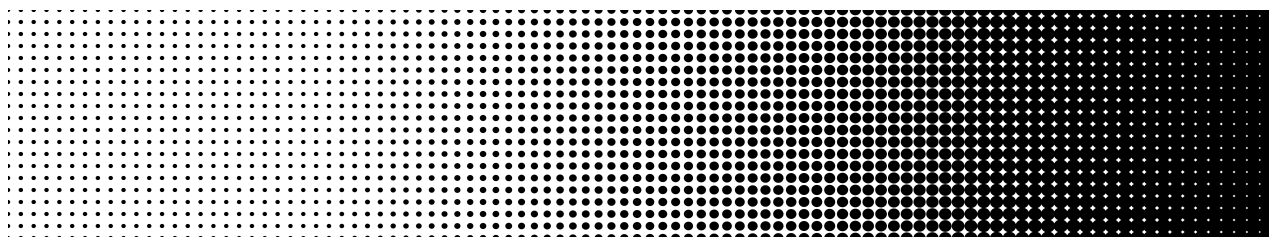


# Mathematical Statistics

pehcy (MurphyShark)    <https://github.com/pehcy>



Based on lectures UECM 3363 Mathematical Statistics, UECM 3253 Nonparametric Statistics,  
UCCM 2263 Applied Statistical Modelling in Universiti Tunku Abdul Rahman in 2019  
Notes taken by MurphyShark.

# Notations in this notes

Here are a list of symbols you will see a lot throughout this notes:

$\mu_X = \mathbb{E}[X]$  represents the mean or expectation of a random variable  $X$ .

$\sigma_X^2 = \text{Var}[X]$  represents the variance of a random variable  $X$ .

$\sigma = \sqrt{\text{Var}[X]}$  represents the standard deviation of a random variable  $X$ .

$\sigma_{XY} = \text{Cov}(X, Y)$  represents the covariance of two random variables  $X$  and  $Y$ .

$\rho_{XY} = \text{Corr}(X, Y)$  represents the correlation coefficient of two random variables  $X$  and  $Y$ .

$\mathbb{P}[X]$  denotes the probability measures of event  $X$ .

$\hat{\beta}$  denotes the estimated value for the unknown parameter  $\beta$ .

# Set Theory and Counting Techniques

Two approaches of the concept of probability will introduced lately in this notes: The classical probability and the experimental probability. The sample space of probability theory is developed using the foundation of set theory.

In set theory, the number of elements in a set has a special name. It is called the **cardinality** of the set. In this notes we write  $\#(A)$  to represent the cardinality of the set  $A$ .

## 1.1 Notion of sets


Set is a basic term in mathematics. One can think of a set to be a *class* or *collection*.

### Definition 1.1 Sets

A set is a collection of objects called elements or numbers.

### Definition 1.2 Empty set

The empty set is the set with no elements, denoted as  $\emptyset$  or  $\{\}$ .

 **Example 1.1.1.** Throughout this notes, we are using the following number systems.

- ❖ The set of all nonzero positive integers, known as natural numbers

$$\mathbb{N} = \{1, 2, 3, \dots\}.$$

- ❖ The set of all integers, regardless positive or negative,

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

- ❖ The set of all rational numbers, the numbers that can be express as ratio of two integers

$$\mathbb{Q} = \left\{ \frac{p}{q} : p, q \in \mathbb{Z} \text{ with } q \neq 0 \right\}$$

- ❖ The set of all real numbers,

$$\mathbb{R} = \{\mathbb{Q} \text{ along with all irrational numbers such as } \sqrt{2}, \pi, e, \dots\}.$$

### Definition 1.3 Subsets

A set  $A$  is a subset of  $B$ ,  $A \subset B$ , if  $a \in A \implies a \in B$ .

 **Example 1.1.2.** What is the cardinality of each of the following sets?

1.  $\emptyset$ .
2.  $\{\emptyset\}$ .
3.  $A = \{\heartsuit, \{\clubsuit\}, \{\heartsuit, \{\heartsuit\}\}\}$ .
4.  $B = \{x | x \text{ is an integer such that } 2x^2 + 3x - 2 = 0\}$ .

⇒ **Solution** 1. Certainly,  $\#(\emptyset) = 0$ .

2. This is a set consists of one element  $\emptyset$ , thus  $\#(\{\emptyset\}) = 1$ .

3. The set  $A$  consists of three elements which are  $\boxed{\heartsuit}$ ,  $\boxed{\{\clubsuit\}}$ , and  $\boxed{\{\heartsuit, \{\heartsuit\}\}}$ . So  $\#(A) = 3$ .

4. Solving the quadratic equation,

$$\begin{aligned} 2x^2 + 3x - 2 = 0 &\implies (2x - 1)(x + 2) = 0 \\ &\implies x = -2 \quad \text{or} \quad x = \frac{1}{2}. \end{aligned}$$

Only  $x = -2$  is integer. Thus  $\#(B) = \#(\{-2\}) = 1$ .



#### Definition 1.4 Power Set

If  $A$  is a set, then  $\mathcal{P}(A) = \{B : B \subset A\}$ .

📌 **Example 1.1.3.** If  $A = \emptyset$ , then  $\mathcal{P}(A) = \{\emptyset\}$ . The power set has only one single element, that is the empty set itself.

## 1.2 Set Operations

### Theorem 1.1 Properties of set union and intersect

For any three sets,  $A$ ,  $B$ , and  $C$ .

S1 [Commutativity]

$$A \cup B = B \cup A,$$

$$A \cap B = B \cap A.$$

S2 [Associativity]

$$A \cup (B \cup C) = (A \cup B) \cup C,$$

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

S3 [Distributive]

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

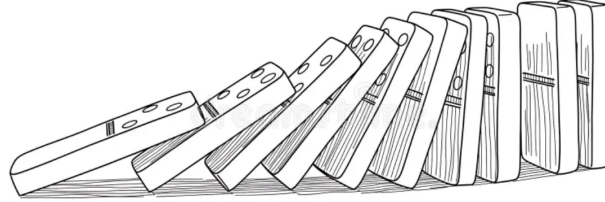
S4 [DeMorgan's law]

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

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

### 1.2.1 Proof by induction

Mathematical induction is a proof technique for proving that a statement is true for all natural numbers. It works by first proving a base case (the first natural number, often 1) and then proving an inductive step which shows that if the statement holds for some arbitrary integer  $k$ .



#### Theorem 1.2 Mathematical Induction

Let  $P(n)$  be a statement relies on  $n \in \mathbb{N}$ . Assume that

1. (Base case)  $P(1)$  is true.
2. (Induction step) If  $P(m)$  is true, then  $P(m + 1)$  is true.

*Proof.* Let

$$S = \{n \in \mathbb{N} : P(n) \text{ is not true}\}.$$

We want to show  $S$  is empty. We will show this by contradiction.

**Intuition:** By assuming  $S \neq \emptyset$  and derive a false statement. The rules of logic  $p \rightarrow q$  (If  $p$  then  $q$ ) say that  $S \neq \emptyset$  is false.

For the sake of contradiction, suppose  $S \neq \emptyset$ . By Well Ordering Principle of natural numbers,  $S$  has a least element  $x \in S$ . Given that the basis step  $P(1)$  is true, then  $1 \notin S \implies x \neq 1$ . In particular,  $x > 1$ .

Since  $x$  is the least element of  $S$ , and  $x - 1 < x$ , thus  $x - 1 \notin S$ . From  $S \neq \emptyset$ , we are now arriving to the conclusion that  $\exists x \in \mathbb{N}$  such that  $x \in S$  and  $x \notin S$  at the same time. This is a contradiction. Therefore  $S$  must be empty.  $\square$

### 1.3 Counting Techniques

#### Theorem 1.3 Multiplication rule of counting

If a choice consists of  $k$  steps, of which the first can be made in  $n_1$  ways, for each of these the second can be made in  $n_2$  ways and so on. For each of these  $k$ -th can be made in  $n_k$  ways, then the whole choice can be made in

$$n_1 \times n_2 \times \cdots \times n_k$$

ways.

*Proof.* In the notion of set theory, we use  $S_i$  to represent the set of outcomes for the  $i$ -th step for all

$i = 1, 2, \dots, k$ . Then  $\#(S_i) = n_i$ . The set of outcomes for the entire job is the Cartesian product

$$S_1 \times S_2 \times \cdots \times S_k = \{(s_1, s_2, \dots, s_k) : s_i \in S_i, \quad 1 \leq i \leq k\}. \quad (1.1)$$

Thus, we just need to show that the number of outcomes is equal to the product of number of choices for each step. That is,

$$\#(S_1 \times S_2 \times \cdots \times S_k) = \#S_1 \cdot \#S_2 \cdots \#S_k \quad (1.2)$$

**[Basis Step]** By Theorem 1.2.5, we have  $\#(S_1 \times S_2) = \#(S_1) \times \#(S_2)$ . Thus, the property is true for  $n = 2$ .

**[Induction Hypothesis]** Suppose

$$\#(S_1 \times S_2 \times \cdots \times S_k) = \#(S_1) \cdot \#(S_2) \cdots \#(S_k)$$

for  $k = 2, 3, \dots, n$ .

**[Induction Step]** We must show

$$\#(S_1 \times S_2 \times \cdots \times S_{n+1}) = \#(S_1) \cdot \#(S_2) \cdots \#(S_{n+1}).$$

To see this, note that there is a one-to-one correspondence between the sets  $S_1 \times S_2 \times \cdots \times S_{n+1}$  and  $(S_1 \times S_2 \times \cdots \times S_n) \times S_{n+1}$  given by  $f(s_1, s_2, \dots, s_n, s_{n+1}) = ((s_1, s_2, \dots, s_n), s_{n+1})$ . See Problem 1.2.18. Thus,

$$\#(S_1 \times S_2 \times \cdots \times S_{n+1}) = \#((S_1 \times S_2 \times \cdots \times S_n) \times S_{n+1}) = \#(S_1 \times S_2 \times \cdots \times S_n) \#(S_{n+1})$$

(by Theorem 1.2.5). Now, applying the induction hypothesis gives

$$\#(S_1 \times S_2 \times \cdots \times S_n \times S_{n+1}) = \#(S_1) \cdot \#(S_2) \cdots \#(S_{n+1}).$$

□

## Tutorials

**Exercise 1.1** Use mathematical induction to show that for any positive integer  $n$ ,  $6^n - 1$  is divisible by 5.

**Exercise 1.2** Show that  $n! > 3^n$  for  $n \geq 7$ .

**Exercise 1.3** Consider the Fibonacci sequence  $\{x_n\}_{n=1}^{\infty}$ , defined the relations  $x_1 = 1, x_2 = 1$  and

$$x_n = x_{n-1} + x_{n-2} \quad \text{for } n \geq 3.$$

Use mathematical induction in order to show that for  $n \geq 1$ .

$$x_n = \frac{1}{\sqrt{5}} \left[ \left( \frac{1 + \sqrt{5}}{2} \right)^n - \left( \frac{1 - \sqrt{5}}{2} \right)^n \right]$$

# Random Variables

## 2.1 Density function

By definition, a random variable  $X$  is a function with domain the sample space and range a subset of the real numbers. For example, in rolling two dice  $X$  might represent the sum of the points on the two dice. Similarly, in taking samples of college students  $X$  might represent the number of hours per week a student studies, a student's GPA, or a student's height. The notation  $X(s) = x$  means that  $x$  is the value associated with the outcome  $s$  by the random variable  $X$ .

There are three types of random variables: discrete random variables, continuous random variables, and mixed random variables.

**Example 2.1.1.** A committee of 4 is selected from a group consisting of 5 men and 5 women. Let  $X$  be the random variable that represents the number of women in the committee. Find the probability mass distribution of  $X$ .

**Solution** For  $x = 0, 1, 2, 3, 4$  we have

$$p_X(x) = \frac{\binom{5}{x} \binom{5}{4-x}}{\binom{10}{4}} \quad x = 0, 1, 2, 3, 4.$$

The probability mass function can be described by the table

$x$	0	1	2	3	4
$p(x)$	$\frac{5}{210}$	$\frac{50}{210}$	$\frac{100}{210}$	$\frac{50}{210}$	$\frac{5}{210}$



## 2.2 Cumulative Distribution

First, we prove that the probability is a continuous set function. In order to do that, we need the following definitions:

### Definition 2.1 Increasing and Decreasing sequence of events

A sequence of sets  $\{E_n\}_{n=1}^{\infty}$  is said to be increasing if

$$E_1 \subset E_2 \subset \dots \subset E_n \subset E_{n+1} \subset \dots$$

whereas it is said to be a decreasing sequence if

$$E_1 \supset E_2 \supset \dots \supset E_n \supset E_{n+1} \supset \dots$$

**Lemma 2.1**

If  $\{E_n\}_{n \geq 1}$  is either an increasing or decreasing sequence of events then

$$\lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n \rightarrow \infty} E_n]. \quad (2.1)$$

that is

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for increasing sequence,} \quad (2.2)$$

and

$$\mathbb{P}\left[\bigcap_{n=1}^{\infty} E_n\right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for decreasing sequence,} \quad (2.3)$$

*Proof.* Firstly, suppose that  $E_n \subset E_{n+1}$  for all  $n \geq 1$ . Define the events

$$\begin{aligned} F_1 &= E_1 \\ F_n &= E_n \cap E_{n-1}^c, \quad n > 1 \end{aligned}$$

Note that for  $n > 1$ ,  $F_n$  consists of those outcomes in  $E_n$  that are not in any of the earlier  $E_n$   $\forall i < n$ . Clearly, for  $i \neq j$  we have  $F_i \cap F_j = \emptyset$ . Also,  $\bigcup_{n=1}^{\infty} F_n = \bigcup_{n=1}^{\infty} E_n$  and for  $n \geq 1$  we have  $\bigcup_{i=1}^n F_i = \bigcup_{i=1}^n E_i$ . From these properties we have

$$\begin{aligned} \mathbb{P}\left[\lim_{n \rightarrow \infty} E_n\right] &= \mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} F_n\right] \\ &= \sum_{n=1}^{\infty} \mathbb{P}[F_n] \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbb{P}[F_i] \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left[\bigcup_{i=1}^n F_i\right] \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left[\bigcup_{i=1}^n E_i\right] \\ &= \lim_{n \rightarrow \infty} \mathbb{P}[E_n]. \end{aligned}$$

On the other hand, now suppose that the sequence  $\{E_n\}_{n \geq 1}$  is a decreasing sequence of events. Then  $\{E_n^c\}_{n \geq 1}$  is an increasing sequence of events. Hence, from the previous part we have

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n^c\right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n^c].$$

By De Morgan's law we have  $\bigcup_{n=1}^{\infty} E_n^c = (\bigcap_{n=1}^{\infty} E_n)^c$ . And

$$\mathbb{P}\left[\left(\bigcap_{n=1}^{\infty} E_n\right)^c\right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n^c].$$



Equivalently,

$$1 - \mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} (1 - \mathbb{P}[E_n]) = 1 - \lim_{n \rightarrow \infty} \mathbb{P}[E_n]$$

or

$$\mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n].$$

□

### Theorem 2.1 Properties of Cumulative Distribution Function

If  $F_X(x)$  is a cumulative distribution function, then

1.  $F_X(-\infty) = \lim_{x \downarrow -\infty} F_X(x) = 0$ .
2.  $F_X(+\infty) = \lim_{x \rightarrow +\infty} F_X(x) = 1$ .
3.  $F_X(x)$  is always *monotonically increasing*. That said, if  $x_1 < x_2$ , then  $F_X(x_1) < F_X(x_2)$ .

*Proof.* 1. Note that  $\lim_{x \downarrow -\infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$  where  $\{x_n\}$  is a decreasing sequence such that  $x_n \downarrow -\infty$ . Define

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain  $E_1 \supseteq E_2 \supseteq \dots$ . Moreover,

$$\emptyset = \bigcap_{n=1}^{\infty} E_n.$$

By previous proposition, we find

$$\lim_{x \rightarrow -\infty} F(x) = \lim_{n \downarrow -\infty} F(x_n) = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P} \left[ \bigcap_{n=1}^{\infty} E_n \right] = \mathbb{P}[\emptyset] = 0.$$

2. In the other hand, suppose that  $\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$  where  $\{x_n\}$  is a increasing sequence such that  $x_n \rightarrow \infty$ . We reuse back the definition of  $E_n$  that is

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain in the opposite direction  $E_1 \subseteq E_2 \subseteq \dots$ . Moreover,

$$\Omega = \bigcup_{n=1}^{\infty} E_n$$

By previous proposition, we find

$$\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} F(x_n) = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P} \left[ \bigcup_{n=1}^{\infty} E_n \right] = \mathbb{P}[\Omega] = 1.$$

3. Consider two real numbers  $a, b$  such that  $a < b$ . Then

$$\{s \in \Omega : X(s) \leq a\} \subset \{s \in \Omega : X(s) \leq b\}.$$


This implies that  $\mathbb{P}[X \leq a] < \mathbb{P}[X \leq b]$ . Hence,  $F(a) < F(b)$ .

□

 **Example 2.2.1.** Let  $X$  be a random variable with probability density function

$$f_X(x) = \begin{cases} 2 - 4|x| & \text{if } \frac{1}{2} < x < -\frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

1. Find the variance of  $X$ .
2. Find the cumulative function  $F(x)$  of  $X$ .

 **Solution** 1. Since the density function  $f(x)$  is odd in  $(-\frac{1}{2}, \frac{1}{2})$ , we have  $\mathbb{E}[X] = 0$ . Therefore

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[X^2] - 0 = \int_{-1/2}^0 x^2(2 + 4x) \, dx + \int_0^{1/2} x^2(2 - 4x) \, dx \\ &= \frac{1}{24}. \end{aligned}$$

2. The cumulative function is


$$\begin{aligned} F(x) &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ \int_{-1/2}^x (2 + 4t) \, dt & \text{if } -\frac{1}{2} \leq x \leq 0 \\ \int_{-1/2}^0 (2 + 4t) \, dt + \int_0^x (2 - 4t) \, dt & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \\ &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ 2x^2 + 2x + \frac{1}{2} & \text{if } -\frac{1}{2} \leq x \leq 0 \\ -2x^2 + 2x + \frac{1}{2} & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \end{aligned}$$

◀


## 2.3 Percentiles and Quantiles

### 2.3.1 Mode

In the discrete case, the mode is the value that is most likely to be sampled. In the continuous case, the mode is where  $f(x)$  is at its peak.

 **Example 2.3.1.** The lifetime of a light bulb has density function,  $f_X$ , where  $f_X(x)$  is proportional to  $\frac{x^2}{1 + x^3}$ ,  $0 < x < 5$ , and 0 otherwise.

Calculate the mode of this distribution.

 **Solution** Given the lifetime of a light bulb  $X$  has density function

$$f_X(x) = \frac{cx^2}{1 + x^3}.$$

Compute the first and second order derivative of  $f$ .

$$\frac{df}{dx} = \frac{(1+x^3) \frac{d}{dx}(cx^2) - cx^2 \frac{d}{dx}(1+x^3)}{(1+x^3)^2} = \frac{2cx - cx^4}{(1+x^3)^2}.$$

$$\begin{aligned} \frac{d^2f}{dx^2} &= \frac{(1+x^3)^2 \frac{d}{dx}(2cx - cx^4) - (2cx - cx^4) \frac{d}{dx}(1+x^3)^2}{(1+x^3)^4} \\ &= \frac{(1+x^3)^2(2c - 4cx^3) - (2cx - cx^4)2(1+x^3)(3x^2)}{(1+x^3)^4} \\ &= \frac{2c(1+x^3)(1-2x^3-3x^5)}{(1+x^3)^4} \\ &= \frac{2c(1-2x^3-3x^5)}{(1+x^3)^3}. \end{aligned}$$

◀

By inspection,  $\frac{d^2f}{dx^2} < 0$ . And so  $\frac{df}{dx} = 0$  is maximum point in  $(0, 5)$ . Solve for  $x$  of the following equation:

$$\frac{2cx - cx^4}{(1+x^3)^2} = 0$$

Since  $(1+x^3)^2 > 0$ , we can remove it safely from the equation. Then

$$\begin{aligned} 2cx - cx^4 = 0 &\implies x^4 - 2x = 0 \\ &\implies x(x^3 - 2) = 0 \\ &\implies x = 0 \text{ or } x = \sqrt[3]{2}. \end{aligned}$$

Since  $x$  cannot be zero, thus the mode is  $\sqrt[3]{2} = 1.26$ .

## 2.4 Expected Value and Moments

For a random variable  $X$ , the expected value is denoted  $\mathbb{E}[X]$ , or  $\mu_X$  or simply  $\mu$ . The expected value is called the expectation of  $X$ , which is the "average" over the range of values that distribution  $X$  can be. You may said the expectation is the "center" of the distribution.

### Definition 2.2 Expectation value

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $E \subseteq \mathbb{R}$  be countable, and let  $X$  be a  $E$ -valued random variable on  $(\Omega, \mathbb{P})$ . The expectation of  $X$ , if it exists, is defined by

$$\mathbb{E}[X] := \sum_{e \in E} e f_X(e). \quad (2.4)$$

### Lemma 2.2

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $E \subseteq \mathbb{R}$  be countable set and let  $X$  be an  $E$ -valued random variable on  $(\Omega, \mathbb{P})$ . Then

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \mathbb{P}[\{\omega\}].$$

*Proof.* Recall that

$$\Omega = \bigcup_{e \in E} \{X = e\}$$

and the events  $\{X = e\}$  are mutually exclusive. Hence

$$\begin{aligned} \sum_{\omega \in \Omega} X(\omega) \mathbb{P}[\{\omega\}] &= \sum_{e \in E} \sum_{\omega \in \{X=e\}} X(\omega) \mathbb{P}[\{\omega\}] \\ &= \sum_{e \in E} e \mathbb{P}\{X = e\} \\ &= \sum_{e \in E} e f_X(e) \end{aligned}$$

as what we expected. □

**Example 2.4.1.** Let  $X$  be a random variable representing the value shown a fair six-sided die is rolled. Then  $X \sim \text{discreteU}(\{1, 2, 3, 4, 5, 6\})$ , and  $f_X(k) = \frac{1}{6}$  for each number.

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

So the expected value of a die rolled is 3.5.

**Example 2.4.2.** If  $f(x) = (k+1)x^2$  for  $0 < x < 1$ , find the moment generating function.

**Solution** Since  $f$  is a density function, thus  $\int_0^1 f(x) dx = 1$ , it follows that

$$(k+1) \times \frac{1}{3} = 1$$

so that  $k = 2$  and now  $f(x) = 3x^2$  for  $0 < x < 1$ . Then the moment generating function is

$$\begin{aligned} M_X(t) &= \int_0^1 e^{tx} (3x^2) dx = \int_0^1 3x^2 d\left(\frac{e^{tx}}{t}\right) \\ &= \frac{3x^2 e^{tx}}{t} \Big|_{x=0}^{x=1} - \int_0^1 \frac{6x e^{tx}}{t} dx \\ &= \frac{3e^t}{t} - \left[ \frac{6x e^{tx}}{t^2} \Big|_{x=0}^{x=1} - \int_0^1 \frac{6x e^{tx}}{t^2} dx \right] \\ &= \frac{3e^t}{t} - \frac{6e^t}{t^2} + \frac{6(e^t - 1)}{t^3} \\ &= \frac{e^t(6 - 6t + 3t^2)}{t^3} - \frac{6}{t^3}. \end{aligned}$$

◀

## 2.4.1 Variance

Variance is a measure of the "dispersion" of  $X$  about the mean.

**Definition 2.3 Variance**

The variance of distribution  $X$  is sum of squared loss

$$Var[X] := \mathbb{E}_X(X_i - \mu_X)^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (2.5)$$

A large variance indicates significant levels of probability or density from points far away from the mean. The variance must be always  $\geq 0$  (Since everything is squared). The variance of  $X$  is equal to zero only if  $X$  is a fixed single point and with probability 1 at that point; In other words, the function of  $X$  is a constant function (For example,  $x \sim f_X(x) = 6$ , then  $Var[X] = 0$ ).

The standard deviation of the random variable  $X$  is the square root

**Theorem 2.2 Chebyshev's Theorem**

Let  $X$  be a random variable with mean  $\mu_X$  and finite variance  $\sigma^2$ . Then,

$$\mathbb{P}[|X - \mu_X| < k\sigma] \geq 1 - \frac{1}{k^2} \quad (2.6)$$

or

$$\mathbb{P}[|X - \mu_X| \geq k\sigma] \leq \frac{1}{k^2} \quad (2.7)$$

for some constant  $k > 0$ .

**2.4.2 The Coefficient of Variation****Definition 2.4 Coefficient of Variation**

The coefficient of variation is

$$CV = \frac{\sigma_X}{\mu_X} = \frac{\sqrt{Var[x]}}{\mathbb{E}[X]}. \quad (2.8)$$

A higher CV implies greater variability, while a lower CV suggests more consistency or reliability of the data. Imagine if we have two datasets:

- ❖ Dataset  $A$  has a mean of 10 and standard deviation of 2, and  $CV_A = 2/10 = 1/5$ .
- ❖ Dataset  $B$  has a mean of 100 and standard deviation of 10, and  $CV_B = 10/100 = 1/10$ .

While dataset  $B$  has higher standard deviation, but it has a lower CV compared to dataset  $A$ . This indicating  $B$  less reliable variation to the mean.

**2.5 Discrete Random Variables****2.5.1 Binomial distribution**

A Bernoulli trial is an experiment with only two outcomes: **Success** and **failure**. The probability of a success is denoted by  $p$  and that of a failure by  $q = 1 - p$ . Moreover,  $p$  and  $q$  are related by the formula

$$p + q = 1.$$

A Bernoulli experiment is a sequence of independent Bernoulli trials. Let  $X$  represent the number of successes that occur in  $n$  independent Bernoulli trials. Then  $X$  is said to be a Binomial random variable  $(n, p)$ . If  $n = 1$ , then  $X$  is said to be a Bernoulli random variable.

### Theorem 2.3

Let  $(\Omega, \mathbb{P})$  be a probability space, let  $p \in [0, 1]$  and let  $X_1, X_2, \dots, X_n : \Omega \rightarrow \{0, 1\}$  be independent random variables such that each  $X_i \sim \text{Bernoulli}(p)$ . Then

$$X_1 + X_2 + \dots + X_n \sim \text{Bin}(n, p).$$

## 2.5.2 Geometric distribution

A geometric random variable with parameter  $p$ ,  $0 < p < 1$  has a probability mass function

$$p_X(n) = \mathbb{P}(X = n) = p(1 - p)^{n-1}, \quad n = 1, 2, \dots \quad (2.9)$$


Note that  $p_X(n) \geq 0$  and

$$\sum_{n=1}^{\infty} p(1 - p)^{n-1} = \frac{p}{1 - (1 - p)} = 1. \quad (2.10)$$

A geometric random variable models the number of successive independent Bernoulli trials that must be performed to obtain the  $r$ -st success. For example, the number of flips of a fair coin until the  $r$ -st head appears follows a geometric distribution.

 **Example 2.5.1.** Consider the experiment of rolling a pair of fair dice.

1. What is the probability of getting a sum of 11?
2. If you roll two dice repeatedly, what is the probability that the first sum of 11 occurs on the 8-th roll?

 **Solution** 1. A sum of 11 occurs when the pair of dice show either (5, 6) or (6, 5) so that the required probability is  $\frac{2}{36} = \frac{1}{18}$ .

2. Let  $X$  be the number of rolls on which the first sum of 11 happened. Then  $X$  is a geometric random variable with probability  $p = \frac{1}{18}$ . Thus

$$\mathbb{P}[X = 8] = \frac{1}{18} \left(1 - \frac{1}{18}\right)^7 = 0.0372.$$



## 2.6 Continuous Probability Distributions

### 2.6.1 Uniform Probability Distribution

**Definition 2.5 Uniform Distribution**

If  $a < b$ , a random variable  $X$  is said to have a continuous uniform distribution if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \quad (2.11)$$

**Theorem 2.4 Mean and variance of Uniform Distribution**

If  $X$  is a continuous uniform distribution on the interval  $[a, b]$ , then the mean is

$$\mu_X = \mathbb{E}[X] = \frac{a+b}{2} \quad (2.12)$$

and

$$\sigma_X^2 = \text{Var}[X] = \frac{(b-a)^2}{12} \quad (2.13)$$

*Proof.* Given  $X$  is a continuous uniform distribution on the interval  $[a, b]$ , with  $a < b$ . Then the expectation of  $X$  is

$$\begin{aligned} \mu_X = \mathbb{E}[X] &= \int_a^b x \frac{1}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_{x=a}^{x=b} \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} \\ &= \frac{a+b}{2}. \end{aligned}$$

Now we continue to work on the variance for  $X$ . But before that, we need to find the expectation of  $X^2$ .

$$\begin{aligned} \mathbb{E}[X^2] &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{x^3}{3(b-a)} \Big|_{x=a}^{x=b} \\ &= \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

The variance of  $X$  is

$$\begin{aligned}
 \sigma_X^2 &= \text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
 &= \frac{4(b^2 + ab + a^2) - 3(a+b)^2}{12} \\
 &= \frac{a^2 - 2ab + b^2}{12} \\
 &= \frac{(b-a)^2}{12}
 \end{aligned}$$

which is what we expected. □

## 2.7 Normal Distribution

## 2.8 Gamma Distribution

Some random variables can yield distributions of data are skewed right and is non-symmetric.

### Definition 2.6 Gamma Distribution

Let  $X$  be a random variable followed *gamma distribution* with parameters  $\alpha > 0$  and  $\beta$ . The density function of  $X$  is

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \quad (2.14)$$

### Theorem 2.5 Mean and variance of Gamma Distribution

If  $X$  is a gamma distribution with parameters  $\alpha > 0$  and  $\beta > 0$ , then the mean and variance are

$$\mu_X = \alpha\beta \quad (2.15)$$

$$\sigma^2 = \alpha\beta^2. \quad (2.16)$$



*Proof.* Using the moment generating function approach to find mean and variance,

$$\begin{aligned}
M_X(t) &= \int_0^{\infty} e^{tx} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^{\alpha} \Gamma(\alpha)} dx \\
&= \frac{1}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x\left(\frac{1}{\beta}-t\right)}}{\Gamma(\alpha)} dx \\
&= \frac{1}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha)} dx \\
&= \frac{\left(\frac{\beta}{1-\beta t}\right)^{\alpha}}{\beta^{\alpha}} \int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha) \left(\frac{\beta}{1-\beta t}\right)^{\alpha}} dx
\end{aligned}$$

Now observe that

$$\int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha) \left(\frac{\beta}{1-\beta t}\right)^{\alpha}} dx = 1$$

as integrating  $x$  along the entire curve will give us 1. Thus,

$$M_X(t) = \frac{\left(\frac{\beta}{1-\beta t}\right)^{\alpha}}{\beta^{\alpha}} \times 1 = \frac{1}{(1-\beta t)^{\alpha}}. \quad (2.17)$$

That is to say, recall that the  $k$ -th derivative of  $M_X(t)$  with respect to  $t$  as at  $t = 0$ , is the  $\mathbb{E}[X^k]$ . From here we compute

$$\mathbb{E}[X] = \dot{M}_X(0) = \alpha\beta(1-\beta t)^{-\alpha-1}\big|_{t=0} = \alpha\beta.$$

$$\mathbb{E}[X^2] = \ddot{M}_X(0) = -\alpha(1+\alpha)\beta^2(1-\beta t)^{-\alpha-2}\big|_{t=0} = \alpha(1+\alpha)\beta^2.$$

And so the variance is

$$\begin{aligned}
Var[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 \\
&= \alpha\beta^2.
\end{aligned}$$

and we are done. □

### 2.8.1 Chi-square distribution

**Definition 2.7**

If  $X$  is a gamma distribution with parameters  $\alpha = \nu/2$  and  $\beta = 2$ , then  $X$  is a  $\chi^2$ -distribution with  $\nu$  degree of freedom. (with  $\nu > 0$ )

# Limiting Distribution

## 3.1 Convergence in distribution


Convergence in distribution state that the sequence of random variables  $X_1, X_2, \dots, X_n$  converges to some distribution  $X$  when  $n$  goes to infinity. It does not require any dependence between the  $X_n$  and  $X$ . Convergence in distribution is consider as the weakest type of convergence.

### Definition 3.1 Convergence in distribution

Let  $\{X_n\}_{n \geq 1}$  be a sequence of random variable and let  $X$  be a random variable. Let  $F_{X_n}$  and  $F_X$  be the cdf of  $X_n$  and  $X$  respectively. And let  $C(F)$  be the set of all continuous points of  $F$ . We say that  $X_n$  converges in distribution to  $X$  if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad (3.1)$$

for all  $x \in C(F)$ . We denote  $X_n \xrightarrow{d} X$ .

 **Example 3.1.1.** Let  $X_1, X_2, X_3, \dots$  be a sequence of random variable such that

$$X_n \sim \text{Geom}(\lambda/n), \quad \forall n = 1, 2, 3, \dots$$

where  $\lambda > 0$  is a constant. Define a new sequence  $Y_n$  as

$$Y_n = \frac{1}{n}X_n, \quad \forall n = 1, 2, 3, \dots$$

Show that  $Y_n$  converges in distribution to  $\text{Exp}(\lambda)$ .

 **Solution** The cdf of  $Y_n$  is

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P} \left[ \frac{1}{n}X_n \leq y \right] \\ &= \mathbb{P} [X_n \leq ny] \\ &= 1 - \left( 1 - \frac{\lambda}{n} \right)^{\text{floor}(ny)}. \end{aligned}$$

and taking limit for  $n \rightarrow +\infty$  we have

$$\begin{aligned}\lim_{n \rightarrow \infty} \left( 1 - \left( 1 - \frac{\lambda}{n} \right)^{\lfloor ny \rfloor} \right) &= 1 - \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{\left( -\frac{n}{\lambda} \right)} \right)^{-n(-y)} \\ &= 1 - \lim_{n \rightarrow \infty} \left( 1 + \frac{1}{\left( -\frac{n}{\lambda} \right)} \right)^{-\frac{n}{\lambda}(-\lambda y)} \\ &= 1 - e^{-\lambda y}\end{aligned}$$

which is the cdf of exponential distribution with parameter  $\lambda$ . ◀

📌 **Example 3.1.2.** Let  $\{X_n\}_{n \geq 1}$  be a sequence of random variables with probability mass function

$$f_{X_n}(x) = \mathbb{P}[X_n = x] = \begin{cases} 1 & \text{if } x = 2 + \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Find the limiting distribution of  $X_n$ .

### 3.1.1 Almost sure converges

We said that  $X_n$  converges to  $X$  **almost surely** if the probability that the sequence  $X_n(s)$  converges to  $X(s)$  is equal to 1.

📌 **Example 3.1.3.** Consider the sample space  $S = [0, 1]$  with uniform probability distribution, for instance,

$$\mathbb{P}([a, b]) = b - a \quad \forall 0 \leq a \leq b \leq 1.$$

Define the sequence  $\{X_n, n = 1, 2, 3, \dots\}$  as

$$X_n(s) = \frac{n}{n+1}s + (1-s)^n.$$

Also, define the random variable  $X$  on the sample space as  $X(s) = s$ . Show that  $X_n$  *almost sure* converges to  $X$ .

🔑 **Solution** For any  $s \in [0, 1]$ , taking the limit when  $n \gg \infty$  we have

$$\begin{aligned}\lim_{n \rightarrow \infty} X_n(s) &= \lim_{n \rightarrow \infty} \left[ \frac{n}{n+1}s + (1-s)^n \right] \\ &= \lim_{n \rightarrow \infty} \frac{n}{n+1}s + \lim_{n \rightarrow \infty} (1-s)^n \\ &= 1 \cdot s + 0 \\ &= s = X(s).\end{aligned}$$

However, if  $s = 0$  then

$$\lim_{n \rightarrow \infty} X_n(0) = \lim_{n \rightarrow \infty} \left[ \frac{n}{n+1}(0) + (1-0)^n \right] = 1.$$

Thus, we conclude that  $\lim_{n \rightarrow \infty} X_n(s) = X(s)$  for all  $s$  lies between 0 and 1. And because  $\mathbb{P}([0, 1]) = 1$ , we conclude that

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

and we are done. ◀

## 3.2 Law of Large Numbers

In this section we will discuss the Weak and Strong Law of Large Numbers. The Law of Large Numbers are considered as a form of convergence in probability.

We will first state the Weak Law of Large Numbers (WLLN),

### Theorem 3.1 Weak Law of Large Numbers (WLLN)

Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variables, each with mean  $\mathbb{E}[X_i] = \mu$  and standard deviation  $\sigma$ , we define

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The Weak Law of Large Numbers (WLLN) states that for all  $\epsilon > 0$ , we have

$$\lim_{n \rightarrow \infty} \mathbb{P}[|\bar{X}_n - \mu| > \epsilon] = 0. \quad (3.2)$$

*Proof.* Suppose that  $\text{Var}[X_i] = \sigma^2 > 0$  for finite  $i$ . Since  $X_1, X_2, \dots, X_n$  are identically independent, there is no correlation between them, thus

$$\begin{aligned} \text{Var}[\bar{X}_n] &= \text{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n^2} \text{Var}[X_1 + X_2 + \cdots + X_n] \\ &= \frac{1}{n^2} [\text{Var}X_1 + \text{Var}X_2 + \cdots + \text{Var}X_n] \\ &= \frac{1}{n^2} (\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ times}}) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Notice that the mean of each  $X_i$  in the sequence is also equal to the mean of the sample average, said  $\mathbb{E}[X_i] = \mu$ . We can now apply Chebyshev's inequality on  $\bar{X}_n$  to get, for all  $\epsilon > 0$ ,

$$\mathbb{P}[|\bar{X}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

So that □

### Theorem 3.2 Strong Law of Large Numbers

Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variables, each with mean  $\mathbb{E}[X_i] = \mu$  and standard deviation  $\sigma$ , then

$$\mathbb{P}[\lim_{n \rightarrow \infty} \bar{X}_n = \mu] = 1. \quad (3.3)$$

*Proof.* By Markov's inequality that

$$\mathbb{P}\left[\frac{1}{n}|\bar{X}_n - n\mu| \geq n^{-\gamma}\right] \leq \frac{\mathbb{E}[(\frac{\bar{X}_n}{n} - \mu)^4]}{n^{-4\gamma}} = Kn^{-2+4\gamma}.$$

Define for all  $\gamma \in (0, \frac{1}{4})$ , and let

$$A_n = \left\{ \frac{1}{n} |\bar{X}_n - n\mu| \geq n^{-\gamma} \right\} \Rightarrow \sum_{n \geq 1} \mathbb{P}[A_n] < \infty \Rightarrow \mathbb{P}[A] = 0$$

by the first Borel-Cantelli lemma, where  $A = \bigcap_{n \geq 1} \bigcup_{m \geq n} A_m$ . But now, the event  $A^c$  happened if and only if

$$\exists N \forall n \geq N \left| \frac{\bar{X}_n}{n} - \mu \right| < n^{-\gamma} \Rightarrow \frac{\bar{X}_n}{n} \xrightarrow{p} \mu.$$

□

### Theorem 3.3 Central Limit Theorem

Let  $\{X_n\}_{n \geq 1}$  be a sequence of i.i.d random variable whose memoment generating function exist in a neighborhood of 0. Let  $\mathbb{E}[X_i] = \mu$  and  $Var[X_i] = \sigma^2 > 0$ . Define  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Then

$$Z = \frac{\sqrt{n}(\bar{X}_i - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (3.4)$$

or

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (3.5)$$

**Example 3.2.1.** Let  $\bar{X}_n$  be the sample mean from a random sampling of size  $n = 100$  from  $\chi_{50}^2$ . Compute approximate value of  $\mathbb{P}(49 < \bar{X} < 51)$ .

**Solution** Because  $\bar{X}$  followed Chi-squared distribution with degree of freedom 50, then the mean and variance are  $\mathbb{E}[X_i] = 50$  and  $Var[X_i] = 2(50) = 100$ . By Central Limit Theorem,

$$\begin{aligned} \mathbb{P}(49 < \bar{X} < 51) &\simeq \mathbb{P} \left[ \frac{\sqrt{100}(49 - \mathbb{E}[X_i])}{\sqrt{Var[X_i]}} < Z < \frac{\sqrt{100}(51 - \mathbb{E}[X_i])}{\sqrt{Var[X_i]}} \right] \\ &\simeq \mathbb{P} \left[ \frac{\sqrt{100}(49 - 50)}{\sqrt{100}} < Z < \frac{\sqrt{100}(51 - 50)}{\sqrt{100}} \right] \\ &\simeq \mathbb{P}[-1 < Z < 1] \\ &\simeq \Phi(1) - \Phi(-1) \\ &\simeq 0.84134 - 0.15866 = 0.68268. \end{aligned}$$

◀

### Theorem 3.4 Slutsky's Theorem

If  $X_n$  converges in distribution to a random variable  $X$ , and  $Y_n$  converges in probability to a constant  $c$ , then

- $Y_n X_n \xrightarrow{d} cX$
- $X_n + Y_n \xrightarrow{d} X + c.$

*Proof.* We will prove both parts of Slutsky's theorem.

**Part 1:** We want to show that  $Y_n X_n \xrightarrow{d} cX$ .

Since  $Y_n \xrightarrow{p} c$ , for any  $\epsilon > 0$ , we have  $\mathbb{P}[|Y_n - c| > \epsilon] \rightarrow 0$  as  $n \rightarrow \infty$ . This implies that  $Y_n$  is

bounded in probability, and we can write  $Y_n = c + (Y_n - c)$  where  $(Y_n - c) \xrightarrow{p} 0$ .

Now,  $Y_n X_n = cX_n + (Y_n - c)X_n$ . Since  $X_n \xrightarrow{d} X$  and multiplication by the constant  $c$  is a continuous operation, we have  $cX_n \xrightarrow{d} cX$ .

For the second term, we need to show that  $(Y_n - c)X_n \xrightarrow{p} 0$ . Since  $Y_n - c \xrightarrow{p} 0$  and  $X_n$  is bounded in probability (as it converges in distribution), their product converges to 0 in probability.

By Slutsky's theorem for sums (which we prove next), we get  $Y_n X_n = cX_n + (Y_n - c)X_n \xrightarrow{d} cX + 0 = cX$ .

**Part 2:** We want to show that  $X_n + Y_n \xrightarrow{d} X + c$ .

We use characteristic functions. Let  $\phi_{X_n}(t)$ ,  $\phi_X(t)$ , and  $\phi_{Y_n}(t)$  denote the characteristic functions of  $X_n$ ,  $X$ , and  $Y_n$ , respectively.

The characteristic function of  $X_n + Y_n$  is given by:

$$\phi_{X_n+Y_n}(t) = \mathbb{E}[e^{it(X_n+Y_n)}] = \mathbb{E}[e^{itX_n} e^{itY_n}]$$

Since  $Y_n \xrightarrow{p} c$ , we have  $e^{itY_n} \xrightarrow{p} e^{itc}$  by the continuous mapping theorem.


Using the fact that  $X_n \xrightarrow{d} X$  implies  $\phi_{X_n}(t) \rightarrow \phi_X(t)$ , and that convergence in probability preserves the limit of expectations for bounded random variables, we get:

$$\begin{aligned} \lim_{n \rightarrow \infty} \phi_{X_n+Y_n}(t) &= \lim_{n \rightarrow \infty} \mathbb{E}[e^{itX_n} e^{itY_n}] \\ &= \mathbb{E}[e^{itX}] \cdot e^{itc} \\ &= \phi_X(t) \cdot e^{itc} \\ &= \phi_{X+c}(t) \end{aligned}$$

Since the characteristic function of  $X_n + Y_n$  converges pointwise to the characteristic function of  $X + c$ , we conclude that  $X_n + Y_n \xrightarrow{d} X + c$ .  $\square$

 **Example 3.2.2.** If the random variable  $X \sim \text{Gamma}(\mu, 1)$ , show that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

 **Solution** Slutsky's theorem stated that If  $X_n$  converges in distribution to a random variable  $X$  and if  $Y_n$  converges in probability to a constant  $c$ . Then  $X_n/Y_n$  converges in distribution to  $X/c$ . By the central limit theorem we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \text{Var}[X_i]).$$

and in this case  $\mathbb{E}X_i = \text{Var}[X_i] = \mu$ , thus we obtained

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \mu).$$

Replacing the theorem denominator  $Y_n$  with  $\bar{X}_n$ , which  $\bar{X}_n$  converges to constant  $\mu$  in probability. Hence

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}\left(\frac{0}{\mu}, \frac{\mu}{\mu}\right) \implies \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

and we are done with the proof.  $\blacktriangleleft$

**Theorem 3.5**

If the random variable  $X_n$  converges to constant  $c$  in probability, then

$$\sqrt{X_n} \xrightarrow{p} \sqrt{c}, \quad c > 0. \quad (3.6)$$

**Theorem 3.6**

If the random variable  $X_n$  converges to constant  $c$  in probability, and  $Y_n$  converges to constant  $d$  in probability, then

- $aX_n + bY_n \xrightarrow{p} ac + bd$ .
- $X_n Y_n \xrightarrow{p} cd$ .
- $\frac{1}{X_n} \xrightarrow{p} \frac{1}{c}$  for all  $c \neq 0$ .

*Proof.* We will prove each part of the theorem.

**Part 1:** We want to show that  $aX_n + bY_n \xrightarrow{p} ac + bd$ .

For any  $\epsilon > 0$ , we have:

$$\begin{aligned} |aX_n + bY_n - (ac + bd)| &= |a(X_n - c) + b(Y_n - d)| \\ &\leq |a||X_n - c| + |b||Y_n - d| \end{aligned}$$

By the triangle inequality, for any  $\delta > 0$ :

$$\begin{aligned} \mathbb{P}[|aX_n + bY_n - (ac + bd)| > \epsilon] &\leq \mathbb{P}[|a||X_n - c| + |b||Y_n - d| > \epsilon] \\ &\leq \mathbb{P}[|a||X_n - c| > \epsilon/2] + \mathbb{P}[|b||Y_n - d| > \epsilon/2] \\ &= \mathbb{P}[|X_n - c| > \frac{\epsilon}{2a}] + \mathbb{P}[|Y_n - d| > \frac{\epsilon}{2b}], \quad \forall a, b > 0 \end{aligned}$$

Since  $X_n \xrightarrow{p} c$  and  $Y_n \xrightarrow{p} d$ , both terms on the right approach 0 as  $n \rightarrow \infty$ .

**Part 2:** We want to show that  $X_n Y_n \xrightarrow{p} cd$ .

We can write:

$$\begin{aligned} X_n Y_n - cd &= X_n Y_n - cY_n + cY_n - cd \\ &= Y_n(X_n - c) + c(Y_n - d) \end{aligned}$$

Since  $Y_n \xrightarrow{p} d$ , the sequence  $\{Y_n\}$  is bounded in probability. That is, for any  $\delta > 0$ , there exists  $M > 0$  such that  $\mathbb{P}[|Y_n| > M] < \delta$  for all  $n$  sufficiently large.

For any  $\epsilon > 0$ :

$$\begin{aligned} |X_n Y_n - cd| &= |Y_n(X_n - c) + c(Y_n - d)| \\ &\leq |Y_n||X_n - c| + |c||Y_n - d| \end{aligned}$$



Given  $\epsilon > 0$ , choose  $\delta > 0$  such that:

$$\begin{aligned} & \mathbb{P}[|X_n Y_n - cd| > \epsilon] \\ & \leq \mathbb{P}[|Y_n||X_n - c| + |c||Y_n - d| > \epsilon] \\ & \leq \mathbb{P}[|Y_n||X_n - c| > \epsilon/2] + \mathbb{P}[|c||Y_n - d| > \epsilon/2] \end{aligned}$$

For the first term, using the boundedness of  $Y_n$  and convergence of  $X_n$ , and for the second term using convergence of  $Y_n$ , both approach 0 as  $n \rightarrow \infty$ .

**Part 3:** We want to show that  $\frac{1}{X_n} \xrightarrow{p} \frac{1}{c}$  for  $c \neq 0$ .

Since  $c \neq 0$ , there exists  $\delta > 0$  such that  $|c| > \delta > 0$ . Because  $X_n \xrightarrow{p} c$ , for any  $\epsilon > 0$ , we have  $\mathbb{P}[|X_n - c| > \epsilon] \rightarrow 0$ .

In particular,  $\mathbb{P}[|X_n - c| > \delta/2] \rightarrow 0$ , which implies  $\mathbb{P}[|X_n| > \delta/2] \rightarrow 1$ . This means  $X_n$  is bounded away from 0 in probability.

Now, for any  $\epsilon > 0$ :

$$\left| \frac{1}{X_n} - \frac{1}{c} \right| = \left| \frac{c - X_n}{X_n c} \right| = \frac{|X_n - c|}{|X_n||c|}$$

On the event  $\{|X_n| > \delta/2\}$ , we have:

$$\left| \frac{1}{X_n} - \frac{1}{c} \right| \leq \frac{2|X_n - c|}{\delta|c|}$$

Therefore:

$$\begin{aligned} \mathbb{P}\left[\left| \frac{1}{X_n} - \frac{1}{c} \right| > \epsilon\right] & \leq \mathbb{P}[|X_n| \leq \delta/2] + \mathbb{P}\left[\frac{2|X_n - c|}{\delta|c|} > \epsilon, |X_n| > \delta/2\right] \\ & \leq \mathbb{P}[|X_n| \leq \delta/2] + \mathbb{P}\left[|X_n - c| > \frac{\epsilon\delta|c|}{2}\right] \end{aligned}$$

As  $n \rightarrow \infty$ , both terms approach 0, completing the proof.  $\square$

### 3.3 Order Statistics

We can ordering the observed random variables based on their magnitudes or ranking. These ordered variables are known as **order statistics**.

Consider  $X_1, X_2, \dots, X_n$  are independent continuous random variables with cdf  $F_X(y)$  and mass function  $f_X(y)$ . We ordered them into order statistics  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  such that

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

In this notion, the maximum random variable is

$$X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$$

and the minimum random variable is

$$X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$$

$X_{(n)}$  is the largest among  $X_1, X_2, \dots, X_n$ , the event  $X_{(n)} \leq y$  will happen only if each  $X_i \leq y$ . Then the joint probability is

$$G_{(n)}(y) = \mathbb{P}[X_{(n)} \leq y] = \mathbb{P}[X_1 \leq y, X_2 \leq y, \dots, X_n \leq y] = \prod_{i=1}^n \mathbb{P}[X_i \leq y]. \quad (\heartsuit)$$

Because  $\mathbb{P}[X_i \leq y] = F_X(y)$  for all  $i = 1, 2, \dots, n$ . It follows that

$$(\heartsuit) \Rightarrow \mathbb{P}[X_1 \leq y] \mathbb{P}[X_2 \leq y] \cdots \mathbb{P}[X_n \leq y] = [F_X(y)]^n.$$

Now letting  $g_{(n)}$  denote the density function of  $Y_{(n)}$ , we see that, on taking derivative on  $G_{(n)}$  with respect to  $y$ .

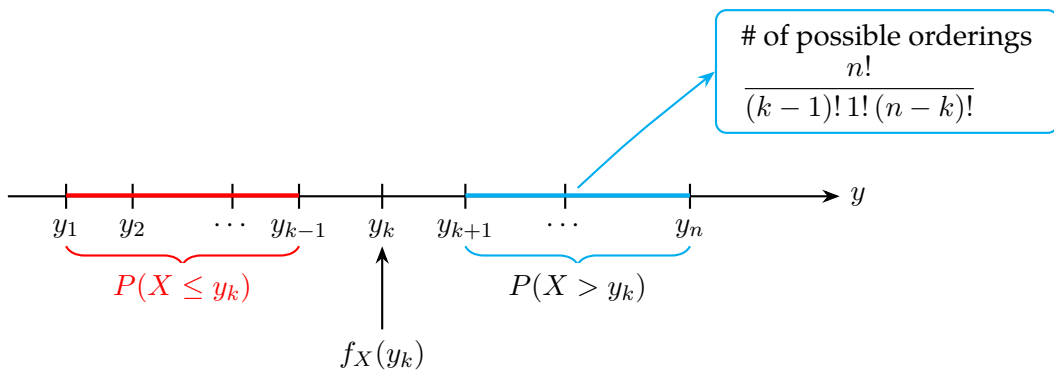
$$\begin{aligned} g_{(n)}(y) &= \frac{d}{dy} [F_X(y)]^n \\ &= n[F_X(y)]^{n-1} \frac{d}{dy} F_X(y) && \text{By Chain rule of derivative} \\ &= n[F_X(y)]^{n-1} f_X(y) \end{aligned}$$

Now we get the maximum variable. For the minimum variable  $X_{(1)}$  can be found using the similar way. The cdf of  $X_{(1)}$  is

$$F_{(1)}(y) = \mathbb{P}[X_{(1)} \leq y] = 1 - \mathbb{P}[X_{(1)} > y].$$

Since  $X_{(1)}$  is the minimum of  $X_1, X_2, \dots, X_n$ , and the event  $Y_i > y$  can occur for  $i = 1, 2, 3, \dots, n$ . In other words, any  $X_i$  in  $X_1, X_2, \dots, X_n$  can be the minimum variable. Hence


$$\begin{aligned} F_{(1)}(y) &= \mathbb{P}[X_{(1)} \leq y] = 1 - \mathbb{P}[X_{(1)} > y] \\ &= 1 - \mathbb{P}[X_1 > y, X_2 > y, \dots, X_n > y] \\ &= 1 - \mathbb{P}[X_1 > y] \mathbb{P}[X_2 > y] \cdots \mathbb{P}[X_n > y] \\ &= 1 - [1 - F_X(y)]^n. \end{aligned}$$




### Theorem 3.7 k-th order statistics

Let  $X_1, X_2, \dots, X_n$  be i.i.d continuous random variable with common cdf  $F_X(y)$  and common density function  $f_X(y)$ . Let  $X_{(k)}$  denote the  $k$ -th order Statistics, then the density function of  $X_{(k)}$  is

$$g_{(n)}(y) = \frac{n!}{(k-1)!(n-k)!} [F_X(y)]^{k-1} [1 - F_X(y)]^{n-k} f_X(y), \quad -\infty < y < \infty. \quad (3.7)$$

 **Example 3.3.1.** Let  $Y \sim \text{Uniform}(0, \theta)$  be the waiting time of bus arrival. A random samples of size  $n = 5$  is taken. Then,

1. Find the distribution of minimum variable.
2. Find the probability that  $Y_{(3)}$  is less than  $\frac{2}{3}\theta$ .
3. Suppose that the waiting time for bus arrival is uniformly distributed on 0 to 15 minutes, find  $\mathbb{P}[Y_{(5)} < 10]$ .

 **Solution** 1. The density of  $X_{(1)}$  is

$$\begin{aligned} Y_{(1)} \sim g_{(1)}(y) &= \frac{5!}{(1-1)!(5-1)!} [F_Y(y)]^{1-1} [1 - F_Y(y)]^{5-1} f_Y(y) \\ &= \frac{5!}{0!4!} [1 - F_Y(y)]^4 f_Y(y) \\ &= 5 \left(1 - \frac{y}{\theta}\right)^4 \left(\frac{1}{\theta}\right) \\ &= \frac{5(\theta - y)^4}{\theta^5}. \end{aligned}$$

Hence compute the mean of  $X_{(1)}$ ,

$$\mathbb{E}[Y_{(1)}] = \int_0^\theta y \left[ \frac{5(\theta - y)^4}{\theta^5} \right] dy = \int_0^\theta \frac{5y(\theta - y)^4}{\theta^5} dy \quad (\clubsuit)$$

using the substitution method and letting  $u = \theta - y$ , and for that

$$y = \theta - u \implies -du = dy$$

substitute back into  $(\clubsuit)$  and we have

$$\begin{aligned} (\clubsuit) &= \int_0^\theta \frac{5(\theta - u)u^4}{\theta^5} (-du) = -\frac{1}{\theta^5} \int_0^\theta (5\theta u^4 - u^5) du \\ &= -\frac{1}{\theta^5} \left[ \theta u^5 - \frac{1}{6} \theta^6 \right]_{u=0}^{u=\theta} \\ &= -\frac{1}{\theta^5} \left[ 0 - \frac{1}{6} \theta^6 \right] \\ &= \frac{\theta}{6} = \mathbb{E}[Y_{(1)}]. \end{aligned}$$

2. First we need to find the probability density function of  $Y_{(3)}$ , that is,

$$\begin{aligned} Y_{(3)} \sim g_{(3)}(y) &= \frac{5!}{(3-1)!(5-3)!} [F_Y(y)]^{3-1} [1 - F_Y(y)]^{5-3} f_Y(y) \\ &= 30 \left(\frac{y}{\theta}\right)^2 \left(1 - \frac{y}{\theta}\right)^2 \frac{1}{\theta}, \quad 0 < y < \theta. \end{aligned}$$

Compute the probability on which that  $Y_{(3)}$  is smaller than  $\frac{2\theta}{3}$ .

$$\begin{aligned}
 \mathbb{P}[Y_{(3)} < \frac{2}{3}\theta] &= \int_0^{\frac{2}{3}\theta} 30 \left(\frac{y}{\theta}\right)^2 \left(1 - \frac{y}{\theta}\right)^2 \frac{1}{\theta} dy \\
 &= \frac{30}{\theta^5} \int_0^{\frac{2}{3}\theta} y^2 (\theta^2 - 2\theta y + y^2) dy \\
 &= \frac{30}{\theta^5} \left[ \frac{1}{3} \theta^2 y^3 - \frac{1}{2} \theta y^4 + \frac{1}{5} y^5 \right]_{y=0}^{y=\frac{2}{3}\theta} \\
 &= 30 \left( \frac{1}{3} \right) \left( \frac{2}{3} \right)^3 - 15 \left( \frac{2}{3} \right)^4 + 6 \left( \frac{2}{3} \right)^5 \\
 &= \frac{64}{81}.
 \end{aligned}$$

3. The probability that  $Y_{(5)}$  less than 10 minutes is equivalent to taking the bus five times. That is

$$\begin{aligned}
 \mathbb{P}[Y_{(5)} < 10] &= \mathbb{P}[Y_{(1)} < 10, Y_{(2)} < 10, \dots, Y_{(5)} < 10] \\
 &= \mathbb{P}[Y_{(1)} < 10] \times \mathbb{P}[Y_{(2)} < 10] \times \dots \times \mathbb{P}[Y_{(5)} < 10] \\
 &= \left( \frac{10}{15} \right)^5 = \frac{32}{243}.
 \end{aligned}$$



## Tutorials

**Exercise 3.1** Let  $X_1, X_2, \dots, X_{2020}$  be a random sample of size 2020 from a Poisson distribution with density function

$$f_{X_i}(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots, \infty.$$

What is the distribution of  $2020\bar{X}$ ?

# Decision theory

For the given observation  $\mathcal{X}$ , we decide to take an action  $a \in \mathcal{A}$ . An action is a map  $a : \mathcal{X} \rightarrow \mathcal{A}$  with  $a(X)$  being the decision taken.


$L(\theta, a)$  denoted as the “loss function”, it is the loss incurred when state is  $\theta$  and an action  $a$  is taken.

$$L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}. \quad (4.1)$$

## 4.1 Conditional Distributions

Recall the definition of conditional probabilities: For two sets  $A$  and  $B$ , with  $P(A) \neq 0$ , the conditional probability of  $B$  given that  $A$  is true is defined as

$$\mathbb{P}(B | A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \quad (4.2)$$

 **Example 4.1.1.** Let  $X$  and  $Y$  be two jointly continuous random variable with joint density function

$$f_{XY}(x, y) = \begin{cases} x^2 + \frac{1}{3}y, & -1 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

For  $0 \leq y \leq 1$ , find the conditional pdf of  $X$  given  $Y = y$ .


 **Solution** First we find the marginal distribution of  $Y$ , which we can obtain by integrating along with  $x$ .

$$\begin{aligned} f_Y(y) &= \int_{-1}^1 f_{XY}(x, y) \, dx = \int_{-1}^1 \left( x^2 + \frac{1}{3}y \right) \, dx \\ &= \frac{1}{3}x^3 + \frac{1}{3}xy \Big|_{-1}^1 \\ &= \frac{2}{3}(1 + y). \end{aligned}$$

The conditional distribution of  $X$  given  $Y = y$  is

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{x^2 + \frac{1}{3}y}{\frac{2}{3}(1 + y)} = \frac{3x^2 + y}{2(1 + y)}, \quad -1 \leq x \leq 1, 0 \leq y \leq 1$$

◀

 **Example 4.1.2 (Two-sample mean problems).** Consider the observations  $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu, \sigma^2)$  response under control treatment. And  $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu + \Delta, \sigma^2)$  are explanatory

data response under test treatment where  $\mu \in \mathbb{R}$ ,  $\sigma^2 \in \mathbb{R}_+$ .  $\sigma^2$  is unknown variance and  $\Delta \in \mathbb{R}$  is unknown treatment effect.

We define two testing hypotheses:

$$H_0 : P \in \{P : \Delta = 0\} = \{P_\theta : \theta \in \Theta_0\}$$

$$H_1 : P \in \{P : \Delta \neq 0\} = \{P_\theta : \theta \notin \Theta_0\}$$

By construct decision rule accepting null hypothesis  $H_0$  if estimate of  $\Delta$  is significantly far away from zero. For instance,  $\hat{\Delta} = \bar{Y} - \bar{X}$  to be the estimate difference in sample means. Since  $\sigma$  is unknown, we use  $\hat{\sigma}$  to estimate true  $\sigma$ . The decision procedure is


$$\delta(X, Y) = \begin{cases} 1 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| < c \\ 0 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| \geq c \end{cases}$$

We again define a zero-one loss function to make decision

$$L(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \Theta_a \quad (\text{correct action}) \\ 1 & \text{if } \theta \notin \Theta_a \quad (\text{wrong action}) \end{cases}.$$

The risk function is linear combination of the loss of correct and wrong actions,

$$\begin{aligned} R(\theta, \delta) &= L(\theta, 0)P_\theta(\delta(X, Y) = 0) + L(\theta, 1)P_\theta(\delta(X, Y) = 1) \\ &= \begin{cases} P_\theta(\delta(X, Y) = 1) & \text{if } \theta \in \Theta_0 \\ P_\theta(\delta(X, Y) = 0) & \text{if } \theta \notin \Theta_0 \end{cases} \end{aligned}$$

 **Example 4.1.3 (Statistical testing).** We are going to use the random variable  $X \sim P_\theta$  with sample space  $\mathcal{X}$  and parameter space  $\Theta$ , we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test  $\delta$  as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that

- ❖ Type I error: the test  $\delta(X)$  rejects  $H_0$  when  $H_0$  is true.
- ❖ Type II error: the test  $\delta(X)$  accepts  $H_0$  when  $H_0$  is false.

The risk under zero-one loss as

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ &= \text{Probability of Type I error.} \end{aligned}$$

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ &= \text{Probability of Type II error.} \end{aligned}$$

✎ **Example 4.1.4** (Statistical testing with two different hypothesis subspace). We are going to use the random variable  $X \sim P_\theta$  with sample space  $\mathcal{X}$  and parameter space  $\Theta$ , we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test  $\delta$  as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that

- ❖ Type I error: the test  $\delta(X)$  rejects  $H_0$  when  $H_0$  is true.
- ❖ Type II error: the test  $\delta(X)$  accepts  $H_0$  when  $H_0$  is false.

The risk under zero-one loss as

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ &= \text{Probability of Type I error.} \end{aligned}$$

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ &= \text{Probability of Type II error.} \end{aligned}$$

## 4.2 Value-at-risk

✎ **Example 4.2.1** (Confidence Interval). We altering the previous decision framework setup:

- ❖  $X$  is a random variable with probability  $P_\theta$ .
- ❖ The parameter of interest is  $\mu(\theta)$ .
- ❖ Define  $\mathfrak{U} = \{\mu = \mu(\theta) : \theta \in \Theta\}$ .
- ❖ Objective: we want to construct an interval estimation of  $\mu(\theta)$ .
- ❖ Action space:  $\mathcal{A} = \{\mathbf{a} = [\underline{a}, \bar{a}] : \underline{a} < \bar{a} \in \mathfrak{U}\}$ .
- ❖ Interval Estimator: define a map  $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$ , that is  $\hat{\mu}(X) = [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)]$

Note that  $\theta$  is not random, the interval is random given a fixed  $\theta$ . We have to use Bayesian models to compute

$$\mathbb{P}[\mu(\theta) \in [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)] \mid X = x].$$

We define the zero-one loss function

$$L(\theta, (\underline{a}, \bar{a})) = \begin{cases} 1 & \text{if } \underline{a} > \mu(\theta) \text{ or } \bar{a} < \mu(\theta) \\ 0 & \text{otherwise.} \end{cases}$$

The risk function under zero-one loss is

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}_X[L(\theta, \hat{\mu}(X)) \mid \theta] \\ &= P_\theta(\hat{\mu}_{\text{Lower}}(X) > \mu(\theta) \text{ or } \hat{\mu}_{\text{Upper}}(X) < \mu(\theta)) \\ &= 1 - P_\theta(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta). \end{aligned}$$

It is said that the interval estimator  $\hat{\mu}(\theta)$  has confidence level  $1 - \alpha$  if

$$P_\theta(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta) \geq (1 - \alpha) \quad \forall \theta \in \Theta.$$

Equivalently, we can said  $R(\theta, \hat{\mu}(X)) \leq \alpha$  for all  $\theta \in \Theta$ .

### 4.3 Admissible

On basis of performance measure by the risk function  $R(\theta, \delta)$ , some rules are obviously bad. We said that a decision procedure  $\delta(\cdot)$  is inadmissible if  $\exists \delta'$  such that

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Theta \quad (4.3)$$

with strict inequality for some  $\theta$ .

**Example 4.3.1.** Suppose, for  $n \geq 2$ , the observations  $X_1, X_2, \dots, X_n$  be i.i.d with mean  $g(\theta) := \mathbb{E}_\theta[X_i] = \mu$ , and  $\text{Var}[X_i] = 1$  for all  $i$ . We take quadratic loss

$$L(\theta, a) := |\mu_X - a|^2.$$

Consider the decision

$$\delta'(X_1, X_2, \dots, X_n) := \bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n},$$

and  $\delta(X_1, X_2, \dots, X_n) := X_1$ . Then for all  $\theta$ , we have

$$R(\theta, \delta') = \frac{1}{n}, \quad R(\theta, \delta) = 1.$$

Therefore  $\delta$  is inadmissible.



# Sampling Distributions

## 5.1 Snedecor's $F$ -distribution

The  $F$ -distribution was named in honor of Sir Ronald Fisher by George Snedecor.  $F$ -distribution arises as the distribution of a ratio of variances. Like, the other two distributions this distribution also tends to normal distribution as  $\nu_1$  and  $\nu_2$  become very large. The following figure illustrates the shape of the graph of this distribution for various degrees of freedom.

### Theorem 5.1

If the random variable  $X$  is  $F$ -distributed with degrees of freedom  $\nu_1$  and  $\nu_2$ , then its mean is

$$\mathbb{E}[X] = \begin{cases} \frac{\nu_2}{\nu_2 - 2} & \text{if } \nu_2 \geq 3 \\ DNE & \text{if } \nu_2 = 1, 2 \end{cases} \quad (5.1)$$

and the variance is

$$Var[X] = \begin{cases} \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} & \text{if } \nu_2 \geq 5 \\ DNE & \text{if } \nu_2 = 1, 2, 3, 4. \end{cases} \quad (5.2)$$

### Theorem 5.2

If a random variable  $X \sim F(\nu_1, \nu_2)$ , then its reciprocal  $\frac{1}{X} \sim F(\nu_2, \nu_1)$ .

### Theorem 5.3

If the random variables  $U \sim \chi^2(\nu_1)$  and  $V \sim \chi^2(\nu_2)$ , and  $U$  and  $V$  are independent, then

$$\frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2). \quad (5.3)$$

**Example 5.1.1.** Let  $X_1, X_2, \dots, X_4$  and  $Y_1, Y_2, \dots, Y_5$  be two random samples of size 4 and 5, respectively, from a standard normal population. What is the variance of the statistic

$$T = \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2}.$$

**Solution** Since the population is standard normal, we have

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4).$$

Similarly,

$$Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2 \sim \chi^2(5).$$

Therefore,

$$\begin{aligned} T &= \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2} \\ &\quad \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{4} \\ &= \frac{4}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2} \\ &\quad 5 \\ &\sim F(4, 5). \end{aligned}$$

Applying theorem, the variance of this statistic is

$$\begin{aligned} Var[T] &= Var[F(4, 5)] \\ &= \frac{2(5)^2(4 + 5 - 2)}{4(5 - 2)^2(5 - 4)} \\ &= \frac{350}{36}. \end{aligned}$$



# Estimation

## Definition 6.1 Estimator

An **estimator** is a formula, that tells how to calculate the value of an estimate based on the observations contained in a sample.

For example, the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a rule that tells us how to calculate the estimate of the population mean  $\mu$  based on the observations in a sample.

## Theorem 6.1 Mean Squared Error

For an estimator  $\hat{\mu}(X)$  of  $\mu = \mu(\theta)$ , the mean-squared error is

$$MSE(\hat{\mu}) = Var[\hat{\mu}(X) | \theta] + Bias(\hat{\mu} | \theta)^2 \quad (6.1)$$

where  $Bias(\hat{\mu} | \theta) = \mathbb{E}_\theta[\hat{\mu}(X) | \theta] - \mu$ .

*Proof.* Consider the following decision framework:

- ❖  $X \sim P_\theta, \theta \in \Theta$ .
- ❖ The parameter of interest,  $\mu(\theta)$  is a certain function.
- ❖ Action space,  $\mathcal{A} = \{\mu = \mu(\theta), \theta \in \Theta\}$ .
- ❖ Decision procedure (or estimator),  $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$ .
- ❖ Squared error loss as loss function:  $L(\theta, a) = [a - \mu(\theta)]^2$ .

with the setup above, the MSE is equal to the risk of decision,

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}[L((\theta, \hat{\mu}(X)) | \theta)] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu(\theta))^2 | \theta] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu)^2 | \theta] \\ &= Var[\hat{\mu}(X) | \theta] + \underbrace{(\mathbb{E}[\hat{\mu}(X) | \theta] - \mu)^2}_{Bias(\hat{\mu}|\theta)} \end{aligned}$$


□

## 6.1 Point Estimators

A **point estimator** is a function of the sample data that provides a single value as an estimate of an unknown population parameter. Since the estimator is calculated from a random sample, it is itself a random variable and has a probability distribution, called the **sampling distribution**.

The sampling distribution of a point estimator describes how the estimator varies from sample to sample. Key properties of the sampling distribution include its mean (which relates to bias) and its variance (which relates to the precision of the estimator). Understanding the sampling distribution is fundamental for assessing the reliability of an estimator, constructing confidence intervals, and performing hypothesis tests.

	Target Parameter	Sample size	Point Estimator	$\mathbb{E}[\theta]$	Standard Error
Population Mean	$\mu$	$n$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	$\mu$	$\frac{\sigma}{\sqrt{n}}$
Proportion	$p$	$n$	$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$	$p$	$\sqrt{\frac{p(1-p)}{n}}$
Difference in Means	$\mu_1 - \mu_2$	$m, n$	$\bar{X} - \bar{Y}$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
Difference in Proportions	$p_1 - p_2$	$m, n$	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$

 **Example 6.1.1.** In a random sample of 80 components of a certain type, 12 are found to be defective.

1. Find a point estimate of the proportion of non-defective components.
2. Find the standard error of the point estimate.


 **Solution** 1. With  $p$  as the proportion of non-defective components, the point estimate for proportion is

$$\hat{p} = \frac{80 - 12}{80} = 0.85.$$

2. The standard error of the point estimate of non-defective proportion is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.85 \times 0.15}{80}} \approx 0.0399.$$

◀

 **Example 6.1.2.** Let  $X$  and  $Y$  denote the strengths of concrete beam and cylinder specimens, respectively. The following data were obtained:

$X$	5.9	7.2	7.3	6.3	8.1	6.8	7.0
	7.6	6.8	6.5	7.0	6.3	7.9	9.0
	8.2	8.7	7.8	9.7	7.4	7.7	9.7
	7.8	7.7	11.6	11.3	11.8	10.7	
$Y$	6.1	5.8	7.8	7.1	7.2	9.2	6.6
	8.3	7.0	8.3	7.8	8.1	7.4	8.5
	8.9	9.8	9.7	14.1	12.6	11.2	

Suppose  $\mathbb{E}[X] = \mu_1$ ,  $Var[X] = \sigma_1^2$ ,  $\mathbb{E}[Y] = \mu_2$ , and  $Var[Y] = \sigma_2^2$ .

1. Show that  $\bar{X} - \bar{Y}$  is an unbiased estimator of  $\mu_1 - \mu_2$ .

2. Find the mean and standard error of the point estimate of  $\mu_1 - \mu_2$ .

**Solution** 1. Since  $X$  and  $Y$  are independent, we have

$$\mathbb{E}[\bar{X} - \bar{Y}] = \mathbb{E}[\bar{X}] - \mathbb{E}[\bar{Y}] = \mu_1 - \mu_2.$$

Thus,  $\bar{X} - \bar{Y}$  is an unbiased estimator of  $\mu_1 - \mu_2$ .

2. The mean of the point estimate is

$$\mathbb{E}[\bar{X} - \bar{Y}] = \bar{x} - \bar{y} = 8.141 - 8.575 = 0.434.$$

The variance of the difference in means is

$$\text{Var}[\bar{X} - \bar{Y}] = \text{Var}[\bar{X}] + \text{Var}[\bar{Y}] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

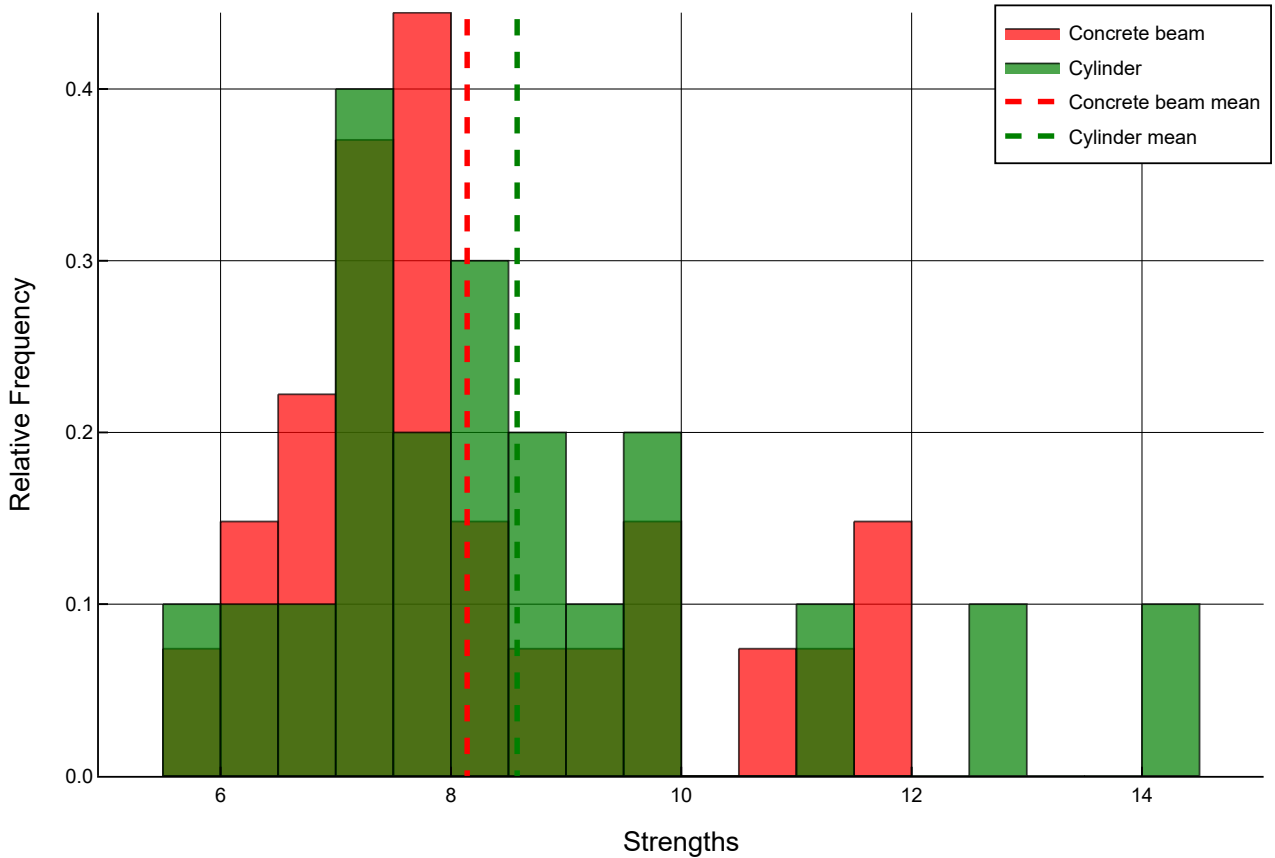
And

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}.$$

Since  $\sigma_1^2$  and  $\sigma_2^2$  are unknown, we use  $s_X^2$  and  $s_Y^2$  to estimate  $\sigma_1^2$  and  $\sigma_2^2$  respectively. Thus,

The standard error of the point estimate is

$$S_{\bar{X}-\bar{Y}} = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = \sqrt{\frac{1.666^2}{27} + \frac{2.104^2}{20}} = 0.5687.$$



**Remark.** Note that  $S_1$  is not an unbiased estimator of  $\sigma_1$ . Similarly,  $S_1/S_2$  is not an unbiased estimator of  $\sigma_1/\sigma_2$ .

## 6.2 Evaluating the Estimators


Suppose  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are two estimators of  $\theta$  that are both unbiased. Then, although the distribution of each estimator is centered at the true value of  $\theta$ , the spreads of the distributions about the true value may be different.

Among all estimators of  $\theta$  that are unbiased, we will always choose the one that has minimum variance. WHY?

The resulting  $\hat{\theta}$  is called the **minimum variance unbiased estimator (MVUE)** of  $\theta$ .

### Definition 6.2 Unbiased estimator

The estimator  $\hat{\mu}$  is unbiased if  $\text{Bias}(\hat{\mu} | \theta) = 0$

 **Example 6.2.1.** Let  $X_1, X_2, X_3$  be a random sample of size 3 from a population with pmf


$$f(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$$

where  $\lambda > 0$  is a parameter. Are the following estimators of  $\lambda$  unbiased?

$$\hat{\lambda}_1 = \frac{1}{4}(X_1 + 2X_2 + X_3), \quad \hat{\lambda}_2 = \frac{1}{9}(4X_1 + 3X_2 + 2X_3)$$

Given,  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  which one is more efficient?

Hence, find an unbiased estimator of  $\lambda$  that is more efficient than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ .

 **Solution** Given the observations  $X_1, X_2, X_3$  are i.i.d with  $X_i \sim \text{Poisson}(\lambda)$ , we have

$$\mathbb{E}[X_i] = \text{Var}[X_i] = \lambda \quad \forall i = 1, 2, 3.$$

It is easy to see that

$$\begin{aligned} \mathbb{E}[\hat{\lambda}_1] &= \frac{1}{4}(\mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_3]) = \frac{1}{4}(\lambda + 2\lambda + \lambda) = \lambda, \\ \mathbb{E}[\hat{\lambda}_2] &= \frac{1}{9}(4\mathbb{E}[X_1] + 3\mathbb{E}[X_2] + 2\mathbb{E}[X_3]) = \frac{1}{9}(4\lambda + 3\lambda + 2\lambda) = \lambda. \end{aligned}$$

Thus, both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  are unbiased estimators of  $\lambda$ . Next, we compute the variances of both estimators,

$$\begin{aligned} \text{Var}[\hat{\lambda}_1] &= \frac{1}{16}(\text{Var}[X_1] + 4\text{Var}[X_2] + \text{Var}[X_3]) = \frac{1}{16}(\lambda + 4\lambda + \lambda) = \frac{3\lambda}{8}, \\ \text{Var}[\hat{\lambda}_2] &= \frac{1}{81}(16\text{Var}[X_1] + 9\text{Var}[X_2] + 4\text{Var}[X_3]) = \frac{1}{81}(16\lambda + 9\lambda + 4\lambda) = \frac{29\lambda}{81}. \end{aligned}$$

By inspection, since  $\frac{3}{8} = 0.375 > \frac{29}{81} \approx 0.358$ , the estimator  $\hat{\lambda}_2$  is more efficient than  $\hat{\lambda}_1$ . We have seen in previous section that the sample mean is always an unbiased estimator of the population mean irrespective of the population distribution. The variance of the sample mean is always equal

to  $\frac{\sigma^2}{n}$ , where  $\sigma^2$  is the population variance and  $n$  is the sample size. Thus

$$\text{Var}[\bar{X}] = \frac{\text{Var}[X_i]}{3} = \frac{1}{3}\lambda.$$

The sample mean has even smaller variance than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ . Thus,  $\bar{X} = \frac{1}{3}\lambda$  is an unbiased estimator of  $\lambda$  that is more efficient than both  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$ . ◀

✎ **Example 6.2.2.** Let  $\hat{\theta}_1$  and  $\hat{\theta}_2$  be unbiased estimators of  $\theta$ . Suppose  $\text{Var}(\hat{\theta}_1) = 1$ ,  $\text{Var}(\hat{\theta}_2) = 2$  and  $\text{Cov}(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{2}$ . What are the values of  $c_1$  and  $c_2$  for which  $c_1\hat{\theta}_1 + c_2\hat{\theta}_2$  is an unbiased estimator of  $\theta$  with minimum variance among unbiased estimators of this type?

⇒ **Solution** We want to find  $c_1$  and  $c_2$  such that  $c_1\hat{\theta}_1 + c_2\hat{\theta}_2$  to be a minimum variance unbiased estimator of  $\theta$ . Then

$$\begin{aligned}\mathbb{E}[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] &= \theta \implies c_1\mathbb{E}[\hat{\theta}_1] + c_2\mathbb{E}[\hat{\theta}_2] = \theta \\ &\implies c_1\theta + c_2\theta = \theta \\ &\implies c_1 + c_2 = 1 \\ &\implies c_2 = 1 - c_1.\end{aligned}$$

Therefore,

$$\begin{aligned}\text{Var}[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] &= c_1^2\text{Var}[\hat{\theta}_1] + c_2^2\text{Var}[\hat{\theta}_2] + 2c_1c_2\text{Cov}[\hat{\theta}_1, \hat{\theta}_2] \\ &= c_1^2(1) + 2(1 - c_1)^2 + 2c_1(1 - c_1)\left(\frac{1}{2}\right) \\ &= 3c_1^2 - 3c_1 + 2.\end{aligned}$$

To find the minimum variance, we differentiate  $\text{Var}[c_1\hat{\theta}_1 + c_2\hat{\theta}_2]$  with respect to  $c_1$  and set it to zero, that is

$$\frac{d}{dc_1}\text{Var}[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] = 6c_1 - 3 = 0 \implies c_1 = \frac{1}{2}.$$

Thus,  $c_2 = 1 - c_1 = \frac{1}{2}$ . Therefore, the minimum variance unbiased estimator of  $\theta$  is

$$\hat{\theta} = \frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_2.$$

In fact, if  $\theta_1$  and  $\theta_2$  are both unbiased estimators of  $\theta$ , then the linear combination  $c_1\theta_1 + c_2\theta_2$  is also an unbiased estimator of  $\theta$  for any  $c_1, c_2$  such that  $c_1 + c_2 = 1$ . Hence

$$\mathcal{C} = \{\hat{\theta} = c\hat{\theta}_1 + (1 - c)\hat{\theta}_2 \mid c \in \mathbb{R}\}$$

◀

**Rule of thumb choosing a good estimator:**


❖ Unbiasedness:  $\mathbb{E}[\hat{\theta}] = \theta$ .

❖ Minimum variance: A good estimator should has smaller  $\text{Var}[\hat{\theta}]$ , the smaller the better.

### 6.2.1 Method of Moments (MoM) Estimator


The method of moments is a technique for estimating population parameters by equating sample moments to theoretical moments. Moments are quantitative measures related to the shape of a distribution, such as the mean (first moment), variance (second moment), skewness (third moment), and kurtosis (fourth moment). The method of moments involves the following steps:

1. Calculate the theoretical (population) moments as functions of the unknown parameters.
2. Calculate the corresponding sample moments from the observed data.
3. Set the population moments equal to the sample moments to create a system of equations.
4. Solve the system of equations for the unknown parameters to obtain the MoM estimators.

 **Example 6.2.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where  $\theta > 0$  is an unknown parameter. Find the method of moments estimator of  $\theta$ .

 **Solution** To find the method of moments estimator, we shall equate the first population moment to the sample moment. The first population moment  $\mathbb{E}[X]$  is given by

$$\begin{aligned} \mathbb{E}[X] &= \int_0^1 x f(x|\theta) dx \\ &= \int_0^1 x(\theta x^{\theta-1}) dx \\ &= \theta \int_0^1 x^{\theta} dx \\ &= \theta \left[ \frac{x^{\theta+1}}{\theta+1} \right]_{x=0}^{x=1} = \frac{\theta}{\theta+1} = M_X(x). \end{aligned}$$

We know that the first moment  $M_X(x) = \bar{X}$ . Now setting  $M_X(x) = \mathbb{E}X$  and solving for  $\theta$ , we have

$$\bar{X} = \frac{\theta}{\theta+1}$$

that is

$$\theta = \frac{\bar{X}}{1 - \bar{X}}.$$

where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  is the sample mean. Thus, the statistic  $\frac{\bar{X}}{1 - \bar{X}}$  is an estimator of parameter  $\theta$ . We write

$$\hat{\theta} = \frac{\bar{X}}{1 - \bar{X}}.$$

Now let say we have the following sample data:

$$0.44, \quad 0.55, \quad 0.60, \quad 0.30$$



we have  $\bar{X} = \frac{0.44 + 0.55 + 0.60 + 0.30}{4} = 0.4725$ , and the estimate of  $\theta$  is

$$\hat{\theta} = \frac{0.4725}{1 - 0.4725} = 0.8957.$$

◀

**Example 6.2.4.** Let  $X \sim \mathcal{N}(\mu, \sigma^2)$ , and  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from  $X$ . Find the method of moments estimators of  $\mu$  and  $\sigma^2$ .

**Solution** The first population moment is

$$\mathbb{E}[X] = \mu.$$

The second population moment is

$$\mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2.$$

The first sample moment is

$$M_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

The estimator of the parameter  $\mu$  is  $\hat{\mu} = \bar{X}$ , that is

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Next, we equate the second population moment to the second sample moment. Note that the variance of the population is

$$\begin{aligned} \sigma^2 &= \mathbb{E}[X^2] - \mu^2 \\ &= M_2 - \mu^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 - (\bar{X})^2. \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2. \end{aligned}$$

The last line follows from the fact that

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

Thus, the estimator of the parameter  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$



**Theorem 6.2**

Let  $X_1, X_2, \dots, X_n$  be a random sample with  $\mathbb{E}[X_i] = \mu$  and  $\text{Var}[X_i] = \sigma^2$ . Then

$$\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

is a *biased estimator* of  $\sigma^2$  but that

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

is an *unbiased estimator* of  $\sigma^2$ .

*Proof.* From previous example, we can see that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2 \dots\dots\dots (\star)$$

Hence, we use this and the fact that

$$\mathbb{E}[X_i^2] = \text{Var}[X_i] + (\mathbb{E}[X_i])^2 = \sigma^2 + \mu^2,$$

and

$$\mathbb{E}[\bar{X}^2] = \text{Var}[\bar{X}] + (\mathbb{E}[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2,$$

and take expectation on both sides of  $(\star)$ , we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \mathbb{E} \left[ \sum_{i=1}^n X_i^2 - n(\bar{X})^2 \right] \\ &= \sum_{i=1}^n \mathbb{E}[X_i^2] - n\mathbb{E}[(\bar{X})^2] \\ &= n(\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \\ &= (n-1)\sigma^2. \end{aligned}$$

It follows that

$$\mathbb{E}[\tilde{S}^2] = \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \frac{n-1}{n} \sigma^2 \neq \sigma^2,$$

and that  $\tilde{S}^2$  is biased since  $\mathbb{E}[\tilde{S}^2] \neq \sigma^2$ . However,

$$\mathbb{E}[S^2] = \frac{1}{n-1} \mathbb{E} \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \sigma^2,$$

thus we can see that  $S^2$  is an unbiased estimator of  $\sigma^2$ . □

### 6.3 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimation (MLE) is a method used to estimate the parameters of a statistical model. The MLE is the parameter value that maximizes the likelihood function, which measures how likely it is to observe the given sample data for different parameter values. Next, we describe this method in detail.

#### Definition 6.3 Likelihood function and MLE

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with pdf/pmf  $f(x|\theta)$ , where  $\theta \in \Theta$  is an unknown parameter. The likelihood function is defined as

$$L(\theta|x) = \prod_{i=1}^n f(x_i|\theta).$$

The maximum likelihood estimator (MLE) of  $\theta$  is the value of  $\theta$  that maximizes the likelihood function  $L(\theta)$ .

This definition states that the likelihood function  $L(\theta|x)$  is the product of the individual pdf evaluated at each observation in the sample, given the parameter  $\theta$ . The likelihood function represents the joint density of a random sample  $X_1, X_2, \dots, X_n$  given the parameter  $\theta$ . The MLE is the value of  $\theta$  that makes the observed data most probable.


The  $\theta$  that maximizes  $L(\theta|x)$  is called the maximum likelihood estimate and is denoted by  $\hat{\theta}$ .

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta|x).$$

In practice, it is often more convenient to work with the natural logarithm of the likelihood function, known as the log-likelihood function:


$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^n \ln f(x_i|\theta).$$

Maximizing the log-likelihood function is equivalent to maximizing the likelihood function itself, as the logarithm is a monotonically increasing function.

 **Example 6.3.1.** Let  $X \sim B(1, p)$ , a Bernoulli random variable with parameter  $p$ , with pmf

$$f(x|p) = \mathbb{P}[X = x|p] = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

where  $0 < p < 1$ . Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from  $X$ . Find the maximum likelihood estimator of  $p$ .

 **Solution** Our goal is to find the value of  $p$  that maximizes the likelihood function based on the observed sample data  $X = (X_1, X_2, \dots, X_n)$ . Note that  $X_1, X_2, \dots, X_n$  are i.i.d. Thus, the

likelihood function is given by

$$\begin{aligned}
L(p|x) &= \prod_{i=1}^n f(x_i|p) \\
&= \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} \\
&= p^{\sum_{i=1}^n x_i} (1-p)^{\sum_{i=1}^n (1-x_i)} \\
&= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}.
\end{aligned}$$

We can simplify the notation by letting  $S_n = \sum_{i=1}^n x_i$ . We want to choose  $p$  such that  $L(p|x)$  is maximized. Take the logarithm of the likelihood function, we have

$$\ell(p|x) = \ln L(p|x) = S_n \ln p + (n - S_n) \ln(1 - p).$$

To find the maximum, we take the derivative of  $\ell(p|x)$  with respect to  $p$  and set it to zero:

$$\begin{aligned}
\frac{\partial \ell(p|x)}{\partial p} &= \frac{S_n}{p} - \frac{n - S_n}{1 - p} = 0 \\
\Rightarrow S_n(1 - p) &= (n - S_n)p \\
\Rightarrow S_n - S_n p &= np - S_n p \\
\Rightarrow S_n &= np \\
\Rightarrow \hat{p} &= \frac{S_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X}.
\end{aligned}$$

The sample mean (proportion)  $\bar{X}$  is the maximum likelihood estimator of  $p$ . ◀

### 6.3.1 MLE based on grouped data

In some cases, data may be grouped into intervals or categories, and we may not have access to the individual data points. In such situations, we can still use the maximum likelihood estimation (MLE) method to estimate parameters based on the grouped data.

In the complete data case, the likelihood is measured by the density (or probability)  $f(x_i)$  at the known data point  $x_i$ . The likelihood function is the product of those densities for the points in the sample. In the interval grouped data case, we measure likelihood of a point as the probability of the interval in which that point occurs. For a data point in the interval  $(c_{j-1}, c_j]$ , the probability of that interval is  $[F(c_j; \theta) - F(c_{j-1}; \theta)]$ . The likelihood function is the product of those interval probabilities for all of the sample points. Since there are  $n_j$  sample points in the interval  $(c_{j-1}, c_j]$ , the likelihood function will include a factor of  $[F(c_j; \theta) - F(c_{j-1}; \theta)]^{n_j}$  for those  $n_j$  points. The overall likelihood function is the product of all of those factors:

#### Definition 6.4 MLE based on grouped data

If the data is grouped into  $k$  intervals with counts  $n_1, n_2, \dots, n_k$  in each interval, the likelihood function for the grouped data is given by

$$L(\theta|x) = \prod_{j=1}^k [F(c_j|\theta) - F(c_{j-1}|\theta)]^{n_j},$$

where  $F(x|\theta)$  is the cumulative distribution function (CDF) of the underlying distribution, and  $[c_{j-1}, c_j)$  is the  $j$ -th interval.

**Example 6.3.2.** For a group of insurance policies, you are given:

1. The losses follow the distribution function

$$F(x|\theta) = 1 - \frac{\theta}{x}, \quad \theta < x < \infty.$$

2. A sample of 20 losses is grouped as follows:

Interval	Number of loss
$x \leq 10$	9
$10 < x \leq 25$	6
$x > 25$	5

Calculate the maximum likelihood estimate of  $\theta$ .

**Solution** The likelihood function is the product of the probabilities of observing the data in each interval, The probability for the interval  $x \leq 10$  is given by

$$F(10|\theta) = 1 - \frac{\theta}{10},$$

the probability for the interval  $10 < x \leq 25$  is given by

$$F(25|\theta) - F(10|\theta) = \left(1 - \frac{\theta}{25}\right) - \left(1 - \frac{\theta}{10}\right) = \frac{\theta}{10} - \frac{\theta}{25} = \frac{3}{50}\theta,$$

and the probability for the interval  $x > 25$  is given by

$$1 - F(25|\theta) = 1 - \left(1 - \frac{\theta}{25}\right) = \frac{\theta}{25}.$$

Then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= [F(10|\theta)]^{n_1} [F(25|\theta) - F(10|\theta)]^{n_2} [1 - F(25|\theta)]^{n_3} \\ &= \left(1 - \frac{\theta}{10}\right)^9 \left(\frac{3\theta}{50}\right)^6 \left(\frac{\theta}{25}\right)^5 \\ &= c(10 - \theta)^9 \theta^{11}, \end{aligned}$$

where  $c = \frac{3^6}{50^6 \times 25^5}$  is a constant. To find the value of  $\theta$  that maximizes  $L(\theta)$ , we take the logarithm of the likelihood function:

$$\ell(\theta) = \ln L(\theta) = \ln c + 9 \ln(10 - \theta) + 11 \ln \theta.$$

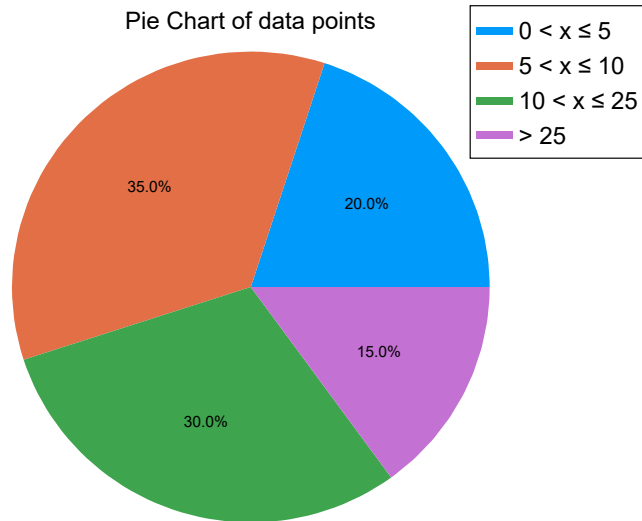
Next, we differentiate  $\ell(\theta)$  with respect to  $\theta$  and set it to zero:

$$\begin{aligned}\frac{d\ell(\theta)}{d\theta} &= \frac{9}{10-\theta}(-1) + \frac{11}{\theta} = 0 \\ \Rightarrow -\frac{9}{10-\theta} + \frac{11}{\theta} &= 0 \\ \Rightarrow 11(10-\theta) &= 9\theta \\ \Rightarrow 110 - 11\theta &= 9\theta \\ \Rightarrow 110 &= 20\theta \\ \Rightarrow \hat{\theta} &= 5.5\end{aligned}$$

Thus, the maximum likelihood estimate of  $\theta$  is  $\hat{\theta} = 5.5$ . ◀

📎 **Example 6.3.3.** A grouped data set has 20 data points grouped into the following intervals:

Interval	$0 < x \leq 5$	$5 < x \leq 10$	$10 < x \leq 25$	$x > 25$
Number of data points	4	7	6	3



Apply the maximum likelihood method to estimate the parameter  $\theta$  of the following two cases:

1. The data follow the exponential distribution with parameter  $\theta$ ,
2. The data follow the uniform distribution on the interval  $(0, \theta)$ .

⇒ **Solution** 1. If we assume that the data follow the exponential distribution with parameter  $\theta$ , then the cdf is given by

$$F(x|\theta) = 1 - e^{-\theta x}, \quad x > 0, \theta > 0.$$

The likelihood function is given by

$$\begin{aligned}
L(\theta) &= [F(5|\theta)]^{n_1} [F(10|\theta) - F(5|\theta)]^{n_2} [F(25|\theta) - F(10|\theta)]^{n_3} [1 - F(25|\theta)]^{n_4} \\
&= (1 - e^{-5\theta})^4 (e^{-5\theta} - e^{-10\theta})^7 (e^{-10\theta} - e^{-25\theta})^6 (e^{-25\theta})^3 \\
&= c(1 - e^{-5\theta})^4 (e^{-5\theta})^7 (1 - e^{-5\theta})^6 (e^{-25\theta})^3 \\
&= c(1 - e^{-5\theta})^{10} (e^{-5\theta})^{10} (e^{-25\theta})^3 \\
&= c(1 - e^{-5\theta})^{10} e^{-50\theta},
\end{aligned}$$

where  $c = 1$  is a constant. To find the value of  $\theta$  that maximizes  $L(\theta)$ , we take the logarithm of the likelihood function:

2. In another case, if we assume that the data follow the uniform distribution on the interval  $(0, \theta)$ , then the cdf is given by

$$F(x|\theta) = \begin{cases} \frac{x}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function is given by

$$\begin{aligned}
L(\theta) &= [F(5|\theta)]^{n_1} [F(10|\theta) - F(5|\theta)]^{n_2} [F(25|\theta) - F(10|\theta)]^{n_3} [1 - F(25|\theta)]^{n_4} \\
&= \left(\frac{5}{\theta}\right)^4 \left(\frac{10}{\theta} - \frac{5}{\theta}\right)^7 \left(\frac{25}{\theta} - \frac{10}{\theta}\right)^6 \left(1 - \frac{25}{\theta}\right)^3 \\
&= c\theta^{-20}(\theta - 25)^3,
\end{aligned}$$

where  $c = 5^4 \times 5^7 \times 15^6$  is a constant. To find the value of  $\theta$  that maximizes  $L(\theta)$ , we take the logarithm of the likelihood function:

$$\ell(\theta) = \ln L(\theta) = \ln c - 20 \ln \theta + 3 \ln(\theta - 25).$$

Next, we differentiate  $\ell(\theta)$  with respect to  $\theta$  and set it to zero:

$$\begin{aligned}
\frac{d\ell(\theta)}{d\theta} &= -\frac{20}{\theta} + \frac{3}{\theta - 25} = 0 \\
\Rightarrow -20(\theta - 25) + 3\theta &= 0 \\
\Rightarrow -20\theta + 500 + 3\theta &= 0 \\
\Rightarrow -17\theta + 500 &= 0 \\
\Rightarrow \hat{\theta} &= \frac{500}{17} \approx 29.41.
\end{aligned}$$

Thus, the maximum likelihood estimate of  $\theta$  is  $\hat{\theta} = \frac{500}{17} \approx 29.41$ .



## Tutorials

**Exercise 6.1** Given 20 observations on breakdown voltage for some materials

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

Assume that after looking at the histogram, we think that the distribution of breakdown voltage is normal with mean value  $\mu$ . What are some point estimators for  $\mu$ ?

**Exercise 6.2** The probability density function of the random variable  $X$  is defined by

$$f(x|\lambda) = 1 - \frac{2}{3}\lambda + \lambda\sqrt{x}, \quad 0 \leq x \leq 1,$$

and zero otherwise. What is the maximum likelihood estimate of the parameter  $\lambda$  based on the two independent observations  $x_1 = \frac{1}{4}$  and  $x_2 = \frac{9}{16}$ ?

**Exercise 6.3** Consider that  $X_1, X_2, \dots, X_n$  is a random sample from a distribution with density function

$$f(x|\beta) = \begin{cases} \frac{x^6 e^{-x/\beta}}{\Gamma(7)\beta^7} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimate of  $\beta$ ?

**Exercise 6.4** Let  $X_1, X_2, \dots, X_n$  be a random sample from the uniform density function

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 2\theta \leq x \leq 3\theta \\ 0 & \text{otherwise} \end{cases}$$

where  $\theta > 0$ . Show that the maximum likelihood estimate of  $\theta$  is

$$\frac{1}{3} \max\{X_1, X_2, \dots, X_n\}.$$

**Exercise 6.5** What is the maximum likelihood estimate of  $\beta$  if five values  $\frac{4}{5}, \frac{2}{3}, 1, \frac{3}{2}, \frac{5}{4}$  were drawn from the population for which  $f(x|\beta) = \frac{1}{2}(1+\beta)^5 \left(\frac{x}{2}\right)^\beta$ ?

**Exercise 6.6** Eight independent trials are conducted of a given system with the following results:

$$S, F, S, F, S, S, S, S$$

where  $S$  denotes the success and  $F$  denotes the failure. What is the maximum likelihood estimate of the probability of successful operation  $p$ ?

**Exercise 6.7** Suppose fertilizer-1 has a mean yield per acre of  $\mu_1$  with variance  $\sigma^2$ , whereas the expected yield for fertilizer-2 is  $\mu_2$  with the same variance  $\sigma^2$ . Let  $S_i^2$  denote the sample variances of yields based on sample sizes  $n_1$  and  $n_2$  respectively, of the two fertilizers.



1. Show that the pooled (combined) estimator

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of  $\sigma^2$ .

2. The measurements for the two types of fertilizers were obtained independently, with testing samples of  $n_1 = n_2 = 100$ , the following sample means and variances were computed:

$$\begin{aligned}\bar{x}_1 &= 179.3 \text{ yield per acre,} & \bar{x}_2 &= 190.0 \text{ yield per acre,} \\ s_1^2 &= 1440.80, & s_2^2 &= 1960,\end{aligned}$$

Estimate the difference mean yield per acre, and the pooled variance of two fertilizers.

**Exercise 6.8** A random sample  $X_1, X_2, \dots, X_n$  of size  $n$  is selected from a normal distribution with variance  $\sigma^2$ . Let  $S^2$  be the unbiased estimator of  $\sigma^2$ , and  $T$  be the maximum likelihood estimator of  $\sigma^2$ . If  $20T - 19S^2 = 0$ , then what is the sample size?

**Exercise 6.9** A box contains 50 red and blue balls out of which are red. A sample of 30 balls is to be selected without replacement. If  $X$  denotes the number of red balls in the sample, then find an estimator for using the *moment method*.

**Exercise 6.10** What is the difference between the parameters and statistics?

**Exercise 6.11** You are given:

1. Losses follow an exponential distribution with mean  $\theta$ .
2. A random sample of 20 losses is observed as follows:

Loss Range	Frequency
[0, 1000]	7
(1000, 2000]	6
(2000, $\infty$ )	7

Calculate the maximum likelihood estimate of  $\theta$ .

# Evaluating the goodness of estimators

## 7.1 Relative efficiency

It is usually possible to retrieve more than one unbiased estimator for the same target parameter  $\theta$ . But we only prefer to use the estimator with the **smaller variance**. That is, if  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are unbiased estimators for the same parameter  $\theta$ , we said that  $\hat{\theta}_1$  is *relatively more efficient* than  $\hat{\theta}_2$  if


$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2).$$

In fact, this can be expressed as the ratio  $\text{Var}(\hat{\theta}_1)/\text{Var}(\hat{\theta}_2)$  to measure the relative efficiency of these two unbiased estimators.

### Definition 7.1 Relative efficiency

Given two unbiased estimators  $\hat{\theta}_1$  and  $\hat{\theta}_2$  for the same parameter  $\theta$ , with variances  $\text{Var}(\hat{\theta}_1)$  and  $\text{Var}(\hat{\theta}_2)$ , respectively. Then the efficiency of  $\hat{\theta}_1$  relative to  $\hat{\theta}_2$ , wrote as  $\text{Eff}(\hat{\theta}_1, \hat{\theta}_2)$ , is defined to be the ratio

$$\text{Eff}(\hat{\theta}_1, \hat{\theta}_2) := \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}. \quad (7.1)$$

 **Example 7.1.1.** If  $Y_1, Y_2, \dots, Y_n$  denote a random sample from the uniform distribution on the interval  $(0, \theta)$ . The two unbiased estimators for  $\theta$  are

$$\hat{\theta}_1 = 2\bar{Y}, \quad \hat{\theta}_2 = \left(\frac{n+1}{n}\right) Y_{(n)},$$

where  $Y_{(n)} = \max\{Y_1, Y_2, \dots, Y_n\}$ . Find the efficiency of  $\hat{\theta}_1$  relative to  $\hat{\theta}_2$ .

 **Solution** Because each  $Y_i \sim U(0, \theta)$ , we have

$$\mu = \mathbb{E}[Y_i] = \frac{\theta}{2}, \quad \sigma^2 = \text{Var}(Y_i) = \frac{\theta^2}{12}.$$

Therefore,

$$\mathbb{E}[\hat{\theta}_1] = \mathbb{E}(2\bar{Y}) = 2\mathbb{E}(\bar{Y}) = 2\mu = \theta,$$

and that  $\hat{\theta}_1$  is unbiased, as claimed. Further we check that

$$\text{Var}(\hat{\theta}_1) = \text{Var}(2\bar{Y}) = 4\text{Var}(\bar{Y}) = 4 \cdot \frac{\sigma^2}{n} = \frac{\theta^2}{3n},$$

The mean of this order statistic is

$$\begin{aligned}\mathbb{E}[Y_{(n)}] &= \int_0^\theta y \cdot n \left(\frac{y}{\theta}\right)^{n-1} \left(\frac{1}{\theta}\right) dy = \frac{n}{\theta^n} \int_0^\theta y^n dy \\ &= \frac{n}{\theta^n} \left[ \frac{y^{n+1}}{n+1} \right]_{y=0}^{y=\theta} \\ &= \frac{n\theta}{n+1}.\end{aligned}$$

and it follows that

$$\mathbb{E}\left[\frac{n+1}{n} Y_{(n)}\right] = \theta;$$

that is,  $\hat{\theta}_2$  is also an unbiased estimator of  $\theta$ . Since

$$\begin{aligned}\mathbb{E}[Y_{(n)}^2] &= \int_0^\theta y^2 \cdot n \left(\frac{y}{\theta}\right)^{n-1} \left(\frac{1}{\theta}\right) dy = \frac{n}{\theta^n} \int_0^\theta y^{n+1} dy \\ &= \frac{n}{\theta^n} \left[ \frac{y^{n+2}}{n+2} \right]_{y=0}^{y=\theta} \\ &= \frac{n\theta^2}{n+2}.\end{aligned}$$

we obtain

$$\begin{aligned}Var[Y_{(n)}] &= \mathbb{E}[Y_{(n)}^2] - (\mathbb{E}[Y_{(n)}])^2 = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \theta^2 \left[ \frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right] \\ &= \theta^2 \left[ \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \right] \\ &= \theta^2 \left[ \frac{n^3 + 2n^2 + n - n^3 - 2n^2}{(n+2)(n+1)^2} \right] \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}.\end{aligned}$$

and hence

$$\begin{aligned}Var[\hat{\theta}_2] &= Var\left[\frac{n+1}{n} Y_{(n)}\right] \\ &= \left(\frac{n+1}{n}\right)^2 Var[Y_{(n)}] \\ &= \left(\frac{n+1}{n}\right)^2 \cdot \frac{n\theta^2}{(n+2)(n+1)^2} \\ &= \frac{\theta^2}{n(n+2)}.\end{aligned}$$

Therefore, the efficiency of  $\hat{\theta}_1$  relative to  $\hat{\theta}_2$  is

$$\text{Eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)} = \frac{\theta^2/[n(n+2)]}{\theta^2/(3n)} = \frac{3}{n+2}.$$

This efficiency is less than 1 if  $n > 1$ . That is, if  $n > 1$ , then  $\hat{\theta}_2$  must have a smaller variance than  $\hat{\theta}_1$ . And in result,  $\theta_2$  is generally more preferable to  $\theta_1$  when estimating  $\theta$ . ◀

## 7.2 Consistency

### Definition 7.2 Consistency

An estimator  $\hat{\theta}_n$  is consistent for parameter  $\theta$  if for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) = 0. \quad (7.2)$$

In other words,  $\hat{\theta}_n$  converges in probability to  $\theta$  as  $n$  goes to infinity, denoted by

$$\hat{\theta}_n \xrightarrow{p} \theta, \quad \text{as } n \rightarrow \infty.$$

### Theorem 7.1


An unbiased estimator  $\hat{\theta}_n$  is consistent for parameter  $\theta$  if

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0. \quad (7.3)$$

 **Example 7.2.1.** Let  $X_1, X_2, \dots, X_n$  be a random sample such that  $\mathbb{E}[X_i] = \mu$ ,  $\mathbb{E}[X_i^2] = \mu'_2$  and  $\mathbb{E}[X_i^4] = \mu'_4$  are all finite. Show that

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

is a consistent estimator of  $\sigma^2 = \text{Var}(X_i)$ .

 **Solution** We are going to use subscript  $n$  on both  $\hat{\sigma}_n^2$  and  $\bar{X}$  to explicitly convey their dependence on the value of the sample size  $n$ .

From previous example, we had derived that

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

The statistic  $\frac{1}{n} \sum_{i=1}^n X_i^2$  is the average of  $n$  independent and identically distributed random variables  $X_1^2, X_2^2, \dots, X_n^2$  with  $\mathbb{E}[X_i^2] = \mu'_2$  and  $V(X_i^2) = \mu'_4 - (\mu'_2)^2$ .

By the weak law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{p} \mu'_2, \quad \text{as } n \rightarrow \infty,$$

and certainly  $\bar{X}_n \xrightarrow{p} \mu$ . And because the function  $g(x) = x^2$  is continuous for all  $x \in \mathbb{R}$ . This implies


that  $\bar{X}_n^2 \xrightarrow{p} \mu^2$ . It follows that

$$\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \xrightarrow{p} \sigma^2.$$

Note that  $\frac{n}{n-1} \rightarrow 1$  when  $n$  goes to infinity. We can conclude that


$$\hat{\sigma}_n^2 \xrightarrow{p} (1)\sigma^2 = \sigma^2, \quad \text{as } n \rightarrow \infty.$$

Equivalently,  $\hat{\sigma}_n^2$ , the sample variance, is a consistent estimator of  $\sigma^2$ , the population variance. ◀

 **Example 7.2.2.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal population  $X$  with mean  $\mu$  and variance  $\sigma^2 > 0$ . Is the likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

of  $\sigma^2$  a consistent estimator of  $\sigma^2$ ?

 **Solution** Since  $\hat{\sigma}^2$  depends on the sample size  $n$ , we denote  $\hat{\sigma}^2$  as  $\hat{\sigma}_n^2$ . Hence

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

The variance of  $\hat{\sigma}_n^2$  is given by

$$\begin{aligned} \text{Var}(\hat{\sigma}_n^2) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{1}{n^2} \text{Var}\left(\sigma^2 \frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^4}{n^2} \text{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^4}{n^2} \text{Var}(\chi^2(n-1)) \\ &= \frac{2(n-1)\sigma^4}{n^2} \\ &= \left[\frac{1}{n} - \frac{1}{n^2}\right] 2\sigma^4. \end{aligned}$$

Hence

$$\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = \lim_{n \rightarrow \infty} \left[\frac{1}{n} - \frac{1}{n^2}\right] 2\sigma^4 = 0.$$

The biased  $B(\hat{\theta}_n, \theta)$  is given by

$$\begin{aligned}
 B(\hat{\theta}_n, \theta) &= E(\hat{\theta}_n) - \sigma^2 = E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) - \sigma^2 \\
 &= \frac{1}{n} E\left(\sigma^2 \frac{(n-1)S^2}{\sigma^2}\right) - \sigma^2 \\
 &= \frac{\sigma^2}{n} E(\chi^2(n-1)) - \sigma^2 \\
 &= \frac{(n-1)\sigma^2}{n} - \sigma^2 \\
 &= -\frac{\sigma^2}{n}.
 \end{aligned}$$

Thus

$$\lim_{n \rightarrow \infty} B(\hat{\theta}_n, \theta) = -\lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0.$$

Hence  $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$  is a consistent estimator of  $\sigma^2$ .


In the last example we saw that the likelihood estimator of variance is a consistent estimator. In general, if the density function  $f(x|\theta)$  of a population satisfies some mild conditions, then the maximum likelihood estimator of  $\theta$  is consistent. Similarly, if the density function  $f(x; \theta)$  of a population satisfies some mild conditions, then the estimator obtained by moment method is also consistent.

Since consistency is a large sample property of an estimator, some statisticians suggest that consistency should not be used alone for judging the goodness of an estimator; rather it should be used along with other criteria. ◀

## 7.3 Sufficiency

### Definition 7.3 Sufficiency

A statistic  $T(X)$  is sufficient for parameter  $\theta \in \Theta$  if the conditional distribution of the sample  $X$  given the statistic  $T(X)$  does not depend on the parameter  $\theta$ . In other words, once we know the value of the sufficient statistic, the sample provides no additional information about the parameter.

 **Example 7.3.1.** If  $X_1, X_2, \dots, X_n$  are i.i.d. random samples from the Bernoulli distribution with parameter  $p$ , then the sum of the samples with density function

$$f_X(x|\theta) = \begin{cases} \theta^x (1-\theta)^{1-x}, & x = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

where  $0 < \theta < 1$ . Show that the statistic  $T(X) = \sum_{i=1}^n X_i$  is a sufficient statistic for  $\theta$ .

 **Solution** First, we find the joint density function of the sample:

$$f_X(x|\theta) = \prod_{i=1}^n f_{X_i}(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}.$$

Since each  $X_i$  is either 0 or 1, the sum  $\sum_{i=1}^n x_i$  counts the number of successes (1s) in the sample. Let  $T(X) = \sum_{i=1}^n X_i$ . Then we can rewrite the joint density function as:

$$Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta),$$

Thus, the joint density function can be expressed as:

$$Y \sim g(y) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \quad y = 0, 1, \dots, n.$$

◀

**Example 7.3.2.** Let  $X_1, \dots, X_n$  be iid  $N(\theta, \sigma_0^2)$  r.v.'s where  $\sigma_0^2$  is known. Evaluate whether  $T(X) = (\sum_{i=1}^n X_i)$  is sufficient for  $\theta$ .

**Solution** We consider the transformation of  $X = (X_1, X_2, \dots, X_n)$  to  $Y = (T, Y_2, Y_3, \dots, Y_n)$  where  $T = \sum X_i$  and  $Y_2 = X_2 - X_1, Y_3 = X_3 - X_1, \dots, Y_n = X_n - X_1$ . The transformation is 1-1, and the Jacobian of the transformation is 1.

The joint distribution of  $X|\theta$  is  $N_n(\mu \times 1, \sigma_0^2 I_n)$ , where  $\mu$  represents the mean parameter  $\theta$  and 1 is the vector of ones. The joint distribution of  $Y|\theta$  is  $N_n(\mu_Y, \Sigma_{YY})$  where  $\mu_Y = (n\theta, 0, 0, \dots, 0)^T$  and the covariance matrix is

$$\Sigma_{YY} = \begin{bmatrix} n\sigma_0^2 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 2\sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & 2\sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & \sigma_0^2 & 2\sigma_0^2 & \cdots & \sigma_0^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \cdots & 2\sigma_0^2 \end{bmatrix}.$$

Since  $T$  and  $(Y_2, \dots, Y_n)$  are independent, it follows that  $(Y_2, \dots, Y_n)$  given  $T = t$  has the unconditional distribution, which means  $T$  is a sufficient statistic for  $\theta$ .

We note that all functions of  $(Y_2, \dots, Y_n)$  are independent of  $\theta$  and  $T$ , which yields independence of  $\bar{X}$  and  $s^2$  where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \left[ \sum_{j=1}^n (X_j - X_j)^2 \right].$$

◀

However, the sufficient statistics are not unique. For example, if  $T(X)$  is a sufficient statistic, then any one-to-one function of  $T(X)$  is also a sufficient statistic. the observation  $X$  itself is always a sufficient for  $\theta$ , but it is not very useful since it does not reduce the data; Said, if we take  $T(X) = X$ , then  $g(t, \theta) = f_X(t|\theta)$  and  $h(x) = 1$ .

But this is not much useful since it does not reduce the data.

For example, if the sample space  $\mathcal{X}^n$  is partitioned into subsets  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_k$  such that the conditional distribution of  $X$  given  $X \in \mathcal{A}_i$  does not depend on  $\theta$ , then the statistic  $T(X)$  that indicates which subset  $X$  belongs to is a sufficient statistic for  $\theta$ .

**Definition 7.4 Minimal sufficient statistic**

A sufficient statistic  $T(X)$  is minimal sufficient if it is a function of every other sufficient statistic. For example, if  $S(X)$  is another sufficient statistic, then

$$S(X) = S(Y) \implies T(X) = T(Y).$$

**Theorem 7.2**

Consider a statistical decision problem with sample space

❖ random variable  $X$  with measure  $\mathbb{P}_\theta$ , the parameter  $\theta \in \Theta$

**Theorem 7.3 Factorization theorem**

A statistic  $T(X)$  with range  $\mathcal{T}$  sufficient for parameter  $\theta \in \Theta$  if and only if there exists functions

$$g(t, \theta) : \mathcal{T} \times \Theta \rightarrow [0, \infty) \quad \text{and} \quad h(x) : \mathcal{X}^n \rightarrow [0, \infty) \quad (7.4)$$

such that the joint density function of the sample can be factored as

$$p(x|\theta) = g(T(x), \theta)h(x), \quad \forall x \in \mathcal{X}^n, \theta \in \Theta. \quad (7.5)$$

*Proof.* ( $\Rightarrow$ ) Consider the *discrete case* where

$$p(x|\theta) = P(X = x|\theta).$$

First of all, suppose  $T$  is sufficient for  $\theta$ . Then, by definition, the conditional distribution of  $X$  given  $T(X) = t$  is independent of  $\theta$  and we can write

$$\begin{aligned} \mathbb{P}_\theta(x) &= \mathbb{P}_\theta(X = x, T = t(x)) \\ &= \mathbb{P}_\theta(X = x|T = t(x))\mathbb{P}_\theta(T = t(x)) \\ &= g(t(x), \theta)h(x) \end{aligned}$$

where  $g(t, \theta) = \mathbb{P}_\theta(T = t)$  and

$$h(x) = \begin{cases} \mathbb{P}_\theta(X = x|T = t(x)), & \text{if } \mathbb{P}_\theta(T = t(x)) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

( $\Leftarrow$ ) Next, suppose that  $\mathbb{P}_\theta(x)$  satisfies the factorization theorem, i.e.,

$$\mathbb{P}_\theta(x) = g(T(x), \theta)h(x).$$

Then, fix a statistic  $t_0$  on  $\mathbb{P}_\theta(T = t_0) > 0$  for some  $\theta \in \Theta$ . Then □

📌 **Example 7.3.3.** Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with density function

$$f_X(x_i|\theta) = \begin{cases} \frac{1}{\theta} e^{-x_i/\theta}, & 0 \leq x < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

where parameter  $\theta > 0$  for  $i = 1, 2, \dots, n$ . Show that  $\bar{X}$  is a sufficient statistic for the parameter  $\theta$ .



⇒ **Solution** The likelihood  $L(\theta)$  of the sample is the joint density of each  $X_i$ , that is

$$\begin{aligned} L(x_1, x_2, \dots, x_n | \theta) &= f(x_1, x_2, \dots, x_n | \theta) \\ &= f(x_1 | \theta) \times f(x_2 | \theta) \times \dots \times f(x_n | \theta) \\ &= \frac{1}{\theta^n} e^{-(x_1 + x_2 + \dots + x_n)/\theta} \\ &= \frac{1}{\theta^n} e^{-n\bar{x}/\theta} \end{aligned}$$

Notice that  $L(\theta)$  is a function only of two parameter:  $\theta$  and  $\bar{x}$ , and that if

$$g(\bar{x}, \theta) = \frac{1}{\theta^n} e^{-n\bar{x}/\theta} \quad \text{and} \quad h(x_1, x_2, \dots, x_n) = 1,$$

then the likelihood can be factored as

$$L(x_1, x_2, \dots, x_n | \theta) = g(\bar{x}, \theta) h(x_1, x_2, \dots, x_n).$$

By the factorization theorem,  $\bar{X}$  is a sufficient statistic for the parameter  $\theta$ . ◀

**Lemma 7.1 Rao Blackwell Theorem**

If  $T(X)$  is a sufficient statistic for parameter  $\theta$ , and  $\hat{\theta}$  is an unbiased estimator of  $\theta$  with  $\mathbb{E}[\hat{\theta}] < \infty$  for all  $\theta \in \Theta$ . Let  $\hat{\theta}^* = \mathbb{E}[\hat{\theta} | T(X)]$ , then  $\hat{\theta}^* = \mathbb{E}[\hat{\theta} | T]$ , then

$$\mathbb{E}[(\hat{\theta}^* - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2]. \quad (7.6)$$

The inequality is strict unless  $\hat{\theta}$  is a function of  $T(X)$ .

*Proof.* By the law of conditional expectation, we have

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta} | T]] = \mathbb{E}[\hat{\theta}] = \theta,$$

so  $\hat{\theta}$  and  $\hat{\theta}^*$  are having the same bias. By the conditional variance formula, we have

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\text{Var}(\hat{\theta} | T)] + \text{Var}(\mathbb{E}[\hat{\theta} | T]) = \mathbb{E}[\text{Var}(\hat{\theta} | T)] + \text{Var}(\hat{\theta}^*).$$

Hence  $\text{Var}[\hat{\theta}^*] \geq \text{Var}[\hat{\theta}]$ , and so  $\text{MSE}[\hat{\theta}^*] \geq \text{MSE}[\hat{\theta}]$ . ◻

## 7.4 Variance of estimators based on sufficient statistics

All estimators can be regarded random variables, and we can compare their variances. therefore the maximum likelihood estimator can also be a random variable.

**Definition 7.5 Fisher information**

The Fisher information of a random variable  $X$  with density function  $f(x|\theta)$  is defined as

$$I_X(\theta) = \mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \right] = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ell(\theta) \right)^2 \right].$$

If  $T(X)$  is a sufficient statistic for  $\theta$ , then the Fisher information contained in  $T(X)$  is equal to the Fisher information contained in the sample  $X$ , i.e.,

$$I_T(\theta) = I_X(\theta).$$

One way to find a minimum variance unbiased estimator for a parameter is to use the Cramér-Rao lower bound or the Fisher information inequality.

**Remark (Multivariate Information Inequality).** A multivariate version of the Information Inequality exists as well. If  $\Theta \subset \mathbb{R}^k$  for some  $k \in \mathbb{N}$ , and if  $T : \mathcal{X} \rightarrow \mathbb{R}^n$  is an  $n$ -dimensional statistic for some  $n \in \mathbb{N}$  for data  $X \sim f(x | \theta)$  taking values in a space  $\mathcal{X}$  of arbitrary dimension, define the mean function  $m : \mathbb{R}^k \rightarrow \mathbb{R}^n$  by  $m(\theta) := \mathbb{E}_\theta T(X)$  and its  $n \times k$  Jacobian matrix by

$$J_{ij} := \frac{\partial m_i(\theta)}{\partial \theta_j}.$$

Then the multivariate Information Inequality asserts that

$$\text{Cov}_\theta[T(X)] \geq J I(\theta)^{-1} J^\top$$

where  $I(\theta) := \text{Cov}_\theta[\nabla_\theta \log f(X | \theta)]$  is the Fisher information matrix, where the notation " $A \geq B$ " for  $n \times n$  matrices  $A, B$  means that  $[A - B]$  is positive semi-definite, and where  $C^\top$  denotes the  $k \times n$  transpose of an  $n \times k$  matrix  $C$ . This gives lower bounds on the variance of  $z^\top T(X)$  for all vectors  $z \in \mathbb{R}^n$  and, in particular, lower bounds for the variance of components  $T_i(X)$ .

#### Theorem 7.4 Cramér-Rao lower bound

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with density function  $f(x|\theta)$  where  $\theta \in \Theta \subseteq \mathbb{R}$ . Suppose the following regularity conditions hold:

1. The support of  $f(x|\theta)$  does not depend on  $\theta$ .
2.  $\frac{\partial}{\partial \theta} \ln f(x|\theta)$  exists for all  $x$  and  $\theta$ .
3.  $\mathbb{E} \left[ \frac{\partial}{\partial \theta} \ln f(X|\theta) \right] = 0$  for all  $\theta$ .
4.  $0 < I_X(\theta) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \ln f(X|\theta) \right)^2 \right] < \infty$  for all  $\theta$ .

If  $T = T(X_1, \dots, X_n)$  is any unbiased estimator of  $\theta$  with finite variance, then

$$\text{Var}(T) \geq \frac{1}{n I_X(\theta)},$$

where  $I_X(\theta)$  is the Fisher information of a single observation. Equality holds if and only if there exists a function  $g(\theta)$  such that

$$T - \theta = g(\theta) \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i|\theta).$$

*Proof.* Let  $S(\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f(X_i|\theta)$  be the score function for the sample. By the regularity conditions, we have  $\mathbb{E}[S(\theta)] = 0$  and  $\text{Var}(S(\theta)) = n I_X(\theta)$ .

Since  $T$  is an unbiased estimator of  $\theta$ , we have  $\mathbb{E}[T] = \theta$ . Taking the derivative with respect to

$\theta$  and using the regularity conditions to interchange the order of differentiation and integration:

$$1 = \frac{d}{d\theta} \mathbb{E}[T] = \mathbb{E} \left[ \frac{\partial T}{\partial \theta} \right] = \mathbb{E} \left[ T \cdot \frac{\partial}{\partial \theta} \ln f(X_1, \dots, X_n | \theta) \right] = \mathbb{E}[T \cdot S(\theta)].$$

Now, since  $\mathbb{E}[T] = \theta$  and  $\mathbb{E}[S(\theta)] = 0$ , we have:

$$\text{Cov}(T, S(\theta)) = \mathbb{E}[T \cdot S(\theta)] - \mathbb{E}[T]\mathbb{E}[S(\theta)] = 1 - \theta \cdot 0 = 1.$$

By the Cauchy-Schwarz inequality:

$$(\text{Cov}(T, S(\theta)))^2 \leq \text{Var}(T) \cdot \text{Var}(S(\theta)),$$

which gives us:

$$1 \leq \text{Var}(T) \cdot nI_X(\theta).$$

Therefore:

$$\text{Var}(T) \geq \frac{1}{nI_X(\theta)}.$$

Equality holds in the Cauchy-Schwarz inequality if and only if  $T - \mathbb{E}[T]$  and  $S(\theta) - \mathbb{E}[S(\theta)]$  are linearly dependent, i.e., there exists a constant  $g(\theta)$  such that:

$$T - \theta = g(\theta)(S(\theta) - 0) = g(\theta)S(\theta).$$

□

**Remark.** An unbiased estimator  $T$  that achieves the Cramér-Rao lower bound is called an efficient estimator or minimum variance unbiased estimator (MVUE). When such an estimator exists, it is unique and coincides with the maximum likelihood estimator under regularity conditions. The Fisher information  $I_X(\theta)$  measures the amount of information about  $\theta$  contained in a single observation, and the Cramér-Rao bound shows that no unbiased estimator can have variance smaller than the reciprocal of the total Fisher information.

### Lemma 7.2 Cramér-Rao lower bound - 1st theorem

Let  $X_1, X_2, \dots, X_n$  be a random sample from a distribution with density function  $f(x|\theta)$ , where  $\theta \in \Theta$  is a scalar parameter. Suppose  $\hat{\theta}$  be any unbiased estimator of  $\theta$  with finite variance. Suppose the likelihood function  $L(\theta)$  is differentiable with respect to  $\theta$  and satisfies

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) L(\theta) dx_1 \dots dx_n = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) \frac{dL(\theta)}{d\theta} dx_1 \dots dx_n. \quad (7.7)$$

for any function  $h(x_1, \dots, x_n)$  with  $\mathbb{E}[h(x_1, \dots, x_n)] < \infty$ . Then

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathbb{E} \left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right]} = \frac{1}{nI_X(\theta)}, \quad (7.8)$$

*Proof.* Since  $L(\theta)$  is the joint density function of the sample, we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(\theta) dx_1 \cdots dx_n = 1. \quad (\clubsuit)$$

Differentiating  $(\clubsuit)$  with respect to  $\theta$ , we get

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{dL(\theta)}{d\theta} dx_1 \cdots dx_n = 0. \quad (\spadesuit)$$

Rewriting  $(\spadesuit)$  as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{L(\theta)} \frac{dL(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n = 0,$$

so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta \frac{d \ln L(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n = 0. \quad (\diamondsuit)$$

Since  $\hat{\theta}$  is an unbiased estimator of  $\theta$ , we can see that

$$\mathbb{E}[\hat{\theta}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta} L(\theta) dx_1 \cdots dx_n = \theta. \quad (\star)$$

Differentiating  $(\star)$  with respect to  $\theta$ , we get

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta} L(\theta) dx_1 \cdots dx_n = 1.$$

Again, using the fact  $(\clubsuit)$  with  $h(X_1, X_2, \dots, X_n) = \hat{\theta}$ , we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta} \frac{dL(\theta)}{d\theta} dx_1 \cdots dx_n = 1.$$

Rewriting the above equation as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta} \frac{1}{L(\theta)} \frac{dL(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n = 1,$$

so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{\theta} \frac{d \ln L(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n = 1. \quad (\heartsuit)$$

From  $(\diamondsuit)$  and  $(\heartsuit)$ , we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{d \ln L(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n = 1.$$

By the Cauchy-Schwarz inequality, we have


$$\begin{aligned}
1^2 &= \left( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\hat{\theta} - \theta) \frac{d \ln L(\theta)}{d\theta} L(\theta) dx_1 \cdots dx_n \right)^2 \\
&\leq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 L(\theta) dx_1 \cdots dx_n \\
&\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left( \frac{d \ln L(\theta)}{d\theta} \right)^2 L(\theta) dx_1 \cdots dx_n \\
1 &= \text{Var}(\hat{\theta}) \mathbb{E} \left[ \left( \frac{d \ln L(\theta)}{d\theta} \right)^2 \right].
\end{aligned}$$

Therefore,

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathbb{E} \left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right]}.$$

and the proof is complete.  $\square$

What this saying is, for any unbiased estimator  $\hat{\theta}$  of  $\theta$ , its variance (MSE) is at least  $\frac{1}{I(\theta)}$ . If we achieve this lower bound, meaning that our variance is exactly equal to  $\frac{1}{I(\theta)}$ , then we have found the best possible unbiased estimator for  $\theta$ . That is, we have found the **minimum variance unbiased estimator (MVUE)** for  $\theta$ .

 **Example 7.4.1.** Suppose that  $Y_1, Y_2, \dots, Y_n$  denote a random sample from the Weibull distribution with pdf

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Find an MVUE for  $\theta$ .

 **Solution** We begin using the factorization criterion to find the sufficient statistic that best summarizes the information about  $\theta$ .

$$\begin{aligned}
L(y_1, y_2, \dots, y_n|\theta) &= f_Y(y_1, y_2, \dots, y_n|\theta) \\
&= \left( \frac{2y_1}{\theta} \right) e^{-y_1^2/\theta} \times \left( \frac{2y_2}{\theta} \right) e^{-y_2^2/\theta} \times \cdots \times \left( \frac{2y_n}{\theta} \right) e^{-y_n^2/\theta} \\
&= \underbrace{\left( \frac{2}{\theta} \right)^n \exp \left\{ -\frac{1}{\theta} \sum_{i=1}^n y_i^2 \right\}}_{g'(\sum y_i|\theta)} \underbrace{(y_1 \times y_2 \times \cdots \times y_n)}_{h(y_1, y_2, \dots, y_n)}
\end{aligned}$$

Thus,  $U = \sum_{i=1}^n Y_i^2$  is the minimal sufficient statistic for  $\theta$  (by Factorization theorem).

We now need to find a function of this statistic that is unbiased for  $\theta$ . Now let  $W = Y_i^2$ . Using the method of transformation,

$$f_W(w) = f_Y(h^{-1}(w)) \frac{dh^{-1}(w)}{dw} = f_Y(\sqrt{w}) \frac{d}{dw}(\sqrt{w}),$$

continue simplify the expression gives

$$f_W(w) = \frac{2}{\theta} \left( \sqrt{w} e^{-w/\theta} \right) \left( \frac{1}{2\sqrt{w}} \right) = \frac{1}{\theta} e^{-w/\theta}, \quad w > 0.$$


That is,  $Y_i^2 \sim \text{Exp}(\theta)$ . As

$$\mathbb{E}[Y_i^2] = \theta \quad (7.9)$$

and

$$\begin{aligned} \mathbb{E} \left[ \sum_{i=1}^n Y_i^2 \right] &= \mathbb{E}[Y_1^2 + Y_2^2 + \cdots + Y_n^2] \\ &= \mathbb{E}[Y_1^2] + \mathbb{E}[Y_2^2] + \cdots + \mathbb{E}[Y_n^2] && \text{linearity or expectation} \\ &= \underbrace{\theta + \cdots + \theta}_{n \text{ times}} \\ &= n\theta. \end{aligned}$$



 **Example 7.4.2 (Simple linear regression).** Simple linear regression studies the relationship between the random response variable  $Y$  and the explanatory variable  $X$  by fitting a linear equation to observed data. The linear model is given by

$$Y = \alpha + \beta X + \epsilon_i,$$

where  $\epsilon \sim N(0, \sigma^2)$  is the error term. The sufficient statistics for the parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$  can be derived from the likelihood function.

The article “Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study” (J. of Automobile Engr., 2009: 565-583) included the following data on  $x$  = iodine value (g) and  $y$  = cetane number for a sample of 14 biofuels (see next slide). The iodine value ( $x$ ) is the amount of iodine necessary to saturate a sample of 100 g of oil.

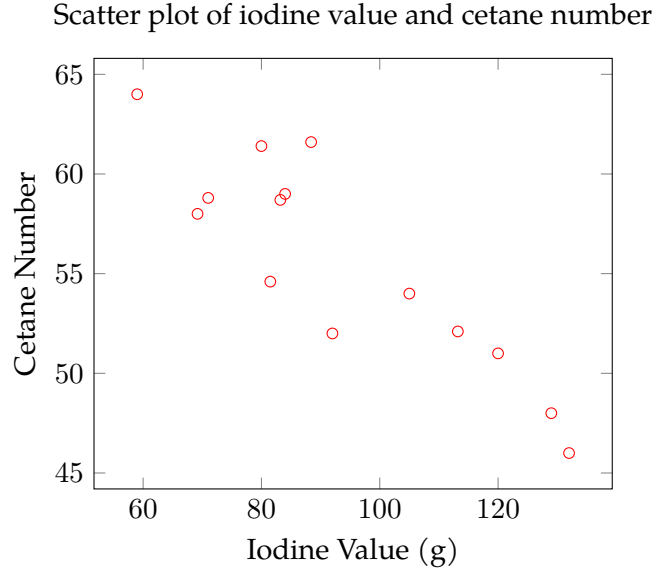
$x$	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
$y$	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Under the additional assumption that  $\epsilon_i$ 's are iid  $N(0, \sigma^2)$ , the likelihood function of the sample is

$$\begin{aligned} L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) &= \prod_{i=1}^n f_Y(y_i | x_i, \alpha, \beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \right). \quad (\spadesuit) \end{aligned}$$

Considering  $(y_1, x_1), \dots, (y_n, x_n)$  as  $n$  pairs of data points plotted in the  $xy$ -plane as the scatterplot in the previous figure.

Think of drawing through this cloud of points a straight line that comes “as close as possible” to all the points, measured by the vertical distances from the points to the straight line. For any



**Figure 7.1:** The scatter plot of iodine value (x) and cetane number (y).

line  $y = \alpha + \beta x$ , the vertical distance from the point  $(x_i, y_i)$  to the line is  $y_i - (\alpha + \beta x_i)$ . Letting

$$\psi(\alpha, \beta) := \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2,$$

Maximizing the likelihood is equivalent to

$$\min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2,$$

Taking the partial derivatives of  $\psi(\alpha, \beta)$  with respect to  $\alpha$  and  $\beta$ , we have

$$\begin{aligned} \frac{\partial \psi(\alpha, \beta)}{\partial \alpha} &= -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0, \\ \frac{\partial \psi(\alpha, \beta)}{\partial \beta} &= -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0. \end{aligned} \quad (\clubsuit)$$

The first equation gives

$$\bar{y} - a - b\bar{x} \implies a = \bar{y} - b\bar{x}.$$

Substituting  $a$  into the second equation  $(\clubsuit)$  by  $\bar{y} - b\bar{x}$  results in

$$\sum_{i=1}^n x_i (y_i - \bar{y}) + b \sum_{i=1}^n x_i (\bar{x} - x_i) = 0.$$

This equation is the same as  $S_{xy} = bS_{xx}$ , that is

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

Therefore, replacing  $y_i$  by the random variable  $Y_i$  for all  $i = 1, 2, \dots, n$  (and we still use  $S_{xy}$  when

$y_i$  is replaced by  $Y_i$ ), we obtain the MLE or LSE as

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \bar{Y} - \frac{S_{xy}}{S_{xx}}\bar{x}. \quad (7.10)$$

We can always assume that  $S_{xx} > 0$ , since  $S_{xx} = 0$  is the trivial case of identical  $x_i$ 's.

We now proceed to show that  $\hat{\alpha}$  and  $\hat{\beta}$  are UMVUE of  $\alpha$  and  $\beta$ , respectively. First of all, we had already shown that they are unbiased estimators of  $\alpha$  and  $\beta$ , respectively.

$$\mathbb{E}[S_{xy}] = \sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}_y(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x}) \beta (x_i - \bar{x}) = \beta S_{xx}.$$

Since  $\hat{\beta}$  is unbiased for  $\beta$  and

$$\mathbb{E}[\hat{\alpha}] = \mathbb{E}[\bar{Y}] - \bar{x} \mathbb{E}[\hat{\beta}] = \alpha + \beta\bar{x} - \bar{x}\beta = \alpha.$$

Continue with ( $\spadesuit$ ), the likelihood function of the sample is

$$\begin{aligned} L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \alpha - \beta\bar{x})^2 + (\beta - \hat{\beta})^2 S_{xx} \right] \right). \end{aligned}$$

where

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2.$$

We still using the notation  $S_{xy}$  and  $S_{xx}$  when  $y_i$  is replaced by  $Y_i$ . From the properties of the exponential family, a complete and sufficient statistic for  $\theta = (\alpha, \beta, \sigma^2)$  is given by  $(\hat{\alpha}, \hat{\beta}, S_{yy})$ .

Since  $\hat{\alpha}$  and  $\hat{\beta}$  are both unbiased estimators and functions of the sufficient and complete statistic, thus they are UMVUE of  $\alpha$  and  $\beta$ , respectively.

**Example 7.4.3 (Simple linear regression (cont.)).** **Question:** What if we remove the normality assumption?  $\hat{\alpha}$  and  $\hat{\beta}$  are still least squares estimators, but they are no longer MLEs. A statistical property that holds for LSE is that it is the best linear unbiased estimator (BLUE) in the sense that  $\hat{\alpha}$  or  $\hat{\beta}$  has the smallest variance among all linear unbiased estimators of the form

$$\sum_{i=1}^n d_i Y_i \quad (7.11)$$

If the estimator of this form is unbiased for  $\beta$ , then

$$\begin{aligned} \beta &= \mathbb{E}_y \left( \sum_{i=1}^n d_i Y_i \right) = \sum_{i=1}^n d_i \mathbb{E}_y[Y_i] \\ &= \sum_{i=1}^n d_i (\alpha + \beta x_i) && \text{from the regression line, } \hat{Y} = \alpha + \beta x \\ &= \alpha \sum_{i=1}^n d_i + \beta \sum_{i=1}^n d_i x_i && \text{linearity of summation.} \end{aligned}$$



holds for all  $\alpha$  and  $\beta$ , which implies that

$$\sum_{i=1}^n d_i = 0, \quad \sum_{i=1}^n d_i x_i = 1.$$

A geometric description of the BLUE of  $\beta$  is given in the figure below.

We want to show the LSE  $\hat{\beta}$  is BLUE. Since

$$\text{Var} \left( \sum_{i=1}^n d_i Y_i \right) = \sigma^2 \sum_{i=1}^n d_i^2,$$

the BLUE of  $\beta$  is the solution of the optimization problem

$$\min_{d_i} \sum_{i=1}^n d_i^2 \tag{7.12a}$$

$$\text{subject to} \quad \sum_{i=1}^n d_i = 0, \tag{7.12b}$$

$$\sum_{i=1}^n d_i x_i = 1. \tag{7.12c}$$

Consider the Lagrange multiplier method by minimizing

$$g(d_1, \dots, d_n, \lambda_1, \lambda_2) = \sum_{i=1}^n d_i^2 + \lambda_1 \left( \sum_{i=1}^n d_i \right) + \lambda_2 \left( \sum_{i=1}^n d_i x_i - 1 \right). \tag{7.13}$$

Taking derivatives with respect to  $d_i$  and setting them to zero, we have

$$\frac{\partial g}{\partial d_i} = 2d_i + \lambda_1 + \lambda_2 x_i = 0.$$

Then

$$\begin{aligned} 0 &= \sum_{i=1}^n (2d_i + \lambda_1 + \lambda_2 x_i) = 2 \sum_{i=1}^n d_i + n\lambda_1 + \lambda_2 \sum_{i=1}^n x_i && \text{linearity of summation} \\ &= \lambda_1 n + \lambda_2 \sum_{i=1}^n x_i && \text{since } \sum d_i = 0 \end{aligned}$$

which yields  $\lambda_1 = -\lambda_2 \bar{x}$  and, hence

$$0 = 2d_i + \lambda_2(x_i - \bar{x}).$$

Then

$$0 = \sum_{i=1}^n (x_i - \bar{x})[2d_i + \lambda_2(x_i - \bar{x})] = 2 + \lambda_2 S_{xx}$$

which gives us  $\lambda_2 = -2/S_{xx}$ . Then

$$d_i = \frac{-(\lambda_1 + \lambda_2 x_i)}{2} = \frac{-\lambda_2(x_i - \bar{x})}{2} = \frac{(x_i - \bar{x})}{S_{xx}}$$

and the BLUE of  $\beta$  is

$$\sum_{i=1}^n d_i Y_i = \sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} Y_i = \frac{S_{xy}}{S_{xx}} = \hat{\beta}.$$

And because

$$\hat{\beta} = \sum_{i=1}^n \frac{(x_i - \bar{x})(\beta x_i + \varepsilon_i)}{S_{xx}} = \beta + \sum_{i=1}^n d_i \varepsilon_i$$

where  $d_i = (x_i - \bar{x})/S_{xx}$ , and the random error  $\varepsilon \sim N(0, 1)$ . We obtain that

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^n d_i^2 \text{Var}(\varepsilon_i) = \frac{\sigma^2}{S_{xx}}.$$

And we are done.

**Corollary 7.1 Cramér-Rao lower bound - 2nd theorem**

If  $L(\theta)$  is twice differentiable with respect to  $\theta$ , then the inequality can be stated equivalently as

$$\text{Var}(\hat{\theta}) \geq \frac{-1}{\mathbb{E} \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right]}. \quad (7.14)$$

*Proof.* If  $L(\theta)$  is twice differentiable, then  $\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}$  exists. And since that  $L(\theta)$  is maximum likelihood so we have

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} < 0.$$

Hence the sign of the right-hand side of the Cramér-Rao inequality must be negative.  $\square$

**Example 7.4.4.** Let  $X_1, X_2, \dots, X_n$  be a random samples from a distribution with density function

$$f(x|\theta) = \begin{cases} 3\theta x^2 e^{-\theta x^3} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the Cramér-Rao lower bound for the variance of unbiased estimator of the parameter  $\theta$ ?

**Solution** Let  $\hat{\theta}$  be an unbiased estimator of  $\theta$ . Cramér-Rao lower bound for the variance of  $\hat{\theta}$  is given by

$$\text{Var}(\hat{\theta}) \geq \frac{-1}{\mathbb{E} \left[ \left( \frac{d \ln L(\theta)}{d\theta} \right)^2 \right]}.$$

First, compute the likelihood function for the sample:

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta) = \prod_{i=1}^n 3\theta x_i^2 e^{-\theta x_i^3} = 3^n \theta^n \left( \prod_{i=1}^n x_i^2 \right) \exp \left( -\theta \sum_{i=1}^n x_i^3 \right).$$

Take the log-likelihood:

$$\ell(\theta) = \ln L(\theta) = n \ln 3 + n \ln \theta + 2 \sum_{i=1}^n \ln x_i - \theta \sum_{i=1}^n x_i^3.$$

Compute the first derivative with respect to  $\theta$ :

$$\frac{\partial}{\partial \theta} \ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^n x_i^3.$$

Compute the second derivative:

$$\frac{\partial^2}{\partial \theta^2} \ell(\theta) = -\frac{n}{\theta^2}.$$

The Fisher information for one observation is


$$I_X(\theta) = -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X|\theta) \right] = \frac{1}{\theta^2}.$$

For  $n$  independent observations, the total Fisher information is  $nI_X(\theta) = \frac{n}{\theta^2}$ .

Therefore, the Cramér-Rao lower bound is


$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI_X(\theta)} = \frac{\theta^2}{n}.$$

where  $L(\theta)$  denotes the likelihood function of the given random sample. ◀

 **Example 7.4.5.** Let  $X_1, X_2, \dots, X_n$  denote a random sample from  $\text{Bin}(1, p)$ . We knew that  $\bar{X}$  is an unbiased estimator of  $p$  and that

$$\text{Var}(\bar{X}) = \frac{p(1-p)}{n}.$$

Find the Cramér-Rao lower bound for the variance of every unbiased estimator of  $p$ .

 **Solution** This is a Bernoulli distribution with parameter  $p$ . The density function for each  $X_i$  is

$$f(x|p) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

Taking logarithm on  $f$ ,

$$\ln f(x|p) = \ln p^x + \ln(1-p)^{1-x} = \boxed{x \ln p + (1-x) \ln(1-p)} \quad (\blacklozenge)$$

Compute the first and second order derivative of  $(\blacklozenge)$  with respect to  $p$ . The second order derivative will be the Fisher information. In this case we have only one parameter which is  $p$ , so the information is just a simple algebraic expression rather than a matrix.

$$\frac{\partial \ln f(x|p)}{\partial p} = \frac{x}{p} + \frac{x-1}{1-p}$$

$$\begin{aligned} \frac{\partial^2 \ln f(x|p)}{\partial p^2} &= -\frac{x}{p^2} + (-1)^2(x-1)(1-p)^{-2} \\ &= -\frac{x}{p^2} + \frac{x-1}{(1-p)^2}. \end{aligned} \quad (\blacktriangle)$$

Hence we find the expectation of  $x$  in (▲), that is

$$\begin{aligned}\mathbb{E}_X \left[ \frac{\partial^2 \ln f(x|p)}{\partial p^2} \right] &= \mathbb{E}_X \left[ -\frac{x}{p^2} + \frac{x-1}{(1-p)^2} \right] \\ &= -\frac{1}{p^2} \mathbb{E}_X[X] + \frac{1}{(1-p)^2} \mathbb{E}_X[X-1] \\ &= -\frac{p}{p^2} + \frac{p-1}{(1-p)^2} \\ &= -\frac{1}{p(1-p)}.\end{aligned}$$

Therefore, the Cramér-Rao lower bound for the variance of unbiased estimator of  $p$  is

$$\text{Var}(\hat{p}) \geq -\frac{1}{-nI_X(p)} = \frac{1}{-n} = \frac{p(1-p)}{n}.$$

◀

### 7.4.1 Delta method – Variance of functions of estimators

The Delta Method (DM) states that we can approximate the asymptotic behaviour of functions over a random variable, if the random variable is itself asymptotically normal. In practice, this theorem tells us that even if we do not know the expected value and variance of the function  $g(X)$  we can still approximate it reasonably. Note that by Central Limit Theorem we know that several important random variables and estimators are asymptotically normal, including the sample mean. We can therefore approximate the mean and variance of some transformation of the sample mean using its variance.

More specifically, suppose that we have some sequence of random variables  $\{X_n\}$ , as  $n \rightarrow \infty$ ,

Given this, if  $g$  is some smooth function (i.e. there are no discontinuous jumps in values) then the Delta Method states that:

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|\dot{g}(\mu)|\sigma} \approx \mathcal{N}(0, 1) \quad (7.15)$$

DM also generalizes to multidimensional functions, where instead of converging on the standard normal the random variable must converge in distribution to a multivariate normal, and the derivatives of  $g$  are replaced with the gradient of  $g$  (a vector of all partial derivatives).

#### Theorem 7.5 Delta method

Suppose that  $g(\theta)$  is a function of estimator. The delta method provides a way to approximate the variance of a function of an estimator. This approximate of the variance is given by

$$\text{Var}(g(\hat{\theta})) \approx [\dot{g}(\theta)]^2 \text{Var}[\hat{\theta}]. \quad (7.16)$$

**Example 7.4.6.** Given  $g(s, t) = \frac{s}{t}$ ,  $h(s, t) = \ln s$  and  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are unbiased estimators of  $\theta_1$  and  $\theta_2$ . Based on a particular sample, the maximum likelihood estimates of  $\theta_1$  and  $\theta_2$  are  $\hat{\theta}_1 = 3.2$  and  $\hat{\theta}_2 = 11.8$ , and the log-likelihood is  $\ell(\theta_1, \theta_2) = -2\theta_1^2\theta_2 - \theta_2^3$ .

⇒ **Solution** We first compute the Fisher information matrix:

$$\frac{\partial^2}{\partial \theta_1^2} \ell(\theta_1, \theta_2) = -4\theta_2, \quad \frac{\partial^2}{\partial \theta_2^2} \ell(\theta_1, \theta_2) = -6\theta_2, \quad \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ell(\theta_1, \theta_2) = -4\theta_1.$$

The information matrix is

$$I_X(\theta) = -\mathbb{E} \begin{bmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell(\theta_1, \theta_2) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ell(\theta_1, \theta_2) \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ell(\theta_1, \theta_2) & \frac{\partial^2}{\partial \theta_2^2} \ell(\theta_1, \theta_2) \end{bmatrix} = \begin{bmatrix} 4\theta_2 & 4\theta_1 \\ 4\theta_1 & 6\theta_2 \end{bmatrix}.$$

The covariance matrix of the MLEs is given by the inverse of the Fisher information matrix evaluated at the MLEs:

$$\Sigma = I_X(\theta)^{-1} = \begin{bmatrix} 4\theta_2 & 4\theta_1 \\ 4\theta_1 & 6\theta_2 \end{bmatrix}^{-1} = \frac{1}{12\theta_2^2 - 8\theta_1^2} \begin{bmatrix} 3\theta_2 & -2\theta_1 \\ -2\theta_1 & 2\theta_2 \end{bmatrix}.$$

The estimated covariance matrix is obtained by substituting the MLEs:

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{12\hat{\theta}_2^2 - 8\hat{\theta}_1^2} \begin{bmatrix} 3\hat{\theta}_2 & -2\hat{\theta}_1 \\ -2\hat{\theta}_1 & 2\hat{\theta}_2 \end{bmatrix} = \frac{1}{12(11.8)^2 - 8(3.2)^2} \begin{bmatrix} 3(11.8) & -2(3.2) \\ -2(3.2) & 2(11.8) \end{bmatrix} \\ &= \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}(\hat{\theta}_1) & \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) \\ \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \text{Var}(\hat{\theta}_2) \end{bmatrix}. \end{aligned}$$

1. Now take partial derivatives of  $g(s, t)$  with respect to  $s$  and  $t$ :

$$\begin{aligned} g_s(s, t) &= \frac{\partial g}{\partial s} = \frac{1}{t} \implies g_s(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{11.8}, \\ g_t(s, t) &= \frac{\partial g}{\partial t} = -\frac{s}{t^2} \implies g_t(\hat{\theta}_1, \hat{\theta}_2) = -\frac{3.2}{(11.8)^2}. \end{aligned}$$

Hence, let  $\mathbf{w} = \begin{bmatrix} g_s(\hat{\theta}_1, \hat{\theta}_2) & g_t(\hat{\theta}_1, \hat{\theta}_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix}$ . the approximate variance of  $g(\hat{\theta}_1, \hat{\theta}_2)$  is

$$\begin{aligned} \text{Var}(g(\hat{\theta}_1, \hat{\theta}_2)) &\approx \mathbf{w} \hat{\Sigma} \mathbf{w}^T \\ &= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{11.8} \\ -\frac{3.2}{11.8^2} \end{bmatrix} \\ &= 0.000208785. \end{aligned}$$

2. Continue with  $h(s, t)$ : take partial derivatives of  $h(s, t)$  with respect to  $s$  and  $t$ :

$$\begin{aligned} h_s(s, t) &= \frac{\partial h}{\partial s} = \frac{1}{s} \implies h_s(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{3.2}, \\ h_t(s, t) &= \frac{\partial h}{\partial t} = 0 \implies h_t(\hat{\theta}_1, \hat{\theta}_2) = 0. \end{aligned}$$

The estimated covariance between  $h(\hat{\theta}_1, \hat{\theta}_2)$  and  $g(\hat{\theta}_1, \hat{\theta}_2)$  is

$$\begin{aligned}\text{Cov}(h(\hat{\theta}_1, \hat{\theta}_2), g(\hat{\theta}_1, \hat{\theta}_2)) &= \begin{bmatrix} \frac{1}{\hat{\theta}_1} & -\frac{\hat{\theta}_1}{\hat{\theta}_2^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\theta}_1} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{3.2} \\ 0 \end{bmatrix} \\ &= 6.2 \times 10^{-3}\end{aligned}$$



## Tutorials

**Exercise 7.1** If  $X$  is uniformly distributed on the interval  $(2\theta, 3\theta)$  where  $\theta > 0$ . And that  $X_1, X_2, \dots, X_n$  is a random sample from the distribution of  $X$ . Find the bias in the maximum likelihood estimator of  $\theta$ .

**Exercise 7.2** Let  $X_1, X_2, \dots, X_n$  be a random sample from a population  $X \sim \text{Poisson}(\lambda)$ , where  $\lambda > 0$  is a parameter. Is the estimator  $\bar{X}$  of  $\lambda$  a consistent estimator of  $\lambda$ ?

**Exercise 7.3** Let  $X_1, X_2, \dots, X_n$  be a random sample from a population  $X$  with density function

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where  $\theta > 1$  is a parameter. Show that

$$-\frac{1}{n} \sum_{i=1}^n \ln(X_i)$$

is a uniform minimum variance unbiased estimator of  $\frac{1}{\theta}$ .

**Exercise 7.4** Let  $Y_1, Y_2, \dots, Y_n$  denote a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ . Consider the following three estimators for  $\mu$ :

$$\hat{\mu}_1 = \frac{1}{2}(Y_1 + Y_2), \quad \hat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + \dots + Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n, \quad \hat{\mu}_3 = \bar{Y}.$$

1. Show that each of the three estimators is unbiased.
2. Find the efficiency of  $\hat{\mu}_3$  relative to  $\hat{\mu}_2$  and  $\hat{\mu}_1$ , respectively.

**Exercise 7.5**

1. State the definition of unbiased estimator.

2. Show that if  $\hat{\Theta}_1$  is an unbiased estimator for  $\theta$ , and  $W$  is a zero mean random variable, then

$$\hat{\Theta}_2 = \hat{\Theta}_1 + W$$

is also an unbiased estimator for  $\theta$ .

3. Show that if  $\hat{\Theta}_1$  is an unbiased estimator for  $\theta$  such that  $\mathbb{E}[\hat{\Theta}_1] = a\theta + b$ , where  $a \neq 0$ . Then

$$\hat{\Theta}_2 = \frac{\hat{\Theta}_1 - b}{a}$$

is an unbiased estimator for  $\theta$ .

**Exercise 7.6** Given  $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$  is a vector of parameters being estimated by maximum likelihood. You are given that the current value of the vector of estimates is  $\hat{\theta} = \begin{bmatrix} 30 \\ 2 \end{bmatrix}$  as well as the estimated information matrix

$$I(\hat{\theta}) = \begin{bmatrix} .075 & -.620 \\ -.620 & 10.0 \end{bmatrix}.$$

Determine the approximate variance of  $\hat{\theta}_1$ .

**Exercise 7.7** Let  $X$  be a random variable, and  $X_n = X + Y_n$ , where

$$\mathbb{E}[Y_n] = \frac{1}{n} \quad \text{and} \quad \text{Var}(Y_n) = \frac{\sigma^2}{n}.$$

where  $\sigma > 0$  is constant. Show that  $X_n \xrightarrow{p} X$  as  $n \rightarrow \infty$ . Hint:  $|Y_n| \leq |Y_n - \mathbb{E}[Y_n]| + \frac{1}{n}$ .

## Confidence Intervals

The reason of using an interval estimator  $[L(\mathbf{X}), U(\mathbf{X})]$  instead of a point estimator  $\hat{\theta}$  is that the interval estimator can have some level of confidence that the unknown parameter  $\theta$  lies within the interval. The certainty of this guarantee is qualified by the following definitions.

### Definition 8.1 Interval Estimator

Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample of size  $n$  from a population with density function  $f(x|\theta)$ , where  $\theta \in \Theta$  is an unknown parameter. The **interval estimator** is a pair of statistics  $[L(\mathbf{X}), U(\mathbf{X})]$  such that  $L(\mathbf{X}) < U(\mathbf{X})$  for all possible samples  $\mathbf{X}$ .

Recall that a sample is a portion of the population usually chosen by some method of random sampling and as such is a set of random variables  $X_1, X_2, \dots, X_n$  with the same probability density function  $f(x|\theta)$  as the population. Once the sampling is done, we will get the sample data

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n.$$

Another well known method for constructing confidence sets is that using the pivotal quantities.

### Definition 8.2 Confidence Coefficient

Consider  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population with density  $f(x|\theta)$ , where  $\theta$  is unknown parameter. An interval estimator  $[L(\mathbf{X}), U(\mathbf{X})]$  of  $\theta$  is called a  $100(1 - \alpha)\%$  **confidence interval** for  $\theta$  if

$$\mathbb{P}_{\theta}[L \leq \theta \leq U] = 1 - \alpha, \quad \forall \theta \in \Theta, \quad (8.1)$$

The random variable  $L$  is called the **lower confidence limit** and the random variable  $U$  is called the **upper confidence limit**. The value  $1 - \alpha$  is called the **confidence coefficient** or **coverage probability** of the interval estimator.

One can find infinitely many pairs of  $L, U$ , such that

$$1 - \alpha = \mathbb{P}_{\theta}[L \leq \theta \leq U]$$

for a given confidence coefficient  $1 - \alpha$ . Thus, there are infinitely many confidence intervals for a given confidence coefficient.

However, we only need the confidence interval that is the shortest in length among all possible confidence intervals. If a confidence interval is constructed by omitting equal tail areas then we obtain what we known as the central confidence interval. In a symmetric distribution, the central confidence interval is also the shortest confidence interval.



## 8.1 Pivotal Quantities

### Definition 8.3 Pivotal Quantity

A function  $Q(X, \theta)$  is called a pivotal quantity (or pivot) if and only if the distribution of  $Q(X, \theta)$  does not depend on any unknown parameter  $\theta$ .

**Remark.** A pivot is not a statistic, although its distribution is known.

With a pivot  $Q(\mathbf{X}, \theta)$ , a confidence set on level  $1 - \alpha$  for any  $\alpha \in (0, 1)$ , can be obtained by finding a Borel set  $\mathcal{A} = [c_1, c_2]$  such that  $\mathbb{P}[Q(\mathbf{X}, \theta) \in \mathcal{A}] \geq 1 - \alpha$ . Then the set

$$C(\mathbf{X}) = \{\theta \in \Theta \mid Q(\mathbf{X}, \theta) \in \mathcal{A}\} \quad (8.2)$$

is a confidence set on level  $1 - \alpha$  since

$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta}(Q(\mathbf{X}, \theta) \in \mathcal{A}) = \mathbb{P}[Q(\mathbf{X}, \theta) \in \mathcal{A}] \geq 1 - \alpha. \quad (8.3)$$

If  $Q(\mathbf{X}, \theta)$  has a continuous cdf, then we can choose  $c_1$  and  $c_2$  such that  $C(x)$  has exact coverage probability  $1 - \alpha$ .

### Definition 8.4 Location-Scale Family

Let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a probability density function. Then for any  $\mu$  and any  $\sigma > 0$ , the family of functions

$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x - \mu}{\sigma}\right) \mid \mu \in (-\infty, \infty), \sigma \in (0, \infty) \right\} \quad (8.4)$$

is called the *location-scale family* with standard probability density  $f(x; \theta)$ . The parameter  $\mu$  is called the *location parameter* and the parameter  $\sigma$  is called the *scale parameter*. If  $\sigma = 1$ , then  $\mathcal{F}$  is called the *location family*. If  $\mu = 0$ , then  $\mathcal{F}$  is called the *scale family*.

It should be noted that each member  $f(x; \mu, \sigma)$  of the location-scale family is a probability density function. If we take  $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ , then the normal density function

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


belongs to the location-scale family. The density function

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases} \quad (8.5)$$

belongs to the scale family. However, the density function

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases} \quad (8.6)$$

does not belong to the location-scale family.

 **Example 8.1.1 (Scale uniform interval estimation).** Let  $X_1, X_2, \dots, X_n$  be a random sample from a uniform distribution  $U(0, \theta)$ , and let  $Y = \max\{X_1, X_2, \dots, X_n\}$ . We are interested in finding an interval estimator of  $\theta$ .

We consider two candidate interval estimators:

Form of pdf	Type of pdf	Pivots
$f(x - \mu)$	Location	$\bar{X} - \mu$
$\frac{1}{\sigma} f(\frac{1}{\sigma})$	Scale	$\frac{\bar{X}}{\sigma}, \frac{S^2}{\sigma^2}, \frac{X_{(n)}}{\sigma}$
$\frac{1}{\sigma} f(\frac{x - \mu}{\sigma})$	Location-scale	$\frac{\bar{X} - \mu}{S}, \frac{S^2}{\sigma^2}$

**Table 8.1:** The location and scale families and some common pivots. Here,  $\bar{X}$  is the sample mean,  $S^2$  is the sample variance, and  $X_{(n)}$  is the maximum order statistic.

1.  $[aY, bY], \quad 1 \leq a < b.$
2.  $[Y + c, Y + d], \quad 0 \leq c < d.$

where  $a, b, c, d$  are specified constants. Note that the parameter  $\theta$  must always larger than  $y$ .

**[First candidate]** For the first candidate, we have

$$\begin{aligned} \mathbb{P}_{\theta}[\theta \in [aY, bY]] &= \mathbb{P}_{\theta}[aY \leq \theta \leq bY] \\ &= \mathbb{P}_{\theta}\left[\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right] \end{aligned}$$

From previous example we know that

$$f_Y(y) = \begin{cases} \frac{ny^{n-1}}{\theta^n} & \text{if } 0 < y < \theta \\ 0 & \text{otherwise.} \end{cases}$$

so the pdf of  $T$  is  $f_T(t) = nt^{n-1}$  for  $0 < t < 1$ . Thus,

$$\mathbb{P}_{\theta}\left[\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right] = \int_{1/b}^{1/a} nt^{n-1} dt = \frac{1}{a^n} - \frac{1}{b^n}.$$

The covergae probability of the first interval estimator is independent of the parameter  $\theta$ , and thus  $\frac{1}{a^n} - \frac{1}{b^n}$  is the confidence coefficient of the interval estimator  $[aY, bY]$ .

**[Second candidate]** On the other hand, for the second candidate, for  $\theta \geq d$  and a similar calculation we obtained

$$\begin{aligned} \mathbb{P}_{\theta}[\theta \in [Y + c, Y + d]] &= \mathbb{P}_{\theta}[Y + c \leq \theta \leq Y + d] \\ &= \mathbb{P}_{\theta}\left[1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right] \\ &= \int_{1-d/\theta}^{1-c/\theta} nt^{n-1} dt \\ &= (1 - c/\theta)^n - (1 - d/\theta)^n \end{aligned}$$

In this case, the coverage probability of the second interval estimator depends on the unknown parameter  $\theta$ . As  $\theta \rightarrow \infty$ , for any fixed  $c$  and  $d$ , the coverage probability is

$$\lim_{\theta \rightarrow \infty} (1 - c/\theta)^n - (1 - d/\theta)^n = 0.$$

Showing that the second interval estimator has zero confidence coefficient.

## 8.2 Confidence Interval for Population Mean

At the outset, we use the pivotal quantity method to construct a confidence interval for the mean of a normal population. First we assume that the population is normal and the population variance is known, but the variance is unknown. Next, we construct the confidence interval for the mean of a population with continuous, symmetric and unimodal probability distribution by applying the central limit theorem.

We know that  $\hat{\mu} = \bar{X}$ . Because each  $X_i$  is identically distributed as  $N(\mu, \sigma^2)$ , the distribution of the sample mean  $\bar{X}$  is given by

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

It is easy to verify that the distribution of the estimator  $\hat{\mu}$  is not independent of the parameter  $\mu$ . If we standardize  $\hat{\mu}$ , we have

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

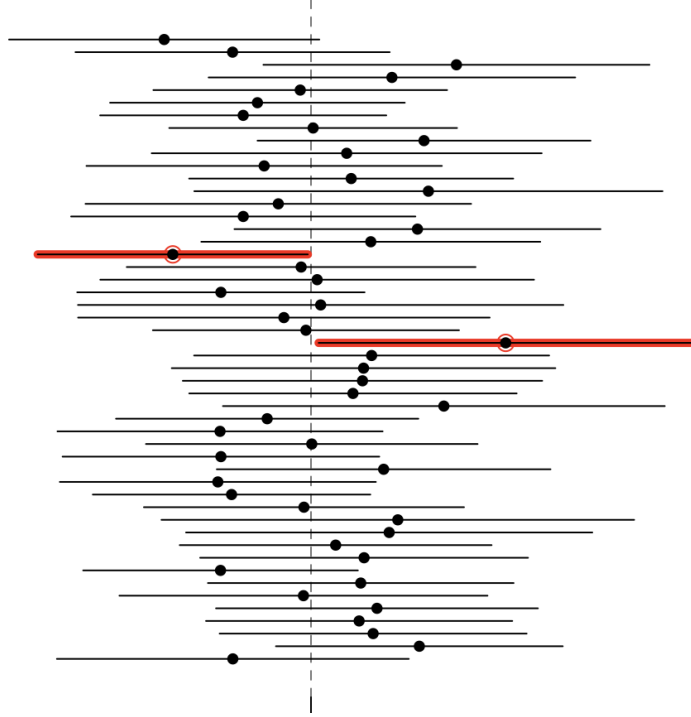
However, the distribution of the standardized variable  $Z$  is independent of the parameter  $\mu$ . Thus,  $Z$  is a pivotal quantity since it is a function of the sample  $X_1, X_2, \dots, X_n$  and parameter  $\mu$ . Using this standardized variable as the pivotal quantity, we can construct a confidence interval for the population mean  $\mu$  as follows:

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_{\mu \sim \bar{X}} \left[ -z_{\alpha/2} \leq Z \leq z_{\alpha/2} \right] \\ &= \mathbb{P}_{\mu \sim \bar{X}} \left[ -z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] \\ &= \mathbb{P}_{\mu \sim \bar{X}} \left[ \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]. \end{aligned}$$

Hence, the  $(1 - \alpha)100\%$  confidence interval for  $\mu$  when the population  $X$  is normal and known variance  $\sigma^2$  is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{8.7}$$

To interpret the confidence interval of  $\mu$ , we say that if you repeating the same experiment process many times, and generate a confidence intervals using the same method. Then approximately  $(1 - \alpha)100\%$  of the intervals will contain the true value of  $\mu$ .



**Figure 8.1:** This is a simulation of the confidence interval for the population mean. We generate 50 samples of size 30 from a normal population with mean 50 and standard deviation 15. The red horizontal line indicates the true mean of the population do not include the true population mean. Observe that 48 out of 50, of the intervals contain the true mean. Thus, the coverage probability is approximately 96%.

### 8.2.1 Confidence interval for small sample mean

Suppose  $X_1, X_2, \dots, X_n$  is random sample from a normal population  $X$  with mean  $\mu$  and variance  $\sigma^2 > 0$ . Let the sample mean and sample variances be  $\bar{X}$  and  $S^2$  respectively. Then

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

and

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1).$$

Therefore, the random variable defined by the ratio of  $\frac{(n-1)S^2}{\sigma^2}$  to  $\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$  has a  $t$ -distribution with  $(n-1)$  degrees of freedom, that is

$$Q(X_1, X_2, \dots, X_n, \mu) = \frac{\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1),$$

where  $Q$  is the pivotal quantity to be used for the construction of the confidence interval for


$\mu$ . Using this pivotal quantity, we construct the confidence interval as follows:


$$\begin{aligned} 1 - \alpha &= P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1)\right) \\ &= P\left(\bar{X} - \left(\frac{S}{\sqrt{n}}\right)t_{\frac{\alpha}{2}}(n-1) \leq \mu \leq \bar{X} + \left(\frac{S}{\sqrt{n}}\right)t_{\frac{\alpha}{2}}(n-1)\right) \end{aligned}$$

Hence, the  $100(1 - \alpha)\%$  confidence interval for  $\mu$  when the population  $X$  is normal with the unknown variance  $\sigma^2$  is given by

$$\left[\bar{X} - \left(\frac{S}{\sqrt{n}}\right)t_{\frac{\alpha}{2}}(n-1), \bar{X} + \left(\frac{S}{\sqrt{n}}\right)t_{\frac{\alpha}{2}}(n-1)\right]. \quad (8.8)$$

**Remark.** For small sample size  $n < 30$ , we should use the  $t$ -distribution to construct the confidence interval for the population mean  $\mu$  as the population variance  $\sigma^2$  is unknown.

 **Example 8.2.1.** Let  $X_1, X_2, \dots, X_{11}$  be a random sample from a normal population with unknown mean  $\mu$  and variance  $\sigma^2 = 9.9$ . Given that  $\sum_{i=1}^{11} x_i = 132$ . Find a 95% confidence interval for  $\mu$ .

 **Solution** From the information above, the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i}{11} = \frac{132}{11} = 12.$$

Furthermore, since  $\mu$  is unknown, we use  $\hat{\mu} = \bar{x} = 12$ . Since each  $X_i \sim N(\mu, \sigma^2 = 9.9)$ , the confidence interval for  $\mu$  at 95% confidence level is

$$\begin{aligned} \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &= 12 \pm z_{\frac{0.05}{2}} \sqrt{\frac{9.9}{11}} \\ &= 12 \pm 1.96\sqrt{0.9}. \end{aligned}$$

That is


$$[10.141, 13.859].$$

#### Definition 8.5 Minimum sample size for estimating population mean

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with unknown mean  $\mu$  and known variance  $\sigma^2$ . For a specified margin of error  $E > 0$  and confidence level  $100(1 - \alpha)\%$ , the minimum sample size  $n$  required to estimate the population mean  $\mu$  is given by

$$n = \left\lceil \left(\frac{z_{\alpha/2} \sigma}{E}\right)^2 \right\rceil. \quad (8.9)$$

The value of  $n$  determined from the above formula is the **minimum** sample size required to ensure that the desired level of confidence is achieved. So we always round up to the next larger integer.

 **Example 8.2.2.** Suppose that we wanted to estimate the true average number of eggs a queen honeybee lays with 95% confidence. The margin of error we are willing to accept is 2 eggs. Suppose

we all know that the sample variance is  $s^2 = 100$ . What is the minimum sample size required?

⇒ **Solution** For a 95% confidence interval, we have  $\alpha = 0.05$  and  $z_{\alpha/2} = z_{0.025} = 1.96$ . The minimum sample size required is

$$n = \left\lceil \left( \frac{1.96 \times \sqrt{100}}{2} \right)^2 \right\rceil = \lceil 96.04 \rceil = 97.$$

Thus, we need a sample size of at least 97 queen honeybees. ◀

### 8.3 Confidence interval for population proportion

Let  $X_1, X_2, \dots, X_n$  be a random sample from a population with unknown proportion  $p$ . The sample proportion is given by

$$f(x|p) = p^x (1-p)^{1-x}, \quad x = 0, 1.$$

We want to construct a  $100(1-\alpha)\%$  approximate confidence interval for the parameter  $p$ .

To do this, we note that the likelihood function of the sample is given by

$$L(p|X) = \prod_{i=1}^n f(X_i|p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i}.$$

Taking the logarithm of the likelihood function, we get

$$\ln L(p) = \sum_{i=1}^n [x_i \ln p + (1-x_i) \ln(1-p)].$$

Differentiating, the above expression, we get

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} \sum_{i=1}^n x_i - \frac{1}{1-p} \sum_{i=1}^n (1-x_i).$$

Setting this equals to zero and solving for  $p$ , we get

$$\frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1-p} = 0,$$

that is

$$(1-p)n\bar{x} = p(n - n\bar{x}),$$

which is

$$n\bar{x} - pn\bar{x} = pn - pn\bar{x}.$$

Hence

$$p = \bar{x}.$$

Therefore, the maximum likelihood estimator of  $p$  is given by

$$\hat{p} = \bar{X}.$$

The variance of  $\bar{X}$  is

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

Since  $X \sim \text{Ber}(p)$ , the variance  $\sigma^2 = p(1 - p)$ , and

$$\text{Var}(\hat{p}) = \text{Var}(\bar{X}) = \frac{p(1 - p)}{n}.$$

Since  $\text{Var}(\hat{p})$  is not free of the parameter  $p$ , we replace  $p$  by  $\hat{p}$  in the expression of  $\text{Var}(\hat{p})$  to get

$$\text{Var}(\hat{p}) \simeq \frac{\hat{p}(1 - \hat{p})}{n}.$$

The  $100(1 - \alpha)\%$  approximate confidence interval for the parameter  $p$  is given by

$$\left[ \hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right].$$

## 8.4 Confidence interval for unknown variance

Consider a random sample  $X_1, X_2, \dots, X_n$  from a normal population with mean  $\mu$  and variance  $\sigma^2$ . When both  $\mu$  and  $\sigma^2$  are unknown, we can use the sample variance  $S^2$  to estimate the population variance  $\sigma^2$ . We know that

$$\frac{(n - 1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \implies \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2.$$

We take  $Q(X_1, \dots, X_n, \sigma^2) = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}$  as a pivotal quantity to construct the confidence interval for  $\sigma^2$ . Hence, we have

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_{\sigma^2} \left[ \frac{1}{\chi_{n-1, \alpha/2}^2} \leq Q \leq \frac{1}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \right] \\ &= \mathbb{P}_{\sigma^2} \left[ \frac{1}{\chi_{n-1, \alpha/2}^2} \leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \leq \frac{1}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \right] \\ &= \mathbb{P}_{\sigma^2} \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right] \end{aligned}$$

Hence, the  $100(1 - \alpha)\%$  confidence interval for  $\sigma^2$  when the population mean is unknown is given by

$$\left[ \frac{(n - 1)S^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \quad \frac{(n - 1)S^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right].$$

### 8.4.1 Confidence Interval for standard deviation

Let  $S^2$  be the sample variance. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Using this random variable as a pivot, we can construct a  $100(1-\alpha)\%$  confidence interval for  $\sigma$  from

$$1 - \alpha = P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right)$$

by suitably choosing the constants  $a$  and  $b$ . Hence, the confidence interval for  $\sigma$  is given by

$$\left[ \sqrt{\frac{(n-1)S^2}{b}}, \sqrt{\frac{(n-1)S^2}{a}} \right].$$

The length of this confidence interval is given by

$$L(a, b) = S\sqrt{n-1} \left[ \frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} \right]$$

In order to find the shortest confidence interval, we should find a pair of constants  $a$  and  $b$  such that  $L(a, b)$  is minimum. Thus, we have a constraint minimization problem. That is

$$\begin{aligned} &\text{Minimize } L(a, b) \\ &\text{Subject to the condition} \\ &\int_a^b f(u) du = 1 - \alpha, \end{aligned} \quad (\text{MP})$$

where

$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} 2^{\frac{1-n}{2}} x^{\frac{n-3}{2}} e^{-\frac{x}{2}}.$$

Differentiating  $L$  with respect to  $a$ , we get

$$\frac{dL}{da} = S\sqrt{n-1} \left( -\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}} \frac{db}{da} \right).$$

From

$$\int_a^b f(u) du = 1 - \alpha,$$

we find the derivative of  $b$  with respect to  $a$  as follows:

$$\frac{d}{da} \int_a^b f(u) du = \frac{d}{da} (1 - \alpha)$$



that is

$$f(b) \frac{db}{da} - f(a) = 0.$$

Thus, we have

$$\frac{db}{da} = \frac{f(a)}{f(b)}.$$

Letting this into the expression for the derivative of  $L$ , we get

$$\frac{dL}{da} = S\sqrt{n-1} \left( -\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}} \frac{f(a)}{f(b)} \right).$$

Setting this derivative to zero, we get

$$S\sqrt{n-1} \left( -\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}} \frac{f(a)}{f(b)} \right) = 0$$

which yields

$$a^{\frac{3}{2}} f(a) = b^{\frac{3}{2}} f(b).$$

Using the form of  $f$ , we get from the above expression

$$\frac{3}{a^{\frac{3}{2}}} a^{\frac{n-3}{2}} e^{-\frac{a}{2}} = \frac{3}{b^{\frac{3}{2}}} b^{\frac{n-3}{2}} e^{-\frac{b}{2}}$$

that is

$$\frac{n}{a^{\frac{3}{2}}} e^{-\frac{a}{2}} = \frac{n}{b^{\frac{3}{2}}} e^{-\frac{b}{2}}.$$

From this we get

$$\ln \left( \frac{a}{b} \right) = \left( \frac{a-b}{n} \right).$$

Hence to obtain the pair of constants  $a$  and  $b$  that will produce the shortest confidence interval for  $\sigma$ , we have to solve the following system of nonlinear equations

$$\begin{cases} \int_a^b f(u) du = 1 - \alpha \\ \ln \left( \frac{a}{b} \right) = \frac{a-b}{n}. \end{cases} \quad (\star)$$

If  $a_o$  and  $b_o$  are solutions of  $(\star)$ , then the shortest confidence interval for  $\sigma$  is given by

$$\left[ \sqrt{\frac{(n-1)S^2}{b_o}}, \sqrt{\frac{(n-1)S^2}{a_o}} \right].$$

Since this system of nonlinear equations is hard to solve analytically, numerical solutions are given in statistical literature in the form of a table for finding the shortest interval for the variance.

## 8.5 Approximate Confidence Interval with MLE

When the sample size  $n$  is large, we can use the asymptotic normality of the maximum likelihood estimator to construct an approximate confidence interval for the parameter  $\theta$ .

$$\frac{\hat{\theta} - \mathbb{E}[\hat{\theta}]}{\sqrt{\text{Var}[\hat{\theta}]}} \sim N(0, 1) \quad \text{as } n \rightarrow \infty. \quad (8.10)$$

Since that for large  $n$ , the maximum likelihood estimator  $\hat{\theta}$  is unbiased, so  $\mathbb{E}[\hat{\theta}] = \theta$ . Thus

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}[\hat{\theta}]}} \sim N(0, 1) \quad \text{as } n \rightarrow \infty. \quad (8.11)$$

Now using

$$Q = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}[\hat{\theta}]}}$$

as the pivotal quantity, we construct an *approximate*  $100(1 - \alpha)\%$  confidence interval for  $\theta$  as follows:

$$\begin{aligned} 1 - \alpha &= \mathbb{P}_{\theta}[-z_{\alpha/2} \leq Q \leq z_{\alpha/2}] \\ &= \mathbb{P}_{\theta}\left[-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}[\hat{\theta}]}} \leq z_{\alpha/2}\right] \\ &= \mathbb{P}_{\theta}\left[\hat{\theta} - z_{\alpha/2}\sqrt{\text{Var}[\hat{\theta}]} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\text{Var}[\hat{\theta}]}\right]. \end{aligned}$$

**Example 8.5.1.** If  $X_1, X_2, \dots, X_n$  is a random sample from an exponential distribution with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where  $\theta > 0$  is an unknown parameter. Find an approximate  $100(1 - \alpha)\%$  confidence interval for  $\theta$  when the sample size  $n$  is large.

**Solution** The log-likelihood function from the previous example is

$$\ell(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^n \ln x_i.$$

The likelihood function  $L(\theta)$  of the sample is

$$L(\theta) = \prod_{i=1}^n \theta x_i^{\theta-1}.$$

Hence

$$\ln L(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^n \ln x_i.$$

The first derivative of the logarithm of the likelihood function is

$$\frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \ln x_i.$$

Setting this derivative to zero and solving for  $\theta$ , we obtain

$$\theta = -\frac{n}{\sum_{i=1}^n \ln x_i}.$$

Hence, the maximum likelihood estimator of  $\theta$  is given by

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^n \ln X_i}.$$

Finding the variance of this estimator is difficult. We compute its variance by computing the Cramér-Rao bound for this estimator. The second derivative of the logarithm of the likelihood function is given by

$$\frac{d^2}{d\theta^2} \ln L(\theta) = \frac{d}{d\theta} \left( \frac{n}{\theta} + \sum_{i=1}^n \ln x_i \right) = -\frac{n}{\theta^2}.$$

Hence

$$E \left( \frac{d^2}{d\theta^2} \ln L(\theta) \right) = -\frac{n}{\theta^2}.$$

Therefore

$$\text{Var}(\hat{\theta}) \geq \frac{\theta}{n}.$$

Thus we take

$$\text{Var}(\hat{\theta}) \simeq \frac{\theta}{n}.$$

Since  $\text{Var}(\hat{\theta})$  has  $\theta$  in its expression, we replace the unknown  $\theta$  by its estimate  $\hat{\theta}$  so that

$$\text{Var}(\hat{\theta}) \simeq \frac{\hat{\theta}^2}{n}.$$

The  $100(1 - \alpha)\%$  approximate confidence interval for  $\theta$  is given by

$$\left[ \hat{\theta} - z_{\frac{\alpha}{2}} \frac{\hat{\theta}}{\sqrt{n}}, \hat{\theta} + z_{\frac{\alpha}{2}} \frac{\hat{\theta}}{\sqrt{n}} \right],$$

which is

$$\left[ -\frac{n}{\sum_{i=1}^n \ln X_i} + z_{\frac{\alpha}{2}} \left( \frac{\sqrt{n}}{\sum_{i=1}^n \ln X_i} \right), -\frac{n}{\sum_{i=1}^n \ln X_i} - z_{\frac{\alpha}{2}} \left( \frac{\sqrt{n}}{\sum_{i=1}^n \ln X_i} \right) \right].$$

⇒ **Remark (Remark 17.7).** In the next section 17.2, we derived the exact confidence interval for  $\theta$  when the population distribution is exponential. The exact  $100(1 - \alpha)\%$  confidence interval for  $\theta$  was given by

$$\left[ -\frac{\chi_{\frac{\alpha}{2}}^2(2n)}{2 \sum_{i=1}^n \ln X_i}, -\frac{\chi_{1-\frac{\alpha}{2}}^2(2n)}{2 \sum_{i=1}^n \ln X_i} \right].$$

Note that this exact confidence interval is not the shortest confidence interval for the parameter  $\theta$ . ◀

**Example 8.5.2.** If  $X_1, X_2, \dots, X_{49}$  is a random sample from a population with density

$$f(x; \theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where  $\theta > 0$  is an unknown parameter, what are 90% approximate and exact confidence intervals for  $\theta$  if  $\sum_{i=1}^{49} \ln X_i = -0.7567$ ?

**Solution** We are given the followings:

$$n = 49$$

$$\sum_{i=1}^{49} \ln X_i = -0.7567$$

$$1 - \alpha = 0.90.$$

Hence, we get

$$z_{0.05} = 1.64,$$

$$\frac{n}{\sum_{i=1}^n \ln X_i} = \frac{49}{-0.7567} = -64.75$$

and

$$\frac{\sqrt{n}}{\sum_{i=1}^n \ln X_i} = \frac{7}{-0.7567} = -9.25.$$

Hence, the approximate confidence interval is given by

$$[64.75 - (1.64)(9.25), 64.75 + (1.64)(9.25)]$$

that is  $[49.58, 79.92]$ .

Next, we compute the exact 90% confidence interval for  $\theta$  using the formula

$$\left[ -\frac{\chi_{\frac{\alpha}{2}}^2(2n)}{2 \sum_{i=1}^n \ln X_i}, -\frac{\chi_{1-\frac{\alpha}{2}}^2(2n)}{2 \sum_{i=1}^n \ln X_i} \right].$$

From chi-square table, we get

$$\chi_{0.05}^2(98) = 77.93 \quad \text{and} \quad \chi_{0.95}^2(98) = 124.34.$$

Hence, the exact 90% confidence interval is

$$\left[ \frac{77.93}{(2)(0.7567)}, \frac{124.34}{(2)(0.7567)} \right]$$

that is  $[51.49, 82.16]$ . ◀

**Example 8.5.3.** If  $X_1, X_2, \dots, X_n$  is a random sample from a population with density

$$f(x; \theta) = \begin{cases} (1 - \theta)\theta^x & \text{if } x = 0, 1, 2, \dots, \infty \\ 0 & \text{otherwise,} \end{cases}$$

where  $0 < \theta < 1$  is an unknown parameter, what is a  $100(1 - \alpha)\%$  approximate confidence interval for  $\theta$  if the sample size is large?

⇒ **Solution** The logarithm of the likelihood function of the sample is

$$\ln L(\theta) = \ln \theta \sum_{i=1}^n x_i + n \ln(1 - \theta).$$

Differentiating we see obtain

$$\frac{d}{d\theta} \ln L(\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n}{1 - \theta}.$$

Equating this derivative to zero and solving for  $\theta$ , we get  $\theta = \frac{\bar{x}}{1 + \bar{x}}$ . Thus, the maximum likelihood estimator of  $\theta$  is given by

$$\hat{\theta} = \frac{\bar{X}}{1 + \bar{X}}.$$

Next, we find the variance of this estimator using the Cramér-Rao lower bound. For this, we need the second derivative of  $\ln L(\theta)$ . Hence

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{n\bar{x}}{\theta^2} - \frac{n}{(1 - \theta)^2}.$$

Therefore

$$\begin{aligned} \mathbb{E} \left( \frac{d^2}{d\theta^2} \ln L(\theta) \right) &= \mathbb{E} \left( -\frac{n\bar{X}}{\theta^2} - \frac{n}{(1 - \theta)^2} \right) \\ &= -\frac{n}{\theta^2} \mathbb{E}(\bar{X}) - \frac{n}{(1 - \theta)^2} \\ &= -\frac{n}{\theta^2} \cdot \frac{1}{(1 - \theta)} - \frac{n}{(1 - \theta)^2} \quad (\text{since each } X_i \sim \text{GEO}(1 - \theta)) \\ &= -\frac{n}{\theta(1 - \theta)} \left[ \frac{1}{\theta} + \frac{\theta}{1 - \theta} \right] \\ &= -\frac{n(1 - \theta + \theta^2)}{\theta^2(1 - \theta)^2}. \end{aligned}$$

Therefore

$$\text{Var}(\hat{\theta}) \simeq \frac{\hat{\theta}^2 (1 - \hat{\theta})^2}{n(1 - \hat{\theta} + \hat{\theta}^2)}.$$

The  $100(1 - \alpha)\%$  approximate confidence interval for  $\theta$  is given by

$$\left[ \hat{\theta} - z_{\frac{\alpha}{2}} \frac{\hat{\theta}(1 - \hat{\theta})}{\sqrt{n(1 - \hat{\theta} + \hat{\theta}^2)}}, \hat{\theta} + z_{\frac{\alpha}{2}} \frac{\hat{\theta}(1 - \hat{\theta})}{\sqrt{n(1 - \hat{\theta} + \hat{\theta}^2)}} \right],$$

where

$$\hat{\theta} = \frac{\bar{X}}{1 + \bar{X}}.$$

## Tutorials

**Exercise 8.1** A machine produces steel rods with lengths that are normally distributed with unknown mean  $\mu$  and standard deviation  $\sigma^2$ .

A quality control inspector uses a gauge to measure the length,  $x$  centimeters, of each rod in a random sample of 100 rods from the machine's production. The summarised data are as follows:

$$\sum_{i=1}^{100} x_i = 1040.0, \quad \sum_{i=1}^{100} x_i^2 = 11102.11.$$

Construct a 95% confidence interval for the mean length of rods produced by the machine.

**Exercise 8.2** In laboratory work, it is desirable to run careful checks on the variability of readings produced on standard samples. In a study of the amount of calcium in drinking water undertaken as part of a water quality assessment, the same standard sample was run through the laboratory six times, yielding the following results (in mg/L):

$$40.1, 39.8, 40.0, 40.2, 39.9, 40.1$$

Construct a 95% confidence interval for the mean calcium content in the water.

**Exercise 8.3** The executives at Aperture Inc. having recently solved their widget crises, have another major problem with one of their products. Many cities are sending complaints that their manhole covers are defective and people are falling into the sewers. Aperture Inc. is pretty sure that only 4% of their manhole covers are defective, but they would like to do a study to confirm this number. They are hoping to construct a 95% confidence interval to get within 0.01 of the true proportion of defective manhole covers. How many manhole covers need to be tested?

**Exercise 8.4** Let  $X_1, X_2, \dots, X_n$  be a random sample of with gamma density function

$$f_{X_i}(x|\theta, \beta) = \frac{1}{\Gamma(\beta) \theta^\beta} x^{\beta-1} e^{-x/\theta}, \quad x > 0.$$

Where  $\theta$  is an unknown parameter and  $\beta$  is a known constant. Show that

$$\left[ \frac{2 \sum_{i=1}^n X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n\beta)}, \frac{2 \sum_{i=1}^n X_i}{\chi^2_{\frac{\alpha}{2}}(2n\beta)} \right]$$

is a  $100(1 - \alpha)\%$  confidence interval for  $\theta$ .

**Exercise 8.5** Let  $X_1, X_2, \dots, X_n$  be a random sample from population with Pareto density function

$$f_{X_i}(x|\theta, \beta) = \begin{cases} \theta \beta^\theta x^{-(\theta+1)} & x \geq \beta \\ 0 & \text{otherwise,} \end{cases}$$

Where  $\theta$  is an unknown parameter and  $\beta$  is a known constant. Show that

$$\left[ \frac{2 \sum_{i=1}^n \ln \left( \frac{X_i}{\beta} \right)}{\chi^2_{1-\frac{\alpha}{2}}(2n\beta)}, \frac{2 \sum_{i=1}^n \ln \left( \frac{X_i}{\beta} \right)}{\chi^2_{\frac{\alpha}{2}}(2n\beta)} \right]$$

is a  $100(1 - \alpha)\%$  confidence interval for  $\frac{1}{\theta}$ .

**Exercise 8.6** Let  $X_1, X_2, \dots, X_4$  be a random sample of sample size  $n = 4$  from population with density function

$$f_{X_i}(x|\theta) = \begin{cases} (1 + \theta)x^\theta & 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

Where  $\theta > 0$  is an unknown parameter. Construct a 90% confidence interval for  $\theta$  if the observed sample data is

$$x_1 = 0.92, x_2 = 0.75, x_3 = 0.85, x_4 = 0.80.$$

# Hypothesis testing

## Definition 9.1 Testing hypotheses

The hypothesis to be tested is called the null hypothesis. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by  $H_0$  and  $H_1$ , respectively.

If  $\theta$  denotes a population parameter, then the general format of the null hypothesis and alternative hypothesis is

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1 \quad (\star)$$

where  $\Theta_0$  and  $\Theta_1$  are disjoint subsets of the parameter space  $\Theta$  such that

$$\Theta_0 \cap \Theta_1 = \emptyset \quad \text{and} \quad \Theta_0 \sqcup \Theta_1 \subseteq \Theta. \quad (9.1)$$

**Remark.** If  $\Theta_0 \cup \Theta_1 = \Theta$ , then the test ( $\clubsuit$ ) becomes

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0 \quad (9.2)$$

## Definition 9.2 Errors in hypothesis

A **type I error** is when the null hypothesis is rejected, but it is true.

A **type II error** is not rejecting null hypothesis  $H_0$ , but in fact  $H_0$  is false.

## 9.1 What is p-value?

One way to think of it is the courtroom example. One day you get arrested as a suspect, you are innocent until proven guilty. The null hypothesis  $H_0$  is that you are innocent. But now all evidence is compiled against you. The question is: given that we are in the world where you are in fact innocent, how likely are we to see this much evidence compiled against you? As opposed to asking “what is the probability that you are innocent?”, that is what  $p$ -value means.

In statistical terminology, the  $p$ -value measures the “extremeness” of the sample.

## Definition 9.3 p-value

The  $p$ -value is the probability we would get the sample we have or something more extreme if the *null hypothesis* were true.

So, the smaller the  $p$ -value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

So what constitutes “sufficiently small” and “extreme enough” to make a decision about the null hypothesis?



## 9.2 Two sample testing

If sample  $X_1$  drawn from  $N(\mu_1, \sigma_1^2)$  and is independent of  $X_2 \sim N(\mu_2, \sigma_2^2)$ , then

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

By Central Limit theorem we obtained the result

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

where  $n_1$  and  $n_2$  are the sample size for  $X_1$  and  $X_2$  correspondingly. Hence the test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \quad (9.3)$$

In general, the test statistic for comparing two normal samples with known variances is as follow:

### Definition 9.4 Two sample means test – Large sample size with different variances

Assume that  $X_1 \sim N(\mu_1, \sigma_1^2)$  and  $X_2 \sim N(\mu_2, \sigma_2^2)$  are two independent samples with  $\sigma_1^2 \neq \sigma_2^2$ . The test statistic value is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \quad (9.4)$$

for null hypothesis

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level $\alpha$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (Two-tailed)

### 9.2.1 Two samples with equal unknown population variances

Back to previous section, the test statistic value is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Since  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , then

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now both sample variances,  $S_1^2$  and  $S_2^2$ , are estimates of  $\sigma^2$ . so this information can be combine to form a *pooled* (or *weighted*) estimate of variance, that is,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Replace  $\sigma^2$  with pooled variance estimator  $S_p^2$ , hence

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\nu=n_1+n_2-2}$$

is a  $t$ -statistic with degrees of freedom given by  $n_1 + n_2 - 2$ .

#### Definition 9.5 Two sample means test – Small sample size with equal variances

Assume that  $X_1$  and  $X_2$  are two independent samples with  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . The test statistic value is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (9.5)$$

where the pooled variance id defined as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \quad (9.6)$$

The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level $\alpha$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$t \geq t_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ (Two-tailed)

### 9.2.2 Paired samples t-test

Assume that we have two dependent samples  $X_{\blacksquare}$  and  $X_{\clubsuit}$ , each of size  $n$ . The paired samples are normally distributed with means  $\mu_1$  and  $\mu_2$ , respectively. From earlier work we define the differences between the paired observations as

$$D_i = X_{\blacksquare i} - X_{\clubsuit i} \quad \text{for } i = 1, 2, \dots, n. \quad (9.7)$$

In fact, a test of  $H_0 : \mu_0 = \mu_1$  is equivalent to a test of  $H_0 : \mu_D = 0$ . Even though the original samples  $X_{\blacksquare}$  and  $X_{\clubsuit}$  are may well be normally distributed. The key assumption is that the differences  $D_i$  are approximately normally distributed with mean  $\mu_D$  and variance  $\sigma_D^2$ .

Let  $\bar{D}$  and  $\hat{\sigma}_D^2$  be the sample mean and sample standard deviation of the differences  $D_i$ , respectively. Where

$$\mu_D = \bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

and

$$\hat{\sigma}_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}.$$

Then by Central Limit Theorem, the sampling distribution of  $\bar{D}$  is approximately normal with mean  $\mu_D$  and variance  $\sigma_D^2/n$ . That is,

$$\bar{D} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

The test statistic is given by

$$t = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

which is a  $t$ -statistic with  $\nu = n - 1$  degrees of freedom.

The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = 0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level $\alpha$
$H_1 : \mu_1 - \mu_2 > 0$	$t \geq t_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < 0$	$t \leq -t_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq 0$	either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ (Two-tailed)

### 9.2.3 Large-Sample Test

Assume that we have a single pooled sample of size  $n_1 + n_2$  rather than having two separate samples of size  $n_1$  and  $n_2$ . For example, we have two different populations  $X \sim \text{BIN}(n_1, p_1)$  and  $Y \sim \text{BIN}(n_2, p_2)$  with  $p_1 = p_2$ . By combining them together we will have a single sample of size  $n_1 + n_2$  from one population with proportion  $p$ , that is,

$$X + Y \sim \text{BIN}(n_1 + n_2, p_1 = p_2 = p)$$

#### Definition 9.6 Large samples test – difference in proportion

Assume that  $X_1$  and  $X_2$  are two populations. The large-sample test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \quad (9.8)$$

The null hypothesis is

$$H_0 : p_1 - p_2 = 0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level $\alpha$
$H_1 : p_1 - p_2 > 0$	$z \geq z_\alpha$ (Upper-tailed)
$H_1 : p_1 - p_2 < 0$	$z \leq -z_\alpha$ (Lower-tailed)
$H_1 : p_1 - p_2 \neq 0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (Two-tailed)

### 9.3 Power of test

#### Definition 9.7 Power of test

For testing null hypothesis  $H_0 : \theta \in \Theta_0$ , the power function of a test  $\phi$  with rejection region  $RR$  is the function of  $\theta$  defined as

$$\beta_\phi(\theta) = \mathbb{P}_\theta[X \in RR]. \quad (9.9)$$

Note that

$$\beta(\theta) = \begin{cases} \text{Type I error probability} & \theta \in \Theta_0 \\ \text{One minus the Type II error probability} & \theta \in \Theta_0^c \end{cases}$$

A good test should has power value near 0 for most  $\theta \in \Theta_0$ , and near 1 for most  $\theta \in \Theta_0^c$ . The power function is similar to the MSE or risk function in estimation in that typically a test is better than another for some  $\theta$ 's but worse for other  $\theta$ 's.

When testing  $H_0 : \theta \leq \theta_0$  with a univariate parameter  $\theta$ , a reasonable test should have the following properties for its power function  $\beta(\theta)$ :

- $\beta(\theta)$  is a increasing function of  $\theta$ .
- $\lim_{n \rightarrow \theta_-} \beta(\theta) = 0$  and  $\lim_{n \rightarrow \theta_+} \beta(\theta) = 1$ , where  $\theta_-$  is the smallest  $\theta$  (might be  $-\infty$ ) and  $\theta_+$  is the largest  $\theta$  (might be  $+\infty$ ).

#### 9.3.1 Neyman-Pearson Lemma and powerful test

##### Lemma 9.1 Neyman-Pearson Lemma

Consider we want to test  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta = \theta_1$ , where  $\theta_0, \theta_1$  are two distinct parameter values.

##### Existence

Any test with the rejection region  $\mathcal{R}$  satisfying two conditions:

$$x \in \mathcal{R} \quad \text{if } f_{\theta_1}(x) > c f_{\theta_0}(x)$$

$$x \notin \mathcal{R} \quad \text{if } f_{\theta_1}(x) < c f_{\theta_0}(x)$$

for some  $c \geq 0$  is a UMP test of size  $\alpha_c$  with

$$\alpha_c = \mathbb{P}_{\theta_0}[X \in \mathcal{R}]. \quad (9.10)$$

And note that nothing will be specified when  $f_{\theta_1}(x) = c f_{\theta_0}(x)$ .

##### Uniqueness

If the previously specified test has a positive  $c$ , then every level  $\alpha_c$  the UMP test has size  $\alpha_c$  and has the same form previously stated except perhaps on a set  $\mathcal{A}$  satisfying

$$\mathbb{P}_{\theta_0}[\mathcal{A}] = \mathbb{P}_{\theta_1}[\mathcal{A}] = 0. \quad (9.11)$$

*Proof.* We consider the continuous case with pdfs, and that every test can be represented by the indicator function of its rejection region.

### [Existence]

Let  $\phi(x)$  be the indicator function of the rejection region of the test in the theorem and  $\psi$  be the indicator function of the rejection region of any other level  $\alpha$  test.

From the construction of  $\phi$  we have

$$[\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] \geq 0 \quad \forall x \in \mathfrak{X}$$

Thus,

$$\begin{aligned} 0 &\leq \int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] dx \\ &= \beta_{\phi}(\theta_1) - \beta_{\psi}(\theta_1) - c[\beta_{\phi}(\theta_0) - \beta_{\psi}(\theta_0)] \\ &= \beta_{\phi}(\theta_1) - \beta_{\psi}(\theta_1) - c[\alpha_c - \beta_{\psi}(\theta_0)] \\ &\leq \beta_{\phi}(\theta_1) - \beta_{\psi}(\theta_1). \end{aligned}$$

This proves  $\beta_{\phi}(\theta_1) \geq \beta_{\psi}(\theta_1)$ .

Since  $\theta_1$  is the only point in  $\Theta_0^c$ , thus  $\phi$  is a UMP test of size  $\alpha_c = \mathbb{P}_{\theta_0}[X \in \mathcal{R}]$ .

We now further continue to proof the uniqueness part.

### [Uniqueness]

Let  $\psi$  be the indicator function of another UMP test of level  $\alpha_c$ . From the previous proof, we know  $\phi$  is a UMP test of size  $\alpha_c$  and hence

$$\beta_{\phi}(\theta_1) = \beta_{\psi}(\theta_1)$$

and

$$0 \leq \int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] dx = -c[\alpha_c - \beta_{\psi}(\theta_0)].$$

Because  $c > 0$ , we have  $\alpha_c - \beta_{\psi}(\theta_0) \leq 0$ . Since  $\psi$  is a level  $\alpha_c$  test,  $\beta_{\psi}(\theta_0) \leq \alpha_c$  and hence  $\beta_{\psi}(\theta_0) = \alpha_c$ . i.e.  $\psi$  has size  $\alpha_c$ , which implies

$$\int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] dx = 0$$

Now let

$$\mathcal{A} := \{x : [\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] > 0\},$$

then

$$\int_{\mathcal{A}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] dx = 0.$$

We pick  $h(x) = [f_{\theta_0}(x) + f_{\theta_1}(x)]/2$ ,  $h(x)$  is a pdf and  $h(x) > 0$  on  $\mathcal{A}$ , which is now

$$\int_{\mathcal{A}} \frac{[\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)]}{h(x)} h(x) dx = 0.$$

From the result we established, this become

$$\mathbb{P}_h(\mathbb{K}_{x \in \mathcal{A}}[\phi(x) - \psi(x)][f_{\theta_1}(x) - cf_{\theta_0}(x)] = 0) = 1.$$

□

**Example 9.3.1.** Suppose that  $X$  represents a single observation from a population with density function given by

$$f_X(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the most powerful test with significance level  $\alpha = 0.05$  to test

$$H_0 : \theta = 2$$

against the alternative

$$H_1 : \theta = 1.$$

**Solution** Neyman-Pearson lemma can be applied to derive this test. In this case we comparing the likelihood ratio between  $H_0$  and  $H_1$ ,

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{f_X(x|\theta_0 = 2)}{f_X(x|\theta_1 = 1)} = \frac{2x}{1x^0} = 2x, \quad 0 < x < 1$$

and the form of the rejection region for the most powerful test is

$$RR = \{x \mid 2x < k\} = \left\{x \mid x < \frac{1}{2}k\right\}$$

for some constant  $k$ . Equivalently,  $k/2$  is a constant. By letting  $k' = k/2$ . The rejection region can be simplify to

$$RR = \{x \mid x < k'\}.$$

At significance level  $\alpha = 0.05$ , the value of  $k'$  is determined by

$$\begin{aligned} 0.05 &= \mathbb{P}[X \in RR \mid \theta = 2] \\ &= \mathbb{P}[X < k' \mid \theta = 2] \\ &= \int_0^{k'} 2x^{2-1} dx \\ &= \boxed{(k')^2}. \end{aligned}$$

Therefore  $(k')^2 = 0.05$ . On solving, the rejection region of the most powerful test is actually

$$RR = \{x \mid x < \sqrt{0.05} = 0.2236\}.$$

What is the actual value for  $\text{power}(\theta)$  when  $\theta = 1$ ?

$$\begin{aligned}\text{power}(1) &= \mathbb{P}[X \in RR \mid \theta = 1] \\ &= \mathbb{P}[X < 0.2236 \mid \theta = 1] \\ &= \int_0^{0.2236} 1 \, dx \\ &= \boxed{0.2236}.\end{aligned}$$

We can see that even though the rejection region gives the maximum value for  $\text{power}(1)$  among all tests with  $\alpha = 0.05$ . But  $\beta(1) = 1 - 0.2236 = 0.7764$  is still very large. ◀

🔗 **Example 9.3.2.** Let  $X_1, X_2, X_3$  denote three independent observations from a population with pdf

$$f_X(x|\theta) = \begin{cases} (1+\theta)x^\theta & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the best critical region of size 0.034 for testing

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta = 2?$$

🔗 **Solution** By Neyman-Pearson lemma, the rejection region of the most powerful test is of the form

$$\begin{aligned}RR &= \{(x_1, x_2, x_3) \in \mathcal{X} \mid \frac{L(\theta_0|x_1, x_2, x_3)}{L(\theta_1|x_1, x_2, x_3)} \leq k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{\prod_{i=1}^3 f_X(x_i|\theta_0 = 1)}{\prod_{i=1}^3 f_X(x_i|\theta_1 = 2)} \leq k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{8x_1x_2x_3}{27x_1^2x_2^2x_3^2} \leq k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{1}{x_1x_2x_3} \leq \frac{27}{8}k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid x_1x_2x_3 \geq k'\}\end{aligned}$$

where  $k'$  is some constant. Hence the most powerful test rejects  $H_0$  is of the form “Reject  $H_0$  if  $x_1x_2x_3 \geq k'$ .”

The value of  $k'$  is determined by the size of the test, that is,  $\alpha = 0.034$ . To evaluate the constant  $k'$ , we have to find the probability distribution of  $X_1X_2X_3$ . The distribution of  $X_1X_2X_3$  is quite challenging to get. But we have shown that

$$-2(1+\theta) \sum_{i=1}^3 \ln X_i \sim \chi_6^2.$$

Now we proceed to find the constant  $k'$ . Since

$$\begin{aligned}
0.034 &= \alpha \\
&= \mathbb{P}[\text{Reject } H_0 \mid H_0 \text{ is true}] \\
&= \mathbb{P}[X_1 X_2 X_3 \geq k' \mid \theta = 1] \\
&= \mathbb{P}[\ln(X_1 X_2 X_3) \geq \ln k' \mid \theta = 1] && \text{Taking logarithm.} \\
&= \mathbb{P}\left[\sum_{i=1}^3 \ln X_i \geq \ln k' \mid \theta = 1\right] \\
&= \mathbb{P}\left[-2(1 + \theta) \sum_{i=1}^3 \ln X_i \geq -2(1 + \theta) \ln k' \mid \theta = 1\right] && \text{Multiply } -2(1 + \theta) \text{ on both sides.} \\
&= \mathbb{P}\left[-4 \sum_{i=1}^3 \ln X_i \geq -4 \ln k'\right] \\
&= \mathbb{P}[\chi_6^2 \geq -4 \ln k']
\end{aligned}$$

From the  $\chi^2$  table, we have

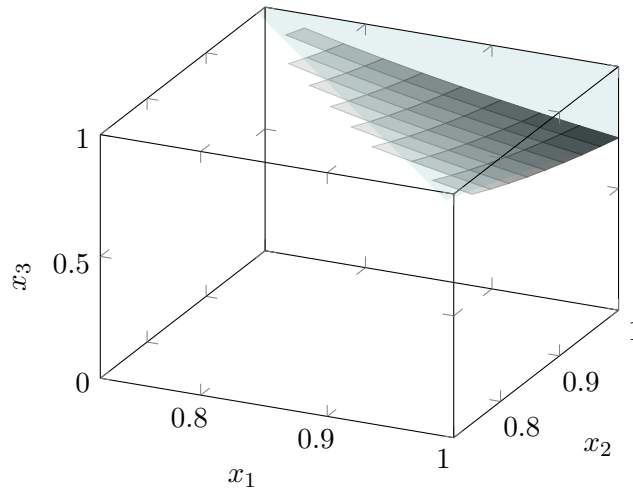
$$-4 \ln k' = 1.4.$$

Therefore

$$k' = e^{-0.35} = \boxed{0.7047}.$$

Thus, the most powerful test is given by “Reject  $H_0$  if  $x_1 x_2 x_3 \geq 0.7047$ .”

The critical region is the region above the surface  $x_1 x_2 x_3 = 0.7047$  in the unit cube  $\mathcal{I}^3 = [0, 1]^3$ . The following figure illustrates the rejection region.



**Figure 9.1:** The rejection region is to the right of the surface  $x_1 x_2 x_3 = 0.7074$ .  
(The shaded volume)



### 9.3.2 Likelihood Ratio Test

#### Definition 9.8 Likelihood Ratio Test

The likelihood ratio test statistic for testing  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_0^c$  is

$$\Lambda(X) = \frac{\sup_{\theta \in \Theta_0} L(\theta|X)}{\sup_{\theta \in \Theta} L(\theta|X)} \quad (9.12)$$

where  $L(\theta|x)$  is the likelihood function based on  $X = x$ . A **likelihood ratio test (LRT)** is any test that has a rejection region of the form

$$\{x \mid \Lambda(x) \leq c\}$$

where  $c$  is a constant satisfying  $0 \leq c \leq 1$ .

The logic behind LRTs is that the likelihood ratio  $\Lambda(x)$  is likely to be small if there are parameter points in  $\Theta_0^c$  for which  $x$  is much more likely than for any parameter in  $\Theta_0$ . Note that in the denominator of the likelihood ratio, the supremum is taken over  $\Theta$ , not  $\Theta_0^c$ .

**Example 9.3.3.** Suppose that  $X_1, X_2, \dots, X_n$  be a random sample from a  $N(\mu, \sigma^2)$  population, where both  $\mu$  and  $\sigma^2$  are unknown. What is the likelihood ratio test of significance of size  $\alpha$  for testing the null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0 ?$$

**Solution** We illustrate the following sets:

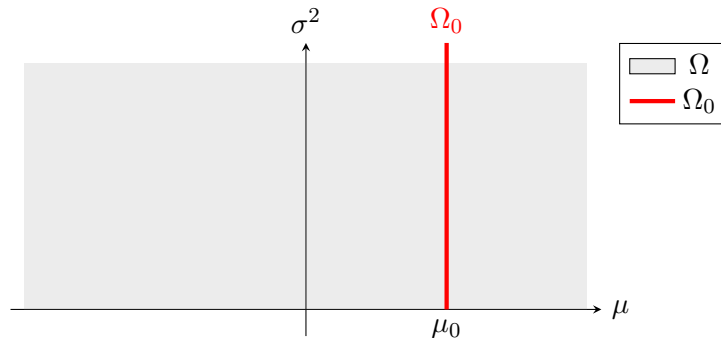
$$\blacksquare \quad \Omega = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 > 0\},$$

$$\color{red}{|} \quad \Omega_0 = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \mu = \mu_0, \sigma^2 > 0\},$$

where  $\mu_0 \in \mathbb{R}$  is a constant. The alternative parameter space is

$$\begin{aligned} \Omega_1 &= \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \mu \neq \mu_0, \sigma^2 > 0\} \\ &= \blacksquare \setminus \color{red}{|} \end{aligned}$$

so that  $\Omega_0 \cup \Omega_1 = \Omega$ .



**Figure 9.2:** The illustrated graphs of the parameter spaces  $\Omega$  and  $\Omega_0$ .

The likelihood function is given by

$$L(\mu, \sigma^2|x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

Next we find the maximum of  $L(\mu, \sigma^2)$  on the set  $\Omega_0$ . Since the set  $\Omega_0$  is equal to the set  $\{(\mu_0, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0\}$ , we have

$$\max_{(\mu, \sigma^2) \in \Omega_0} L(\mu, \sigma^2|x) = \max_{\sigma^2 > 0} L(\mu_0, \sigma^2|x).$$

Since both  $L(\mu_0, \sigma^2|x)$  and  $\ln L(\mu, \sigma^2|x)$  are maximized at the same point  $\sigma$ , we determine the value of  $\sigma$  where  $\ln L(\mu_0, \sigma^2|x)$  is maximized. Taking the natural logarithm of the likelihood function, we have

$$\ln L(\mu_0, \sigma^2|x) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2. \quad (\clubsuit)$$

differentiating with respect to  $\sigma^2$  and setting the result equal to zero, we have

$$\frac{d}{d\sigma^2} \ln L(\mu_0, \sigma^2|x) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu_0)^2 = 0.$$

Solving for  $\sigma$ , we obtain

$$\hat{\sigma}^2 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2}. \quad (\spadesuit)$$

Thus  $\ln L(\mu, \sigma^2)$  attains maximum at  $(\spadesuit)$ . Since this value of  $\sigma$  is also yield maximum value of  $L(\mu, \sigma^2)$ , we have

$$\max_{\substack{(\mu, \sigma^2) \in \Omega_0, \\ \sigma > 0}} L(\mu, \sigma^2|x) = \left(\frac{2\pi}{n} \sum_{i=1}^n (x_i - \mu_0)^2\right)^{-n/2} e^{-\frac{n}{2}}. \quad (\heartsuit)$$

Next, we need to determine the maximum of  $L(\mu, \sigma^2|x)$  on the set  $\Omega$ . As before, we consider the natural logarithm of the likelihood function is already achieves its maximum at the same point as the likelihood function itself. Taking the partial derivatives of  $(\clubsuit)$  with respect to  $\mu$  and  $\sigma^2$  respectively, we have

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2|x) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0,$$

and

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2|x) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (x_i - \mu)^2 = 0.$$

Solving for  $\mu$  and  $\sigma^2$ , we obtain

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

where  $s^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  is the sample variance.

Plugging these optimal values of  $\mu$  and  $\sigma$  into the likelihood function  $L(\mu, \sigma)$ , we have

$$\max_{\substack{(\mu, \sigma^2) \in \Omega, \\ \sigma > 0}} L(\mu, \sigma^2 | x) = \left( \frac{2\pi}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-n/2} e^{-\frac{n}{2}}. \quad (\heartsuit)$$

The likelihood ratio test statistic is given by

$$\Lambda = \frac{(\heartsuit)}{(\diamondsuit)} = \frac{\max_{\substack{(\mu, \sigma^2) \in \Omega_0, \\ \sigma > 0}} L(\mu, \sigma^2 | x)}{\max_{\substack{(\mu, \sigma^2) \in \Omega, \\ \sigma > 0}} L(\mu, \sigma^2 | x)} = \frac{\left( \frac{2\pi}{n} \sum_{i=1}^n (x_i - \mu_0)^2 \right)^{-n/2} e^{-\frac{n}{2}}}{\left( \frac{2\pi}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-n/2} e^{-\frac{n}{2}}} = \left[ \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^{-n/2} \quad (9.13)$$

Since

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

and

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2.$$

We have

$$\Lambda(x_1, x_2, \dots, x_n) = \left[ \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^{-n/2} = \left( 1 + \frac{n}{n-1} \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2}.$$

Now the inequality  $\Lambda(x_1, x_2, \dots, x_n) \leq k$  becomes

$$\left( 1 + \frac{n}{n-1} \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2} \leq k \quad (9.14)$$

and which can be rewritten in the form of

$$\left( \frac{\bar{x} - \mu_0}{s} \right)^2 \geq \frac{n-1}{n} \left( k^{-\frac{2}{n}} - 1 \right)$$

or

$$\left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq K$$

where  $K := \sqrt{(n-1) \left( k^{-\frac{2}{n}} - 1 \right)}$ . In view of the above inequality, the critical region can be described as

$$RR = \left\{ (x_1, x_2, \dots, x_n) : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq K \right\}.$$

That is, the best likelihood ratio test is described as: “ Reject  $H_0$  if  $\left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq K$  “. Since we are given  $\alpha$ , the size of the critical region. We can determine the constant  $K$ .

In order to find  $K$ , we need the density function of the statistic  $\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$  when the population  $X$  is normally distributed and the null hypothesis  $H_0 : \mu = \mu_0$  is true.

Since the population is normal with mean  $\mu$  and variance  $\sigma^2$ , so the statistic

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{\nu=n-1}$$

where  $S$  is the sample standard deviation that equals to  $\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$ . Hence

$$K = t_{\frac{\alpha}{2}; n-1} \frac{s}{\sqrt{n}},$$

where  $t_{\frac{\alpha}{2}; n-1}$  is a real number such that the integral of the  $t$ -distribution with  $n - 1$  degrees of freedom from  $t_{\frac{\alpha}{2}; n-1}$  to  $+\infty$  with area under curve equals  $\alpha/2$ .

Therefore, the likelihood ratio test is given by “Reject  $H_0$  if

$$|\bar{X} - \mu_0| \geq t_{\frac{\alpha}{2}; n-1} \frac{s}{\sqrt{n}}.”$$

if we denote

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

then the above inequality becomes

$$|T| \geq t_{\frac{\alpha}{2}; n-1}.$$

Thus the critical region is now simplified to

$$RR = \{(x_1, x_2, \dots, x_n) \mid |t| \geq t_{\frac{\alpha}{2}; n-1}\}. \quad (9.15)$$

This tells us that the null hypothesis must be rejected when the absolute value of  $t$  takes on a value greater than or equal to  $t_{\frac{\alpha}{2}; n-1}$ . ◀

**Remark.** In the example above, if we had right-sided alternative

$$H_0 : \mu \neq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.$$

Then the critical region becomes

$$RR = \{(x_1, x_2, \dots, x_n) \mid t \geq t_{\alpha; n-1}\}.$$

Similarly, if the alternative would have been left-sided, that is,

$$H_0 : \mu \neq \mu_0 \quad \text{versus} \quad H_1 : \mu < \mu_0.$$

then the critical region would have been

$$RR = \{(x_1, x_2, \dots, x_n) \mid t \leq -t_{\alpha; n-1}\}.$$

We summarize the three cases of hypotheses test of the mean of the normal population (with unknown variance) in the following table.

$H_0$	$H_1$	Critical region
$\mu = \mu_0$	$\mu > \mu_0$	$t \geq t_{\alpha; n-1}$
$\mu = \mu_0$	$\mu < \mu_0$	$t \leq -t_{\alpha; n-1}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$ t  \geq t_{\frac{\alpha}{2}; n-1}$

## 9.4 Hypothesis testing for variance

Let  $X_1, X_2, \dots, X_n$  be a random sample from a normal population with mean  $\mu$  and variance  $\sigma^2$ . What is the likelihood ratio test of significance of size  $\alpha$  for testing the null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1 : \sigma^2 \neq \sigma_0^2?$$

We illustrate the following example.

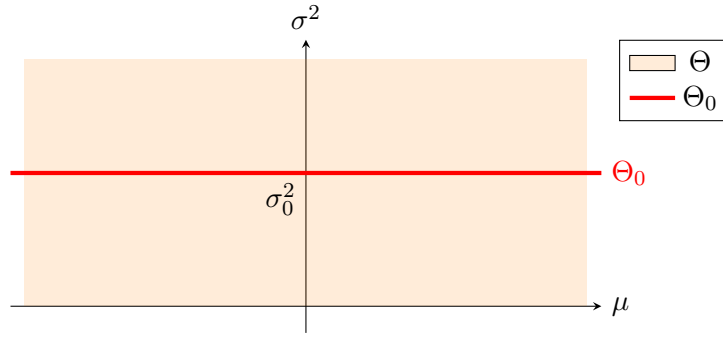
$$\text{---} \quad \Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 > 0\}$$

$$\text{---} \quad \Theta_0 = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 = \sigma_0^2\}$$

where  $\sigma_0^2 > 0$  is a constant. The alternative parameter space is

$$\begin{aligned} \Theta_1 &= \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 \neq \sigma_0^2\} \\ &= \text{---} \setminus \text{---} \end{aligned}$$

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



**Figure 9.3:** The illustrated graphs of the parameter spaces  $\Theta$  and  $\Theta_0$ .

The likelihood function is given by

$$L(\mu, \sigma^2 | x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

Next we find the maximum of  $L(\mu, \sigma^2)$  on the set  $\Theta_0$ . Since the set  $\Theta_0$  is equal to  $\{(\mu, \sigma_0^2) \in \mathbb{R}^2 \mid \mu \in \mathbb{R}\}$ , we have

$$\max_{(\mu, \sigma^2) \in \Theta_0} L(\mu, \sigma^2 | x) = \max_{\mu \in \mathbb{R}} L(\mu, \sigma_0^2 | x).$$

From here we can see that both  $L(\mu, \sigma^2 | x)$  and  $\ln L(\mu, \sigma^2 | x)$  are achieving at the same value of  $\mu$ . We further determine the value of  $\mu$  that maximizes  $\ln L(\mu, \sigma_0^2 | x)$ . Taking the natural logarithm of the likelihood function and differentiating with respect to  $\mu$ , we have

$$\begin{aligned} \frac{d \ln L(\mu, \sigma_0^2 | x)}{d\mu} &= \frac{d}{d\mu} \left[ -\frac{n}{2} \ln(\sigma_0^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 \right] \\ &= \frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \mu) \end{aligned} \quad (\heartsuit)$$

Setting (♥) to zero and solving for  $\mu$ , we have

$$\frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - \mu) = 0 \implies \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

Hence, we continue to maximize  $L(\bar{x}, \sigma_0^2 | x)$  with respect to  $\sigma^2$ . Let  $\sigma^2 = \varsigma$ , we have

$$\begin{aligned} \frac{d \ln L(\bar{x}, \varsigma | x)}{d \varsigma} &= \frac{d}{d \varsigma} \left[ -\frac{n}{2} \ln(\varsigma) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= -\frac{n}{2\varsigma} + \frac{1}{2(\varsigma)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned} \quad (\clubsuit)$$

Setting the above equation (♣) to zero and solving for  $\varsigma$ , we have

$$\begin{aligned} -\frac{n}{2\varsigma} + \frac{1}{2(\varsigma)^2} \sum_{i=1}^n (x_i - \bar{x})^2 &= 0 \\ \implies \frac{1}{2(\varsigma)^2} \sum_{i=1}^n (x_i - \bar{x})^2 &= \frac{n}{2\varsigma} \\ \implies \hat{\varsigma} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ \implies \hat{\varsigma} &= \frac{n-1}{n} \cdot \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n-1}{n} s^2. \end{aligned}$$

where  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  is the sample variance. Therefore, we obtain

$$\sup_{(\mu, \sigma^2) \in \Theta} L(\mu, \sigma^2 | x) = L(\bar{x}, \hat{\varsigma} | x) = \left( \frac{n}{2\pi(n-1)s^2} \right)^{n/2} \exp \left[ -\frac{n}{2(n-1)s^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right].$$

Thus using the optimal values we found, the likelihood ratio test statistic is

$$\begin{aligned} \Lambda(x_1, x_2, \dots, x_n) &= \frac{\sup_{(\mu, \sigma^2) \in \Theta_0} L(\mu, \sigma^2 | x)}{\sup_{(\mu, \sigma^2) \in \Theta} L(\mu, \sigma^2 | x)} \\ &= \frac{\left( \frac{n}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right]}{\left( \frac{n}{2\pi(n-1)s^2} \right)^{n/2} \exp \left[ -\frac{n}{2(n-1)s^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right]} \\ &= n^{-n/2} e^{n/2} \left( \frac{(n-1)s^2}{\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{1}{2} \left( \frac{1}{\sigma_0^2} - \frac{n}{(n-1)s^2} \right) \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= n^{-n/2} e^{n/2} \left( \frac{(n-1)s^2}{\sigma_0^2} \right)^{n/2} \exp \left[ -\frac{(n-1)s^2}{2\sigma_0^2} \right] \leq k. \end{aligned}$$

Now this inequality can be rearranged to

$$\left( \frac{(n-1)s^2}{\sigma_0^2} \right)^n \exp \left[ -\frac{(n-1)s^2}{\sigma_0^2} \right] \leq \left[ k \left( \frac{n^{n/2}}{e} \right) \right]^2 := K_0.$$

where  $K_0$  is some constant. Now let  $H$  be a function defined by

$$H(w) := w^n e^{-w} \quad \text{for } w > 0.$$

With this notation, we see that the above inequality is equivalent to

$$H\left(\frac{(n-1)s^2}{\sigma_0^2}\right) \leq K_0.$$

From this it follows that

$$\frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2.$$

In view of these inequalities, the rejection region is of the form

$$RR = \left\{ (x_1, x_2, \dots, x_n) \mid \frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2 \right\} \quad (9.16)$$

and the best likelihood ratio test can be described as follows: "Reject  $H_0$  if

$$\frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2."$$

Since we are given the size of the test  $\alpha$ , the values of  $K_1$  and  $K_2$  can be determined. As the sample  $X_1, X_2, \dots, X_n$  is drawn from a normal population with mean  $\mu$  and variance  $\sigma^2$ , so

$$\frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2. \quad (9.18)$$

when the null hypothesis  $H_0 : \sigma^2 = \sigma_0^2$  is true. Therefore, the likelihood ratio test of size  $\alpha$  rejects  $H_0$  if

$$RR = \left\{ (x_1, x_2, \dots, x_n) \mid \frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{\frac{\alpha}{2}, n-1}^2 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{1-\frac{\alpha}{2}, n-1}^2 \right\} \quad (9.19)$$

where  $\chi_{\frac{\alpha}{2}, n-1}^2$  and  $\chi_{1-\frac{\alpha}{2}, n-1}^2$  are the lower and upper  $\frac{\alpha}{2}$  points of the chi-square distribution with  $n-1$  degrees of freedom, respectively.

**Example 9.4.1.** A random sample of 16 recorded deaths in the city of Urbana was compiled. The sample average is 71.8 years old and the sample standard deviation is 9 years. Assuming that life expectancy is normally distributed but with no known standard deviation, can we claim that the standard deviation is equal to 7 years? Or is it different than that? Use a 5% level of significance.

**Solution** We want to test

$$H_0 : \sigma^2 = 49 \quad \text{versus} \quad H_1 : \sigma^2 \neq 49,$$

where  $\sigma^2$  is the variance of life expectancy in Urbana. The sample size is  $n = 16$ , and the test statistic is

$$\chi_0^2 = \frac{(16-1) \times 9^2}{49} = \boxed{24.796}.$$

The corresponding critical region is

$$RR = \{ \chi_0^2 \leq \chi_{0.025, 15}^2 = 27.488 \quad \text{or} \quad \chi_0^2 \geq \chi_{0.975, 15}^2 = 6.262 \}.$$

Hence, we do not reject  $H_0$  at the 5% level of significance, as

$$\chi_{0.975,15}^2 \leq \chi_0^2 \leq \chi_{0.025,15}^2.$$



## Tutorials

**Exercise 9.1** A firm obtains its supply of steel wire of a particular gauge from each of two manufacturers  $A$  and  $B$ . The firm suspects that the mean breaking strength, in newtons (N), of wire from manufacturer  $A$  differs from that supplied to manufacturer  $B$ .

The table below shows the breaking strengths of random samples of wire

A	80.5	83.1	73.6	70.4	68.9	71.6	82.3	78.6	73.4
B	71.4	86.2	81.4	72.3	78.9	80.3	81.4	78.0	

Assuming all such breaking strengths to be normally distributed with a standard deviation of 5N. Test, at the 5% significance level, the firm's suspicion.

**Exercise 9.2** A microbiologist wishes to determine whether there is any difference in the time it takes to make yoghurt from two different starters; *Lactobacillus acidophilus* (A) and *Lactobacillus bulgaricus* (B). Seven batches of yoghurt were made with each of the starters. The table below shows the time taken, in hours, to make each batch.

Starter A	6.8	6.3	7.4	6.1	8.2	7.3	6.9
Starter B	6.1	6.4	5.7	5.5	6.9	6.3	6.7

Assuming that both sets of times may be considered to be random samples from normal populations with the same variance, test the hypothesis that the mean time taken to make yoghurt is the same for both starters.

**Exercise 9.3** A new chemical process is developed for the manufacture of nickel-cadmium batteries. The company believes that this new process will increase the mean lifetime of a battery by 5 hours as compared to that of batteries produced by the old process. Sixteen batteries produced by the old process were randomly selected and the mean and the standard deviation of the lifetimes of these batteries were 105.2 hours and 9.1 hours, respectively. Fifteen batteries produced by the new process were also randomly selected and calculations gave corresponding values of 112.4 and 8.3 hours.

Assuming all battery lifetimes to be normally distributed, test at the 5% significance level whether there is

1. a difference in the variability of the two processes,
2. an increase of 5 hours in the mean lifetime of batteries produced by the new process as compared to that of batteries produced by the old process.



**Exercise 9.4** What is the difference between simple and composite hypothesis?

**Exercise 9.5** If an observation  $X$  is drawn from a population with probability mass function

$$f_X(x|\theta) = \begin{cases} \frac{2x}{\theta^2} & \text{for } 0 \leq x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Use Neyman-Pearson lemma to find the most powerful test for testing

$$H_0 : \theta = 4 \quad \text{versus} \quad H_1 : \theta = 5.$$

Hence, find the power of the test.

**Exercise 9.6** Let  $X$  be a random variable whose pdf under  $H_0$  and  $H_1$  are given by

$x$	1	2	3	4	5	6	7
$f(x H_0)$	.01	.01	.01	.01	.01	.01	.94
$f(x H_1)$	.06	.05	.04	.03	.02	.01	.79

Use the Neyman-Pearson lemma to find the most powerful test for testing  $H_0$  against  $H_1$  at level of significance  $\alpha = 0.04$ . Hence, compute the probability of type II error for this test.

**Exercise 9.7** Let  $X$  be a random sample from Bernoulli distribution with probability of success  $\theta$ . It is proposed to test

$$H_0 : \theta = 0.5 \quad \text{against} \quad H_1 : \theta = 0.3$$

based on sample of size 5.

1. Show that the rejection region for the test is

$$RR = \left\{ \sum_{i=1}^5 X_i > 3 \right\}.$$

2. Find the probabilities of type I and type II errors, as well as the power of test.

**Exercise 9.8** Construct the UMP test for testing

$$H_0 : \lambda = 2 \quad \text{against} \quad H_1 : \lambda > 3$$

when observations  $X$  of size  $n$  are drawn from population with density function

$$f_X(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

Considering level of significance  $\alpha = 0.02$ .

**Exercise 9.9** Find the likelihood ratio test for testing  $H_0 : \mu = \mu_0$  versus  $H_1 : \mu \neq \mu_0$  when a

random sample is drawn from  $N(\mu, 625)$ .

**Exercise 9.10** The daily output of 9 randomly selected operators was recorded before and after a two-week training programme:

Operator	A	B	C	D	E	F	G	H	I
Before	52	72	58	55	67	63	56	69	57
After	50	90	62	56	80	72	58	84	60

Assuming the change in the daily output follows a normal distribution, examine the hypothesis that the two-week training programme results in a significant increase in the mean daily output of the operators at the 5% significance level.

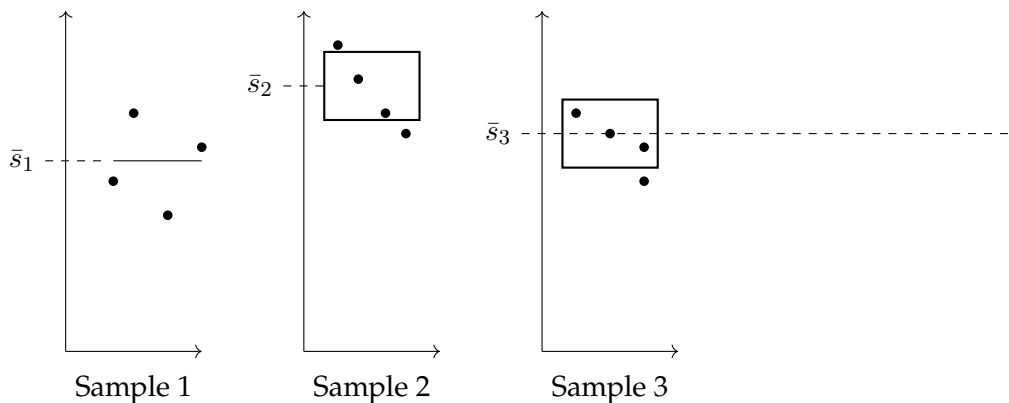
# Analysis of Variance

The main idea of an ANOVA test is that if researchers ask if a set of sample means gives evidence of differences in the population means, what matters is not how far apart the sample means are, but how far apart they are *relative to the variability of individual observations*.

## 10.1 One-way ANOVA

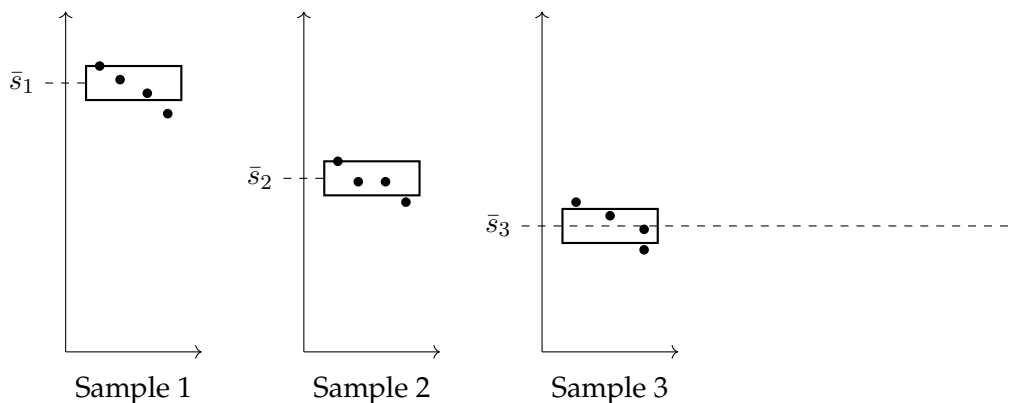
### Case 1

In this case the variation within samples is roughly on a par with that occurring between samples.



### Case 2

In this case the variation within samples is considerably less than that occurring between samples.



**Figure 10.1:** Comparison of within-sample and between-sample variation in ANOVA

**Lemma 10.1**

The total sum of squares is equal to the *sum of within* and *between sum of squares*, that is

$$SS_T = SS_W + SS_B. \quad (10.1)$$

*Proof.* Rewriting we have

$$\begin{aligned} SS_T &= \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2 \\ &= \sum_{i=1}^m \sum_{j=1}^n [(Y_{ij} - \bar{Y}_{i.}) + (Y_{i.} - \bar{Y}_{..})]^2 \\ &= \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^m \sum_{j=1}^n (Y_{i.} - \bar{Y}_{..})^2 + 2 \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})(Y_{i.} - \bar{Y}_{..}) \end{aligned}$$

Hence we obtain the asserted result

$$SS_T = SS_W + SS_B$$

and the proof of the lemma is complete.  $\square$

**Theorem 10.1**

Suppose the one-way ANOVA model is given by the equation where the  $\varepsilon_{ij}$ 's are independent and normally distributed random variables with mean zero and variance  $\sigma^2$  for  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n$ .

The null hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_m = \mu$$

is rejected whenever the statistics  $\mathcal{F}$  satisfies

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/[m(n-1)]} > F_\alpha(m-1, m(n-1)). \quad (10.2)$$

where  $\alpha$  is the significance level of the hypothesis test and  $F_\alpha(m-1, m(n-1))$  denotes the  $100(1-\alpha)$ -th percentile of the  $F$ -distribution with  $m-1$  numerator and  $m(n-1)$  denominator degrees of freedom.

*Proof.* Under the null hypothesis  $H_0 : \mu_1 = \mu_2 = \dots = \mu_m = \mu$ , the likelihood function takes the form

$$\begin{aligned} L(\mu, \sigma^2 | Y) &= \prod_{i=1}^m \prod_{j=1}^n \left\{ \frac{1}{\sqrt{2\sigma^2}} \exp \left[ -\frac{(Y_{ij} - \mu)^2}{2\sigma^2} \right] \right\} \\ &= \left( \frac{1}{\sqrt{2\sigma^2}} \right)^{nm} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \mu)^2 \right] \end{aligned} \quad (\heartsuit)$$

Maximizing the natural logarithm of the likelihood function (♥), we obtain

$$\hat{\mu} = \bar{Y}_{\bullet\bullet} \quad \text{and} \quad \widehat{\sigma^2_{H_0}} = \frac{1}{mn} SS_T$$

as the maximum likelihood estimators of  $\mu$  and  $\sigma^2$ , respectively. Plugging these estimators back into (♥), we have the maximum likelihood function, that is,

$$\max L(\mu, \sigma^2 | Y) = \left( \frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}} \right)^{nm} \exp \left[ -\frac{1}{2\widehat{\sigma^2_{H_0}}} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \right].$$

Simplifying the above expression, we see that

$$\begin{aligned} \max L(\mu, \sigma^2 | Y) &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}} \right)^{nm} \exp \left[ -\left( \frac{2}{nm} SS_T \right)^{-1} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \right] \\ &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}} \right)^{nm} \exp \left[ -\frac{nm}{2 SS_T} SS_T \right] \\ &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}} \right)^{nm} e^{-\frac{nm}{2}} \end{aligned} \quad (\clubsuit)$$

When no restrictions imposed, we obtain the maximum of the likelihood function from [lemma 10.1](#) as

$$\max L(\mu_1, \mu_2, \dots, \mu_m, \sigma^2 | Y) = \left( \frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}} \right)^{nm} \exp \left[ -\frac{1}{2\widehat{\sigma^2_{H_0}}} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i\bullet})^2 \right] \quad (\star)$$

notice that the grand mean  $\bar{Y}_{\bullet\bullet}$  now replace by  $\bar{Y}_{i\bullet}$ , and  $\widehat{\sigma^2_{H_0}}$  being replace by  $\widehat{\sigma^2}$ . Again simplifying the expression above and

$$\begin{aligned} \max L(\mu_1, \mu_2, \dots, \mu_m, \sigma^2 | Y) &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2}}} \right)^{nm} \exp \left[ -\left( \frac{2}{nm} SS_W \right)^{-1} \sum_{i=1}^m \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i\bullet})^2 \right] \\ &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2}}} \right)^{nm} \exp \left[ -\frac{nm}{2 SS_W} SS_W \right] \\ &= \left( \frac{1}{\sqrt{2\widehat{\sigma^2}}} \right)^{nm} e^{-\frac{nm}{2}}. \end{aligned}$$

Next, we are going to find the likelihood ratio statistic  $\Lambda$  for testing the null hypothesis  $H_0$ . Recall that the likelihood ratio statistic  $\Lambda$  can be found by evaluating

$$\Lambda = \frac{\max L(\mu, \sigma^2)}{\max L(\mu_1, \mu_2, \dots, \mu_m, \sigma^2)}.$$

Using (♣) divide (★), we have

$$\Lambda = \left( \frac{\widehat{\sigma^2}}{\widehat{\sigma_{H_0}^2}} \right)^{\frac{nm}{2}}.$$

Recall that the likelihood ratio test to reject the null hypothesis is

$$\Lambda < k_0 \implies \left( \frac{\widehat{\sigma^2}}{\widehat{\sigma_{H_0}^2}} \right)^{\frac{nm}{2}} < k_0 \implies \frac{\widehat{\sigma_{H_0}^2}}{\widehat{\sigma^2}} > \left( \frac{1}{k_0} \right)^{\frac{2}{nm}}$$

Applying lemma 10.1,

$$\frac{SS_W + SS_B}{SS_W} > \left( \frac{1}{k_0} \right)^{\frac{2}{nm}} \implies \frac{SS_B}{SS_W} > k' \quad (\spadesuit)$$

where  $k' := \left( \frac{1}{k_0} \right)^{\frac{2}{nm}} - 1$ . In order to find the cutoff point  $k'$  in (♠). We apply lemma 10.1. Thus

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/(m(n-1))} > \frac{m(n-1)}{m-1} k'.$$

□

Source of Variation	Sums of squares	Degree of freedom	Mean squares	$\mathcal{F}$ -statistic
Between	$SS_B$	$m-1$	$MS_B = \frac{SS_B}{m-1}$	$\mathcal{F} = \frac{MS_B}{MS_W}$
Within	$SS_W$	$N-m$	$MS_W = \frac{SS_W}{N-m}$	
Total	$SS_T$	$N-1$		

**Table 10.1:** One-way ANOVA table with unequal sample size

## 10.2 Test for the Homogeneity of Variances

One of the assumptions behind the ANOVA test is that the variances of each samples under consideration should be the same for all population.

The test statistic  $B_c$  is given by

$$B_c = \frac{(N-m) \ln S_p^2 - \sum_{i=1}^m (n_i-1) \ln s_i^2}{1 + \frac{1}{3(m-1)} \left[ \sum_{i=1}^m \frac{1}{n_i-1} - \frac{1}{N-m} \right]} \quad (10.3)$$

In the formula above,

- $s_i^2$  is the sample variance of the  $i$ -th group.
- $n_i$  is the sample size of  $i$ -th group.

- $N = \sum n_i$  is the total sample size.
- $m$  is the number of groups.

and the pooled variance  $S_p^2$  is given by

$$S_p^2 = \frac{\sum_{i=1}^m (n_i - 1) s_i^2}{N - m} = MS_W. \quad (10.4)$$

The sampling distribution of  $B_c$  is approximately chi-square with  $m - 1$  degrees of freedom, that is,

$$B_c \sim \chi^2(m - 1)$$

when  $(n_i - 1) \geq 3$ . Therefore the Barlett test rejects the null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$  at a significance level  $\alpha$  if

$$B_c \sim \chi_{1-\alpha}^2(m - 1)$$

where  $\chi_{1-\alpha}^2(m - 1)$  denotes the upper  $(1 - \alpha) \times 100$  percentile of the chi-square with  $m - 1$  degrees of freedom.

#### Definition 10.1 Barlett's test

The test statistic for Barlett's test is

$$B_c = \frac{(N - m) \ln S_p^2 - \sum_{i=1}^m (n_i - 1) \ln S_i^2}{1 + \frac{1}{3(m - 1)} \left[ \sum_{i=1}^m \frac{1}{n_i - 1} - \frac{1}{N - m} \right]}. \quad (10.5)$$

Reject  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$  if  $B_c > \chi_{\alpha}^2(m - 1)$ .

Barlett's test is the uniformly most powerful (UMP) test for the homogeneity of variances problem under the assumption that each treatment population is normally distributed. Bartlett's test is useful whenever the assumption of equal variances is made. In particular, this assumption is made for the frequently used one-way analysis of variance.

However, Barlett's test has crucial weaknesses if the normality assumption is not met:

- The tests reliability is sensitive (not robust) to non-normality.
- If the treatment populations are not approximately normal, the true significance level can be very different from the nominal significance level (say,  $\alpha = 0.05$ ). This difference depends on the kurtosis (4th moment) of the distribution.

In this case, Bartlett's or Levene's test should be applied to verify the assumption.

### 10.2.1 Levene's test

The Bartlett's test assumes that the grouped samples should be taken from normal populations. Thus Bartlett test is sensitive to departures from normality. The Levene's test is an alternative to the Bartlett's test that is less sensitive to departures from normality. Levene (1960) proposed a test for the homogeneity of population variances that considers the random variables

#### Definition 10.2 Levene's Test

To perform Levene's Test:

1. Calculate each  $Z_{ij} = |Y_{ij} - \bar{Y}_{i\bullet}|$ .
2. Run an ANOVA on the set of  $Z_{ij}$  values.
3. If  $\mathcal{F} > F_{\alpha}(m-1, N-m)$ , reject null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$  and conclude that the variances are not all equal.

Brown and Forsythe (1974) proposed using the transformed variables based on the absolute deviations from the median, that is


$$Z_{ij} = |Y_{ij} - \tilde{Y}_{i\bullet}|, \quad (10.6)$$

where  $\tilde{Y}_{i\bullet}$  denotes the median of  $i$ -th group. Again if the  $F$ -test is significant, the homogeneity of variances is rejected.

### Definition 10.3 Brown-Forsythe Test

To perform Brown-Forsythe's Test:

1. Calculate each  $Z_{ij} = |Y_{ij} - \tilde{Y}_{i\bullet}|$ , where  $\tilde{Y}_{i\bullet}$  is the  $i$ -th median of the treatment.
2. Run an ANOVA on the set of  $Z_{ij}$  values.
3. If  $\mathcal{F} > F_{\alpha}(m-1, N-m)$ , reject null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2$  and conclude that the variances are not all equal.

 **Example 10.2.1.** For the following data set contains 10 measurements of gear diameter (in centimeter) for five different batches for a total of 50 measurements.

Batches				
1	2	3	4	5
1.006	0.998	0.991	1.005	0.998
0.996	1.006	0.987	1.002	0.998
0.998	1.000	0.997	0.994	0.982
1.000	1.002	0.999	1.000	0.990
0.992	0.997	0.995	0.995	1.002
0.993	0.998	0.994	0.994	0.984
1.002	0.996	1.000	0.998	0.996
0.999	1.000	0.999	0.996	0.993
0.994	1.006	0.996	1.002	0.980
1.000	0.988	0.996	0.996	1.018

The Cochran's  $C$  test statistic is popular for test equivalence of variances among multiple sample groups. It is a one-sided upper limit variance outlier test, and it is simple to apply with the following assumptions:

- The data in each group are normally distributed.
- The data set is balanced design, i.e., each group has the same sample size.

The Cochran's  $C$ -test has been used as an alternative to Bartlett, Levene and Brown Forsythe's



tests in the evaluation of homoscedasticity (same variance) such as in a linear regression model. Note that this Cochran's  $C$  test is totally different with the Cochran's  $Q$  test, which the last one is being used in the analysis of two-way randomized block designs with different treatments in a design of experiments.

$$C = \frac{\max_{1 \leq i \leq m} s_i^2}{\sum_{i=1}^m s_i^2}. \quad (10.7)$$

The test statistic is the ratio of the maximum variance among the data set, and the sum of all the variances.

## Tutorials

**Exercise 10.1** Given 20 observations on breakdown voltage for some materials

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

# Simple Linear Regression

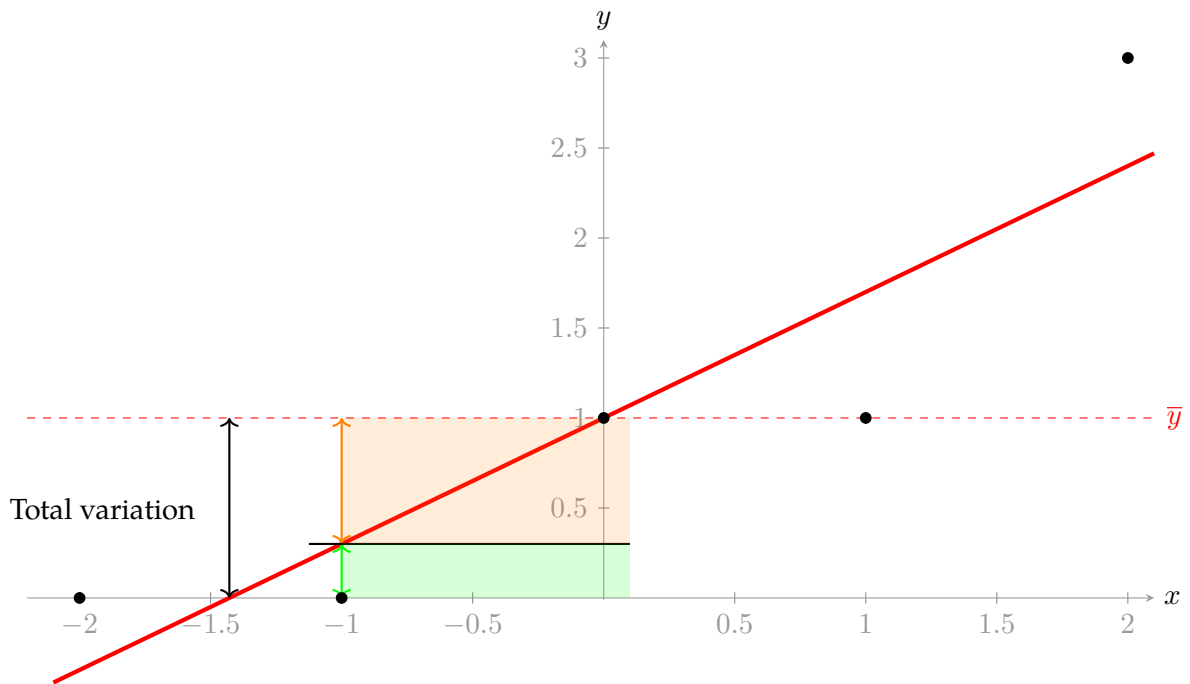
The regression model is to explain variability in independent variable by means of one or more of independent pr control variables.

A **simple regression model** can be expressed as

$$\text{Value of dependent variable} = y\text{-intercept} + (\text{Slope} \times \text{Value of Indep. variable}) + \text{Error}$$

$$y = \beta_0 + \beta x + \varepsilon.$$

Now consider that we observe five pairs of  $x$  and  $y$  observations as follow:  $(-2, 0)$ ,  $(-1, 0)$ ,  $(0, 1)$ ,  $(1, 1)$  and  $(2, 3)$ .



**Figure 11.1:** The illustrated graphs of the parameter spaces  $\Theta$  and  $\Theta_0$ .

## Definition 11.1 R-square

The statistic

$$R^2 = \frac{SS_B}{SS_T} = 1 - \frac{SS_R}{SS_T} = \frac{S_{xy}}{SS_T} = \frac{\hat{\beta}^2 S_{xx}}{S_{yy}}. \quad (11.1)$$

is called the  $R$ -square statistic or coefficient of determination, which measures the proportion of the total variation in  $Y_i$ 's that is explained by the fitted line and, hence, it measures how well the simple linear regression fitted the sample points.

## 11.1 Least squares Method

Let  $\mathcal{X} \times \mathcal{Y} = \{(x_i, y_i) \mid i = 1, 2, \dots, n\}$  be a set of data with sample size  $n$ . Consider we have a linear regression

$$\mathbb{E}_x[Y_i|x_i] = \beta_0 + \beta_1 x_i, \quad (11.2)$$

that is a simple straight line

$$y_i = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \dots, n. \quad (11.3)$$

Take the distance of each data points to the straight line and summing up. The sum of the squares of the error is given by

$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2. \quad (11.4)$$

The least squares estimates of the fitting parameters  $\beta_0$  and  $\beta_1$  are said to be those values which minimize the sum of squares error. That is,

$$\underline{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \mathcal{E}(\beta_0, \beta_1). \quad (11.5)$$

## 11.2 Making Inference

### Lemma 11.1

Given the simple linear regression  $Y_i = \beta_0 + \beta_1 x_i$ , and each  $x_i$  to be a set of arbitrary but fixed real numbers, then

1.  $SS_E = SS_T - \hat{\beta}_1 S_{xY}$ .
2.  $\text{Cov}(X, Y) = S_{xx} \hat{\beta}_1$ .
3.  $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x} \sigma^2}{S_{xx}}$ .

*Proof.* 1.

$$\begin{aligned} SS_T &= SS_E + SS_R \implies SS_E = SS_T - SS_R \\ &\implies SS_E = SS_T - SS_R \end{aligned}$$

□

### 11.2.1 Estimate $\sigma^2$

The error sum of squares (equivalently, *residual sum of squares*), denoted by  $SS_E$ , is

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad (11.6)$$

and the estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2. \quad (11.7)$$

The divisor  $n - 2$  in  $S^2$  is the number of degrees of freedom associated with  $SS_E$  and the estimate  $s^2$ . This is because to obtain the sample variance  $s^2$ , the two parameters  $\beta_0$  and  $\beta_1$  must first be estimated, which results in a loss of 2 degree of freedom, just as  $\mu$  had to be estimated in one sample problems, resulting in an estimated variance based on  $n - 1$  degree of freedom in our previous  $t$ -tests assumption. Indeed,  $S^2$  is an unbiased estimator for  $\sigma^2$ .

**Theorem 11.1 Unbiased Estimator for  $\sigma^2$**

Given the simple linear regression  $Y_i = \beta_0 + \beta_1 x_i$ , then

$$MS_E = S^2 = \frac{1}{n-2} SS_E = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (11.8)$$

is an unbiased estimator for population variance  $\sigma^2$ .

*Proof.* To show that  $S^2$  is an unbiased estimator for  $\sigma^2$ , we want verify that

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{SS_E}{n-2}\right] = \frac{1}{n-2} \mathbb{E}[SS_E] \stackrel{?}{=} \sigma^2,$$

it is important to find  $\mathbb{E}[SS_E]$  in order to verify the identity above.

Notice that

$$\begin{aligned} \mathbb{E}[SS_E] &= \mathbb{E}\left[\sum_{i=1}^n (Y_i - \hat{Y}_i)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n (Y_i - \bar{Y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^n ((Y_i - \bar{Y}) + \hat{\beta}_1 (\bar{x} - x_i))^2\right] \end{aligned}$$

continue to expand the squares and we have

$$\mathbb{E}[SS_E] = \mathbb{E}\left[\sum (Y_i - \bar{Y})^2 + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(Y_i - \bar{Y})\right]. \quad (\clubsuit)$$

Since  $\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \hat{\beta}_1 S_{xx}$ , and also  $\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$ . Combining these two results and replace the first and last term of  $(\clubsuit)$ . We have

$$\begin{aligned} \mathbb{E}[SS_E] &= \mathbb{E}\left[\sum Y_i^2 - n\bar{Y}^2 + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 (\hat{\beta}_1 S_{xx})\right] \\ &= \mathbb{E}\left[\sum Y_i^2 - n\bar{Y}^2 - \hat{\beta}_1^2 S_{xx}\right] \\ &= \sum \mathbb{E}[Y_i^2] - n\mathbb{E}[\bar{Y}^2] - S_{xx} \mathbb{E}[\hat{\beta}_1^2]. \quad (\heartsuit) \end{aligned}$$

Applying the theorem that for any random variable  $U$ ,  $\mathbb{E}[U^2] = \text{Var}[U] + \mathbb{E}[U]^2$ . We can see that

$$\begin{aligned}
 (\heartsuit) = \mathbb{E}[SS_E] &= \sum_{i=1}^n [V(Y_i) - \mathbb{E}[Y_i]^2] - n\mathbb{E}[V(\bar{Y}) - \mathbb{E}[\bar{Y}]^2] \\
 &\quad - S_{xx} \left( \frac{\sigma^2}{S_{xx}} + \beta_1^2 \right) \\
 &= n\sigma^2 + \sum_{i=1}^n (\beta_0 + \beta_1 x_i)^2 - n \left[ \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2 \right] \\
 &\quad - S_{xx} \left( \frac{\sigma^2}{S_{xx}} + \beta_1^2 \right) \\
 &= n\sigma^2 + n(\beta_0 + \beta_1 \bar{x})^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2 - \sigma^2 - S_{xx}\beta_1^2 \\
 &= \sigma^2(n-2).
 \end{aligned}$$

□

### 11.2.2 Inference about the Slope parameter $\beta_1$

#### Lemma 11.2

In normal regression, the distributions of the slope parameters  $\hat{\beta}_1$  and  $\hat{\beta}_0$  are given by

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{and} \quad \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}\right). \quad (11.9)$$

*Proof.* On the other hand, now we need to determine the distribution of intercept  $\hat{\beta}_0$ . Since  $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ , the distribution of  $\bar{Y}$  is given by

$$\bar{Y} \sim N\left(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n}\right).$$

We proved that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Multiply it by  $\bar{x}$  and the distribution of  $\bar{x}\hat{\beta}_1$  is given by

$$\bar{x}\hat{\beta}_1 \sim N\left(\bar{x}\beta_1, \bar{x}^2 \frac{\sigma^2}{S_{xx}}\right).$$

Notice that  $\hat{\beta}_0 = \bar{Y} - \bar{x}\hat{\beta}_1$ , and  $\bar{Y}$  and  $\bar{x}\hat{\beta}_1$  both are normal random variables.  $\hat{\beta}_0$  is also normal with mean equal to  $\beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$  and variance is equal to

$$\text{Var}[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}.$$

That is,

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}\right)$$

provided the distribution of  $Y_i|x_i$  is normal. The proof of the theorem is now complete. □

**Theorem 11.2**

The test statistic to test the hypotheses

$$H_0 : \beta_1 = \beta_1^* \quad \text{against} \quad H_0 : \beta_1 \neq \beta_1^*$$


is

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1^*}{\hat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}}, \quad (11.10)$$

which has a  $t$ -distribution with  $n - 2$  degree of freedom. Reject  $H_0$  if

$$|T| < t_{n-2, \frac{\alpha}{2}}.$$

The  $p$ -value based on  $n - 2$  df can be computed as was done previously for  $t$ -tests.

 **Example 11.2.1.** The frequency of chirping of a cricket is thought to be related to temperature. This study suggests the possibility that the temperature can be estimated through the chirp frequency. The following data shown the number of chirping per second,  $x$  by the striped ground cricket and the temperature,  $y$  in Fahrenheit:

$x$	20	16	20	18	17	16	15	17	15	16
$y$	89	72	93	84	81	75	70	82	69	83

Assuming the regression of temperature on the number of chirps per second is normal random variable. Test the hypotheses

$$H_0 : \beta_1 = 4 \quad \text{against} \quad H_1 : \beta_1 \neq 4$$

at the significance level 0.1.

## 11.3 Prediction

### 11.3.1 Prediction vs Estimation

What is the main reason we fitting a model to data? It is often to accomplish one of two goals. We can either use a model to estimate the relationship between response and the predictors, or to predict the response based on the predictors. Often, a good model can do both, but here we'll discuss both goals separately since the process of finding models for explaining and predicting are slightly different. In most cases, prediction is less precise than estimation with a bigger standard error.

Rather than calculate an interval estimate for  $\mu_{Y, x^*}$ , an investigator may wish to obtain a range or an interval of possible values of  $Y$  associated with some future observation when the independent variable has value  $x^*$ . Consider, for example, relating vocabulary size  $y$  to age of a child  $x$ . The CI with  $x^* = 6$  would provide a range that covers with 95% confidence the true average vocabulary size for all 6-year-old children.

Alternatively, we might wish an interval of plausible values for the vocabulary size of a particular 6-year-old child. How can you tell that a child is “off the chart” for example?

A **confidence interval** refers to a *parameter*, or population characteristic, whose value is *fixed but unknown* to us.

In contrast, a **future value** of  $Y$  is not a parameter but instead a *random variable*; for this reason we refer to an *interval of plausible values for a future  $Y$*  as a **prediction interval** rather than a confidence interval. Determining a prediction interval for  $Y$  requires that we model the error involved in the prediction of the  $Y$  variable.

One important application in linear regression is the prediction of a future value  $Y_0$  at the covariate value  $x_0$ ,

$$Y_0 = \beta_0 + \beta x_0 + \varepsilon_0$$

Since the mean of  $Y_0$  is  $\beta_0 + \beta x_0$ , the commonly used predict value for  $Y$  is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}x_0$$

which has mean  $\beta_0 + \beta x_0$  and variance

$$\begin{aligned} \text{Var}(\hat{Y}_0) &= \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}) + \text{Var}(\varepsilon_0) \\ &= \frac{\sigma^2}{S_{xx}} \left( \frac{1}{n} S_{xx} + \bar{x}^2 + x_0^2 - 2x_0\bar{x} \right) \\ &= \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \end{aligned}$$

The prediction error  $Y_0 - \hat{Y}_0$  must have zero mean, which means it is an unbiased prediction and with variance

$$\sigma^2 + \text{Var}(\hat{Y}_0) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]. \quad (11.11)$$

We would like to construct an interval  $C$  based on samples data such that

$$\mathbb{P}[Y_0 \in C] = 1 - \alpha$$

This is called a prediction interval, where  $\mathbb{P}$  is the probability measure with respect to both  $Y_0$  and  $Y_1, \dots, Y_n$ , then we can apply the fact that, under the normality assumption on  $\varepsilon$ 's,

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{\nu=n-2} \quad (11.12)$$

The resulting interval is  $C = [\hat{Y}_-, \hat{Y}_+]$ , where

$$\hat{Y}_{\pm} = \hat{Y}_0 \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (11.13)$$

### Theorem 11.3 Prediction interval for Future value $Y$

The prediction interval of  $Y$  is

$$\hat{Y}_{\pm} = \hat{Y}_0 \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (11.14)$$

Suppose that we want to construct a confidence interval for  $\theta = \alpha + \beta x_0$ , the mean of  $Y$ -values at  $x_0$ , instead of a prediction interval for the future value  $Y_0$ .

We can use  $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$  as an estimator of  $\theta$  and we still have

$$\text{Var}(\hat{Y}_0) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Under the normality assumption on  $\varepsilon$ 's,

$$\frac{\hat{Y}_0 - \theta}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t\text{-distribution with degrees of freedom } n - 2$$

Hence, the resulting confidence interval is  $C_0 = [\hat{\theta}_-, \hat{\theta}_+]$ , where

$$\hat{\theta}_{\pm} = \hat{Y}_0 \pm t_{n-2, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

This interval is **shorter** than the prediction interval for a future  $Y_0$ , because of the additional variability from  $Y_0$  in the prediction interval. Without the normality assumption,  $C_0$  is still asymptotically valid but  $C$  is not asymptotically valid.

The interpretation of the prediction level  $100(1 - \alpha)\%$  is similar to that of previous confidence levels—if is used repeatedly, in the long run the resulting interval will actually contain **the observed  $y$  values**  $100(1 - \alpha)\%$  of the time.

Notice that the 1 underneath the initial square root symbol makes the PI wider than the CI, though the intervals are both centered at  $\hat{\beta}_0 + \hat{\beta}_1 x^*$ .

Also, as  $n \rightarrow \infty$ , the width of the CI approaches 0, whereas the width of the PI does not (because even with perfect knowledge of  $\beta_0$  and  $\beta_1$ , there will still be randomness in prediction).

## 11.4 Correlation

Recall that if  $X$  and  $Y$  are bivariate random variables with means  $\mu_X$  and  $\mu_Y$  and standard deviations  $\sigma_X$  and  $\sigma_Y$ , respectively, then the correlation coefficient between  $X$  and  $Y$  is defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\mathbb{E}[(X - \mu_X)^2] \mathbb{E}[(Y - \mu_Y)^2]}}. \quad (11.15)$$

Here is the correlation coefficient for a sample data set.

### Definition 11.2 Sample Correlation Coefficient

Given a set of data  $\{(x_i, y_i) \mid i = 1, 2, \dots, n\}$ , the sample correlation coefficient  $r$  is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (11.16)$$

We can associate this sample data set into two vectors  $\vec{x} = (x_1, x_2, \dots, x_n)$  and  $\vec{y} = (y_1, y_2, \dots, y_n)$  in  $\mathbb{R}^n$ , the real  $n$ -dimensional Euclidean space. Let

$$\mathfrak{G} = \{\lambda \mathbf{e} \mid \lambda \in \mathbb{R}\}$$

to be a subset of  $\mathbb{R}^n$ , where  $\mathbf{e} = (1, 1, \dots, 1)$  is a vector in  $\mathbb{R}^n$  with all entries equal to 1.

Consider the quotient space  $V = \mathbb{R}^n / \mathfrak{G}$ , which is the set of all cosets of the form  $\vec{x} + \mathfrak{G}$ .



The analysis of variance consists of calculations that provide information about levels of variability within a regression model and form a basis for testing the

Source of Variation	Sums of squares	Degree of freedom	Mean squares	$\mathcal{F}$ -statistic
Residual	$SS_R$	1	$MS_R = \frac{SS_R}{1}$	$\mathcal{F} = \frac{MS_R}{MS_E}$
Error	$SS_E$	$n - 2$	$MS_E = \frac{SS_E}{n - 2}$	
Total	$SS_T$	$n - 1$		

**Table 11.1:** ANOVA table for simple linear regression

The final  $\mathcal{F}$ -value provides a statistic for testing the null hypothesis that there is no linear relationship between  $X$  and  $Y$  in the population. If the null hypothesis is rejected, it suggests that there is a significant linear relationship between the variables. That is, we are testing the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

This test statistic is the ratio  $SS_R/SS_E$ , the mean square residual divided by the mean square error. Under the null hypothesis, this statistic has an  $\mathcal{F}$ -distribution with 1 and  $n - 2$  degrees of freedom. When the  $SS_R$  term is large relative to the  $SS_E$  term, then the ratio will be large and indicates that there is evidence against the null hypothesis.

## Tutorials

**Exercise 11.1** Let  $Y_1, Y_2, \dots, Y_n$  be independent random variables of size  $n$ , such that for each  $Y_i \stackrel{\text{iid}}{\sim} \text{EXP}(\beta x_i)$ , where  $\beta$  is an unknown parameter. If

$$\{(x_i, y_i) \mid i = 1, 2, \dots, n\}$$

is a data set where  $y_1, y_2, \dots, y_n$  are the observed values based on  $x_1, x_2, \dots, x_n$ . Find the least squares estimator of  $\hat{\beta}$ .

**Exercise 11.2** Let  $Y_1, Y_2, \dots, Y_n$  be independent random variables of size  $n$ , such that for each  $Y_i \stackrel{\text{iid}}{\sim} \text{POI}(\beta x_i)$ , where  $\beta$  is an unknown parameter. If

$$\{(x_i, y_i) \mid i = 1, 2, \dots, n\}$$

is a data set where  $y_1, y_2, \dots, y_n$  are the observed values based on  $x_1, x_2, \dots, x_n$ . Find the least squares estimator of  $\hat{\beta}$ .

**Exercise 11.3** Given the five pairs of data points  $(x_i, y_i)$  shown below:

$x$	10	20	30	40	50
$y$	50.071	0.078	0.112	0.120	0.131

What is the curve of the form  $y = a + bx + cx^2$  best fits the data by method of least squares?

## Design of Experiments

Source of Variation	Sums of squares	Degree of freedom	Mean squares	$\mathcal{F}$ -statistic
Blocks	$SS_{Block}$	$m - 1$	$MS_B = \frac{SS_B}{m - 1}$	$\mathcal{F} = \frac{MS_B}{MS_W}$
Treatment	$SS_{Treatment}$	$N - m$	$MS_W = \frac{SS_W}{N - m}$	
Error	$SS_E$	$N - m$	$MS_E = \frac{SS_E}{N - m}$	
Total	$SS_T$	$N - 1$		

**Table 12.1:** One-way ANOVA table with unequal sample size

## Bayesian Models

## Nonparametric Statistics II – Procedures that utilize data from three or more related samples

Frequently, we can greatly improve the ability to detect group differences in the variable of interest by dividing subjects into homogeneous subgroups, called blocks, and then making comparisons among subjects within the subgroups. We can do this by using randomized complete block design (two-way ANOVA). This technique extends the two-sample paired comparison model discussed in Chapter 4 to the case in which several samples are available for analysis. Thus, for three or more samples, a block is composed of three or more subjects, more generally referred to as experimental units, who are more homogeneous with respect to each other than with respect to subjects in another block. We could form blocks on the basis of age, education and physical condition. In certain situations a single subject may be a block.

### 14.1 Friedman Two-Way Analysis Of Variance by ranks

This test is a nonparametric analogue of the parametric two-way analysis of variance. We perform calculations on ranks, which may be derived from observations measured on a higher scale or may be the original observations themselves. The procedure may be used when for one reason or another it is undesirable to use the parametric two-way ANOVA. For example, the investigator may be unwilling to assume that the sampled populations are normally distributed, a requirement for the valid use of the parametric test. Also, in some cases only ranks may be available for analysis.

The objective is to determine if we may conclude from sample evidence that there is a difference among treatment effects. We reason that if treatments do not differ in their effects, the median response of a population of subjects receiving a given treatment will be the same as the median response of a population of subjects receiving any one of the other treatments under study, after the effect of the blocking variable has been removed. Thus, if we are comparing  $k$  samples

#### 14.1.1 Ties

Theoretically, no ties should occur, since the variable whose values are ranked is assumed to be continuous. In practice, however, ties do occur, and we give tied observations the mean of the rank positions for which they are tied. Note that only ties within a given block are of concern.

$$W = \frac{12 \sum_{j=1}^k R_j^2 - 3b^2k(k+1)^2}{b^2k(k^2-1) - b(\sum t^3 - \sum t)} \quad (14.1)$$

The Friedman test is based on  $b$  sets of ranks, and the treatments are ranked separately in each set. Such a ranking scheme allows for intrablock comparisons only, since interblock comparisons

are not meaningful. When the number of treatments is small, this may pose a disadvantage. When situations arise in which comparability among blocks is desirable, the method of aligned ranks may be employed.

1. Subtract from each observation within a block some measure of location such as the block mean or median. The resulting differences, called aligned observation, which keep their identities with respect to the block and treatment combination to which they belong, are then ranked from 1 to  $[kb]$  relative to each other (the same as the **Kruskal-Wallis Test**).
2. If there is no treatment effect, we would expect each of the blocks to receive approximately the same sequence of aligned ranks. We would expect the treatment rank totals to be about equal. In the absence of ties, the aligned-ranks test statistic for the randomized complete block design is

$$T = \frac{(k-1) \left[ \sum_{j=1}^k \hat{R}_j^2 - \frac{kb^2}{4}(kb+1)^2 \right]}{\left[ \frac{kb(kb+1)(2kb+1)}{6} \right] - \frac{1}{k} \sum_{i=1}^b \hat{R}_i^2} \sim \chi_{k-1}^2. \quad (14.2)$$

3. If ties are present, replace the denominator of  $T$  with

$$\sum_{i=1}^b \sum_{j=1}^k \hat{R}_{ij}^2 - \frac{1}{k} \sum_{i=1}^b R_i^2. \quad (14.3)$$

## 14.2 Page's Test for Ordered Alternatives

### Definition 14.1 Page's test for Ordered Alternatives

The testing hypotheses are

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_k$$

versus

$$H_1 : \text{The treatment effects } \tau_1, \tau_2, \dots, \tau_k \text{ are ordered in the form of } \tau_1 \leq \tau_2 \leq \dots \leq \tau_k.$$

The test statistic is

$$L = \sum_{j=1}^k jR_j = R_1 + 2R_2 + 3R_3 + \dots + kR_k$$

where  $R_j$  for  $j = 1, 2, 3, \dots, k$  are treatment rank sums obtained in the manner

### Theorem 14.1 Large Sample Approximation for Page's Test

For large sample size, we use the test statistic:

$$z = \frac{L - \frac{bk(k+1)^2}{4}}{\sqrt{\frac{b(k^3 - k)^2}{144(k-1)}}} \sim N(0, 1) \quad (14.4)$$

Note that the large sample statistic follows standard normal instead of chi-square distribution.

### 14.3 Durbin's Test for Incomplete Block Designs

In designing an experiment, the investigator may find that it is impossible or impractical to construct a randomized complete block design of the type discussed so far. It may be impossible or impractical to apply all treatments to each block. This becomes an important problem when the number of treatments is large and the size of the blocks is limited. For example, we are going to compare the effects of seven treatments by administering the treatments to laboratory animals, with litters serving as block. Because the subjects must meet certain criteria, we can use only three animals from each litter. These conditions suggest that we use an incomplete block design, since we cannot administer each treatment to an animal from each litter.

The particular type of incomplete block design with which we are concerned is the balanced incomplete block design. In this design every possible pair of treatments appears the same number of times. Further, the balanced incomplete block design requires that each block contain the same number of subjects and that each treatment occur the same number of times.

# Answers for some questions

## Answer of exercise 6.10

A parameter is a number that describes the population, while a statistic is a number that describes the sample.

## Answer of exercise 8.1

The sample mean is

$$\bar{X} = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{1040.0}{100} = 10.4.$$

and the sample variance is

$$\begin{aligned} s^2 &= \frac{1}{100 - 1} \left( \sum_{i=1}^{100} x_i^2 - 100\bar{X}^2 \right) \\ &= \frac{1}{99} (11102.11 - 100(10.4)^2) \\ &= 2.895. \end{aligned}$$

Since the population variance  $\sigma^2$  is unknown, we use the sample variance  $s^2$  to estimate  $\sigma^2$ . The unbiased estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2 = \frac{100(8.8209)}{99} = 2.895.$$