) = \psi(\theta)$, then if the two regularity conditions are satisfied with $I(\theta)$ being the Fisher Information of θ based on the sample, and also*

$$Var[T(\boldsymbol{X})] = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)}$$

$T(\boldsymbol{X})$ is the UMVUE of $\psi(\theta)$

Example 31.4.1: Let $X_1, \dots X_n$ be a random sample from $N(\theta, 1)$, and $\psi(\theta) = \theta$, verify that $\bar{X}$ is the UMVUE of $\psi(\theta)$

Solution: On one hand, Theorem 19.9 has shown that $\bar{X} \sim N\left(\theta, \frac{1}{n}\right)$, in other words,

$$E(\bar{X}) = \theta, \qquad Var(\bar{X}) = \frac{1}{n}$$

On the other hand, the Fisher Information of the θ in $N(\theta, 1)$ based on one single X is that:

$$I_1(\theta) = Var\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right] = Var\left[\frac{\partial}{\partial\theta}\log\left((2\pi)^{-\frac{1}{2}}e^{-\frac{(x-\theta)^2}{2}}\right)\right]$$

$$= Var\left[\frac{\partial}{\partial\theta}\left(-\frac{1}{2}\log(2\pi) - \frac{(x-\theta)^2}{2}\right)\right] = Var\left[\frac{\partial}{\partial\theta}\left(-\frac{1}{2}(x-\theta)^2\right)\right]$$

$$= Var\left[-\frac{1}{2}\cdot 2(x-\theta)\cdot(-1)\right] = Var(X-\theta) = Var(X) = 1$$

Then, from Corollary 31.2 and 31.3, we have,

$$C.R.L. = \frac{1}{I(\theta)} = \frac{1}{nI_1(\theta)} = \frac{1}{n}$$

which is equal to $Var(\bar{X})$. Therefore, $\bar{X}$ is the UMVUE of θ by Proposition 31.4. ■

Example 31.4.2: Let $X_1, \dots X_n$ be a random sample from $Bernoulli(\theta)$, and $\psi(\theta) = \theta$, verify that $\bar{X}$ is the UMVUE of $\psi(\theta)$

Solution: On one hand, Theorem 19.5 has given that $E(\bar{X}) = \theta,\ Var(\bar{X}) = \frac{\theta(1-\theta)}{n}$

On the other hand, the Fisher Information of the θ in $Bernoulli(\theta)$ based on the random sample has been specified in Example 30.3.1, that is:

$$I(\theta) = \frac{n}{\theta(1-\theta)}$$

Then, from Corollary 31.2, we have,

$$C.R.L. = \frac{1}{I(\theta)} = \frac{1}{n/\theta(1-\theta)} = \frac{\theta(1-\theta)}{n}$$

which is equal to $Var(\bar{X})$. Therefore, $\bar{X}$ is the UMVUE of θ by Proposition 31.4. ■

Example 31.4.3: Let $X_1, \dots X_n$ be a random sample from $Exp(\theta)$, and $\psi(\theta) = 1/\theta$, verify that $\bar{X}$ is the UMVUE of $\psi(\theta)$

Solution: On one hand, from Theorem 19.5 and property of exponential distribution, we have

$$E(\bar{X}) = E(X) = \frac{1}{\theta}, \qquad Var(\bar{X}) = \frac{Var(X)}{n} = \frac{1/\theta^2}{n} = \frac{1}{n\theta^2}$$

On the other hand, from the pdf of $Exp(\theta)$, which is $f_\theta(x) = \begin{cases}\theta e^{-\theta x}, x > 0\\ 0,\ O.W.\end{cases}$, we have the Fisher Information of θ based on one single X is that:

$$\begin{aligned} I_1(\theta) &= Var\left[\frac{\partial}{\partial\theta}\log f_\theta(x)\right] = Var\left[\frac{\partial}{\partial\theta}\log\left(\theta e^{-\theta x}\right)\right] = Var\left[\frac{\partial}{\partial\theta}(\log\theta - \theta x)\right] \\ &= Var\left[\frac{1}{\theta} - x\right] = Var\left(X - \frac{1}{\theta}\right) = Var(X) = \frac{1}{\theta^2} \end{aligned}$$

Then, from Corollary 31.2 and 31.3, we have,

$$C.R.L. = \frac{(\psi'(\theta))^2}{I(\theta)} = \frac{\left(\frac{d}{d\theta}\left(\frac{1}{\theta}\right)\right)^2}{I(\theta)} = \frac{\left(-\frac{1}{\theta^2}\right)^2}{nI_1(\theta)} = \frac{1}{n\theta^4 I_1(\theta)} = \frac{1}{n\theta^4\cdot\frac{1}{\theta^2}} = \frac{1}{n\theta^2}$$

which is equal to $Var(\bar{X})$. Therefore, $\bar{X}$ is the UMVUE of $1/\theta$ by Proposition 31.4. ■

Lecture 32

Recall from last lecture, Proposition 31.4 states that the unbiased estimator of $\psi(\theta)$, $T(\boldsymbol{X})$, which reaches the Cramer-Rao lower bound must be the UMVUE of $\psi(\theta)$. However, it is possible for a UMVUE to have variance strictly larger than the Cramer-Rao lower bound.

Example 32.1:

$\mathbf{Example\ 6.1}: X\ is\ \mathcal{P}(\theta). Let\ \psi(\theta) = e^{-\theta}, find\ the\ UMVUE\ of\ \psi(\theta), and\ show\ that\ its\ variance\ is\ strictly\ larger\ than\ CRL$

$\mathbf{Solution}: \psi(\theta) = e^{-\theta}, P(X = x) = \frac{e^{-\theta}\theta^x}{x!}, x = 0,1,2, \dots \dots$

$\psi(\theta) = e^{-\theta} = P(X = 0) = E\left[I_{\{X=0\}}\right]$

$Since\ there\ is\ only\ one\ observation, let\ T(x) = I_{\{X=0\}}$

$E\left[I_{\{X=0\}}\right] = e^{-\theta} \Longrightarrow T(x)\ is\ an\ unbiased\ estimator\ for\ e^{-\theta}$

$It\ has\ been\ shown\ that\ Possion\ r.v. X\ is\ complete\ sufficient\ and\ T(x) = I_{\{X=0\}}, which\ is\ also\ a\ function\ of\ X \xRightarrow{Lehmann-Scheffe\ Theorem} T(x) = I_{\{X=0\}}\ is\ UMVUE$

$$Var\big(T(x)\big) = E[T^2(x)] - E^2[T(x)] = P(X = 0) - \big(P(X = 0)\big)^2$$
$$= e^{-\theta} - e^{-2\theta} = e^{-\theta}\left(1 - e^{-\theta}\right)$$

$$CRL = \frac{\left(\psi'(\theta)\right)^2}{I(\theta)} = \frac{\left(\frac{d}{d\theta}\left(e^{-\theta}\right)\right)^2}{I_1(\theta)} = \frac{e^{-2\theta}}{I_1(\theta)} = \theta e^{-2\theta}$$

$Then, show\ that\ e^{-\theta}\left(1 - e^{-\theta}\right) > \theta e^{-2\theta}$

$\Rightarrow equivalently\ to\ show\ that\ e^{\theta} - 1 > \theta$

$One\ way\ to\ show\ it:$

$$e^{\theta} = \sum_{k=0}^{\infty} \frac{\theta^x}{x!} = 1 + \theta + \sum_{k=2}^{\infty} \frac{\theta^x}{x!} > 1 + \theta$$

$\Rightarrow e^{\theta} - 1 > \theta$

$The\ other\ way\ to\ show\ it:$

$Let\ g(\theta) = e^{\theta} - 1 - \theta \Rightarrow g'(\theta) = e^{\theta} - 1, as\ \theta \geq 0 \Rightarrow g'(\theta) > 1 - 1 = 0$

$\Rightarrow g(\theta) \geq g(0) = e^{0} - 1 - 0 = 0 \Rightarrow e^{\theta} - 1 - \theta > 0$

$\Rightarrow e^{\theta} - 1 > \theta$

Example 6.2: $X\ is\ \mathcal{P}(\theta). Let\ \psi(\theta) = e^{-2\theta}, find\ the\ UMVUE\ of\ \psi(\theta), and\ show\ that\ it\ is\ bad.$

Solution:

$Since\ the\ Possion\ r.v.X\ is\ complete\ sufficient, so\ we\ look\ for\ h(X), s.t. E[h(X)] = e^{-2\theta}$

$$E[h(X)] = \sum_{x=0}^{\infty} h(x)P(X = x) = e^{-2\theta}$$

$$\Rightarrow \sum_{x=0}^{\infty} h(x) \cdot \frac{e^{-\theta}\theta^x}{x!} = e^{-2\theta} \Rightarrow \sum_{x=0}^{\infty} \frac{h(x)\theta^x}{x!} = e^{-\theta}$$

$$Since\ e^{-\theta} = \sum_{x=0}^{\infty} \frac{(-\theta)^x}{x!} = \sum_{x=0}^{\infty} \frac{(-1)^x\theta^x}{x!}$$

$$\Rightarrow \sum_{x=0}^{\infty} \frac{(-1)^x\theta^x}{x!} = \sum_{x=0}^{\infty} \frac{h(x)\theta^x}{x!}$$

$\Rightarrow h(X) = (-1)^X$

$By\ Lehmann - Scheffe, h(X) = (-1)^X\ is\ UMVUE\ of\ e^{-2\theta}$

$If\ X = 2k + 1, h(X) = -1;$

$\quad X = 2k, h(X) = 1$

$However, e^{-2\theta}\ is\ always\ strictly\ larger\ than\ 0, indicating\ that\ (-1)^X\ is\ a\ bad\ estimator$

<u>Theorem of C.R.L.:</u>

If the regularity conditions are satisfied and $E(T) = \psi(\theta)$, then:

$$Var\big(T(\boldsymbol{X})\big) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)}(CRL) \Leftrightarrow \frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}) = k(\theta)[T(\boldsymbol{X}) - \psi(\theta)], where\ k(\theta) = \frac{I(\theta)}{\psi'(\theta)}$$

Proof in direction of ⟸:

$$Suppose\ \frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}) = k(\theta)[T(\boldsymbol{X}) - \psi(\theta)]$$

$$(i)\ E\left[\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right] = k(\theta)\big[E\big(T(\boldsymbol{X})\big) - \psi(\theta)\big]$$

$$\Rightarrow 0 = k(\theta)\big[E\big(T(\boldsymbol{X})\big) - \psi(\theta)\big] \Rightarrow E\big(T(\boldsymbol{X})\big) = \psi(\theta)$$

$$(ii)\ Want\ to\ show\ Var\big(T(\boldsymbol{X})\big) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)}$$

$$\Rightarrow \left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right)^2 = k^2(\theta)[T(\boldsymbol{X}) - \psi(\theta)]^2$$

$$\Longrightarrow E\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right)^2\right] = k^2(\theta)E[T(\boldsymbol{X}) - \psi(\theta)]^2$$

$$\Longrightarrow I(\theta) = k^2(\theta)E[T(\boldsymbol{X}) - \psi(\theta)]^2 = k^2(\theta)Var\big(T(\boldsymbol{X})\big)$$

$$\Longrightarrow I(\theta) = \left(\frac{I(\theta)}{\psi'(\theta)}\right)^2 Var\big(T(\boldsymbol{X})\big)$$

$$\Longrightarrow Var\big(T(\boldsymbol{X})\big) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)}$$

Proof in direction of ⟹:

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right] = \int\ldots\int\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\,d\boldsymbol{x} = \frac{\partial}{\partial\theta}\int\ldots\int\log f_\theta(\boldsymbol{x})\,d\boldsymbol{x} = \frac{\partial}{\partial\theta}1 = 0$$

$$\big(\psi'(\theta)\big)^2 = \left(\frac{\partial}{\partial\theta}\psi(\theta)\right)^2 = \left(\frac{\partial}{\partial\theta}E(T)\right)^2 = \left(\frac{\partial}{\partial\theta}\int\ldots\int T(\boldsymbol{x})\,f_\theta(\boldsymbol{x})d\boldsymbol{x}\right)^2$$

$$= \left(\int\ldots\int\frac{\partial}{\partial\theta}T(\boldsymbol{x})\,f_\theta(\boldsymbol{x})\,d\boldsymbol{x}\right)^2\ (parameter\ space\ has\ nothing\ to\ do\ with\ sample\ space)$$

$$= \left(\int\ldots\int T(\boldsymbol{x})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\cdot f_\theta(\boldsymbol{x})\,d\boldsymbol{x}\right)^2$$

$$= \left[E\left(T(\boldsymbol{x})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)\right]^2$$

$$Cov\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}), T(\boldsymbol{X})\right) = E\left[T(\boldsymbol{X})\cdot\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right] - E\big(T(\boldsymbol{X})\big)\,E\left[\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right]$$

$$= E\left[T(\boldsymbol{X})\cdot\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right]$$

$$\Longrightarrow \big(\psi'(\theta)\big)^2 = \left[E\left(T(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right)\right]^2 = \left[Cov\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}), T(\boldsymbol{X})\right)\right]^2$$

$$Given\ that\ Var(T) = \frac{\big(\psi'(\theta)\big)^2}{I(\theta)}$$

$$\Longrightarrow \big(\psi'(\theta)\big)^2 = Var(T)\cdot I(\theta) = Var(T)Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)$$

$$\Longrightarrow \left[Cov\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}), T(\boldsymbol{X})\right)\right]^2 = Var(T)Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{x})\right)$$

$$\Longrightarrow \frac{\left[Cov\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}), T(\boldsymbol{X})\right)\right]^2}{Var(T)Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right)} = 1 \Longrightarrow \left|\frac{Cov\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}), T(\boldsymbol{X})\right)}{\sqrt{Var(T)Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right)}}\right| = 1$$

$$\Longrightarrow \frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}) = aT(\boldsymbol{X}) + b\ (*)$$

$$\Rightarrow \begin{cases} E\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right) = E(aT(\boldsymbol{X})+b) \\ Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right) = Var(aT(\boldsymbol{X})+b) \end{cases} \Rightarrow \begin{cases} E\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right) = aE(T)+b \\ Var\left(\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X})\right) = a^2Var(T) \end{cases}$$

$Given\ that\ \ E(T)=\psi(\theta)\ and\ Var(T)=\frac{\left(\psi\,'(\theta)\right)^2}{I(\theta)}, (a\ and\ \psi'(\theta)\ have\ same\ sign)$

$$\Rightarrow \begin{cases} aE(T)+b=0 \\ a^2Var(T)=I(\theta) \end{cases} \Rightarrow \begin{cases} a\psi(\theta)+b=0 \\ a^2\frac{\left(\psi'(\theta)\right)^2}{I(\theta)}=I(\theta) \end{cases} \Rightarrow \begin{cases} a=\sqrt{\frac{\left(I(\theta)\right)^2}{\left(\psi'(\theta)\right)^2}}=\frac{I(\theta)}{\psi'(\theta)} \\ b=-\frac{I(\theta)}{\psi'(\theta)}\psi(\theta) \end{cases}$$

$Plug\ a\ and\ b\ in\ (*),$

$$\frac{\partial}{\partial\theta}\log f_\theta(\boldsymbol{X}) = \frac{I(\theta)}{\psi'(\theta)}T(\boldsymbol{X}) - \frac{I(\theta)}{\psi'(\theta)}\varphi(\theta) = \frac{I(\theta)}{\psi'(\theta)}[T(\boldsymbol{X})-\psi(\theta)]$$

$$= k(\theta)[T(\boldsymbol{X})-\psi(\theta)], k(\theta)=\frac{I(\theta)}{\psi'(\theta)}$$

Notice: $Var\big(T(\boldsymbol{X})\big) = \frac{\left(\psi'(\theta)\right)^2}{I(\theta)} = \frac{\psi'(\theta)}{\frac{I(\theta)}{\psi'(\theta)}} = \frac{\psi'(\theta)}{k(\theta)}$

Example 6.3: $X_1,\dots,X_n\ be\ i.i.d.\mathcal{P}(\theta).\psi(\theta)=\theta, find\ the\ UMVUE\ of\ \psi(\theta)\ by\ Theorem\ of\ CRL$

Solution: $f_\theta(x)=\frac{e^{-\theta}\theta^x}{x!}, x=0,1,2,\dots\dots$

$$\Rightarrow f_\theta(\boldsymbol{x}) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(\log f_\theta(\boldsymbol{x})\right) = \frac{\sum_{i=1}^n x_i}{\theta} - n = \frac{n}{\theta}[\bar{X}-\theta]$$

$According\ to\ the\ Theorem, CRL\ can\ be\ reached\ by\ \bar{X}, and\ further\ by$

$Cramer-Rao\ Theorem, \bar{X}\ is\ an\ UMVUE\ for\ \theta, with\ Var(\bar{X}) = \frac{\psi'(\theta)}{k(\theta)} = \frac{1}{{n}/{\theta}} = \frac{\theta}{n}$

Example 6.4: $X_1,\dots,X_n\ be\ i.i.d.Ber(\theta).\psi(\theta)=\theta, find\ the\ UMVUE\ of\ \psi(\theta)\ by\ Theorem\ of\ CRL$

Solution: $f_\theta(x)=\theta^x(1-\theta)^{1-x}, x=0,1$

$$\Rightarrow f_\theta(\boldsymbol{x}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(\log f_\theta(\boldsymbol{x})\right) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n-\sum_{i=1}^n x_i}{1-\theta} = \frac{n}{\theta(1-\theta)}[\bar{X}-\theta]$$

$According\ to\ the\ Theorem, CRL\ can\ be\ reached\ by\ \bar{X}, and\ further\ by$

$Cramer - Rao\ Theorem, \bar{X}\ is\ an\ UMVUE\ for\ \theta, with\ Var(\bar{X}) = \dfrac{\psi'(\theta)}{k(\theta)} = \dfrac{1}{\dfrac{n}{\theta(1-\theta)}} = \dfrac{\theta(1-\theta)}{n}$

Example 6.5: $X_1, \ldots, X_n\ be\ i.i.d.\ Exp(\theta).\ \psi(\theta) = \dfrac{1}{\theta}, find\ the\ UMVUE\ of\ \psi(\theta)\ by\ Theorem\ of\ CRL$

Solution: $f_\theta(x) = \theta^x(1-\theta)^{1-x}, x = 0,1$

$$\Rightarrow f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(\log f_\theta(\boldsymbol{x})\right) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n-\sum_{i=1}^n x_i}{1-\theta} = \frac{n}{\theta(1-\theta)}[\bar{X}-\theta]$$

$According\ to\ the\ Theorem, CRL\ can\ be\ reached\ by\ \bar{X}, and\ further\ by$

$Cramer - Rao\ Theorem, \bar{X}\ is\ an\ UMVUE\ for\ \theta, with\ Var(\bar{X}) = \dfrac{\psi'(\theta)}{k(\theta)} = \dfrac{1}{\dfrac{n}{\theta(1-\theta)}} = \dfrac{\theta(1-\theta)}{n}$

Example 6.6: $X_1, \ldots, X_n\ be\ i.i.d.\ N(\theta, 1).\ \psi(\theta) = \theta, find\ the\ UMVUE\ of\ \psi(\theta)\ by\ Theorem\ of\ CRL$

Solution: $f_\theta(x) = (2\pi)^{-\frac{1}{2}}e^{-\frac{(x-\theta)^2}{2}}$

$$\Rightarrow f_\theta(\boldsymbol{x}) = \prod_{i=1}^{n}(2\pi)^{-\frac{1}{2}}e^{-\frac{(x_i-\theta)^2}{2}} = (2\pi)^{-\frac{1}{2}n}e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2}}$$

$$\Rightarrow \log f_\theta(\boldsymbol{x}) = -\frac{1}{2}n\log(2\pi) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\left(\log f_\theta(\boldsymbol{x})\right) = \frac{\partial}{\partial\theta}\left[-\frac{1}{2}n\log(2\pi) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}\right] = -\frac{1}{2}\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}(x_i-\theta)^2\right]$$

$Since\ 1,2, \ldots., n\ does\ not\ depend\ on\ the\ parameter\ space, we\ have:$

$$-\frac{1}{2}\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}(x_i-\theta)^2\right] = -\frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}(x_i-\theta)^2 = -\frac{1}{2}\sum_{i=1}^{n}2(x_i-\theta)\cdot(-1) = \sum_{i=1}^{n}(x_i-\theta)$$

$$\Rightarrow \frac{\partial}{\partial\theta}(\log f_\theta(\boldsymbol{x})) = \sum_{i=1}^{n}(x_i-\theta) = \sum_{i=1}^{n}x_i - n\theta = n\left[\frac{\sum_{i=1}^n x_i}{n} - \theta\right] = n(\bar{X}-\theta)$$

$According\ to\ the\ Theorem, CRL\ can\ be\ reached\ by\ \bar{X}, and\ further\ by$

$Cramer - Rao\ Theorem, \bar{X}\ is\ an\ UMVUE\ for\ \theta, with\ Var(\bar{X}) = \dfrac{\psi'(\theta)}{k(\theta)} = \dfrac{1}{n}$

Lecture 33

Cramer-Rao Lower bound and Exponential Family (1-parameter)

Theorem A).:

If $\{\boldsymbol{f_{\theta}(x)}, \boldsymbol{\theta} \in \boldsymbol{\Omega}\}$ satisfies the regularity conditions and if $\boldsymbol{T(x)}$ is an unbiased estimator of $\boldsymbol{\psi(\theta)}$, with $\boldsymbol{Var(T)} = {\left(\boldsymbol{\psi'(\theta)}\right)^{\mathbf{2}}}\big/{\boldsymbol{I(\theta)}}$. Then $\boldsymbol{f_{\theta}(x)} = \boldsymbol{exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A(x)}$, where A doesn't depend on $\boldsymbol{\theta}$

Proof: *According to the theorem of C.R.L., we have*:

$$\Rightarrow \frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x}) = k(\theta)[T(\boldsymbol{x}) - \psi(\theta)]\,, k(\theta) = \frac{I(\theta)}{\psi'(\theta)}$$

$$\Rightarrow \int \frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x})\,d\theta = \int \frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x})\,d\theta + 0 = \int k(\theta)[T(\boldsymbol{x}) - \psi(\theta)]d\theta + S(\boldsymbol{x})$$

$$= \int k(\theta)T(\boldsymbol{x})d\theta - \int k(\theta)\psi(\theta)d\theta + S(\boldsymbol{x})$$

$$= T(\boldsymbol{x})\int k(\theta)d\theta - \int k(\theta)\psi(\theta)d\theta + S(\boldsymbol{x})$$

$$Let\ c(\theta) = \int k(\theta)d\theta\,, d(\theta) = -\int k(\theta)\psi(\theta)d\theta$$

$$\Rightarrow \log f_{\theta}(\boldsymbol{x}) = T(\boldsymbol{x})c(\theta) + d(\theta) + S(\boldsymbol{x})$$

$$\Rightarrow f_{\theta}(\boldsymbol{x}) = \exp\{T(\boldsymbol{x})c(\theta) + d(\theta) + S(\boldsymbol{x})\}$$

$$If\ \boldsymbol{x} \in A, then\ I_A(\boldsymbol{x}) = 1$$

$$\Rightarrow f_{\theta}(\boldsymbol{x}) = \exp\{T(\boldsymbol{x})c(\theta) + d(\theta) + S(\boldsymbol{x})\} \cdot 1 = \exp\{T(\boldsymbol{x})c(\theta) + d(\theta) + S(\boldsymbol{x})\} \cdot I_A(\boldsymbol{x})$$

Theorem B). :

If $f_{\theta}(x) = \boldsymbol{exp\{c(\theta)T}(x) + \boldsymbol{S}(x) + \boldsymbol{d(\theta)\} \cdot I_A}(x)$ and if $c'(\theta) \neq 0$, and $E[T(\boldsymbol{x})] = \psi(\theta)$.

Then $Var(T) = {\left(\boldsymbol{\psi'(\theta)}\right)^{\mathbf{2}}}\big/{\boldsymbol{I(\theta)}}$

Proof: $\log f_{\theta}(\boldsymbol{x}) = \boldsymbol{c(\theta)T}(x) + \boldsymbol{S}(x) + \boldsymbol{d(\theta)} + \log \boldsymbol{I_A}(x)$

$$\Rightarrow \frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x}) = \frac{\partial}{\partial\theta}\{\boldsymbol{c(\theta)T}(x) + \boldsymbol{S}(x) + \boldsymbol{d(\theta)} + \log \boldsymbol{I_A}(x)\} = c'(\theta)T(\boldsymbol{x}) + d'(\theta)$$

$$Since\ E\left[\frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x})\right] = 0 \Rightarrow E[c'(\theta)T(\boldsymbol{x}) + d'(\theta)] = c'(\theta)E[T(\boldsymbol{x})] + d'(\theta) = 0$$

$$\Rightarrow \psi(\theta) = E[T(\boldsymbol{x})] = -\frac{d'(\theta)}{c'(\theta)} \Rightarrow d'(\theta) = -c'(\theta)\psi(\theta)$$

$$\Rightarrow \frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x}) = c'(\theta)T(\boldsymbol{x}) - \ c'(\theta)\psi(\theta) = c'(\theta)[T(\boldsymbol{x}) - \psi(\theta)]\ (1)$$

$$\Rightarrow \frac{\partial^2}{\partial\theta^2}\log f_{\theta}(\boldsymbol{x}) = \frac{\partial}{\partial\theta}\left[\frac{\partial}{\partial\theta}\log f_{\theta}(\boldsymbol{x})\right] = \frac{\partial}{\partial\theta}\left[c'(\theta)[T(\boldsymbol{x}) - \psi(\theta)]\right]$$

$$= c''(\theta)[T(\boldsymbol{x}) - \psi(\theta)] + \frac{\partial}{\partial\theta}\big(T(\boldsymbol{x}) - \psi(\theta)\big)c'(\theta)$$

$$= c''(\theta)[T(\boldsymbol{x}) - \psi(\theta)] - \psi'(\theta)c'(\theta)$$

$$\Longrightarrow E\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(\boldsymbol{x})\right] = E\{c''(\theta)[T(\boldsymbol{x}) - \psi(\theta)] - \psi'(\theta)c'(\theta)\} = c''(\theta)E[T(\boldsymbol{x}) - \psi(\theta)] - \psi'(\theta)c'(\theta)$$

$$\Longrightarrow -I(\theta) = 0 - \psi'(\theta)c'(\theta) = -\psi'(\theta)c'(\theta)$$

$$\Longrightarrow c'(\theta) = \frac{I(\theta)}{\psi'(\theta)}(2)$$

$By\ the\ C.R.L.Theorem, (1)\ and\ (2)\ gives\ that\ Var(T) = {\left(\boldsymbol{\psi}'(\boldsymbol{\theta})\right)^{\mathbf{2}}}\Big/_{\boldsymbol{I}(\boldsymbol{\theta})}$

Lecture 34

Maximum Likelihood Estimation

$\boldsymbol{\theta}$ is to be estimated by its most likely value $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ (a function of the data), where $\widehat{\boldsymbol{\theta}}$ is a value of $\boldsymbol{\theta}$ which maximizes the likelihood function:

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = \prod_{i=1}^{n} f(\boldsymbol{\theta}, x_i)$$

$\widehat{\boldsymbol{\theta}}$ is the MLE of θ if $L(\boldsymbol{\theta}, \boldsymbol{x}) = \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{x})$

Remark: If $\widehat{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta}, \boldsymbol{x})$, then it also maximizes $\log L(\boldsymbol{\theta}, \boldsymbol{x})$

Likelihood Equation

If $\log L(\boldsymbol{\theta}, \boldsymbol{x})$ is differentiable w.r.t. $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots \theta_k)$, then θ must satisfy the **likelihood equation**:

$$\frac{\partial}{\partial\theta_i}\log L(\boldsymbol{\theta}, \boldsymbol{x}) = 0, i = 1,2, \ldots \ldots k$$

Example 8.1: $X_1, X_2, \ldots, X_n\ are\ i.i.d.Ber(\theta). Find\ the\ MLE\ of\ \theta$

Solution: $f_\theta(x) = \theta^x(1-\theta)^{1-x}$

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}$$

$$\Longrightarrow \frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\sum_{i=1}^{n}x_i}{\theta} - \frac{n - \sum_{i=1}^{n}x_i}{1-\theta} = 0$$

$$\Longrightarrow \hat{\theta} = \frac{\sum_{i=1}^{n}X_i}{n} = \bar{X}, is\ the\ MLE\ of\ \theta$$

Example 8.2: $X_1, X_2, \ldots, X_n\ are\ i.i.d.N(\theta, 1). Find\ the\ MLE\ of\ \theta$

Solution: $f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2}}$$

$$\Rightarrow \frac{\partial}{\partial\theta} \log L(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^{n} 2(x_i - \theta) = \sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta = 0$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}, is\ the\ MLE\ of\ \theta$$

Example 8.3: $X_1, X_2, \dots, X_k$ *are multinomial random variables with parameters* $n, \theta_1, \dots, \theta_k$ $\left(\sum_1^k x_i = n, \sum_1^k \theta_i = 1\right)$. *Find the MLE of* θ_i

Solution:

Recall : n independent trials where each trial results in one of the outcomes $S_1, \dots, S_k$

X_i: *# of times that* S_i *occurs in n trials*, $i = 1, \dots, k$ *and* $\sum_1^k x_i = n$

Suppose that S_i *occurs in any trial is* $\theta_i, i = 1, .., k$ *and* $\sum_1^k \theta_i = 1$

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = P(\theta_1, \dots, \theta_k; x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} {\theta_1}^{x_1} {\theta_2}^{x_2} \dots {\theta_k}^{x_k}$$

$$\Rightarrow \log L(\boldsymbol{\theta}, \boldsymbol{x}) = \log \frac{n!}{x_1! \dots x_k!} + x_1 \log\theta_1 + \dots + x_k \log\theta_k$$

$$= \log \frac{n!}{x_1! \dots x_k!} + x_1 \log\theta_1 + \dots + x_k \log(1 - \theta_1 - \dots - \theta_{k-1})$$

$$\Rightarrow \frac{\partial \log L}{\partial\theta_i} = \frac{x_i}{\theta_i} - \frac{x_k}{1 - \sum_{j=1}^{k-1} \theta_j} = \frac{x_i}{\theta_i} - \frac{x_k}{\theta_k} = 0, i = 1, 2, \dots, k$$

$$i.e.\ \frac{x_1}{\theta_1} = \frac{x_2}{\theta_2} = \dots = \frac{x_k}{\theta_k} = \frac{\sum_1^k x_i}{\sum_1^k \theta_i} = \frac{n}{1} = n$$

$$\Rightarrow \widehat{\theta_i} = \frac{x_i}{n}, is\ the\ MLE\ of\ \theta_i$$

Example 8.4: $X_1, X_2, \dots, X_n$ *are i.i.d.* $N(\mu, \sigma^2)$, *where* μ *and* σ^2 *are both unknown. Find a joint MLE of* $\boldsymbol{\theta} = (\mu, \sigma^2)$

Solution: $f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\Rightarrow \frac{\partial}{\partial\mu} \log L(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \Rightarrow \sum_{i=1}^{n} x_i - n\mu = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

$$\Rightarrow \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\theta}, \boldsymbol{x}) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^4} = 0 \Rightarrow \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

$$\Rightarrow \widehat{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

$$Therefore, \left(\bar{X}, \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}\right) \ is\ the\ joint\ MLF\ for\ (\mu, \sigma^2)$$

<u>Claim (relationship with sufficiency):</u>

In the case of 1-parameter, $\hat{\theta}$ is the MLE of θ if $L(\theta, \boldsymbol{x}) = \max_{\theta} L(\theta, \boldsymbol{x})$. The MLE of θ must be a function of a ***sufficient statistics*** for θ.

Proof: $by\ factorization\ theorem, if\ T(\boldsymbol{x})\ is\ sufficient, then$

$$L(\theta, \boldsymbol{x}) = \ g(T(\boldsymbol{x}), \theta) \cdot h(\boldsymbol{x}), Find\ \hat{\theta}\ that\ maximize\ L(\theta, \boldsymbol{x})$$

$$\Rightarrow \log L(\theta, \boldsymbol{x}) = \log\ g(T(\boldsymbol{x}), \theta) + \log h(\boldsymbol{x})$$

$$\Rightarrow \frac{\partial}{\partial \theta} L(\theta, \boldsymbol{x}) = \frac{\partial}{\partial \theta} \log\ g(T(\boldsymbol{x}), \theta) = \frac{g'(T(\boldsymbol{x}), \theta)}{g(T(\boldsymbol{x}), \theta)} = 0$$

$$\Rightarrow g'(T(\boldsymbol{x}), \theta) = 0 \Rightarrow g(T(\boldsymbol{x}), \theta) = C$$

$$\Rightarrow \hat{\theta}\ must\ be\ a\ function\ of\ T(\boldsymbol{x})$$

<u>Theorem (functional invariance of MLE):</u>

If $\hat{\theta}$ is an MLE of θ, then $h(\hat{\theta})$ is an MLE of $h(\theta)$, where h is bijective within the support

Proof: $Let\ \Omega\ be\ the\ parameter\ space,$

$$Since\ h\ is\ bijective, then\ P[h(x)|x] = P[x|h(x)] = 1$$

$$Given\ that\ \hat{\theta}\ \ is\ an\ MLE\ of\ \theta \Rightarrow \hat{\theta}\ = \arg\max_{\theta} \log \prod_{i=1}^{n} f(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i|\theta)$$

$$\Rightarrow \sum_{i=1}^{n} \log f\left(x_i\middle|\hat{\theta}\right) \geq \sum_{i=1}^{n} \log f(x_i|\theta)\,, \forall \theta \in \Omega$$

$$\Rightarrow \sum_{i=1}^{n} \log f\left(x_i\middle|\hat{\theta}\right) + n \cdot \log 1 \geq \sum_{i=1}^{n} \log f(x_i|\theta) + n \cdot \log 1$$

$$\Rightarrow \sum_{i=1}^{n} \log f\left(x_i\middle|\hat{\theta}\right) + n \cdot \log P\left[\hat{\theta}\middle|h\left(\hat{\theta}\right)\right] \geq \sum_{i=1}^{n} \log f(x_i|\theta) + n \cdot \log P[\theta|h(\theta)]$$

$$\Rightarrow \sum_{i=1}^{n} \left[\log f\left(x_i\middle|\hat{\theta}\right) + \log P\left[\hat{\theta}\middle|h\left(\hat{\theta}\right)\right]\right] \geq \sum_{i=1}^{n} [\log f(x_i|\theta) + \log P[\theta|h(\theta)]]$$

$$\Rightarrow \sum_{i=1}^{n} \log\left[f\left(x_i\middle|\hat{\theta}\right) \cdot P\left[\hat{\theta}\middle|h\left(\hat{\theta}\right)\right]\right] \geq \sum_{i=1}^{n} \log[f(x_i|\theta) \cdot P[\theta|h(\theta)]]$$

$$\Rightarrow \sum_{i=1}^{n} \log f\left(x_i \middle| h(\hat{\theta})\right) \geq \sum_{i=1}^{n} \log f\big(x_i \big| h(\theta)\big) \Rightarrow \log f\left(\boldsymbol{x} \middle| h(\hat{\theta})\right) \geq \log f\big(\boldsymbol{x} \big| h(\theta)\big)$$

$$\Rightarrow f\left(\boldsymbol{x} \middle| h(\hat{\theta})\right) \geq f\big(\boldsymbol{x} \big| h(\theta)\big), \forall \theta \in \Omega$$

$By\ definition, h(\hat{\theta})\ is\ the\ MLE\ of\ h(\theta)$

Example of application: $Let\ X_1, \dots, X_n\ be\ i.i.d.\ Ber(\theta). Find\ the\ MLE\ of\ \theta^2$

Solution: $i). Find\ the\ MLE\ of\ \theta \Longrightarrow\ \bar{X}\ (shown\ in\ example 1)$

$ii). MLE\ of\ \theta^2 \Longrightarrow\ \bar{X}^2\ (by\ theorem)$

<u>Cases when $\log L(\boldsymbol{\theta}, \boldsymbol{x})$ is not differentiable</u>

Example 8.5: $Let\ X_1, \dots, X_n\ be\ i.i.d.\ Unif(0, \theta)\ with\ pdf\ f_\theta(x) = \begin{cases} \frac{1}{\theta}, 0 < x < \theta \\ 0, \quad\quad O.W. \end{cases}.$

$Find\ the\ MLE\ of\ \theta.$

Solution:

$$L(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f_\theta(x_i) \cdot I_{\{0<x_i<\theta, i=1,2,\dots,n\}} = \left(\frac{1}{\theta}\right)^n \cdot I_{\{0<x_i<\theta, i=1,2,\dots,n\}}$$

$$\max_\theta \left(\frac{1}{\theta}\right)^n \cdot I_{\{0<x_i<\theta, i=1,2,\dots,n\}} = \max_\theta \theta^{-n} \cdot I_{\{0<x_i<\theta, i=1,2,\dots,n\}}$$

$Where\ \ \theta^{-n}\ is\ an\ decreasing\ function, which\ is\ maximized$

$at\ the\ lower\ bound\ of\ \theta\ (1)\ and\ indicator\ function$

$I_{\{0<x_i<\theta, i=1,2,\dots,n\}}\ is\ maximized\ at\ 1\ (2)$

$In\ order\ to\ satisfy\ (2)\ , we\ should\ have\ I_{\{0<x_i<\theta, i=1,2,\dots,n\}} = 1$

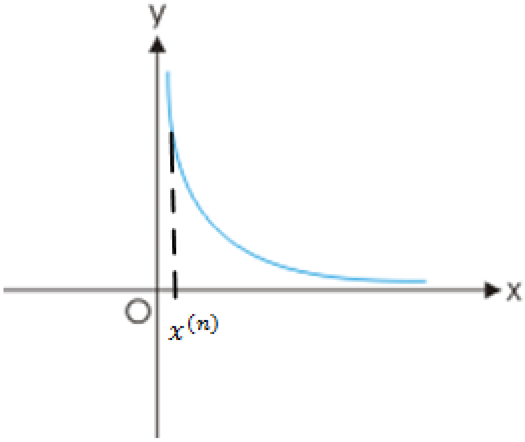
$$\Rightarrow\ I_{\{0<x^{(1)}<\dots<x^{(n)}<\theta\}} = 1 \Rightarrow I_{\{\theta>x^{(n)}\}} = 1 \Rightarrow \theta > x^{(n)}, a.s.$$

$\Rightarrow x^{(n)}\ \ is\ the\ lower\ bound\ of\ \theta, which\ at\ the\ same\ time$

$satisfy\ (1),\ \ as\ shown\ in\ the\ graph$

$\Rightarrow\ \ x^{(n)}\ maximizes\ L(\boldsymbol{x}; \theta)$

$\Rightarrow x^{(n)}\ is\ the\ MLE\ of\ \theta$



Example 8.6: $Let\ X_1, \dots, X_n\ be\ i.i.d.\ Unif(\theta, \theta+2)\ with\ pdf\ f_\theta(x) = \begin{cases} \frac{1}{2}, \theta < x < \theta + 2 \\ \quad 0, \quad\quad O.W. \end{cases}.$

$Find\ the\ MLE\ of\ \theta$

Solution:

$$L(\boldsymbol{x};\theta)=\prod_{i=1}^{n} f_\theta(x_i)\cdot I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}=\left(\frac{1}{2}\right)^n\cdot I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}$$

$$\max_\theta\left(\frac{1}{2}\right)^n\cdot I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}=\left(\frac{1}{2}\right)^n\cdot\max_\theta I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}$$

$Where\ the\ indicator\ function\ I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}\ is\ maximized\ at\ 1\ (1)$

$In\ order\ to\ satisfy\ (1)\ ,we\ should\ have\ I_{\{\theta<x_i<\theta+2,i=1,2,\dots,n\}}=1$

$\Longrightarrow\ I_{\{\theta<x_i<\theta+2\}}=1\Longrightarrow Any\ x_i between\ \left(x^{(n)}-2,x^{(1)}\right)\ is\ the\ MLE\ of\ \theta$

Example 8.7: $Let\ X_1,\dots,X_n\ be\ i.i.d.Unif(\alpha,\beta)\ with\ pdf\ f_\theta(x)=\begin{cases}\dfrac{1}{\alpha-\beta},\alpha<x<\beta\\ 0,\qquad O.W.\end{cases}.$

$Find\ the\ MLE\ of\ \boldsymbol{\theta}=(\alpha,\beta)$

Solution:

$$L(\boldsymbol{x};\boldsymbol{\theta})=\prod_{i=1}^{n} f_{\boldsymbol{\theta}}(x_i)\cdot I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}=\left(\frac{1}{\beta-\alpha}\right)^n\cdot I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}$$

$$\max_{\boldsymbol{\theta}}\left(\frac{1}{\beta-\alpha}\right)^n\cdot I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}=\max_{\boldsymbol{\theta}}(\beta-\alpha)^{-n}\cdot I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}$$

$Where\ \ (\beta-\alpha)^{-n}\ is\ an\ decreasing\ function, which\ maximized\ at\ the\ lower\ bound\ of\ \beta-\alpha\ (1)$

$and\ indicator\ function\ I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}\ is\ maximized\ at\ 1\ (2)$

$In\ order\ to\ satisfy\ (2)\ ,we\ should\ have\ I_{\{\alpha<x_i<\beta,i=1,2,\dots,n\}}=1$

$\Longrightarrow\ I_{\{\alpha<x^{(1)}<\dots<x^{(n)}<\beta\}}=1\Longrightarrow I_{\{\beta>x^{(n)}\}}\cdot I_{\{\alpha<x^{(1)}\}}=1\Longrightarrow\beta>x^{(n)}and\ \alpha<x^{(1)},a.s.$

$\Longrightarrow\beta-\alpha>x^{(n)}-x^{(1)}$

$\Longrightarrow x^{(n)}-x^{(1)}which\ is\ the\ lower\ bound\ of\ \beta-\alpha, which\ satisfying\ (1)$

$\Longrightarrow\widehat{\boldsymbol{\theta}}=\left(x^{(1)},x^{(n)}\right)\ is\ the\ MLE\ of\ \boldsymbol{\theta}=(\alpha,\beta)$

Lecture 35

<u>Theorem (Consistency of MLE):</u>

If $\boldsymbol{T_n}$ is MLE of $\boldsymbol{\theta}$, then $\boldsymbol{T_n}$ is a consistent estimator

Proof: $First, recall\ the\ definition\ of\ consistent\ estimator$

$\Longrightarrow T_n\ is\ a\ consistent\ estimator\ of\ \theta\ if\ T_n\xrightarrow{in\ probability}\theta$

$$\forall\epsilon>0,P\{|T_n(X)-\theta|\geq\epsilon\}\leq\frac{E(T_n(X)-\theta)^2}{\epsilon^2}=\frac{E\left[T_n(X)-E\left(T_n(X)\right)+E\left(T_n(X)\right)-\theta\right]^2}{\epsilon^2}$$

$$= \frac{E\left[T_n(X) - E\big(T_n(X)\big) + E\big(T_n(X)\big) - \theta\right]^2}{\epsilon^2}$$

$$= \frac{E\left[T_n(X) - E\big(T_n(X)\big)\right]^2 + \left[E\big(T_n(X)\big) - E\big(T_n(X)\big)\right]\left[E\big(T_n(X)\big) - \theta\right] + E\left[E\big(T_n(X)\big) - \theta\right]^2}{\epsilon^2}$$

$$= \frac{Var\big(T_n(X)\big) + \left[E\big(T_n(X)\big) - \theta\right]^2}{\epsilon^2}$$

Central Limit Theorem

$$X_1, X_2, \dots, X_n \ are\ i.i.d., E(X) = \mu, Var(X) = \sigma^2 < \infty,$$

$$Then\ \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{In\ distribution} N(0,1);\ or\ \overline{X_n} \xrightarrow{In\ distribution} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Theorem (Asymptotical Normal of MLE):

(i) Under regularity conditions and if T_n is MLE of θ, then T_n is asymptotically normal with mean θ and variance ${}^{1}\!/_{nI_1(\theta)}$

(ii) Under regularity conditions and if T_n is MLE of $\psi(\theta)$, then T_n is asymptotically normal with mean $\psi(\theta)$ and variance ${}^{\left(\psi'(\theta)\right)^2}\!\Big/_{nI_1(\theta)}$

Proof of (i):

$Let\ T_n\ be\ the\ MLE, and\ \theta_0\ be\ the\ real\ parameter, and\ \boldsymbol{X}\ is\ a\ observed\ sample\ of\ size\ n$

$X_i\ are\ i.i.d. with\ pdf\ f(X|\theta)$

$$Let\ l(\theta) = \log f(\boldsymbol{X}|\theta) \Longrightarrow l'(\theta) = \frac{\partial}{\partial\theta} l(\theta)$$

$By\ first - order\ Taylor\ expansion\ about\ \theta_0\ (\theta_0\ is\ singular)$

$l'(\theta) \approx l'(\theta) + (\theta - \theta_0) l''(\theta_0)$

$$Since\ T_n \in \Theta, the\ parameter\ space \Longrightarrow l'(T_n) = l'(\theta_0) + (T_n - \theta_0) l''(\theta_0) + o(n), o(n) \xrightarrow{n\to\infty} 0$$

$$By\ the\ definition\ of\ MLE \Longrightarrow 0 = l'(T_n) = l'(\theta_0) + (T_n - \theta_0) l''(\theta_0) \Longrightarrow T_n - \theta_0 = -\frac{l'(\theta_0)}{l''(\theta_0)}$$

$$\Longrightarrow \sqrt{n}(T_n - \theta_0) = -\sqrt{n}\frac{l'(\theta_0)}{l''(\theta_0)} = -\frac{l'(\theta_0)/\sqrt{n}}{l''(\theta_0)/n} = -\frac{l_1(\theta_0)}{l_2(\theta_0)}, l_1(\theta_0) = l'(\theta_0)/\sqrt{n}, l_2(\theta_0) = -l''(\theta_0)/n$$

$$(1)\ l_1(\theta_0) = \frac{l'(\theta_0)}{\sqrt{n}} = \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\,|_{\theta=\theta_0}$$

$$(2)\ E[l_1(\theta_0)] = E\left[\frac{l'(\theta_0)}{\sqrt{n}}\right] = -\frac{1}{\sqrt{n}}E[l'(\theta_0)] = \frac{1}{\sqrt{n}}E\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\right]|_{\theta=\theta_0}$$

$$= \frac{1}{\sqrt{n}} E\left[\frac{\partial}{\partial\theta}\prod_{i=1}^{n} \log f(X_i|\theta)\right]|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} E\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n} \log f(X_i|\theta)\right]|_{\theta=\theta_0}$$

$$= \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right]|_{\theta=\theta_0}(regularity\ condition)$$

$$= \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} E\left(\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right)\right]|_{\theta=\theta_0}(independence)$$

$$= \sqrt{n}E\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)|_{\theta=\theta_0} = 0\ (i.i.d.) < \infty$$

$$(3)\ Var\big(l_1(\theta_0)\big) = Var\left[\frac{l'(\theta_0)}{\sqrt{n}}\right] = \frac{1}{n}Var\left[\frac{\partial}{\partial\theta}\prod_{i=1}^{n} \log f(X_i|\theta)\right]|_{\theta=\theta_0}$$

$$= \frac{1}{n}Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n} \log f(X_i|\theta)\right]|_{\theta=\theta_0} = \frac{1}{n}\left[\sum_{i=1}^{n} Var\left(\frac{\partial}{\partial\theta} \log f(X_i|\theta)\right)\right]|_{\theta=\theta_0}$$

$$= Var\left(\frac{\partial}{\partial\theta} \log f(X|\theta)\right)|_{\theta=\theta_0} = I_1(\theta_0) < \infty$$

$$(4) l_2(\theta_0) = -l''(\theta_0)/n = -\frac{1}{n}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right]|_{\theta=\theta_0} = -\frac{1}{n}\left[\frac{\partial^2}{\partial\theta^2}\prod_{i=1}^{n}\log f(X_i|\theta)\right]|_{\theta=\theta_0}$$

$$= -\frac{1}{n}\left[\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right]|_{\theta=\theta_0}$$

$$(5) E\left[-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right]|_{\theta=\theta_0} = \frac{1}{n}E\left[-\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\,|_{\theta=\theta_0}\right] = I_1(\theta_0)\ \text{(by definition)}$$

$$(6) Var\left[-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X}|\theta)\right]|_{\theta=\theta_0} = \frac{1}{n^2}\ Var\left[\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\,|_{\theta=\theta_0}\right] = \frac{S(\theta_0)}{n} < \infty$$

$Apply\ CLT\ to\ l_1(\theta_0)\ and\ l_2(\theta_0), by\ (1)-(6)$

$$l_1(\theta_0) \Rightarrow \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\,|_{\theta=\theta_0} \sim N\big(0, I_1(\theta_0)\big)$$

$$l_2(\theta_0) \Rightarrow\ -\frac{1}{n}\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}|\theta)\,|_{\theta=\theta_0} \sim N\left(I_1(\theta_0), \frac{S(\theta_0)}{n}\right)$$

$By\ Chebyshev's\ Inequality,$

$$\Rightarrow \forall\epsilon > 0, P\left\{\left|\frac{1}{n}\left[-\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right]|_{\theta=\theta_0} - \big(I_1(\theta_0)\big)\right| > \epsilon\right\} \leq \frac{S(\theta_0)}{n\epsilon} \to 0, as\ n \to \infty$$

$$\Rightarrow l_2(\theta_0) = \frac{1}{n}\left[\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right]|_{\theta=\theta_0} \xrightarrow{in\ probability} I_1(\theta_0) \Rightarrow \frac{1}{l_2(\theta_0)} \xrightarrow{in\ probability} \frac{1}{I_1(\theta_0)}$$

$$By\ Slustky\ Theorem, as\ l_1(\theta_0) \overset{\mathcal{D}}{\Rightarrow} N\big(0, I_1(\theta_0)\big), \frac{1}{l_2(\theta_0)} \overset{\mathcal{P}}{\rightarrow} \frac{1}{I_1(\theta_0)}$$

$$\Longrightarrow l_1(\theta_0) \cdot \frac{1}{l_2(\theta_0)} \overset{\mathcal{D}}{\Rightarrow} \frac{1}{I_1(\theta_0)} N\big(0, I_1(\theta_0)\big) = N\left(0, \frac{1}{I_1(\theta_0)}\right)$$

$$\Longrightarrow \sqrt{n}(T_n - \theta_0) \overset{\mathcal{D}}{\Rightarrow} N\left(0, \frac{1}{I_1(\theta_0)}\right) \Longrightarrow T_n \sim N\left(\theta_0, \frac{1}{nI_1(\theta_0)}\right)$$

Proof of (ii): $If\ T_n\ is\ the\ MLE\ of\ \theta, T_n'\ denotes\ the\ MLE\ of\ \psi(\theta), and\ \psi(\theta)\ is\ differentiable,$ $with\ \psi(\theta_0)\ be\ real\ parameter$

$\Longrightarrow By\ functional\ invariance\ of\ MLE, so\ that\ T_n' = \psi(T_n)$

$By\ the\ result\ in\ (i), T_n\ is\ the\ MLE\ of\ \theta, and\ as\ n \to \infty, T_n \sim N\left(\theta_0, \frac{1}{nI_1(\theta_0)}\right)$

$By\ Taylor\ expansion\ about\ \theta_0, T_n \in \Theta$

$$\psi(\theta) \approx \psi(\theta_0) + \psi'(\theta_0)(\theta - \theta_0)$$

$$\Longrightarrow \psi(T_n) \approx \psi(\theta_0) + \psi'(\theta_0)(T_n - \theta_0) = \psi'(\theta_0)T_n + [\psi(\theta_0) - \theta_0\psi'(\theta_0)]\ (*)$$

$As\ \psi'(\theta_0), \psi(\theta_0), and\ \theta_0\psi'(\theta_0)\ are\ constant, so\ that (*)\ is\ a\ linear\ tranformation\ of\ T_n$

$$\Longrightarrow E[\psi(T_n)] = E[\psi(\theta_0) + \psi'(\theta_0)(T_n - \theta_0)] = \psi(\theta_0);$$

$$Var[\psi(T_n)] = Var[\psi(\theta_0) + \psi'(\theta_0)(T_n - \theta_0)] = \psi'(\theta_0)^2 Var(T_n) = \frac{\psi'(\theta_0)^2}{nI_1(\theta_0)}$$

$$\Longrightarrow T_n' = \psi(T_n) \sim N\left(\psi(\theta_0), \frac{\psi'(\theta_0)^2}{nI_1(\theta_0)}\right)$$

Implications:

a) Regardless of distribution of T_n, the MLE will have a ***normal distribution***, as $n \to \infty$
b) When n is large, the MLE is asymptotically ***unbiased***
c) When n is large, the MLE is an ***UMVUE***, which will achieve ***CRL***

Proposition (functional invariance):

If g is continuous and differentiable and if T_n is a MLE of θ, then:

i). $g(T_n)$ is an MLE of $g(\theta)$

Proof: $Shown\ in\ the\ theorem\ of\ 'functional\ invariance\ of\ MLE'\ in\ Lecture\ 8$

ii). $g(T_n)$ has asymptotically normal distribution $N\left(g(\theta), {\big(g'(\theta)\big)^2}\big/{nI_1(\theta)}\right)$

Proof: $Same\ as\ the\ proof\ of\ (ii)\ of\ Theorem\ of\ asymtotical\ normality\ of\ MLE$

Uniformly Minimal Variance Unbiased Estimators for more than 1 parameter

Example 9.1: $Let\ X_1, \dots, X_n\ i.i.d.\ N(\mu, \sigma^2). Let\ \boldsymbol{\theta} = (\mu, \sigma^2). Find\ the\ UMVUE\ of\ \boldsymbol{\theta}$

$i). Using\ Lehmann - Scheffe$

$ii). Using\ Cramer - Rao\ Lower\ Bound$

Solution:

$$(i). \widetilde{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} X_i\,, \sum_{i=1}^{n} {X_i}^2\right)\ \ has\ been\ shown\ that\ it\ is\ complete\ sufficient$$

$By\ Lehmann - Scheffe\ Theorem,$

$$If\ \begin{cases} E\left[h_1\left(\widetilde{\boldsymbol{\theta}}\right)\right] = \mu \\ E\left[h_2\left(\widetilde{\boldsymbol{\theta}}\right)\right] = \sigma^2 \end{cases} \Rightarrow h_1\left(\widetilde{\boldsymbol{\theta}}\right)\ is\ the\ UMVUE\ of\ \mu, and\ h_2\left(\widetilde{\boldsymbol{\theta}}\right)\ is\ the\ UMVUE\ of\ \sigma^2$$

$$\Rightarrow Since\ \mu = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \Rightarrow h_1\left(\sum_{i=1}^{n} X_i\,, \sum_{i=1}^{n} {X_i}^2\right) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$Also\ \sigma^2 = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] \Rightarrow h_2\left(\sum_{i=1}^{n} X_i\,, \sum_{i=1}^{n} {X_i}^2\right) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left({X_i}^2 - 2\bar{X}X_i + \bar{X}^2\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n} {X_i}^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2\right] = \frac{1}{n-1}\sum_{i=1}^{n} {X_i}^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} X_i\right)^2$$

$$Therefore, \widehat{\boldsymbol{\theta}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\,, \frac{1}{n-1}\sum_{i=1}^{n} {X_i}^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} X_i\right)^2\right)\ is\ the\ UMVUE\ for\ \boldsymbol{\theta}$$

$$ii). \log f\,(\boldsymbol{X}|\theta) = \log\prod_{i=1}^{n} f(X_i|\theta) = \log\left[(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

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

$$\Rightarrow \frac{\partial}{\partial\mu}\log f\,(\boldsymbol{X}|\theta) = \sum_{i=1}^{n}\frac{(x_i - \mu)}{\sigma^2};\ \frac{\partial}{\partial\sigma^2}\log f\,(\boldsymbol{X}|\theta) = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^4}$$

$$\psi(\boldsymbol{\theta}) = \boldsymbol{\theta} = (\mu, \sigma^2) \Rightarrow \psi'(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial\mu}\mu & \frac{\partial}{\partial\sigma^2}\mu \\ \frac{\partial}{\partial\mu}\sigma^2 & \frac{\partial}{\partial\sigma^2}\sigma^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$I(\boldsymbol{\theta}) = E\begin{bmatrix} -\frac{\partial^2}{\partial\mu^2}\log f\,(\boldsymbol{X}|\theta) & -\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f\,(\boldsymbol{X}|\theta) \\ -\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f\,(\boldsymbol{X}|\theta) & -\frac{\partial^2}{\partial(\sigma^2)^2}\log f\,(\boldsymbol{X}|\theta) \end{bmatrix}$$

$$= E\begin{bmatrix} -\frac{\partial}{\partial\mu}\sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^2} & -\frac{\partial}{\partial\sigma^2}\sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^2} \\ -\frac{\partial}{\partial\mu}\left[-\frac{n}{2\sigma^2}+\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^4}\right] & -\frac{\partial}{\partial\sigma^2}\left[-\frac{n}{2\sigma^2}+\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^4}\right] \end{bmatrix}$$

$$= E\begin{bmatrix} \frac{n}{\sigma^2} & \sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^4} \\ \sum_{i=1}^{n}\frac{(X_i-\mu)}{\sigma^4} & -\frac{n}{2\sigma^4}+\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^6} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4}+\frac{n}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

$$k(\boldsymbol{\theta}) = \psi^{-1}(\boldsymbol{\theta})I(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

$$\Rightarrow I^{-1}(\boldsymbol{\theta})\begin{bmatrix} \sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2}+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}\begin{bmatrix} \sum_{i=1}^{n}\frac{(x_i-\mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2}+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^4} \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n}\frac{(x_i-\mu)}{n} \\ -\sigma^2+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} \end{bmatrix}, if\ exists\ UMVUE\ that\ reaching\ CRL, then\ iff$$

$$\Rightarrow g\big(\boldsymbol{T}(\boldsymbol{X})\big) - \boldsymbol{\theta} = \begin{bmatrix} \sum_{i=1}^{n}\frac{(x_i-\mu)}{n} \\ -\sigma^2+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} \end{bmatrix} \Rightarrow g\big(\boldsymbol{T}(\boldsymbol{X})\big) = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n}\frac{(x_i-\mu)}{n} \\ -\sigma^2+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} \end{bmatrix}$$

$$Where\ \mu+\sum_{i=1}^{n}\frac{(X_i-\mu)}{n} = \mu+\frac{\sum_{i=1}^{n}X_i}{n}-\mu = \frac{\sum_{i=1}^{n}X_i}{n} = \bar{X};$$

$$\sigma^2-\sigma^2+\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n}$$

$$\Rightarrow g(\boldsymbol{T}) = \begin{bmatrix} \frac{\sum_{i=1}^{n}X_i}{n} \\ \sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n} \end{bmatrix}$$

Lecture 36

Bayesian vs. Classical

Example:

$$\text{Coin flipping}\begin{cases} \boldsymbol{Classical\,View}: \boldsymbol{observed\ heads}\big/\boldsymbol{Total\ flips} = \boldsymbol{P}(\boldsymbol{H}) \\ \boldsymbol{Bayesian\,View}: \boldsymbol{use\ prior\ information\ before\ flipping, then\ determine\ P}(\boldsymbol{H}) \end{cases}$$

Loss Function

Let $\widehat{g(\theta)}$ be estimator of $g(\theta)$, one of the loss function is squared error loss (S.E.L.):

$$L\left(\widehat{g(\theta)}, g(\theta)\right) = \left[\widehat{g(\theta)} - g(\theta)\right]^2 \qquad (10.1)$$

Other loss functions:

i) $\omega(\theta)\left[\widehat{g(\theta)} - g(\theta)\right]^2$, where $\omega(\theta)$ is the **cost function**

ii) $\left|\widehat{g(\theta)} - g(\theta)\right|$

……

In general, $L\left(\widehat{g(\theta)}, g(\theta)\right)$ is a convex function of the difference between $\widehat{g(\theta)}$ and $g(\theta)$

Bayesian Estimation ("Bayes")

$\mathbf{Example\ 10.1}$: $Let\ X_1, \dots, X_n\ be\ i.i.d. Ber(\theta). The\ pdf\ is\ expressed\ differently\ by\ Bayesian$
$\Longrightarrow f(x|\theta) = \theta^x(1-\theta)^{1-x}, x = 0,1\ (compared\ with\ Classical\ f(x) = \theta^x(1-\theta)^{1-x}, x = 0,1)$
$where\ \theta\ is\ treated\ as\ random\ variable\ but\ fixed$

$\mathbf{Step\ I}$: $Choose\ the\ right\ \boldsymbol{prior\ information}\ \theta: \theta \sim \beta(a,b)$

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)}, 0 < \theta < 1$$

$\mathbf{Step\ II}$: $Posterior\ density\ (Updated\ prior)\ of\ \theta\ given\ X_1, \dots, X_n$

$$\pi[\theta|X_1, X_2, \dots, X_n] = \frac{f(X_1, X_2, \dots, X_n|\theta)\pi(\theta)}{m(X_1, X_2, \dots, X_n)} = \frac{f(X_1, X_2, \dots, X_n|\theta)\pi(\theta)}{\int_0^1 f(X_1, X_2, \dots, X_n|\theta)\pi(\theta)d\theta}\ (Bayes'Law)$$

$$\Longrightarrow \underbrace{\pi[\theta|X_1, X_2, \dots, X_n]}_{\text{Posterior}} \propto \underbrace{f(X_1, X_2, \dots, X_n|\theta) \cdot \pi(\theta)}_{\text{Likelihood} \times \text{Prior}}$$

Remark: Bayesian statisticians consider the classical view as when prior information is uniform

$\mathbf{Step\ III}$: $f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$

$$\Rightarrow \pi[\theta|X_1,X_2,\ldots,X_n] = \frac{\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}\cdot\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)\cdot m(X_1,X_2,\ldots,X_n)}$$
$$= \frac{\theta^{a+\sum_{i=1}^{n}x_i-1}(1-\theta)^{n+b+\sum_{i=1}^{n}x_i-1}}{\beta(a,b)\cdot m(X_1,X_2,\ldots,X_n)}$$

Claim: $\theta|X_1,\ldots,X_n \sim \beta\left(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i\right)$

Proof: $m(X_1,X_2,\ldots,X_n) = \int_0^1 f(X_1,X_2,\ldots,X_n|\theta)\pi(\theta)d\theta = \int_0^1 \frac{\theta^{a+\sum_{i=1}^{n}x_i-1}(1-\theta)^{n+b-\sum_{i=1}^{n}x_i-1}}{\beta(a,b)}d\theta$

$$= \frac{\beta(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i)}{\beta(a,b)}\int_0^1 \frac{\theta^{a+\sum_{i=1}^{n}x_i-1}(1-\theta)^{n+b-\sum_{i=1}^{n}x_i-1}}{\beta(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i)}d\theta$$
$$= \frac{\beta(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i)}{\beta(a,b)}$$

$$\Rightarrow \pi[\theta|X_1,X_2,\ldots,X_n] = \frac{\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}\cdot\theta^{a-1}(1-\theta)^{b-1}}{\beta(a,b)\cdot\frac{\beta\left(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i\right)}{\beta(a,b)}} \sim \beta\left(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i\right)$$

$Hence, the\ posterior\ density\ of\ \theta\ given\ X_1,\ldots,X_n, is\ \beta\left(a+\sum_{i=1}^{n}x_i\,, n+b-\sum_{i=1}^{n}x_i\right).$

Conjugate Family

$\pi(\theta)$ is a conjugate prior for f if $\pi(\theta)$ and the posterior density $\pi(\theta|X_1,\ldots,X_n)$ are from the same family

i) The Beta distribution is a conjugate family for Bernoulli distribution
ii) The Gamma distribution is a conjugate family for Poisson distribution
iii) The Gamma distribution is a conjugate family for Exponential distribution
iv) The Normal distribution is a conjugate family for Normal distribution

Example 10.2: $Let\ X_1,\ldots,X_n\ be\ i.i.d. Poisson\ r.v. and\ \theta \sim \gamma(a,p), find\ the\ posterior\ density\ of\ \theta$

Solution: $f(X|\theta) = \frac{e^{-\theta}\theta^x}{x!}, x = 0,1,2,\ldots; \pi(\theta) = \frac{p^a e^{-p\theta}\theta^{a-1}}{\Gamma(a)}$

$$\pi[\theta|X_1,X_2,\ldots,X_n] = \frac{f(X_1,X_2,\ldots,X_n|\theta)\pi(\theta)}{m(X_1,X_2,\ldots,X_n)} = \frac{\prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}\cdot\frac{p^a e^{-p\theta}\theta^{a-1}}{\Gamma(a)}}{m(X_1,X_2,\ldots,X_n)}$$
$$\propto \prod_{i=1}^{n}\frac{e^{-\theta}\theta^{x_i}}{x_i!}\cdot\frac{p^a e^{-p\theta}\theta^{a-1}}{\Gamma(a)} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}\cdot\frac{p^a e^{-p\theta}\theta^{a-1}}{\Gamma(a)} = \frac{p^a e^{-(n+p)\theta}\theta^{a+\sum_{i=1}^{n}x_i-1}}{\prod_{i=1}^{n}x_i!\cdot\Gamma(a)}$$

$$\propto \ (n+p)^{a+\sum_{i=1}^{n} x_i} e^{-(n+p)\theta} \theta^{a+\sum_{i=1}^{n} x_i - 1} \sim \gamma\left(a + \sum_{i=1}^{n} x_i \, , n+p\right)$$

Example 10.3: $Let\ X_1, \dots, X_n\ be\ i.i.d.\ N(\theta, 1)\ r.v.\ and\ \theta \sim N(\mu, \sigma^2), find\ the\ posterior\ density\ of\ \theta$

Solution: $\pi(\theta) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}; f(X_1, \dots, X_n|\theta) = (2\pi)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2}}$

$$\pi[\theta|X_1, \dots, X_n] = (2\pi)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} = (2\pi)^{-\frac{n+1}{2}} (\sigma^2)^{-\frac{1}{2}} e^{\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{1}{2\sigma^2}(\theta-\mu)^2\right]}$$

$$\propto \ e^{\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 - \frac{1}{2\sigma^2}(\theta-\mu)^2\right]} = e^{-\frac{1}{2}\sum_{i=1}^{n} x_i{}^2 + \theta \sum_{i=1}^{n} x_i - \frac{1}{2} n\theta^2 - \frac{1}{2\sigma^2}\theta^2 + \frac{\mu}{\sigma^2}\theta - \frac{\mu^2}{2\sigma^2}}$$

$$\propto \ e^{-\frac{1}{2\sigma^2}\left[(n\sigma^2+1)\theta^2 - 2(\sigma^2 \sum_{i=1}^{n} x_i + \mu)\theta\right]} \propto \ e^{-\frac{(n\sigma^2+1)}{2\sigma^2}\left[\theta^2 - 2\frac{(\sigma^2 \sum_{i=1}^{n} x_i + \mu)}{(n\sigma^2+1)}\theta\right]}$$

$$\propto \exp\left\{-\frac{\left(\theta - \frac{(\sigma^2 \sum_{i=1}^{n} x_i + \mu)}{(n\sigma^2+1)}\right)^2}{2\sigma^2 / (n\sigma^2+1)}\right\} \sim N\left(\frac{(\sigma^2 \sum_{i=1}^{n} x_i + \mu)}{(n\sigma^2+1)}, 2\sigma^2 / (n\sigma^2+1)\right)$$

Lecture 37

Risk Function

Let $X_1, \dots, X_n$ be i.i.d.,$l(\theta, T(\boldsymbol{X}))$ be loss function, and $R(\theta, T)$ be **risk function**
In Bayesian: $f(\boldsymbol{x}|\theta) = \theta^{\boldsymbol{x}}(1-\theta)^{1-\boldsymbol{x}}$ ("For a given value of θ, X is Bernoulli random variable")
By giving $\theta \sim \pi(\theta)$ as the prior information, the risk function $R(\theta, T)$ is defined as:

$$R(\theta, T) = E^{X|\theta}\left[l(\theta, T(\boldsymbol{X}))\right] = \int \dots \int l(\theta, T(\boldsymbol{x})) \cdot f(x_1, \dots, x_n|\theta) dx_1 \dots dx_n$$

$$= \int \dots \int l(\theta, T(\boldsymbol{x})) \cdot \prod_{i=1}^{n} f(x_i|\theta)\, dx_1 \dots dx_n$$

Note: $R(\theta, T_1), R(\theta, T_2)$ are random variables, so they are not comparable.

Bayes Risk

Bayes risk is defined as $r(\pi, T)$:

$$r(\pi, T) = E^{\pi}[R(\theta, T)] = E^{\pi}\left[E^{X|\theta}\left[l(\theta, T(\boldsymbol{X}))\right]\right] \quad (11.1)$$

$$= \int_{\Omega} \left[\int_{\mathbb{R}^n} l(\theta, T(\boldsymbol{x})) \cdot f(\boldsymbol{x}|\theta) d\boldsymbol{x}\right] \pi(\theta) d\theta \quad (11.2)$$

i) With respect to **joint density** $J(\theta, \boldsymbol{X}) = f(\theta, \boldsymbol{X}) = f(\boldsymbol{X}|\theta)\pi(\theta)$, the Bayes risk can be expressed as:

$$r(\pi,T) = E^J\left[l\big(\theta,T(\boldsymbol{X})\big)\right] \tag{11.3}$$

Proof: $r(\pi,T) = \int\limits_{\Omega}\left[\int\limits_{\mathbb{R}^n} l\big(\theta,T(\boldsymbol{x})\big)\cdot f(\boldsymbol{x}|\theta)d\boldsymbol{x}\right]\pi(\theta)d\theta = \int\limits_{\Omega}\int\limits_{\mathbb{R}^n} l\big(\theta,T(\boldsymbol{x})\big)\cdot[f(\boldsymbol{x}|\theta)\pi(\theta)]d\boldsymbol{x}d\theta$

$$= \int\limits_{\Omega}\int\limits_{\mathbb{R}^n} l\big(\theta,T(\boldsymbol{x})\big)\cdot J(\theta,\boldsymbol{X})d\boldsymbol{x}d\theta = E^J\left[l\big(\theta,T(\boldsymbol{X})\big)\right]$$

ii) With respect to **marginal density** $m(\boldsymbol{X})$ **and posterior density** $\pi[\theta|\boldsymbol{X}]$, the Bayes risk can be expressed as:

$$r(\pi,T) = E^m\left[E^{\theta|\boldsymbol{X}}\left(l\big(\theta,T(\boldsymbol{X})\big)\right)\right] \tag{11.4}$$

Proof: $\pi[\theta|\boldsymbol{X}] = \dfrac{J(\theta,\boldsymbol{X})}{m(\boldsymbol{X})} \Longrightarrow J(\theta,\boldsymbol{X}) = \pi[\theta|\boldsymbol{x}]\cdot m(\boldsymbol{x})$

$$r(\pi,T) = \int\limits_{\Omega}\int\limits_{\mathbb{R}^n} l\big(\theta,T(\boldsymbol{x})\big)\cdot J(\theta,\boldsymbol{X})d\boldsymbol{x}d\theta = \int\limits_{\mathbb{R}^n}\int\limits_{\Omega} l\big(\theta,T(\boldsymbol{x})\big)\cdot\pi[\theta|\boldsymbol{x}]\cdot m(\boldsymbol{x})\,d\theta d\boldsymbol{x}$$

$$= \int\limits_{\mathbb{R}^n}\left[\int\limits_{\Omega} l\big(\theta,T(\boldsymbol{x})\big)\cdot\pi[\theta|\boldsymbol{x}]d\theta\right]m(\boldsymbol{x})d\boldsymbol{x} = E^m\left[E^{\theta|\boldsymbol{X}}\left(l\big(\theta,T(\boldsymbol{X})\big)\right)\right]$$

Bayes Estimator

A Bayes estimator T^* is the estimator that **minimizes** $r(\pi,T)$:

$$r(\pi,T^*) = \min_T r(\pi,T)$$

Posterior Expected Loss (P.E.L.): $E^{\theta|\boldsymbol{X}}\left(l\big(\theta,T(\boldsymbol{X})\big)\right)$

Squared Error Loss (S.E.L.): $(T(\boldsymbol{X})-\theta)^2$

Let the loss function be the S.E.L., by (3) the Bayes risk becomes:

$$r(\pi,T) = E^m\left[E^{\theta|\boldsymbol{X}}(T(\boldsymbol{X})-\theta)^2\right] = E^m\left[\int\limits_{\Omega}(T(\boldsymbol{x})-\theta)^2\cdot\pi(\theta|\boldsymbol{x})d\theta\right]$$

Claim: T^* that minimizes $E^m\left[\int_\Omega(T(\boldsymbol{x})-\theta)^2\cdot\pi(\theta|\boldsymbol{x})d\theta\right]$ is **the same** T^* that minimizes $\int_\Omega(T(\boldsymbol{x})-\theta)^2\cdot\pi(\theta|\boldsymbol{x})d\theta$

Note: Bayes estimator is the estimator that minimizes **the posterior expected loss**

Theorem (Bayes Estimator for Square Error Loss function):

Loss function is S.E.L., then the Bayes estimator is the $T^*(\boldsymbol{X}) = E[\theta|\boldsymbol{X}]$

Proof: $\dfrac{\partial}{\partial T}\left[\int\limits_{\Omega}(T(\boldsymbol{x})-\theta)^2\cdot\pi(\theta|\boldsymbol{x})d\theta\right] = 0$

$$\Rightarrow \int_{\Omega} \frac{\partial}{\partial T}[(T(\boldsymbol{x})-\theta)^2 \cdot \pi(\theta|\boldsymbol{x})]d\theta = 0 \Rightarrow \int_{\Omega} 2(T(\boldsymbol{x})-\theta)\cdot\pi(\theta|\boldsymbol{x})d\theta = 0$$

$$\Rightarrow \int_{\Omega} T(\boldsymbol{x})\cdot\pi(\theta|\boldsymbol{x})d\theta = \int_{\Omega} \theta\cdot\pi(\theta|\boldsymbol{x})d\theta \Rightarrow T^*(\boldsymbol{x})\int_{\Omega}\pi(\theta|\boldsymbol{x})d\theta = E[\theta|\boldsymbol{X}]$$

$$\Rightarrow T^*(\boldsymbol{X})\cdot 1 = E[\theta|\boldsymbol{X}] \Rightarrow\ T^*(\boldsymbol{X}) = E[\theta|\boldsymbol{X}]$$

Then, the **minimized Posterior Error Loss** and **minimized Bayes Risk** are:

$$r(\pi,T^*) = E^m\left[E^{\theta|\boldsymbol{X}}(T(\boldsymbol{X})-\theta)^2\right] = E^m\left[E^{\theta|X}(E^2(\theta|X)-2\theta E(\theta|X)+\theta^2)\right]$$

$$= E^m[E^2(\theta|X)-2E^2(\theta|X)+E(\theta^2|X)] = E^m[E(\theta^2|X)-E^2(\theta|X)] = E^m[Var(\theta|\boldsymbol{X})]$$

Example 11.1: $Given\ \theta, X_1, \dots, X_n\ be\ i.i.d. Ber(\theta), and\ l(\theta,T) = (T-\theta)^2, \theta \sim \beta(a_0,b_0)$

$Question$: 1). $Find\ the\ Bayes\ estimator\ of\ \theta$

2). $Find\ the\ minimal\ Bayes\ risk$

Solution:

1). $T^* = E[\theta|\boldsymbol{X}]$

$In\ lecture 9, it\ has\ been\ shown\ that\ the\ posterior\ density\ is\ \theta|\boldsymbol{X} \sim \beta(a_n,b_n),$

$$where\ a_n = a_0 + \sum_{i=1}^{n} x_i\ , b_n = b_0 + n - \sum_{i=1}^{n} x_i$$

$$\Rightarrow T^* = E[\theta|\boldsymbol{X}] = \frac{a_n}{a_n+b_n} = \frac{(a_0+\sum_{i=1}^n X_i)}{(a_0+\sum_{i=1}^n X_i)+(b_0+n-\sum_{i=1}^n X_i)} = \frac{a_0+\sum_{i=1}^n X_i}{a_0+b_0+n}$$

2). $minimal\ Bayes\ risk\ \ r(\pi,T^*) = E^m[Var(\theta|\boldsymbol{X})], given\ that\ \theta|\boldsymbol{X} \sim \beta(a_n,b_n)$

$$\Rightarrow r(\pi,T^*) = E^m\left[\frac{a_nb_n}{(a_n+b_n)^2(a_n+b_n+1)}\right] = E^m\left[\frac{(a_0+\sum_{i=1}^n X_i)(b_0+n-\sum_{i=1}^n X_i)}{(a_0+b_0+n)^2(a_0+b_0+n+1)}\right]$$

$$= \frac{1}{(a_0+b_0+n)^2(a_0+b_0+n+1)}E^m\left[\left(a_0+\sum_{i=1}^n X_i\right)\left(b_0+n-\sum_{i=1}^n X_i\right)\right]$$

$$= \frac{1}{(a_0+b_0+n)^2(a_0+b_0+n+1)}E^m\left[a_0b_0+b_0\sum_{i=1}^n X_i+na_0+n\sum_{i=1}^n X_i-a_0\sum_{i=1}^n X_i-\left(\sum_{i=1}^n X_i\right)^2\right]$$

$The\ expectation\ part\ becomes$:

$$\Rightarrow a_0b_0+na_0+(b_0-a_0+n)E^m\left[\sum_{i=1}^n X_i\right]-E^m\left[\left(\sum_{i=1}^n X_i\right)^2\right]\ (*), where$$

$$E^m\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E^m(X_i) = nE^m(X) = nE^\theta[E(X|\theta)] = nE(\theta) = \frac{na_0}{a_0+b_0}$$

$$Var\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n Var(X_i) = nVar(X) = nE^m\left[\left(X-E(X)\right)^2\right] = nE^\theta[E(X^2-2XE(X)+E^2(X)|\theta)]$$

$$= nE[E(X^2|\theta-2XE(X)|\theta+E^2(X)|\theta)] = nE[E(X^2|\theta)-2E(X)E(X|\theta)+E^2(X)]$$

$$
\begin{aligned}
&= nE^{\theta}[\theta(1-\theta)+\theta^2-2\theta E(X)+E^2(X)] = nE^{\theta}[\theta-2\theta E(X)+E^2(X)] \\
&= nE^{\theta}\left[\left(\theta-E(X)\right)^2\right] = nE^{\theta}\left[\left(E[X|\theta]-E(E[X|\theta])\right)^2\right] \\
&= nE^{\theta}[Var(X|\theta)] = nE^{\theta}[\theta(1-\theta)] \\
&= n\int\frac{\theta(1-\theta)\theta^{a_0-1}(1-\theta)^{b_0-1}}{B(a_0,b_0)}d\theta = n\frac{B(a_0+1,b_0+1)}{B(a_0,b_0)}\int\frac{\theta^{a_0}(1-\theta)^{b_0}}{B(a_0+1,b_0+1)}d\theta \\
&= n\frac{B(a_0+1,b_0+1)}{B(a_0,b_0)} = n\frac{\Gamma(a_0+1)\Gamma(b_0+1)\Big/\Gamma(a_0+b_0+2)}{\Gamma(a_0)\Gamma(b_0)\Big/\Gamma(a_0+b_0)} = n\frac{a_0b_0}{(a_0+b_0+1)(a_0+b_0)}
\end{aligned}
$$

$$
E^m\left[\left(\sum_{i=1}^n X_i\right)^2\right] = Var\left[\sum_{i=1}^n X_i\right] + \left(E\left(\sum_{i=1}^n X_i\right)\right)^2 = n\frac{a_0b_0}{(a_0+b_0+1)(a_0+b_0)} + \left(\frac{na_0}{a_0+b_0}\right)^2
$$

$Therefore, (*)\ becomes$:

$$
\begin{aligned}
&a_0b_0+na_0+(b_0-a_0+n)E^m\left[\sum_{i=1}^n X_i\right]-E^m\left[\left(\sum_{i=1}^n X_i\right)^2\right] \\
&= a_0b_0+na_0+\frac{na_0(b_0-a_0+n)}{a_0+b_0}-n\frac{a_0b_0}{(a_0+b_0+1)(a_0+b_0)}-\left(\frac{na_0}{a_0+b_0}\right)^2 \\
&= \frac{na_0b_0}{a_0+b_0}-\frac{na_0b_0}{(a_0+b_0)}\cdot\frac{1}{(a_0+b_0+1)}+\frac{na_0(n-a_0)}{a_0+b_0}-\left(\frac{na_0}{a_0+b_0}\right)^2+a_0b_0+na_0 \\
&= \frac{na_0b_0}{a_0+b_0}\left[1-\frac{1}{(a_0+b_0+1)}\right]+\frac{na_0}{a_0+b_0}\left[(n-a_0)-\frac{na_0}{a_0+b_0}\right]+a_0b_0+na_0 \\
&= \frac{na_0b_0}{a_0+b_0+1}-\frac{na_0({a_0}^2-nb_0+a_0b_0)-na_0(a_0+b_0)^2-a_0b_0(a_0+b_0)^2}{(a_0+b_0)^2} \\
&= \frac{na_0b_0}{a_0+b_0+1}+\frac{na_0b_0(a_0+b_0+n)+a_0b_0(a_0+b_0)^2}{(a_0+b_0)^2} \\
&= na_0b_0\left(\frac{(a_0+b_0)^2\frac{(a_0+b_0+n+1)}{n}+(a_0+b_0+n)(a_0+b_0+1)}{(a_0+b_0+1)(a_0+b_0)^2}\right) \\
&= \frac{a_0b_0(a_0+b_0+n+1)}{(a_0+b_0+1)}+\frac{na_0b_0(a_0+b_0+n)}{(a_0+b_0)^2}
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow\ r(\pi,T^*) &= \frac{1}{(a_0+b_0+n)^2(a_0+b_0+n+1)}\left[\frac{a_0b_0(a_0+b_0+n+1)}{(a_0+b_0+1)}+\frac{na_0b_0(a_0+b_0+n)}{(a_0+b_0)^2}\right] \\
&= \frac{a_0b_0}{(a_0+b_0+n)^2(a_0+b_0+1)}+\frac{na_0b_0}{(a_0+b_0)^2(a_0+b_0+n)(a_0+b_0+n+1)} \\
&= \frac{a_0b_0}{(a_0+b_0+n)}\left[\frac{1}{(a_0+b_0+n)(a_0+b_0+1)}+\frac{n}{(a_0+b_0)^2(a_0+b_0+n+1)}\right]
\end{aligned}
$$

Remark:

$i).\ E[\theta|\boldsymbol{X}] = \dfrac{a_0+\sum_{i=1}^n X_i}{a_0+b_0+n} = \dfrac{a_0+n\bar{X}}{a_0+b_0+n};$

$If\ n\to\infty, Bayes\ estimator \to \bar{X}, which\ is\ UMVUE; Or\ if\ a_0, b_0 \to 0, Bayes\ estimator \to \bar{X}$

$ii). In\ particular, when\ a_0 = b_0 = 1 \Longrightarrow \theta \sim Unif(0,1) \Longrightarrow \pi(\theta) = \begin{cases} 1, 0 < \theta < 1 \\ 0, \quad\quad O.W. \end{cases}$

$$In\ this\ case, T^* = \frac{1+n\bar{X}}{2+n}, where\ \boldsymbol{n}\ makes\ a\ difference\ \left(E.g. If\ n = 5 \Longrightarrow T^* = \frac{1+5\bar{X}}{7} \neq \bar{X}\right)$$

Lecture 38

Generalized Squared Error Loss

$$l(\theta, T(\boldsymbol{X})) = \omega(\theta)[T(\boldsymbol{X}) - \theta]^2$$

T^* is a Bayes Estimator if it minimizes Posterior Error Loss:

$$PEL = \int_\Omega l(\theta, T(\boldsymbol{X})) \cdot \pi(\theta|\boldsymbol{X}) d\theta = \int_\Omega \omega(\theta)[T(\boldsymbol{X}) - \theta]^2 \cdot \pi(\theta|\boldsymbol{X}) d\theta$$

$$\Longrightarrow \frac{\partial}{\partial T} \int_\Omega \omega(\theta)[T(\boldsymbol{X}) - \theta]^2 \cdot \pi(\theta|\boldsymbol{X}) d\theta = 0, \forall\theta$$

$$\Longrightarrow \int_\Omega \frac{\partial}{\partial T} \omega(\theta)[T(\boldsymbol{X}) - \theta]^2 \cdot \pi(\theta|\boldsymbol{X}) d\theta = \int_\Omega 2\omega(\theta)[T(\boldsymbol{X}) - \theta] \cdot \pi(\theta|\boldsymbol{X}) d\theta = 0$$

$$\Longrightarrow T(\boldsymbol{X}) \int_\Omega \omega(\theta)\pi(\theta|\boldsymbol{X}) d\theta = \int_\Omega \omega(\theta)\theta\pi(\theta|\boldsymbol{X}) d\theta \Longrightarrow T(\boldsymbol{X}) \cdot E(\theta|\boldsymbol{X}) = E(\theta\omega(\theta)|\boldsymbol{X})$$

$$\Longrightarrow T^*(\boldsymbol{X}) = \frac{E(\theta\omega(\theta)|\boldsymbol{X})}{E(\omega(\theta)|\boldsymbol{X})}\ (Bayes\ Estimator\ under\ the\ generalized\ squared\ error\ loss)$$

In particular, when $\omega(\theta) = 1 \Longrightarrow T^*(\boldsymbol{X}) = E(\theta|\boldsymbol{X})$

Case: $\boldsymbol{\omega(\theta) = \theta^{-1}}$

$In\ this\ case, the\ loss\ function\ becomes:$

$$l(\theta, T) = \frac{(T-\theta)^2}{\theta}$$

$Bayes\ Estimator\ becomes:$

$$T^*(\boldsymbol{X}) = \frac{E(\theta\omega(\theta)|\boldsymbol{X})}{E(\omega(\theta)|\boldsymbol{X})} = \frac{E(\theta\theta^{-1}|\boldsymbol{X})}{E(\theta^{-1}|\boldsymbol{X})} = \frac{1}{E\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)}$$

$Minimal\ Bayes\ risk\ becomes:$

$$r(\pi, T^*) = E^J(l(\theta, T^*)) = E^J\left(\frac{(T^*-\theta)^2}{\theta}\right) = E^J\left(\frac{1}{\theta}\left(T^{*2} - 2T^*\theta + \theta^2\right)\right)$$

$$= E^J(\theta) - 2E^J(T^*) + E^J\left(\frac{T^{*2}}{\theta}\right)$$

$$E^J\left(\frac{T^{*2}}{\theta}\right) = E^J\left(\frac{1}{\theta E^2\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)}\right) = E\left[E\left(\frac{1}{\theta E^2\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)}\middle|\boldsymbol{X}\right)\right] \ (Smoothing\ Theorem)$$

$$= E\left[\frac{1}{E^2\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)}\cdot E\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)\right] = E^J\left(\frac{1}{E\left(\frac{1}{\theta}\middle|\boldsymbol{X}\right)}\right) = E^J(T^*)$$

$$\Longrightarrow r(\pi, T^*) = E^J(\theta) - 2E^J(T^*) + E^J(T^*) = E^J(\theta) - E^J(T^*)$$

Example 12.1: *Given that $\theta, X_1, \dots, X_n$ are i.i.d. $Exp(\theta), \pi(\theta) \sim \gamma(a,p)$. Find the Bayes estimator and minimal Bayes risk of $\frac{1}{\theta}$ for the loss function $l(\theta, T) = \frac{(T-\theta)^2}{\theta}$*

Solution:

$$l\left(\frac{1}{\theta}, T\right) = \frac{\left(T - {}^1\!/_\theta\right)^2}{{}^1\!/_\theta} = \theta\left(T - \frac{1}{\theta}\right)^2$$

$$\Longrightarrow \frac{\partial}{\partial T}\int_\Omega \theta\left[T(\boldsymbol{X}) - \frac{1}{\theta}\right]^2 \cdot \pi(\theta|\boldsymbol{X})d\theta = 0, \forall\theta$$

$$\Longrightarrow \int_\Omega \frac{\partial}{\partial T}\theta\left[T(\boldsymbol{X}) - \frac{1}{\theta}\right]^2 \cdot \pi(\theta|\boldsymbol{X})d\theta = \int_\Omega 2\theta\left[T(\boldsymbol{X}) - \frac{1}{\theta}\right]\cdot\pi(\theta|\boldsymbol{X})d\theta = 0$$

$$\Longrightarrow T(\boldsymbol{X})\int_\Omega \theta\pi(\theta|\boldsymbol{X})d\theta = \int_\Omega \theta\cdot\frac{1}{\theta}\pi(\theta|\boldsymbol{X})d\theta \Longrightarrow T(\boldsymbol{X})\cdot E(\theta|\boldsymbol{X}) = E(1|\boldsymbol{X}) = 1$$

$$\Longrightarrow T^*(\boldsymbol{X}) = \frac{1}{E(\theta|\boldsymbol{X})}$$

$$r(\pi, T^*) = E^J\left(l\left(\frac{1}{\theta}, T^*\right)\right) = E^J\left(\theta\left(T^* - \frac{1}{\theta}\right)^2\right) = E^J\left(\theta T^{*2}\right) - 2E^J(T^*) + E^J\left(\frac{1}{\theta}\right)$$

$$E^J\left(\theta T^{*2}\right) = E^J\left(\frac{\theta}{E^2(\theta|\boldsymbol{X})}\right) = E\left[E\left(\frac{\theta}{E^2(\theta|\boldsymbol{X})}\middle|\boldsymbol{X}\right)\right] = E\left[\frac{1}{E^2(\theta|\boldsymbol{X})}\cdot E(\theta|\boldsymbol{X})\right] = E^J\left(\frac{1}{E(\theta|\boldsymbol{X})}\right) = E^J(T^*)$$

$$\Longrightarrow r(\pi, T^*) = E^J\left(\frac{1}{\theta}\right) - E^J(T^*)$$

$$Since\ X_i\ has\ pdf\ f(x|\theta) = \begin{cases}\theta e^{-\theta x}, x > 0\\ 0, \qquad O.W.\end{cases}$$

$$\pi(\theta|\boldsymbol{X}) \propto f(\boldsymbol{x}|\theta)\cdot\pi(\theta) = \theta^n e^{-\theta\sum_{i=1}^n x_i}\cdot e^{-p\theta}\theta^{a-1} = e^{-(\sum_{i=1}^n x_i + p)\theta}\theta^{n+a-1}$$

$$\Longrightarrow \theta|\boldsymbol{X} \sim \gamma(a_n, p_n)\,, a_n = a+n, p_n = p + \sum_{i=1}^n X_i \Longrightarrow E(\theta|\boldsymbol{X}) = \frac{a_n}{p_n} = \frac{a+n}{p + \sum_{i=1}^n X_i}$$

$$\Longrightarrow\ T^*(\boldsymbol{X}) = \frac{1}{E(\theta|\boldsymbol{X})} = \frac{p + \sum_{i=1}^n X_i}{a+n}$$

$$
\begin{aligned}
r(\pi,T^*) &= E^J\left(\frac{1}{\theta}\right) - E^J(T^*) = E^m\left[E\left(\frac{1}{\theta}\middle|X\right)\right] - E^m[E(T^*)] \\
&= E^m\left(\int \frac{1}{\theta}\cdot\frac{{p_n}^{a_n}e^{-p_n\theta}\theta^{a_n-1}}{\Gamma(a_n)}\,d\theta\right) - E^m\left(\frac{p+\sum_{i=1}^n X_i}{a+n}\right) \\
&= E^m\left[\frac{p_n\Gamma(a_n-1)}{\Gamma(a_n)}\right] - E^m\left(\frac{p+\sum_{i=1}^n X_i}{a+n}\right) = E^m\left[\frac{p_n}{a_n-1}\right] - E^m\left(\frac{p+\sum_{i=1}^n X_i}{a+n}\right) \\
&= E^m\left[\frac{p+\sum_{i=1}^n X_i}{a+n-1}\right] - E^m\left(\frac{p+\sum_{i=1}^n X_i}{a+n}\right) = \frac{p+nE^m(X)}{a+n-1} - \frac{p+nE^m(X)}{a+n} \\
&= \frac{p+nE^\theta[E(X|\theta)]}{a+n-1} - \frac{p+nE^\theta[E(X|\theta)]}{a+n} = \frac{p+nE^\theta\left[\frac{1}{\theta}\right]}{a+n-1} - \frac{p+nE^\theta\left[\frac{1}{\theta}\right]}{a+n} \\
&= \frac{p+nE^\theta\left[\frac{1}{\theta}\right]}{a+n-1} - \frac{p+nE^\theta\left[\frac{1}{\theta}\right]}{a+n} = \frac{p+nE^\theta\left[\frac{1}{\theta}\right]}{(a+n-1)(a+n)} \\
&= \frac{p+n\frac{p\Gamma(a-1)}{\Gamma(a)}\int\frac{p^{a-1}e^{-p\theta}\theta^{a-2}}{\Gamma(a-1)}\,d\theta}{(a+n-1)(a+n)} = \frac{p+n\frac{p}{a-1}}{(a+n-1)(a+n)} \\
&= \frac{a+n-1}{(a+n-1)(a+n)(a-1)}p = \frac{p}{(a+n)(a-1)}
\end{aligned}
$$

Lecture 39

Hypothesis Testing

Problem: Suppose $\boldsymbol{X} = (X_1, \dots, X_n)$ has a p.d.f. belonging to $\mathcal{P} = \{p(x,\theta), \theta \in \Omega\}$

Where Ω denotes the parameter space, for example:

$X \sim Ber(\theta) \Longrightarrow \Omega = [0,1]$

$X \sim N(\mu,\theta) \Longrightarrow \Omega = [0,\infty)$

$X \sim N(\theta_1,\theta_2) \Longrightarrow \Omega = \{\mathbb{R}, [0,\infty)\}$

We wish to test that:

Null hypothesis $H_0\colon \theta \in \Omega_0$; against the

Alternative hypothesis $H_a\colon \theta \in \Omega - \Omega_0 = \overline{\Omega_0}$

In order to test H_0 $vs.$ H_a, we are given an observed value of $\boldsymbol{X}$

Initiative approach: Assume that H_0 is correct (true) and see if the observed $\boldsymbol{X}$ agrees or disagrees with H_0. If it disagrees, then we reject H_0.

More precisely, we split the sample space S of all values of $\boldsymbol{X}$ into 2 region $C, S - C$ such that:

a) C contains those values of X which “deviate most” from H_0. C is called “critical region”, or “rejection region” .
b) The probability that X lies in C when H_0 being true is small

Example 13.1: Let X denotes the grades, which are normally distributed. $X \sim N(\mu, \sigma^2)$, μ is unknown and $\sigma = 7, n = 7$. Want to test $H_0: \mu = 70 \;\; vs. \;\; H_a > 70$.

Take a random sample: $X_1, \dots, X_n$. $\bar{X}$ has been shown to be a good estimator of μ.

If $\bar{X} = 80$, statisticians may state: "maybe μ is actually larger then 70". In order to verify this statement, some cut-off points x_c will be needed, such that $\begin{cases} If\ \bar{X} > x_c \Longrightarrow reject\ H_0 \\ If\ \bar{X} \le x_c \Longrightarrow Do\ not\ reject\ H_0 \end{cases}$

In order to determine x_c,

$$P[reject\ H_0, when\ H_0\ is\ true] = P[\bar{X} > x_c, when\ \mu = 70] = \alpha$$

We want to α as small as possible, so fix it at 0.05

$Under\ H_0: \mu = 70, \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N(70,7)$

$\Longrightarrow P(\bar{X} > x_c) = P\left(\frac{\bar{X} - 70}{\sqrt{7}} > \frac{x_c - 70}{\sqrt{7}}\right) = 0.05 \Longrightarrow P\left(Z > \frac{x_c - 70}{\sqrt{7}}\right) = 0.05$

$\Longrightarrow \frac{x_c - 70}{\sqrt{7}} = 1.64 \Longrightarrow x_c \approx 74.34$

The sample space is split into $C = (74.34, +\infty), and\ \bar{C} = (-\infty, 74.34]$

<u>Two Types of Error:</u>

Type I error: Reject H_0, when H_0 is true

$i).\ P(Type\ I\ error) = P_\theta(\boldsymbol{X} \in C), \theta \in \Omega_0$

Type II error: Accept H_0, when H_0 is false

$$ii).\ P(Type\ II\ error) = 1 - P_\theta(reject\ H_0\ when\ H_0\ is\ false)$$
$$= 1 - P_\theta(X \in C), \theta \in \overline{\Omega_0}$$

We should find a test whose error probabilities as small as possible—cannot construct a test where simultaneously minimizes the error probabilities

The classical approach to constructing optimum is to set an **upper bound** for the **type I error probability**, and to select a test which **minimizes the type II error probability**.

Begin by specifying α $(0 \le \alpha \le 1, usually\ \alpha = 0.05, 0.01, 0.001)$. And consider only tests for which:

$$P_\theta(\boldsymbol{X} \in C) \le \alpha, \forall \theta \in \Omega_0$$

α is called the level of significance

The **size** of a test is defined as:

$$\sup_{\theta \in \Omega_0} P_\theta(\boldsymbol{X} \in C) \tag{13.1}$$

Restrict our search to tests whose size does not exceed α

<u>Power of a test</u>

The **power of the test** is defined as:

$$P_\theta(X \in C), \theta \in \overline{\Omega_0} = 1 - P(Type\ II\ error) \tag{13.2}$$

Considered as a continuous function of θ for all $\theta \in \Omega$, the power is called the power function:

$$\beta(\theta) \overset{\text{def}}{=} P_\theta(\boldsymbol{X} \in C, \theta \in \Omega) \tag{13.3}$$

Lecture 40

The Most Powerful Test

Define $\beta(\theta) = P_\theta[reject\ H_0]$, we look for tests satisfying:

i) $\sup\limits_{\theta \in \Omega_0} \beta(\theta) \le \alpha$

ii) Maximize $\beta(\theta), \theta \in \Omega_1$

Example 14. 1: $X_1, \dots, X_n\ be\ i.i.d. N(\theta, 1). Want\ to\ test\ \mathrm{H}_0\!: \theta = 0\ \text{vs.}\ \mathrm{H}_1\!: \theta \neq 0$

$Set\ the\ reject\ region\ at\ C = \left\{X\!: |\bar{X}| \ge \frac{1.96}{\sqrt{n}}\right\}$

$By\ satifying\ condition\ i).:$

$\sup\limits_{\theta \in \Omega_0} \beta(\theta) = \beta(0) = P[reject\ H_0, when\ H_0\ is\ true] = P\left[|\bar{X}| \ge \frac{1.96}{\sqrt{n}}, when\ \theta = 0\right]$

$under\ H_0\!: \theta = 0, X_1, \dots, X_n \sim N(0,1)$

$\Longrightarrow \bar{X} \sim N\left(0, \frac{1}{n}\right) \Longrightarrow \sqrt{n}\bar{X} \sim N(0,1)$

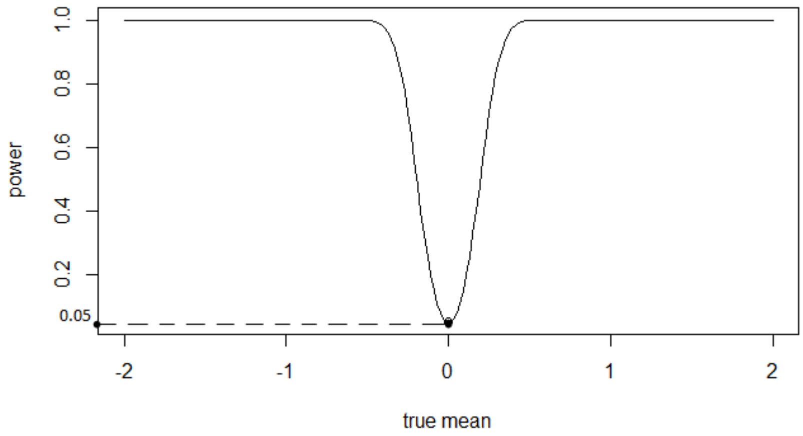
$\Longrightarrow P\left[|\bar{X}| \ge \frac{1.96}{\sqrt{n}}\right] = P[|Z| \ge 1.96] = 0.05$

$The\ power\ function\ is\ \beta(\theta) = P_\theta(\boldsymbol{X} \in C) = P_\theta\left(|\bar{X}| \ge \frac{1.96}{\sqrt{n}}\right)$

$X_1, \dots, X_n \sim N(\theta, 1) \Longrightarrow \bar{X} \sim N\left(\theta, \frac{1}{n}\right) \Longrightarrow \sqrt{n}(\bar{X} - \theta) \sim N(0,1)$

$$\begin{aligned}
\Longrightarrow \beta(\theta) &= P\big(\sqrt{n}|\bar{X}| \ge 1.96\big) = 1 - P\big(\sqrt{n}|\bar{X}| \le 1.96\big) = 1 - P\big(-1.96 \le \sqrt{n}\bar{X} \le 1.96\big) \\
&= 1 - P\big(-1.96 - \sqrt{n}\theta \le \sqrt{n}(\bar{X} - \theta) \le 1.96 - \sqrt{n}\theta\big) \\
&= 1 - P\big(-1.96 - \sqrt{n}\theta \le Z \le 1.96 - \sqrt{n}\theta\big) \\
&= 1 - \Phi\big(1.96 - \sqrt{n}\theta\big) + \Phi\big(-1.96 - \sqrt{n}\theta\big)
\end{aligned}$$

Graph the power function under $H_0\!: \theta = 0$



Remarks: Intuitively, we can see that the further $\bar{X}$ is from θ_0, the more easier to reject H_0

Critical function

The critical function is defined based on the reject region:

$$\psi(\boldsymbol{X}) = \begin{cases} 1, \boldsymbol{X} \in C \; (reject\ H_0) \\ 0, \boldsymbol{X} \in \bar{C} \; (Not\ reject\ H_0) \end{cases}$$

Therefore, by definition, the relationship between critical function and power function is:

$$E[\psi(X)] = P[X \in C] = \beta(\theta)$$

Problem: Want to select a critical function ψ in order to maximize the power $\beta(\theta)$, which as defined is

$$\beta_\psi(\theta) = E_\theta\big(\psi(X)\big), \theta \in \Omega_1,$$

Subject to the condition:

$$\beta_\psi(\theta) = E_\theta\big(\psi(X)\big) \le \alpha, for\ all\ \theta \in \theta_0$$

Simple vs. Composite Hypothesis

Simple hypothesis:

$$H_0 \colon \theta = \theta_0$$

Composite hypothesis:

$$H_a \colon \theta \neq \theta_0 \text{ or } H_a \colon \theta > \theta_0 \text{ or } H_a \colon \theta < \theta_0$$

Uniformly Most Powerful Test

A test that maximizes the power for all $\theta \in \Omega_1$ is called a uniformly most powerful test (UMP) In particular, under simple null and alternative hypothesis, such test is called most powerful test

Neymann-Pearson Lemma:

Let $\boldsymbol{X}$ have pdf $f(\boldsymbol{X}, \theta), \theta \in \Omega = \{\theta_0, \theta_1\}$:

Under simple null and alternative hypothesis:

$$H_0: \theta = \theta_0 \text{ vs. } H_1: \theta = \theta_1$$

The likelihood when H_0 being true is: $L_0(\boldsymbol{X}) = f(\boldsymbol{X}, \theta_0)$

The likelihood when H_1 being true is: $L_1(\boldsymbol{X}) = f(\boldsymbol{X}, \theta_1)$

The most powerful test of size α is a test that rejects H_0 is

$$\frac{L_0}{L_1} < k \left(or \ \frac{L_1}{L_0} > k^* \right) \tag{14.1}$$

where k is determined through the fact that the size of the test is α

More specifically,

$$\psi(\boldsymbol{X}) = \begin{cases} 1, \dfrac{L_0}{L_1} < k \\ 0, \dfrac{L_0}{L_1} \geq k \end{cases} \quad v.s. \ \psi^*(\boldsymbol{X}) = \begin{cases} 1, \boldsymbol{X} \in C^* \\ 0, \boldsymbol{X} \notin C^* \end{cases}$$

Under the condition: $\beta_\psi(\theta_0) \leq \alpha \ and \ \beta_{\psi^*}(\theta_0) \leq \alpha$, the theorem claims that:

$$\beta_\psi(\theta_1) \geq \beta_{\psi^*}(\theta_1)$$

$\mathbf{Example\ 14.2}$: $Let\ X_1, \dots, X_n\ be\ i.i.d.\ N(\theta, 1).\ Find\ the\ most\ powerful\ test\ for$ $H_0: \theta = 0\ vs. H_1: \theta = 1, when\ \alpha = 0.05, n = 25$

$\mathbf{Solution}$: $By\ N - P\ Lemma, the\ most\ powerful\ test\ is\ \psi(\boldsymbol{X}) = \begin{cases} 1, \dfrac{L_0}{L_1} < k \\ 0, \dfrac{L_0}{L_1} \geq k \end{cases}$

$$L_0(\boldsymbol{X}) = f(\boldsymbol{X}, \theta_0) = \prod_{i=1}^{n} f(X_i, \theta = 0) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} X_i^{\,2}\right\};$$

$$L_1(\boldsymbol{X}) = f(\boldsymbol{X}, \theta_1) = \prod_{i=1}^{n} f(X_i, \theta = 1) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (X_i - 1)^2\right\}$$

$$\Rightarrow Reject\ H_0, if\ \frac{L_0}{L_1} = \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} X_i^{\,2}\right\}}{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} (X_i - 1)^2\right\}} = \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n} X_i^{\,2} - \sum_{i=1}^{n} (X_i - 1)^2\right]\right\} < k$$

$$\Leftrightarrow Reject\ H_0, if \sum_{i=1}^{n} X_i^{\,2} - \sum_{i=1}^{n} (X_i - 1)^2 = 2\sum_{i=1}^{n} X_i - n > k^*$$

$$\Leftrightarrow Reject\ H_0, if\ \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} > C$$

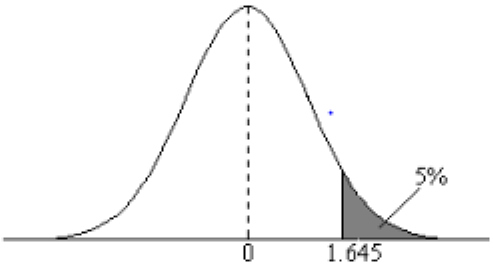
$By\ satisfying\ the\ condition\ P_{\theta=0}[Reject\ H_0] = P_{\theta=0}[\bar{X} > C] = \alpha$

$\Longrightarrow\ \bar{X} \sim N\left(0, \frac{1}{n}\right) \Longrightarrow \sqrt{n}\bar{X} \sim N(0,1)$

$\Longrightarrow P\left(\sqrt{n}\bar{X} > \sqrt{n}C\right) = P\left(Z > \sqrt{n}C\right) = \alpha$
$= 0.05$

$\Longrightarrow \sqrt{n}C = 5C = 1.645$

$\Longrightarrow C = \frac{1.645}{5} = 0.329$



$Therefore\ the\ most\ powerful\ test\ in\ this\ case\ is\ to\ reject\ H_0\ if\ \bar{X} > 0.329$

Lecture 41

<u>Proof of Neymann-Pearson Lemma</u>

Revisit and compare the definition of the “size of test” and type -I error (α):

$i).\ \alpha = P[type\ I\ error] = P[reject\ H_0\ when\ H_0\ is\ true]$

$ii).\ size\ of\ test = \sup_{\theta \in \Omega_0} \beta(\theta) \leq \alpha\ , when\ \theta \in \Omega_0$

When θ has only one single value $\Longrightarrow \boldsymbol{size\ of\ test = \alpha}$

Let $X_1, \ldots, X_n$ be identical independent distribution with p.d.f. $f(x_i, \theta)$

Want to test $H_0\colon \theta = \theta_0$ vs. $H_1\colon \theta = \theta_1$

Let $\alpha = (0,1)$ be fixed, $k^* > 0$, and C be a subset of $\mathbb{R}$, which satisfy:

$i).\ P_{\theta_0}(\boldsymbol{X} \in C) = \alpha$

$ii).\ \lambda = \frac{L_0(\theta_1, \boldsymbol{X})}{L_1(\theta_1, \boldsymbol{X})} = \frac{L_0}{L_1} < k, if\ \boldsymbol{X} \in C$

Let test T (defined in ii).), which corresponding to the critical region C; and

Let test T^* be any other test, which corresponding to the critical region C'

To prove the theorem is equivalently to show that:

$$\beta_{T^*}(\theta_1) \geq \beta_T(\theta_1), for\ which\ \beta_T(\theta_0) \leq \alpha$$

$\beta_{T^*}(\theta_1) = P[\boldsymbol{X} \in C, \theta = \theta_1, T^*] = \int_C L_1\, d\boldsymbol{x}$

$\beta_T(\theta_1) = P[\boldsymbol{X} \in C', \theta = \theta_1, T] = \int_{C'} L_1\, d\boldsymbol{x}$

$\Longrightarrow\ \beta_{T^*}(\theta_1) - \beta_T(\theta_1) = \int_C L_1\, d\boldsymbol{x} - \int_{C'} L_1\, d\boldsymbol{x} = \int_{(C\cap C')\cup(C\cap \overline{C'})} L_1\, d\boldsymbol{x} - \int_{(C\cap C')\cup(\bar{C}\cap C')} L_1\, d\boldsymbol{x}$

$$= \int_{C\cap\overline{C'}} L_1\,d\boldsymbol{x} - \int_{\bar{C}\cap C'} L_1\,d\boldsymbol{x}\ (*)$$

$On\ C\colon \dfrac{L_0}{L_1} < k \Longrightarrow\ \dfrac{L_1}{L_0} > \dfrac{1}{k} \Longrightarrow\ L_1 > \dfrac{1}{k}L_0\ \ (1)$

$On\ \bar{C}\colon \dfrac{L_0}{L_1} \geq k \Longrightarrow\ \dfrac{L_1}{L_0} \leq \dfrac{1}{k} \Longrightarrow\ L_1 \leq \dfrac{1}{k}L_0 \Longrightarrow -L_1 \geq -\dfrac{1}{k}L_0\ (2)$

$Plug\ (1),(2)\ into\ (*),$

$$\int_{C\cap\overline{C'}} L_1\,d\boldsymbol{x} - \int_{\bar{C}\cap C'} L_1\,d\boldsymbol{x} \geq \int_{C\cap\overline{C'}} \frac{1}{k}L_0\,d\boldsymbol{x} - \int_{\bar{C}\cap C'} \frac{1}{k}L_0\,d\boldsymbol{x} = \frac{1}{k}\left[\int_{C\cap\overline{C'}} L_0\,d\boldsymbol{x} - \int_{\bar{C}\cap C'} L_0\,d\boldsymbol{x}\right]$$

$$= \frac{1}{k}\left[\left(\int_{C\cap\overline{C'}} L_0\,d\boldsymbol{x} + \int_{C\cap C'} L_0\,d\boldsymbol{x}\right) - \left(\int_{C\cap C'} L_0\,d\boldsymbol{x} + \int_{\bar{C}\cap C'} L_0\,d\boldsymbol{x}\right)\right]$$

$$= \frac{1}{k}\left[\int_{C} L_0\,d\boldsymbol{x} - \int_{C'} L_0\,d\boldsymbol{x}\right] = \frac{1}{k}\left[\int_{C} L_0\,d\boldsymbol{x} - \int_{C'} L_0\,d\boldsymbol{x}\right]$$

$$= \frac{1}{k}\big(\alpha - \beta_T(\theta_0)\big) \geq 0$$

Extend the most powerful (MP) test to the uniformly most powerful (UMP) tests when H_1 is not simple

Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$

Want to test: $H_0\colon \theta = \theta_0$ vs. $H_1\colon \theta = \theta_1\ (assuming\ \theta_1 > \theta_0\ )$

For each θ_1, the most powerful test of size α is

$$C = \left\{\boldsymbol{X}\colon \frac{L_0}{L_1} < k\right\}$$

Where k will be determined through

$$P_{\theta_0}(\boldsymbol{X} \in C) = \alpha$$

$$L_0 = \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta_0)^2\right\};\ L_1 = \frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta_1)^2\right\}$$

$$\frac{L_0}{L_1} = \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(X_i - \theta_0)^2 - \sum_{i=1}^{n}(X_i - \theta_1)^2\right]\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[2(\theta_1 - \theta_0)\sum_{i=1}^{n}X_i + n(\theta_1 - \theta_0)(\theta_1 + \theta_0)\right]\right\} < k$$

$$\Longrightarrow 2(\theta_1 - \theta_0)\sum_{i=1}^{n}X_i + n(\theta_1 - \theta_0)(\theta_1 + \theta_0) > k^*\ (**)$$

$By\ assuming\ that\ \theta_1 > \theta_0 \Longrightarrow \theta_1 - \theta_0 > 0\ \big(direction\ of\ the\ inequality\ not\ change\ in\ (**)\big)$

$$\Rightarrow \sum_{i=1}^{n} X_i > \frac{k^*}{\theta_1 - \theta_0} - \frac{1}{2}(\theta_1 + \theta_0) = K$$

$$\Rightarrow The\ reject\ region\ C = \left\{\sum_{i=1}^{n} X_i > K\right\}, for\ all\ \theta > \theta_0$$

<u>How to determine K?</u>

K is determined by using the fact that the test is of size α:

$$P\left[\sum_{i=1}^{n} X_i > K, \theta = \theta_0\right] = \alpha\ or\ P[\bar{X} > K^*, \theta = \theta_0] = \alpha$$

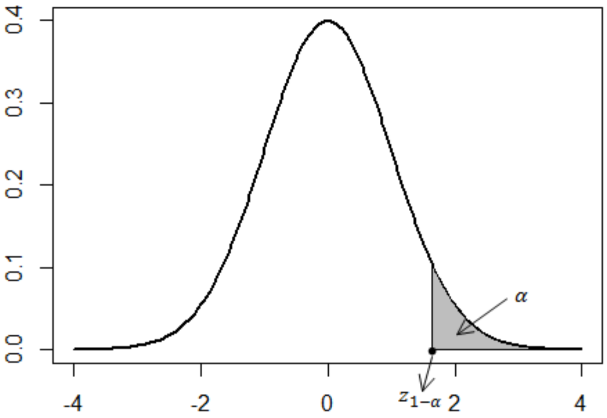
$Under\ H_0: X_1, \dots, X_n \sim N(\theta_0, 1) \Rightarrow \bar{X} \sim N\left(\theta_0, \frac{1}{n}\right) \Rightarrow \sqrt{n}(\bar{X} - \theta_0) \sim N(0,1)$

$\Rightarrow P[\bar{X} > K^*, \theta = \theta_0] = P\left[\sqrt{n}(\bar{X} - \theta_0) > \sqrt{n}(K^* - \theta_0)\right] = \alpha$

$\Rightarrow P\left[Z > \sqrt{n}(K^* - \theta_0)\right] = \alpha$

$\Rightarrow \sqrt{n}(K^* - \theta_0) = z_{1-\alpha}$

$\Rightarrow K^* = \theta_0 + \frac{z_{1-\alpha}}{\sqrt{n}}$



Example 15.1: $r.v. X \sim N(\theta, 1), z_{1-\alpha} = 1.645. Test\ H_0: \theta = \theta_0\ vs. H_1: \theta > \theta_0$

$Take\ a\ random\ sample\ X_1, \dots, X_n\ and\ find\ the\ sample\ mean\ \bar{X}$

$Based\ on\ Neymann - Pearson\ Lemma,$

$The\ UMP\ test\ is\ to\ reject\ H_0\ only\ if\ \bar{X} > \theta_0 + \frac{1.645}{\sqrt{n}}$

Remarks:

i) The critical region of decision to reject H_0 does not depend θ_1, therefore the test is the uniformly most powerful, when we test $H_0: \theta = \theta_0$ vs. $H_1: \theta > \theta_0$

ii) When n is very large, $\bar{X}$ is close to true θ, giving a better decision

Example 15.2: $Find\ the\ UMP\ test\ of\ size\ \alpha = 0.05, when\ we\ test\ the\ following\ hypothesis$:

$$H_0: \theta = \theta_0\ \ vs.\ \ H_1: \theta < \theta_0,\ \ where\ X_1, \dots, X_n \sim N(\theta, 1)$$

Solution: $By\ previous\ result, we\ have$

$$\frac{L_0}{L_1} = \exp\left\{-\frac{1}{2}\left[2(\theta_1-\theta_0)\sum_{i=1}^{n} X_i + n(\theta_1-\theta_0)(\theta_1+\theta_0)\right]\right\} < k$$

$$\Rightarrow 2(\theta_1-\theta_0)\sum_{i=1}^{n} X_i + n(\theta_1-\theta_0)(\theta_1+\theta_0) > k^* \ (**)$$

$By\ assuming\ that\ \theta_1 < \theta_0 \Rightarrow \theta_1 - \theta_0 < 0\ \big(direction\ of\ the\ inequality\ will\ change\ in\ (**)\big)$

$$\Rightarrow \sum_{i=1}^{n} X_i < \frac{k^*}{\theta_1-\theta_0} - \frac{1}{2}(\theta_1+\theta_0) = K$$

$$\Rightarrow The\ reject\ region\ C = \left\{\sum_{i=1}^{n} X_i < K\right\}, for\ all\ \theta > \theta_0$$

K is determined by using the fact that the test is of size α:

$$P\left[\sum_{i=1}^{n} X_i < K, \theta = \theta_0\right] = \alpha\ or\ P[\bar{X} < K^*, \theta = \theta_0] = \alpha$$

$$Under\ H_0: X_1, \ldots, X_n \sim N(\theta_0, 1) \Rightarrow \bar{X} \sim N\left(\theta_0, \frac{1}{n}\right) \Rightarrow \sqrt{n}(\bar{X}-\theta_0) \sim N(0,1)$$

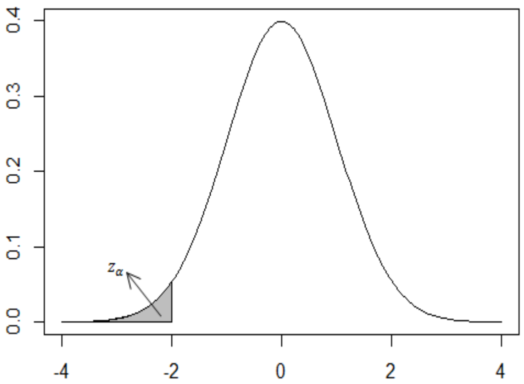
$$\Rightarrow P[\bar{X} < K^*, \theta = \theta_0]$$

$$= P\big[\sqrt{n}(\bar{X}-\theta_0) < \sqrt{n}(K^*-\theta_0)\big] = \alpha$$

$$\Rightarrow P\big[Z < \sqrt{n}(K^*-\theta_0)\big] = \alpha$$

$$\Rightarrow \sqrt{n}(K^*-\theta_0) = z_\alpha$$

$$\Rightarrow K^* = \theta_0 + \frac{z_\alpha}{\sqrt{n}} = \theta_0 - \frac{z_{1-\alpha}}{\sqrt{n}}$$



Remark: $the\ UMP\ test\ does\ \textbf{not}\ exist\ when\ test\ H_0: \theta = \theta_0\ \ vs.\ \ H_1: \theta \neq \theta_0$

Lecture 42

<u>Monotone Likelihood Ratio</u>

A family $\boldsymbol{f(X;\theta)}$ when $\boldsymbol{\theta \in \Omega}$, has a monotone likelihood ratio (LR) in $\boldsymbol{T(X)}$ if, for all $\boldsymbol{\theta' \leq \theta''}$:

i) $L(\theta', \boldsymbol{X})/L(\theta'', \boldsymbol{X})$ is an increasing function in $T(\boldsymbol{X})$, OR

ii) $L(\theta', \boldsymbol{X})/L(\theta'', \boldsymbol{X})$ is a decreasing function in $T(\boldsymbol{X})$

Example 16.1: $Let\ X_1, \ldots, X_n\ be\ i.i.d.\ Exp(\theta), check\ if\ f(x,\theta)\ has\ an\ increasing\ or\ decreasing\ LR$

Solution:

$$\frac{L(\theta', \boldsymbol{X})}{L(\theta'', \boldsymbol{X})} = \frac{\theta'^n e^{-\theta' \sum_{i=1}^n x_i}}{\theta''^n e^{-\theta'' \sum_{i=1}^n x_i}} = \left(\frac{\theta'}{\theta''}\right)^n e^{(\theta''-\theta')\sum_{i=1}^n x_i}$$

$$For\ \theta' \leq \theta'' \Rightarrow \theta'' - \theta' \geq 0, so\ that\ \left(\frac{\theta'}{\theta''}\right)^n e^{(\theta''-\theta')\sum_{i=1}^n x_i}\ is\ increasing\ in\ \sum_{i=1}^{n} x_i$$

$$Therefore, f(x,\theta)\ has\ an\ increasing\ LR\ in\ \sum_{i=1}^{n} x_i$$

Example 16.2: $Let\ X_1, \dots, X_n\ be\ i.i.d. Ber(\theta), check\ if\ f(x,\theta)\ has\ an\ increasing\ or\ decreasing\ LR$

Solution:

$$\Rightarrow \frac{L(\theta', \boldsymbol{X})}{L(\theta'', \boldsymbol{X})} = \frac{\theta'^{\sum_{i=1}^{n} x_i}(1-\theta')^{n-\sum_{i=1}^{n} x_i}}{\theta''^{\sum_{i=1}^{n} x_i}(1-\theta'')^{n-\sum_{i=1}^{n} x_i}} = \left[\frac{\theta'(1-\theta'')}{\theta''(1-\theta')}\right]^{\sum_{i=1}^{n} x_i} \left(\frac{1-\theta'}{1-\theta''}\right)^n$$

$$For\ \theta' \le \theta'' \Rightarrow \theta'(1-\theta'') - \theta''(1-\theta') = \theta' - \theta'\theta'' - \theta'' + \theta'\theta'' = \theta' - \theta'' \le 0$$

$$\Rightarrow \theta'(1-\theta'') \le \theta''(1-\theta') \Rightarrow \frac{\theta'(1-\theta'')}{\theta''(1-\theta')} \le 1$$

$$So\ that\ \left[\frac{\theta'(1-\theta'')}{\theta''(1-\theta')}\right]^{\sum_{i=1}^{n} x_i} \left(\frac{1-\theta'}{1-\theta''}\right)^n\ is\ decreasing\ in\ \sum_{i=1}^{n} x_i$$

$$Therefore, f(x,\theta)\ has\ a\ decreasing\ LR\ in\ \sum_{i=1}^{n} x_i$$

Example 16.3: $Let\ X_1, \dots, X_n\ be\ i.i.d. \mathcal{P}(\theta), check\ if\ f(x,\theta)\ has\ an\ increasing\ or\ decreasing\ LR$

Solution:

$$\frac{L(\theta', \boldsymbol{X})}{L(\theta'', \boldsymbol{X})} = \frac{\theta'^{\sum_{i=1}^{n} x_i} e^{-n\theta'} / \prod_{i=1}^{n} x_i!}{\theta''^{\sum_{i=1}^{n} x_i} e^{-n\theta''} / \prod_{i=1}^{n} x_i!} = \left(\frac{\theta'}{\theta''}\right)^{\sum_{i=1}^{n} x_i} e^{n(\theta''-\theta')}$$

$$For\ \theta' \le \theta'' \Rightarrow \frac{\theta'}{\theta''} \le 1, so\ that\ \left(\frac{\theta'}{\theta''}\right)^{\sum_{i=1}^{n} x_i} e^{n(\theta''-\theta')}\ is\ decreasing\ in\ \sum_{i=1}^{n} x_i$$

$$Therefore, f(x,\theta)\ has\ a\ decreasing\ LR\ in\ \sum_{i=1}^{n} x_i$$

Example 16.4: $Let\ X_1, \dots, X_n\ be\ i.i.d. N(\theta, 1), check\ if\ f(x,\theta)\ has\ an\ increasing\ or\ decreasing\ LR$

Solution:

$$\frac{L(\theta', \boldsymbol{X})}{L(\theta'', \boldsymbol{X})} = \frac{(2\pi)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta')^2}{2}}}{(2\pi)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta'')^2}{2}}} = e^{\frac{1}{2}\left[\sum_{i=1}^{n}(x_i-\theta'')^2 - \sum_{i=1}^{n}(x_i-\theta')^2\right]} = e^{\frac{1}{2}\left[2(\theta'-\theta'')\sum_{i=1}^{n} x_i + n\left(\theta''^2 - \theta'^2\right)\right]}$$

$$= e^{(\theta'-\theta'')\sum_{i=1}^{n} x_i} e^{\frac{n}{2}\left(\theta''^2 - \theta'^2\right)}$$

$$For\ \theta' \le \theta'' \Rightarrow \theta' - \theta'' \le 0, so\ that\ e^{(\theta'-\theta'')\sum_{i=1}^{n} x_i} e^{\frac{n}{2}\left(\theta''^2 - \theta'^2\right)}\ is\ decreasing\ in\ \sum_{i=1}^{n} x_i$$

$$Therefore, f(x,\theta)\ has\ a\ decreasing\ LR\ in\ \sum_{i=1}^{n} x_i$$

Likelihood Ratio for One Parameter Family

Theorem: For $\theta' \le \theta''$, and let $f(x,\theta)$ be of one parameter family with pdf:

$$f(x,\theta) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A\ , A\ does\ not\ depend\ on\ \theta$$

$$\frac{L(\theta',\boldsymbol{x})}{L(\theta'',\boldsymbol{x})}=\frac{\exp\{c(\theta')\sum_{i=1}^{n}T(x_i)+S(\boldsymbol{x})+d(\theta')\}}{\exp\{c(\theta'')\sum_{i=1}^{n}T(x_i)+S(\boldsymbol{x})+d(\theta'')\}}=\exp\left\{[c(\theta')-c(\theta'')]\sum_{i=1}^{n}T(x_i)\right\}\cdot g(\boldsymbol{x},\theta)$$

Case 1: $If\ c(\theta)\ is\ increasing, then\ f(x,\theta)\ has\ a\ decreasing\ LR\ in\ \ \sum_{i=1}^{n}T(x_i)$

Case 2: $If\ c(\theta)\ is\ decreasing, then\ f(x,\theta)\ has\ an\ increasing\ LR\ in\ \ \sum_{i=1}^{n}T(x_i)$

Example 16.5: $Revisit\ the\ case\ of\ Bernoulli\ distribution\ by\ using\ this\ theorem$

Solution: $f(x,\theta)=\theta^x(1-\theta)^{1-x}=\exp\{x\log\theta+(1-x)\log(1-\theta)\}$

$$=\exp\left\{x\log\left(\frac{\theta}{1-\theta}\right)+\log(1-\theta)\right\},\theta\in(0,1)$$

$By\ the\ theorem\ above,$

$$T(x)=x, c(\theta)=\log\left(\frac{\theta}{1-\theta}\right)\Rightarrow c'(\theta)=\frac{1-\theta}{\theta}\cdot\frac{(1-\theta)+\theta}{(1-\theta)^2}=\frac{1}{\theta(1-\theta)}>0$$

$$So\ that, c(\theta)\ is\ increasing\Rightarrow f(\theta,x)\ has\ a\ decreasing\ LR\ in\ \sum_{i=1}^{n}X_i$$

Theorem 1:

In general, in order to test $H_0{:}\,\theta\le\theta_0$ vs. $H_1{:}\,\theta>\theta_0$. If $f(x,\theta)$ has an increasing LR in $T(\boldsymbol{x})$, then the **uniformly most powerful** test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X})<k,$$

$$where\ k\ is\ determined\ by\ \sup P[T(X)<k,\theta\le\theta_0]=P[T(X)<k,\theta=\theta_0]=\alpha$$

Theorem 2:

In general, in order to test $H_0{:}\,\theta\le\theta_0$ vs. $H_1{:}\,\theta>\theta_0$. If $f(x,\theta)$ has an decreasing LR in $T(\boldsymbol{x})$, then the **uniformly most powerful** test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X})>k,$$

$$where\ k\ is\ determined\ by\ \sup P[T(X)<k,\theta\le\theta_0]=P[T(X)>k,\theta=\theta_0]=\alpha$$

Theorem 3:

In general, in order to test $H_0{:}\,\theta\ge\theta_0$ vs. $H_1{:}\,\theta<\theta_0$. If $f(x,\theta)$ has an increasing LR in $T(\boldsymbol{x})$, then the **uniformly most powerful** test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X})>k,$$

$$where\ k\ is\ determined\ by\ \sup P[T(X)<k,\theta\ge\theta_0]=P[T(X)>k,\theta=\theta_0]=\alpha$$

Theorem 4:

In general, in order to test $H_0{:}\,\theta\ge\theta_0$ vs. $H_1{:}\,\theta<\theta_0$. If $f(x,\theta)$ has an decreasing LR in $T(\boldsymbol{x})$, then the uniformly most powerful test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X})<k,$$

$$where\ k\ is\ determined\ by\ \sup P[T(X)<k,\theta\ge\theta_0]=P[T(X)<k,\theta=\theta_0]=\alpha$$

Lemma:

Let $f(x,\theta)$ be of one parameter family with pdf:

$$\boldsymbol{f(x,\theta) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\} \cdot I_A \;, A\ does\ not\ depend\ on\ \theta}$$

1) Test $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$, and $c(\theta)$ is decreasing, then the UMP test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X}) < k, where\ k\ is\ determined\ by\ P[T(X) < k, \theta = \theta_0] = \alpha$$

2) Test $H_0: \theta \leq \theta_0$ vs. $H_1: \theta > \theta_0$, and $c(\theta)$ is increasing, then the UMP test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X}) > k, where\ k\ is\ determined\ by\ P[T(X) > k, \theta = \theta_0] = \alpha$$

3) Test $H_0: \theta \geq$ vs. $H_1: \theta < \theta_0$, and $c(\theta)$ is decreasing, then the UMP test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X}) > k, where\ k\ is\ determined\ by\ P[T(X) > k, \theta = \theta_0] = \alpha$$

4) Test $H_0: \theta \geq \theta_0$ vs. $H_1: \theta < \theta_0$, and $c(\theta)$ is increasing, then the UMP test of size α is given by:

$$Reject\ H_0, if\ T(\boldsymbol{X}) < k, where\ k\ is\ determined\ by\ P[T(X) < k, \theta = \theta_0] = \alpha$$

Lecture 43

Uniformly Most Powerful Test for Non-Exponential Family

Example 17. 1: $Let\ X_1, \dots, X_n\ be\ \mathcal{U}(0,\theta), and\ test\ H_0: \theta \leq \theta_0\ vs. H_1: \theta > \theta_0.$ $Find\ the\ UMP\ test\ of\ size\ \alpha$

Solution: $f(\theta,x) = \begin{cases} \frac{1}{\theta}, 0 < x \leq \theta \\ 0, \quad\quad O.W. \end{cases}, for\ \theta' \leq \theta''$

$$L(\theta',\boldsymbol{x}) = \left(\frac{1}{\theta'}\right)^n \cdot I_{\{x^{(n)} \leq \theta'\}};\ L(\theta'',\boldsymbol{x}) = \left(\frac{1}{\theta''}\right)^n \cdot I_{\{x^{(n)} \leq \theta''\}}$$

$$\Rightarrow \frac{L(\theta',\boldsymbol{x})}{L(\theta'',\boldsymbol{x})} = \frac{\left(\frac{1}{\theta'}\right)^n \cdot I_{\{x^{(n)} \leq \theta'\}}}{\left(\frac{1}{\theta''}\right)^n \cdot I_{\{x^{(n)} \leq \theta''\}}} = \left(\frac{\theta''}{\theta'}\right)^n \cdot \frac{I_{\{x^{(n)} \leq \theta'\}}}{I_{\{x^{(n)} \leq \theta''\}}} = \begin{cases} \left(\frac{\theta''}{\theta'}\right)^n, 0 < x^{(n)} \leq \theta' \\ 0, \quad\quad \theta' < x^{(n)} \leq \theta'' \end{cases}, where\ \left(\frac{\theta''}{\theta'}\right)^n > 0$$

$$So\ that\ \frac{L(\theta',\boldsymbol{x})}{L(\theta'',\boldsymbol{x})}\ is\ decreasing\ in\ x^{(n)}$$

$By\ the\ Theorem\ 2, the\ UMP\ test\ is$:

$$Reject\ H_0, when\ X^{(n)} > C$$

$where\ P\left[X^{(n)} > C, when\ \theta = \theta_0\right] = \alpha$

$$f\left(\theta_0, X^{(n)} = t\right) = n\left(\frac{t}{\theta_0}\right)^{n-1} \cdot \frac{1}{\theta_0} = \frac{n}{{\theta_0}^n} t^{n-1} \Rightarrow P\left[X^{(n)} > C, when\ \theta = \theta_0\right] = \int_C^\infty \frac{n}{{\theta_0}^n} t^{n-1} dt = \alpha$$

$$\Rightarrow 1 - \int_0^C \frac{n}{{\theta_0}^n} t^{n-1} dt = 1 - \frac{n}{{\theta_0}^n} \cdot \frac{C^n}{n} = 1 - \left(\frac{C}{\theta_0}\right)^n = \alpha \Rightarrow C = \theta_0(1-\alpha)^{\frac{1}{n}}$$

$Therefore, the\ UMP\ test\ of\ size\ \alpha\ is$:

$$Reject\ H_0, when\ X^{(n)} > \theta_0(1-\alpha)^{\frac{1}{n}}$$

Likelihood Ratio Test

In order to test $H_0: \theta \in \Omega_0$ vs. $H_1: \theta \notin \Omega_0$, when UMP test does not exist in this case, we will:

$$Reject\ H_0, if\ \lambda = \frac{\max L(\theta), \theta \in \Omega_0}{\max L(\theta), \theta \in \Omega} = \frac{\max L(\theta), \theta \in \Omega_0}{L(\hat{\theta})} < k,$$

$where\ k\ is\ determined\ by\ \sup P[\lambda < k, \theta \in \Omega_0] = \alpha$

Example 17.2: $Let\ X_1, \dots, X_n$ be i.i.d. $N(\mu, 1)$. Test $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$. Find LRT with size $\alpha = 0.05$

Solution: $X_1, \dots, X_n\ has\ i.i.d. pdf\ of\ f(x, \mu) = \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\mu)^2}{2}\right\}$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(x_i-\mu)^2}{2}\right\} = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$\Rightarrow l(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \Rightarrow \frac{dl(\mu)}{d\mu} = \sum_{i=1}^{n}(x_i-\mu) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\max L(\mu) = L(\hat{\mu}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i-\hat{\mu})^2}{2}\right\} = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}$$

$$\max L(\mu_0) = L(\mu_0) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i-\mu_0)^2}{2}\right\} = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}$$

$$\Rightarrow \lambda(\boldsymbol{X}) = \frac{\max L(\mu_0)}{\max L(\mu)} = \frac{\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}}{\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}} = \exp\left\{\frac{1}{2}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 - \sum_{i=1}^{n}(X_i-\mu_0)^2\right]\right\}$$

$$= \exp\left\{(\mu_0-\bar{X})\sum_{i=1}^{n} X_i - \frac{1}{2}n(\mu_0{}^2 - \bar{X}^2)\right\} = \exp\left\{(\mu_0-\bar{X}) \cdot n\bar{X} - \frac{1}{2}n(\mu_0{}^2 - \bar{X}^2)\right\}$$

$$= \exp\left\{-\frac{1}{2}n\bar{X}^2 + n\mu_0\bar{X} - \frac{1}{2}n\mu_0{}^2\right\} = \exp\left\{-\frac{1}{2}n(\bar{X}-\mu_0)^2\right\} < k$$

$$\Rightarrow (\bar{X}-\mu_0)^2 > K \Rightarrow |\bar{X}-\mu_0| > C \Rightarrow \bar{X}-\mu_0 > C, or\ \bar{X}-\mu_0 < -C$$

$Therefore, the\ critical\ region\ of\ LRT\ in\ this\ case\ is$:

$$\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0\ (*)$$

$where\ the\ constant\ C\ is\ determined\ by\ \sup P[\lambda < k, \mu = \mu_0] = \alpha, in\ this\ case,$

$$\Longrightarrow \sup P[\lambda < k, \mu = \mu_0] = P[\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0, \mu = \mu_0] = \alpha$$

$$Under\ H_0, \bar{X} \sim N\left(\mu_0, \frac{1}{n}\right) \Longrightarrow \sqrt{n}(\bar{X} - \mu_0) \sim N(0,1)$$

$$\Longrightarrow P[\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0] = P\left[\sqrt{n}(\bar{X} - \mu_0) > \sqrt{n}C, or\ \sqrt{n}(\bar{X} - \mu_0) < -\sqrt{n}C\right] = \alpha$$

$$\Longrightarrow P\left[Z > \sqrt{n}C, or\ Z < -\sqrt{n}C\right] = \alpha$$

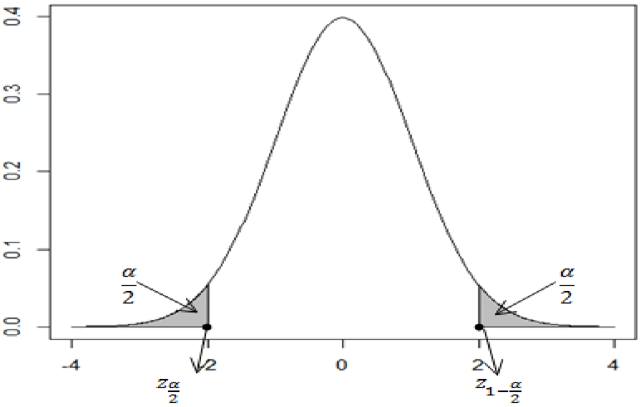
$Since\ standard\ normal\ is\ symmetric\ about\ 0, so\ that,$

$$P\left[Z > \sqrt{n}C, or\ Z < -\sqrt{n}C\right] = P\left(Z > \sqrt{n}C\right) + P\left(Z < -\sqrt{n}C\right) = 2P\left(Z > \sqrt{n}C\right) = 2\left(Z < -\sqrt{n}C\right) = \alpha$$

$$\Longrightarrow P\left(Z > \sqrt{n}C\right) = P\left(Z < -\sqrt{n}C\right) = \frac{\alpha}{2}$$

$$\Longrightarrow -\sqrt{n}C = z_{\frac{\alpha}{2}}, \sqrt{n}C = z_{1-\frac{\alpha}{2}}$$

$$\Longrightarrow C = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} = -\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$$



$The\ LRT\ defined\ in\ (*)\ becomes:$

$$To\ reject\ H_0, when\ \bar{X} > \mu_0 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, or\ \bar{X} < \mu_0 + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} = \mu_0 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

$Specifically, when\ \alpha = 0.05, reject\ H_0,$

$$when\ \bar{X} > \mu_0 + \frac{z_{0.975}}{\sqrt{n}}, or\ \bar{X} < \mu_0 - \frac{z_{0.975}}{\sqrt{n}} \Longleftrightarrow\ \bar{X} > \mu_0 + \frac{1.96}{\sqrt{n}}, or\ \bar{X} < \mu_0 - \frac{1.96}{\sqrt{n}}$$

Lecture 44

LRT for μ of the Normal Distribution

$Let\ X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Test $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$. Find LRT with size $\alpha = 0.05$.

Case 1: $When\ \sigma^2\ is\ known$

Solution: $X_1, \ldots, X_n\ has\ i.i.d.pdf\ of\ f(x, \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$\Longrightarrow l(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \Longrightarrow \frac{dl(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \Longrightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}X_i = \bar{X}$$

$$\max L(\mu) = L(\hat{\mu}) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i - \hat{\mu})^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}$$

$$\max L(\mu_0) = L(\mu_0) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i - \mu_0)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}$$

$$\Longrightarrow \lambda(\boldsymbol{X}) = \frac{\max L(\mu_0)}{\max L(\mu)} = \frac{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}}{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right]\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right]\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{X} - \mu_0)^2\right]\right\} < k$$

$$\Longrightarrow (\bar{X} - \mu_0)^2 > K \Longrightarrow |\bar{X} - \mu_0| > C \Longrightarrow \bar{X} - \mu_0 > C, or\ \bar{X} - \mu_0 < -C$$

$Therefore, the\ critical\ region\ of\ LRT\ in\ this\ case\ is:$

$$\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0\ (*)$$

$where\ the\ constant\ C\ is\ determined\ by\ \sup P[\lambda < k, \mu = \mu_0] = \alpha, in\ this\ case,$

$$\Longrightarrow \sup P[\lambda < k, \mu = \mu_0] = P[\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0, \mu = \mu_0] = \alpha$$

$Under\ H_0, \bar{X} \sim N\left(\mu_0, \frac{1}{n}\right) \Longrightarrow \sqrt{n}(\bar{X} - \mu_0) \sim N(0,1)$

$$\Longrightarrow P[\bar{X} > C + \mu_0, or\ \bar{X} < -C + \mu_0] = P\left[\sqrt{n}(\bar{X} - \mu_0) > \sqrt{n}C, or\ \sqrt{n}(\bar{X} - \mu_0) < -\sqrt{n}C\right] = \alpha$$

$$\Longrightarrow P\left[Z > \sqrt{n}C, or\ Z < -\sqrt{n}C\right] = \alpha$$

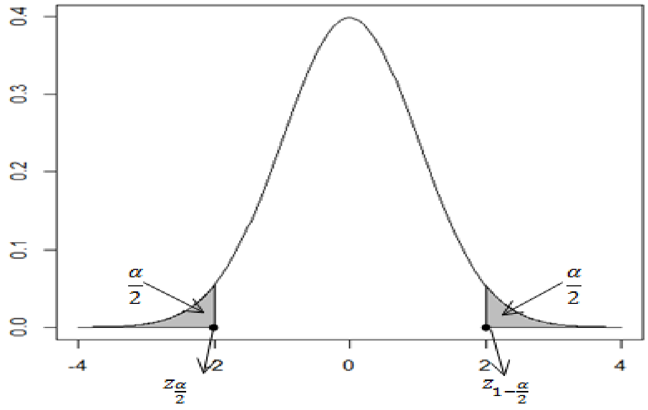
$Since\ standard\ normal\ is\ symmetric\ about\ 0, so\ that,$

$$P\left[Z > \sqrt{n}C, or\ Z < -\sqrt{n}C\right] = P\left(Z > \sqrt{n}C\right) + P\left(Z < -\sqrt{n}C\right) = 2P\left(Z > \sqrt{n}C\right) = 2\left(Z < -\sqrt{n}C\right) = \alpha$$

$$\Longrightarrow P\left(Z > \sqrt{n}C\right) = P\left(Z < -\sqrt{n}C\right) = \frac{\alpha}{2}$$

$$\Longrightarrow -\sqrt{n}C = z_{\frac{\alpha}{2}}, \sqrt{n}C = z_{1-\frac{\alpha}{2}}$$

$$\Longrightarrow C = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} = -\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$$



$The\ LRT\ defined\ in\ (*)\ becomes:$

$$To\ reject\ H_0, when\ \bar{X} > \mu_0 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, or\ \bar{X} < \mu_0 + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} = \mu_0 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

$Specifically, when\ \alpha = 0.05, reject\ H_0,$

$$when\ \bar{X} > \mu_0 + \frac{z_{0.975}}{\sqrt{n}}, or\ \bar{X} < \mu_0 - \frac{z_{0.975}}{\sqrt{n}} \Longleftrightarrow \bar{X} > \mu_0 + \frac{1.96}{\sqrt{n}}, or\ \bar{X} < \mu_0 - \frac{1.96}{\sqrt{n}}$$

Case 2: $When\ \sigma^2\ is\ unknown$

Solution: $X_1, \dots, X_n\ has\ i.i.d.pdf\ of\ f(x,\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left\{-\frac{(x_i-\mu)^2}{2}\right\} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

$$\Longrightarrow l(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Longrightarrow \frac{dl(\mu)}{d\sigma^2} = -\frac{n}{2}\cdot\frac{1}{2\pi\sigma^2}\cdot 2\pi + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\Longrightarrow \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \Longrightarrow under\ H_0, \widehat{\sigma^2}_0 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2\ ;$$

$$under\ H_1, \widehat{\sigma^2}_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\max L(\mu) = L(\hat{\mu}) = \prod_{i=1}^{n} \frac{1}{\left(2\pi\widehat{\sigma^2}_1\right)^{\frac{1}{2}}} \exp\left\{-\frac{(X_i-\hat{\mu})^2}{2\widehat{\sigma^2}_1}\right\} = \frac{1}{\left(2\pi\widehat{\sigma^2}_1\right)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\widehat{\sigma^2}_1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}$$

$$= \frac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\right)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}$$

$$= \frac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\right)^{\frac{n}{2}}}\exp -\frac{n}{2}$$

$$\max L(\mu_0) = L(\mu_0) = \prod_{i=1}^{n}\frac{1}{(2\pi\widehat{\sigma^2}_0)^{\frac{1}{2}}}\exp\left\{-\frac{(X_i-\mu_0)^2}{2\widehat{\sigma^2}_0}\right\} = \frac{1}{(2\pi\widehat{\sigma^2}_0)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\widehat{\sigma^2}_0}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}$$

$$= \frac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2\right)\right)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2\right)}\sum_{i=1}^{n}(X_i-\mu_0)^2\right\}$$

$$= \frac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2\right)\right)^{\frac{n}{2}}}\exp -\frac{n}{2}$$

$$\Rightarrow \lambda(\boldsymbol{X}) = \frac{\max L(\mu_0)}{\max L(\mu)} = \frac{\dfrac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2\right)\right)^{\frac{n}{2}}}\exp -\dfrac{n}{2}}{\dfrac{1}{\left(2\pi\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\right)^{\frac{n}{2}}}\exp -\dfrac{n}{2}} = \left[\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{-\frac{n}{2}} \quad (*)$$

Rewrite the nominator in $(*)$,

$$\sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu_0) + n(\bar{X}-\mu_0)^2$$

where the cross term $\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu_0) = (\bar{X}-\mu_0)\sum_{i=1}^{n}(X_i-\bar{X}) = (\bar{X}-\mu_0)\left(\sum_{i=1}^{n}X_i - n\bar{X}\right)$

$$= (\bar{X}-\mu_0)(n\bar{X}-n\bar{X}) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2\ ,\textit{put it into } (*)$$

$$\lambda(\boldsymbol{X}) = \left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{-\frac{n}{2}} = \left[1+\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{-\frac{n}{2}} < k$$

$$\Rightarrow \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} > K \Rightarrow \frac{(\bar{X}-\mu_0)^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} > K \Rightarrow \frac{|\bar{X}-\mu_0|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}} > C$$

$$\Rightarrow \left|\frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2(n-1)}}}\right| > C^*\ (*), as\ \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \sim \mathcal{X}_{n-1}{}^2, and\ S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$(*)\ becomes\ \left|\frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\mathcal{X}_{n-1}{}^2}{(n-1)}}}\right| = \left|\frac{Z}{\sqrt{\frac{\mathcal{X}_{n-1}{}^2}{(n-1)}}}\right| > C^*, where\ \frac{Z}{\sqrt{\frac{\mathcal{X}_{n-1}{}^2}{(n-1)}}} \sim t_{n-1}(student\ t-distribution)$$

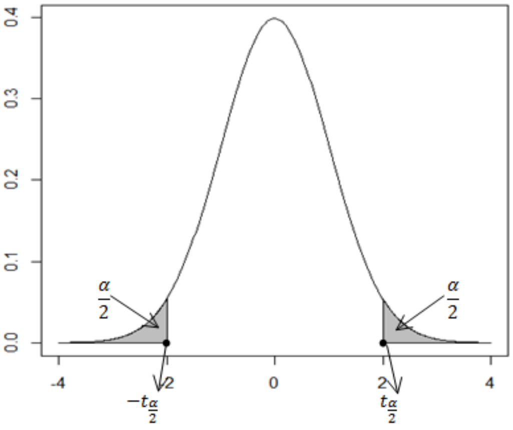
$$And\ also\ \left|\frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2(n-1)}}}\right| = \left|\frac{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}{S\sqrt{\frac{1}{\sigma^2}}}\right|$$

$$= \left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > C^*$$

$Therefore, \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$

$\Rightarrow P\left[\left|\frac{\bar{X}-\mu_0}{S/\sqrt{n}}\right| > C^*, \mu = \mu_0\right] = \alpha$

$$\Rightarrow Reject\ H_0\ if\ \frac{\bar{X}-\mu_0}{S/\sqrt{n}} > t_{\frac{\alpha}{2}}\ or\ \frac{\bar{X}-\mu_0}{S/\sqrt{n}} < -t_{\frac{\alpha}{2}}$$



$$\Rightarrow Reject\ H_0, if\ \bar{X} > \mu_0 + \frac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}},$$

$$OR, \bar{X} < \mu_0 - \frac{S}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$$

$Specifically, when\ \alpha = 0.05, reject\ H_0,$

$$when\ \bar{X} > \mu_0 + \frac{S}{\sqrt{n}} t_{n-1,0.975}, or\ \bar{X} < \mu_0 - \frac{S}{\sqrt{n}} t_{n-1,0.975}$$

$$where\ S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}\ , t\ is\ determined\ by\ student-t\ distribution\ with\ known\ n$$

Lecture 45

Hypothesis testing for two samples

Example 19. 1: $Let\ X_1, \ldots, X_m \sim N(\mu_1, \sigma_1{}^2), and\ Y_1, \ldots, Y_n \sim N(\mu_2, \sigma_2{}^2)\ be\ two\ independent\ samples$ $propose\ a\ good\ test\ for\ H_0: \mu_1 = \mu_2 = \mu\ vs. H_1: \mu_1 \neq \mu_2\ of\ size\ \alpha$

Assume that $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$

Solution: *By LRT*,

$$\lambda(\boldsymbol{X}) = \frac{\max L(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega_0}{\max L(\boldsymbol{\theta})\,, \boldsymbol{\theta} \in \Omega} = \frac{\max\limits_{\mu,\sigma^2 \in \Omega_0} L(\mu,\mu,\sigma^2,\sigma^2)}{\max\limits_{\widehat{\mu_1},\widehat{\mu_2},\hat{\sigma}^2 \in \Omega} L(\widehat{\mu_1},\widehat{\mu_2},\hat{\sigma}^2,\hat{\sigma}^2)}$$

$In\ order\ to\ solve\ for \max\limits_{\mu,\sigma^2 \in \Omega_0} L(\mu,\mu,\sigma^2,\sigma^2)$

$$\Rightarrow \frac{\partial}{\partial \mu} \ln L(\mu,\mu,\sigma^2,\sigma^2) = \frac{\partial}{\partial \mu}\left[-\frac{m+n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(x_i-\mu)^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right]\right] = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\left[\sum_{i=1}^{m}(x_i-\mu) + \sum_{i=1}^{n}(y_i-\mu)\right] = 0 \Rightarrow \hat{\mu} = \frac{1}{m+n}\left(\sum_{i=1}^{m} x_i + \sum_{i=1}^{n} y_i\right)$$

$$\Rightarrow \frac{\partial}{\partial \sigma^2} \ln L(\mu,\mu,\sigma^2,\sigma^2) = \frac{\partial}{\partial \sigma^2}\left\{-\frac{m+n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(x_i-\mu)^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right]\right\} = 0$$

$$\Rightarrow -\frac{m+n}{2\sigma^2} + \frac{\left[\sum_{i=1}^{m}(x_i-\mu)^2 + \sum_{i=1}^{n}(y_i-\mu)^2\right]}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\hat{\mu})^2 + \sum_{i=1}^{n}(y_i-\hat{\mu})^2\right]$$

$$\Rightarrow \max_{\mu,\sigma^2 \in \Omega_0} L(\mu,\mu,\sigma^2,\sigma^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{\frac{m+n}{2}}} e^{-\frac{m+n}{2}} = \left(\frac{e^{-1}}{2\pi\hat{\sigma}^2}\right)^{\frac{m+n}{2}}$$

$In\ order\ to\ solve\ for \max\limits_{\widehat{\mu_1},\widehat{\mu_2},\hat{\sigma}^2 \in \Omega} L(\widehat{\mu_1},\widehat{\mu_2},\hat{\sigma}^2,\hat{\sigma}^2)$

$$\Rightarrow \frac{\partial}{\partial \mu_1} \ln L(\mu_1,\mu_2,\sigma^2,\sigma^2) = \frac{\partial}{\partial \mu_1}\left\{-\frac{m+n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(x_i-\mu_1)^2 + \sum_{i=1}^{n}(y_i-\mu_2)^2\right]\right\} = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu_1) = 0 \Rightarrow \widehat{\mu_1} = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}$$

$$\Rightarrow \frac{\partial}{\partial \mu_2} L(\mu_1,\mu_2,\sigma^2,\sigma^2) \Rightarrow \widehat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

$$\Rightarrow \frac{\partial}{\partial \sigma^2} \ln L(\mu_1,\mu_2,\sigma^2,\sigma^2) = \frac{\partial}{\partial \sigma^2}\left\{-\frac{m+n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{m}(x_i-\mu_1)^2 + \sum_{i=1}^{n}(y_i-\mu_2)^2\right]\right\} = 0$$

$$\Rightarrow -\frac{m+n}{2\sigma^2} + \frac{\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2\right]}{2\sigma^4} = 0 \Rightarrow \hat{\sigma}^{*2} = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$\Rightarrow \max_{\mu,\sigma^2 \in \Omega_0} L(\mu_1,\mu_2,\sigma^2,\sigma^2) = \frac{1}{(2\pi\hat{\sigma}^{*2})^{\frac{m+n}{2}}} e^{-\frac{m+n}{2}} = \left(\frac{e^{-1}}{2\pi\hat{\sigma}^{*2}}\right)^{\frac{m+n}{2}}$$

$$\Longrightarrow \lambda(\boldsymbol{X}) = \frac{\left(\frac{e^{-1}}{2\pi\hat{\sigma}^2}\right)^{\frac{m+n}{2}}}{\left(\frac{e^{-1}}{2\pi\hat{\sigma}^{*2}}\right)^{\frac{m+n}{2}}} = \left(\frac{2\pi\hat{\sigma}^2}{2\pi\hat{\sigma}^{*2}}\right)^{-\frac{m+n}{2}} < k \Longrightarrow \frac{\hat{\sigma}^2}{\hat{\sigma}^{*2}} > K$$

$$\hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\hat{\mu})^2 + \sum_{i=1}^{n}(y_i-\hat{\mu})^2 + n(\bar{y}-\hat{\mu})^2\right]$$

$$= \frac{m(\bar{x}-\hat{\mu})^2 + n(\bar{y}-\hat{\mu})^2}{m+n} + \hat{\sigma}^{*2} = \frac{m\left(\bar{x}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2 + n\left(\bar{y}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2}{m+n} + \hat{\sigma}^{*2}$$

$$= \frac{mn}{(m+n)^2}(\bar{x}-\bar{y})^2 + \hat{\sigma}^{*2}$$

$$\Longrightarrow \frac{\hat{\sigma}^2}{\hat{\sigma}^{*2}} = \frac{\frac{mn}{(m+n)^2}(\bar{x}-\bar{y})^2 + \hat{\sigma}^{*2}}{\hat{\sigma}^{*2}} = 1 + \frac{\frac{mn}{m+n}(\bar{x}-\bar{y})^2}{\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2} > K$$

$$\Longrightarrow \frac{\frac{mn}{m+n}(\bar{x}-\bar{y})^2}{\frac{1}{m+n-2}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2\right]} = \frac{\frac{mn}{m+n}(\bar{x}-\bar{y})^2}{\frac{1}{m+n-2}\left[(m-1){s_x}^2 + (n-1){s_y}^2\right]} > C$$

$$\Longrightarrow \left|\frac{(\bar{X}-\bar{Y})\Big/\sigma/\left(\frac{mn}{m+n}\right)}{\sqrt{\left(\frac{(m-1){s_x}^2}{\sigma^2} + \frac{(n-1){s_y}^2}{\sigma^2}\right)\Big/ m+n-2}}\right| = \left|\frac{Z}{\sqrt{\frac{{\mathcal{X}_{m-1}}^2 + {\mathcal{X}_{n-1}}^2}{m+n-2}}}\right| = \left|\frac{Z}{\sqrt{\frac{{\mathcal{X}_{m+n-2}}^2}{m+n-2}}}\right| \sim t_{m+n-2}$$

$$\Longrightarrow \left|\frac{\bar{x}-\bar{y}}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right){s_{pl}}^2}}\right| > C^*, where\ {s_{pl}}^2 = \frac{1}{m+n-2}\left[(m-1){s_x}^2 + (n-1){s_y}^2\right]$$

$$\frac{\bar{x}-\bar{y}}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right){s_{pl}}^2}}\ follows\ t-distribution, with\ m+n-2\ degree\ of\ freedom$$

Therefore, the decision rule is:

$$Reject\ H_0, when\ \frac{\bar{x}-\bar{y}}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right){s_{pl}}^2}} > t_{\frac{\alpha}{2},m+n-2}, or\ \frac{\bar{x}-\bar{y}}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right){s_{pl}}^2}} < -t_{\frac{\alpha}{2},m+n-2}$$

Example 19.2: *Let* $X_1, \dots, X_m \sim N(\mu_1, {\sigma_1}^2)$, *and* $Y_1, \dots, Y_n \sim N(\mu_2, {\sigma_2}^2)$ *be two independent samples propose a good test for* $H_0: {\sigma_1}^2 = {\sigma_2}^2 = \sigma^2\ vs.\ H_1: {\sigma_1}^2 \neq {\sigma_2}^2$ *of size* α

Solution:

$$\lambda(\boldsymbol{X}) = \frac{\max L(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega_0}{\max L(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega} = \frac{\max\limits_{\mu_1,\mu_2,\sigma^2 \in \Omega_0} L(\mu_1,\mu_2,\sigma^2,\sigma^2)}{\max\limits_{\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2 \in \Omega} L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2)}$$

$In\ \Omega_0, to\ solve\ the\ maximization\ of\ L(\mu_1,\mu_2,\sigma^2,\sigma^2) = (2\pi\sigma^2)^{-\frac{m+n}{2}} e^{-\frac{1}{2\sigma^2}[\sum_{i=1}^m (x_i-\mu_1)^2+\sum_{i=1}^n (y_i-\mu_2)^2]}$

$$\frac{\partial}{\partial\mu_1}\ln L(\mu_1,\mu_2,\sigma^2,\sigma^2) = 0 \Longrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu_1) = 0 \Longrightarrow \mu_1 = \frac{\sum_{i=1}^m x_i}{m} = \bar{x}$$

$$\frac{\partial}{\partial\mu_2}\ln L(\mu_1,\mu_2,\sigma^2,\sigma^2) = 0 \Longrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_2) = 0 \Longrightarrow \mu_2 = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

$$\frac{\partial}{\partial\sigma^2}\ln L(\mu_1,\mu_2,\sigma^2,\sigma^2) = 0 \Longrightarrow -\frac{m+n}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_{i=1}^{m}(x_i-\mu_1)^2 + \sum_{i=1}^{n}(y_i-\mu_2)^2\right] = 0$$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{m+n}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$\Longrightarrow \max_{\mu_1,\mu_2,\sigma^2\in\Omega_0} L(\mu_1,\mu_2,\sigma^2,\sigma^2) = \left(2\pi(\hat{\sigma}^2)\right)^{-\frac{m+n}{2}} e^{-\frac{m+n}{2}}$$

$In\ \Omega_0, to\ maximize\ L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = (2\pi{\sigma_1}^2)^{-\frac{m}{2}}(2\pi{\sigma_2}^2)^{-\frac{n}{2}} e^{-\frac{1}{2{\sigma_1}^2}[\sum_{i=1}^m (x_i-\mu_1)^2]}\ e^{-\frac{1}{2{\sigma_2}^2}[\sum_{i=1}^n (y_i-\mu_2)^2]}$

$$\frac{\partial}{\partial\mu_1}\ln L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = 0 \Longrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu_1) = 0 \Longrightarrow \mu_1 = \frac{\sum_{i=1}^m x_i}{m} = \bar{x}$$

$$\frac{\partial}{\partial\mu_2}\ln L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = 0 \Longrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu_2) = 0 \Longrightarrow \mu_2 = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

$$\frac{\partial}{\partial{\sigma_1}^2} L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = 0 \Longrightarrow -\frac{m}{2{\sigma_1}^2} + \frac{1}{2{\sigma_1}^4}\left[\sum_{i=1}^{m}(x_i-\mu_1)^2\right] = 0 \Longrightarrow \widehat{\sigma_1}^2 = \frac{1}{m}\left[\sum_{i=1}^{m}(x_i-\bar{x})^2\right]$$

$$\frac{\partial}{\partial{\sigma_2}^2} L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = 0 \Longrightarrow -\frac{n}{2{\sigma_2}^2} + \frac{1}{2{\sigma_2}^4}\left[\sum_{i=1}^{n}(y_i-\mu_2)^2\right] = 0 \Longrightarrow {\hat{\sigma}_2}^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

$$\Longrightarrow \max_{\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2\in\Omega} L(\widehat{\mu_1},\widehat{\mu_2},{\sigma_1}^2,{\sigma_2}^2) = (2\pi\widehat{\sigma_1}^2)^{-\frac{m}{2}}(2\pi\widehat{\sigma_2}^2)^{-\frac{n}{2}} e^{-\frac{m+n}{2}}$$

$$\Longrightarrow \lambda(X) = \frac{\left(2\pi(\hat{\sigma}^2)\right)^{-\frac{m+n}{2}} e^{-\frac{m+n}{2}}}{\left(2\pi\widehat{\sigma_1}^2\right)^{-\frac{m}{2}}\left(2\pi\widehat{\sigma_2}^2\right)^{-\frac{n}{2}} e^{-\frac{m+n}{2}}} = \frac{(\hat{\sigma}^2)^{-\frac{m+n}{2}}}{\left(\widehat{\sigma_1}^2\right)^{-\frac{m}{2}}\left(\widehat{\sigma_2}^2\right)^{-\frac{n}{2}}}$$

$$= \frac{\left(\frac{1}{m}\left[\sum_{i=1}^m (x_i-\bar{x})^2\right]\right)^{m/2}\left(\frac{1}{n}\left[\sum_{i=1}^n (y_i-\bar{y})^2\right]\right)^{n/2}}{\left(\frac{1}{m+n}\left[\sum_{i=1}^m (x_i-\bar{x})^2 + \sum_{i=1}^n (y_i-\bar{y})^2\right]\right)^{\frac{m+n}{2}}}$$

$$= \frac{\left(\frac{m-1}{m}{s_x}^2\right)^{m/2}\left(\frac{n-1}{n}{s_y}^2\right)^{n/2}}{\left(\frac{1}{m+n}\left[(m-1){s_x}^2 + (n-1){s_y}^2\right]\right)^{\frac{m+n}{2}}} < k$$

$$\Rightarrow \frac{\left(\frac{{s_x}^2}{{s_y}^2}\right)^{m/2}\left({s_y}^2\right)^{n/2}\left({s_y}^2\right)^{m/2}}{\left((n-1)\left[\left(\frac{m-1}{n-1}\right){s_x}^2+{s_y}^2\right]\right)^{\frac{m+n}{2}}} = \frac{\left(\frac{{s_x}^2}{{s_y}^2}\right)^{m/2}}{\left((n-1)\left[\left(\frac{m-1}{n-1}\right)\left(\frac{{s_x}^2}{{s_y}^2}\right)+1\right]\right)^{\frac{m+n}{2}}} < K$$

$$\Rightarrow \frac{\left(\left(\frac{m-1}{n-1}\right)\frac{{s_x}^2}{{s_y}^2}\right)^{m/2}}{\left(\left[\left(\frac{m-1}{n-1}\right)\left(\frac{{s_x}^2}{{s_y}^2}\right)+1\right]\right)^{\frac{m+n}{2}}} < C \Rightarrow g(x) = \frac{x^{\frac{m}{2}}}{(1+x)^{\frac{m+n}{2}}}, x = \left(\frac{m-1}{n-1}\right)\frac{{s_x}^2}{{s_y}^2}$$

$$\Rightarrow g'(x) = \frac{\frac{m}{2}x^{\frac{m}{2}-1}(1+x)^{\frac{m+n}{2}} - \frac{m+n}{2}(1+x)^{\frac{m+n}{2}-1}x^{\frac{m}{2}}}{(1+x)^{m+n}} = 0 \Rightarrow \frac{m}{2x} - \frac{m+n}{2(1+x)} = 0 \Rightarrow x_0 = \frac{m}{n}$$

$When\ x > x_0, g'(x) < 0, and\ x < x_0, g'(x) > 0$

$\Rightarrow The\ critical\ region\ is\ \left(\frac{m-1}{n-1}\right)\frac{{s_x}^2}{{s_y}^2} < x_1\ or\ \left(\frac{m-1}{n-1}\right)\frac{{s_x}^2}{{s_y}^2} > x_2 \Rightarrow \frac{{s_x}^2}{{s_y}^2} < c_1\ or\ \frac{{s_x}^2}{{s_y}^2} > c_2$

$Where\ \frac{{s_x}^2/\sigma^2}{{s_y}^2/\sigma^2} = \frac{\frac{{\mathcal{X}_{m-1}}^2}{m-1}}{\frac{{\mathcal{X}_{n-1}}^2}{n-1}} \sim F_{m-1,n-1}, and\ c_1, c_2\ will\ be\ determined\ through\ \alpha$

$\Rightarrow \alpha = P\left(\frac{{s_x}^2}{{s_y}^2} < c_1\right) + P\left(\frac{{s_x}^2}{{s_y}^2} > c_2\right), assigning\ \alpha\ equally\ to\ two\ tails\ of\ the\ F\ distribution$

$\Rightarrow c_1 = F_{1-\frac{\alpha}{2},m-1,n-1}, c_2 = F_{\frac{\alpha}{2},m-1,n-1}$

Therefore, the decision rule is:

$$Reject\ H_0, when\ \frac{{s_x}^2}{{s_y}^2} < F_{1-\frac{\alpha}{2},m-1,n-1}\ or\ \frac{{s_x}^2}{{s_y}^2} > F_{\frac{\alpha}{2},m-1,n-1}$$

<u>Approximation of generalized likelihood ratio</u>

$$\boldsymbol{X} \sim f_{\boldsymbol{\theta}}(\boldsymbol{x}); \boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$$

Test $H_0: \theta_1 = \theta_1^0, \theta_2 = \theta_2^0 \dots, \theta_r = \theta_r^0, \theta_{r+1}, \theta_{r+2}, \dots, \theta_k$ vs.

$H_1: \theta_1 \neq \theta_1^0, \theta_2 \neq \theta_2^0 \dots, \theta_r \neq \theta_r^0, \theta_{r+1}, \theta_{r+2}, \dots, \theta_k$

According to the likelihood ratio test, we will reject H_0 if:

$$\lambda(\boldsymbol{X}) = \frac{\max L(\boldsymbol{\theta}, \boldsymbol{X}), \boldsymbol{\theta} \in \Omega_0}{\max L(\boldsymbol{\theta}, \boldsymbol{X}), \boldsymbol{\theta} \in \Omega} < k$$

Where Ω_0 has dimension of $k-r$, and Ω is the full parameter space with dim = k

Since $\lambda(\boldsymbol{X})$ is always smaller than 1, so $\log \lambda(\boldsymbol{X}) < 0$, therefore,

$$\lambda(\boldsymbol{X}) < k \Rightarrow \log \lambda(\boldsymbol{X}) < -C \Rightarrow -\log \lambda(\boldsymbol{X}) > C, for\ C > 0$$

The LRT, which reject $\lambda(\boldsymbol{X})$ when it is very small, equivalently it is to reject H_0 if $-\log\lambda(\boldsymbol{X})$ is large, i.e.:

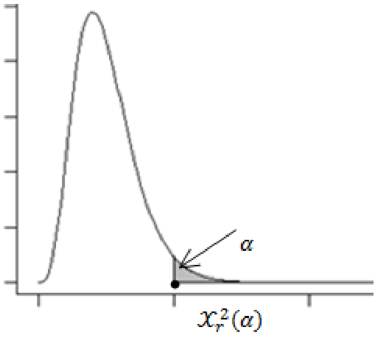
$$-2\log\lambda(\boldsymbol{X}) > C$$

When n is very large, $-2\log\lambda(\boldsymbol{X})$ has an approximately chi-square distribution:

$$-2\log\lambda(\boldsymbol{X}) \sim \mathcal{X}_r^2$$

So, for the large n case, we reject H_0 if:

$$-2\log\lambda(\boldsymbol{X}) > \mathcal{X}_{r,\alpha}^2$$



Revisit Example 1 by using Approximation approach:

Solution: $It\ is\ much\ easier\ to\ solve\ this\ by\ using\ the\ approaximation\ approach.$

$First\ reparameterize$: $\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_2, \theta_3 = {\sigma_1}^2, \theta_4 = {\sigma_2}^2, k = 4$

Test $H_0: \theta_1 = {\theta_1}^0 = 0, \theta_2, \theta_3, \theta_4\ unspecified$

$H_1: \theta_1 \neq {\theta_1}^0 \neq 0, \theta_2, \theta_3, \theta_4\ unspecified$

$r = 1\ in\ this\ case$

Therefore, the decision rule is:

$$Reject\ H_0, when -2\log\lambda(\boldsymbol{X}) > {\mathcal{X}_1}^2(\alpha)$$

$By\ result\ from\ Example\ 1$

$$\lambda(\boldsymbol{X}) = \frac{\left(\frac{e^{-1}}{2\pi\hat{\sigma}^2}\right)^{\frac{m+n}{2}}}{\left(\frac{e^{-1}}{2\pi\hat{\sigma}^{*2}}\right)^{\frac{m+n}{2}}} = \left(\frac{2\pi\hat{\sigma}^2}{2\pi\hat{\sigma}^{*2}}\right)^{-\frac{m+n}{2}}$$

$$\Rightarrow -2\log\lambda(\boldsymbol{X}) = -2\left(-\frac{m+n}{2}\log\left(\frac{\hat{\sigma}^2}{\hat{\sigma}^{*2}}\right)\right)$$

$$= (m+n)\left[\log\left(1 + \frac{\frac{mn}{m+n}(\bar{x}-\bar{y})^2}{\sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2}\right)\right] > {\mathcal{X}_1}^2(\alpha)$$

$$\Rightarrow Reject\ H_0, when \log\left(1 + \frac{(\bar{x}-\bar{y})^2}{\frac{(m-1)}{m}{s_x}^2 + \frac{(n-1)}{n}{s_y}^2}\right) > \frac{{\mathcal{X}_1}^2(\alpha)}{m+n}$$

Lecture 46

Confidence Intervals for mean

Let X has pdf $f_\theta(x)$, where θ is unkown, classical Confidence Interval of θ is defined as

$$\left(T_1(\boldsymbol{X}), T_2(\boldsymbol{X})\right)$$

Where $P[T_1(\boldsymbol{X}), \leq \theta \leq T_2(\boldsymbol{X})] = \gamma = 1 - \alpha$

Remarks:

- γ is always fixed, and is called confidence level, $\gamma \in [0,1]$
- E.g., $\gamma = 90\%$ means "I am 90% confident that θ is between $T_1(\boldsymbol{X}), T_2(\boldsymbol{X})$"

Pivotal Quantity

Q is a function of $\boldsymbol{X}, \theta$, but its distribution is independent of θ

$\mathbf{Example\ 20.1}$: $Let\ X_1, \dots, X_n\ to\ be\ N(\theta, \sigma^2), where\ \sigma^2\ is\ known$

$$\Rightarrow Z = \frac{\bar{X} - \theta}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), not\ dependent\ on\ \theta, which\ is\ a\ Pivotal\ Quantity$$

$\mathbf{Example\ 20.2}$: $Let\ X_1, \dots, X_n\ to\ be\ N(\theta, \sigma^2), where\ \sigma^2\ is\ unknown$

$$\Rightarrow T = \frac{\bar{X} - \theta}{\frac{S}{\sqrt{n}}} \sim t_{n-1}, not\ dependent\ on\ \theta, which\ is\ a\ Pivotal\ Quantity$$

$\mathbf{Example\ 20.3}$: $Let\ X_1, \dots, X_n\ to\ be\ N(\mu, \theta^2), where\ \mu\ is\ unknown$

$$\Rightarrow Q = \frac{(n-1)S^2}{\theta^2} \sim \mathcal{X}_{n-1}{}^2, not\ dependent\ on\ \theta, which\ is\ a\ Pivotal\ Quantity$$

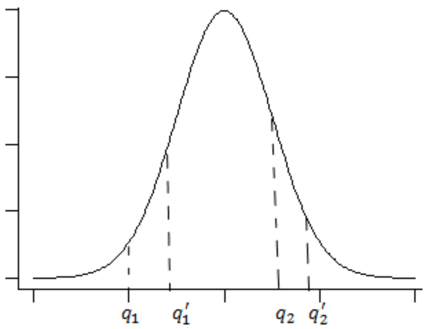
Let q_1, q_2 be the cutoff points for the Pivotal Quantity, we have:

$$P[q_1 \leq Q(\boldsymbol{X}, \theta) \leq q_2] = \gamma \Leftrightarrow P[T_1(\boldsymbol{X}) \leq \theta \leq T_2(\boldsymbol{X})]\gamma$$

$In\ Example\ 1\ above, we\ will\ have\ the\ construction\ of\ Confidence\ Interval$:

$$P[q_1 \leq Q(\boldsymbol{X}, \theta) \leq q_2] = P\left[q_1 \leq \frac{\bar{X} - \theta}{\frac{\sigma}{\sqrt{n}}} \leq q_2\right] = \gamma \Rightarrow P\left[\bar{X} - q_2\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{X} - q_1\frac{\sigma}{\sqrt{n}}\right] = \gamma\ (*)$$

How to determine $q_1$and q_2?



For $\gamma = 0.8$, there can be many choices of q_1, q_2 (as in the graph, q_1, q_2 and q_1', q_2').

But for the symmetric distribution like standard normal distribution, the best choice exists!

Shortest Confidence Interval

Continue with Example 1, from the construction of Confidence Interval above, we have the CI length:

$$L = \left(\bar{X} - q_1 \frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - q_2 \frac{\sigma}{\sqrt{n}}\right) = \frac{\sigma}{\sqrt{n}}(q_2 - q_1)$$

$$In\ order\ to\ minimize\ L, with\ constraint\ \int_{q_1}^{q_2} f_Z(t)dt = \gamma$$

$$Since\ \gamma\ is\ a\ constant, \frac{\partial}{\partial q_1}\left[\int_{q_1}^{q_2} f_Z(t)dt\right] = 0 \Longrightarrow \frac{\partial}{\partial q_1}[F_Z(q_2) - F_Z(q_1)] = 0$$

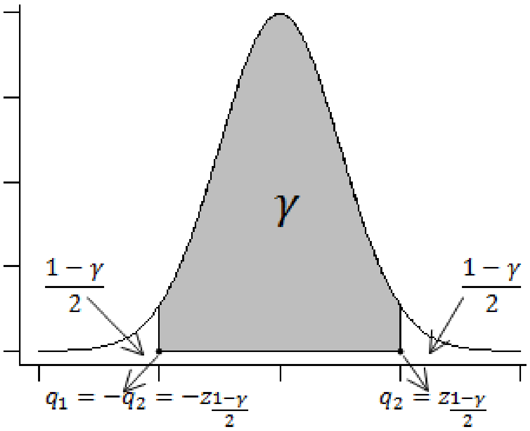
$$\Longrightarrow \frac{\partial}{\partial q_1} F_Z(q_2) - f_Z(q_1) = 0 \Longrightarrow f_Z(q_2)\frac{\partial q_2}{\partial q_1} - f_Z(q_1) = 0$$

$$Combine\ with\ the\ condition\ \min L, we\ will\ have\ the\ function\ sets:$$

$$\Longrightarrow \begin{cases} \dfrac{\partial}{\partial q_1}\left[\displaystyle\int_{q_1}^{q_2} f_Z(t)dt\right] = 0 \Longrightarrow f_Z(q_2)\dfrac{\partial q_2}{\partial q_1} - f_Z(q_1) = 0 \Longrightarrow \dfrac{\partial q_2}{\partial q_1} = \dfrac{f_Z(q_1)}{f_Z(q_2)} \\ \dfrac{\partial}{\partial q_1} L = 0 \Longrightarrow \dfrac{\partial}{\partial q_1}\left[\dfrac{\sigma}{\sqrt{n}}(q_2 - q_1)\right] = 0 \Longrightarrow 1 - \dfrac{\partial q_2}{\partial q_1} = 0 \Longrightarrow \dfrac{\partial q_2}{\partial q_1} = 1 \end{cases}$$

$$\Longrightarrow \frac{f_Z(q_1)}{f_Z(q_2)} = 1 \Longrightarrow f_Z(q_2) = f_Z(q_1)$$

$$For\ symmetric\ distribution, we\ have\ the\ property\ that\ f_Z(-x) = f_Z(x)$$



$$\Longrightarrow f_Z(q_2) = f_Z(q_1) = f_Z(-q_2) = f_Z(-q_1)$$

$$\Longrightarrow q_1 = -q_2$$

Base on the relationship of q_1, q_2, we have:

Therefore, the γ CI for θ is:

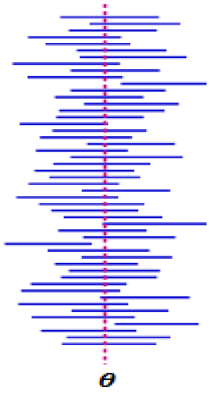
$$\left[\bar{X} - z_{\frac{1-\gamma}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{1-\gamma}{2}} \frac{\sigma}{\sqrt{n}}\right] \ (*)$$

Example 20.4: $n = 9, \bar{x} = 70, \sigma = 5, \gamma = 0.9, find\ CI\ for\ this\ practical\ problem$

Solution: Plug the number into $(*)$, we have the CI for θ is:

$$\left[70 - 1.645 \cdot \frac{5}{3}, 70 + 1.645 \cdot \frac{5}{3}\right] = [67,73] = 70 \pm 3$$

<u>Interpretation of Confidence Interval</u>



As shown in the graph besides:

Each horizontal line represents a CI from one sample, if we repeat sampling for many times , than $\gamma \cdot 100\%$ of times the intervals will contain the true parameter θ

Example: $Let\ X_1, \dots, X_n\ to\ be\ N(\theta, \sigma^2), where\ \sigma^2\ is\ unknown, find\ the\ CI$

Solution: $Shown\ in\ example\ 2, T = \dfrac{\bar{X} - \theta}{\frac{S}{\sqrt{n}}}\ is\ Pivotal\ Quantity$

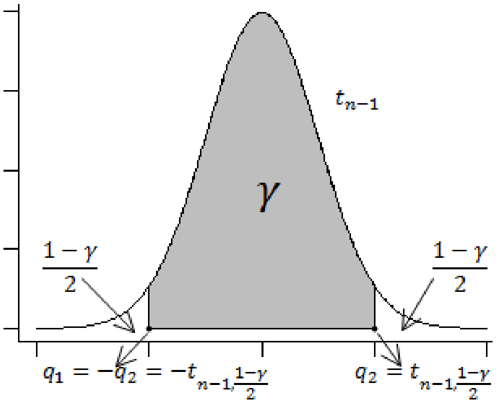
$$\Rightarrow P[q_1 \le Q(\boldsymbol{X}, \theta) \le q_2] = P\left[q_1 \le \frac{\bar{X} - \theta}{\frac{S}{\sqrt{n}}} \le q_2\right] = \gamma$$

$$\Rightarrow P\left[\bar{X} - q_2 \frac{S}{\sqrt{n}} \le \theta \le \bar{X} - q_1 \frac{S}{\sqrt{n}}\right] = \gamma$$

$$Minimize\ L = \frac{S}{\sqrt{n}}(q_2 - q_1), with\ constraint\ \int_{q_1}^{q_2} f_T(t)dt = \gamma$$

$$\Rightarrow \begin{cases} \dfrac{\partial}{\partial q_1}\left[\displaystyle\int_{q_1}^{q_2} f_T(t)dt\right] = 0 \Rightarrow f_T(q_2)\dfrac{\partial q_2}{\partial q_1} - f_T(q_1) = 0 \Rightarrow \dfrac{\partial q_2}{\partial q_1} = \dfrac{f_T(q_1)}{f_T(q_2)} \\ \dfrac{\partial}{\partial q_1} L = 0 \Rightarrow \dfrac{\partial}{\partial q_1}\left[\dfrac{\sigma}{\sqrt{n}}(q_2 - q_1)\right] = 0 \Rightarrow 1 - \dfrac{\partial q_2}{\partial q_1} = 0 \Rightarrow \dfrac{\partial q_2}{\partial q_1} = 1 \end{cases}$$

$$\Rightarrow \frac{f_Z(q_1)}{f_Z(q_2)} = 1 \Rightarrow f_Z(q_2) = f_Z(q_1)$$



$Student\ t\ is\ also\ a\ symmetric\ distribution, we\ have\ the\ property\ that\ f_Z(-x) = f_Z(x)$

$\Longrightarrow f_Z(q_2) = f_Z(q_1) = f_Z(-q_2) = f_Z(-q_1)$

$\Longrightarrow q_1 = -q_2$

Therefore, the γ CI for θ is:

$$\left[\bar{X} - t_{n-1,\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{1-\gamma}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Lecture 47

Confidence Intervals for variance

$Let\ X_1, \dots, X_n\ be\ i.i.d. N(\theta, \sigma^2), where\ \sigma^2\ is\ unknown. Construct\ Confidence\ Interval\ for\ \sigma^2$ $with\ \gamma\ confidence$

Solution: $Shown\ in\ example\ 3, \frac{(n-1)S^2}{\sigma^2} \sim \mathcal{X}_{n-1}{}^2\ is\ Pivotal\ Quantity$

$$\Longrightarrow P[q_1 \le Q(\boldsymbol{X}, \theta) \le q_2] = P\left[q_1 \le \frac{(n-1)S^2}{\sigma^2} \le q_2\right] = \gamma \Longrightarrow P\left[\frac{(n-1)S^2}{q_2} \le \theta \le \frac{(n-1)S^2}{q_1}\right] = \gamma$$

Since Chi-square distribution is not symmetric, instead of following the shortest CI rule, we generally choose q_1 and q_2 to be:

$$P\left[\mathcal{X}_{n-1}{}^2 > q_2\right] = \frac{1-\gamma}{2};\ P\left[\mathcal{X}_{n-1}{}^2 < q_1\right] = \frac{1-\gamma}{2}$$

However, if we follow the shortest CI rule,

$$Minimize\ L = \frac{(n-1)S^2}{q_1} - \frac{(n-1)S^2}{q_2}, with\ constraint \int_{q_1}^{q_2} f_Q(t)dt = \gamma$$

$$\Longrightarrow \begin{cases} \frac{\partial}{\partial q_1}\left[\int_{q_1}^{q_2} f_Q(t)dt\right] = 0 \Longrightarrow f_Q(q_2)\frac{\partial q_2}{\partial q_1} - f_Q(q_1) = 0 \Longrightarrow \frac{\partial q_2}{\partial q_1} = \frac{f_Q(q_1)}{f_Q(q_2)} \\ \frac{\partial}{\partial q_1}L = 0 \Longrightarrow \frac{\partial}{\partial q_1}\left[\frac{1}{q_1} - \frac{1}{q_2}\right] = 0 \Longrightarrow \frac{\partial q_2}{\partial q_1} = \frac{q_2{}^2}{q_1{}^2} \end{cases}$$

$\Longrightarrow f_Q(q_1)q_1{}^2 - f_Q(q_2)q_2{}^2 = 0$

Therefore, the best choices of q_1, q_2 is to choose them such that:

$$f_Q(q_1)q_1{}^2 - f_Q(q_2)q_2{}^2 = 0\ subject\ to \int_{q_1}^{q_2} f_Q(t)dt = \gamma$$

Which can be solved by numerical analysis