Statistical Inference and Multivariate Analysis (MA 324)

Class Notes
January – May, 2022

Course Instructor
Ayon Ganguly
Department of Mathematics
IIT Guwahati

Contents

Chapter 1

Review

In this chapter, we will recall some of the concepts and theorems that were covered in the course Probability Theory and Random Processes (MA 225). We will use these concepts and theorems in this course.

1.1 Transformation Techniques

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector and $g : \mathbb{R}^n \to \mathbb{R}^m$ be a Borel measurable function. Clearly, $\boldsymbol{Y} = g(\boldsymbol{X})$ is a m-dimensional random vector. In this section, we will discuss different methods to find the distribution of the random vector $\boldsymbol{Y} = g(\boldsymbol{X})$, when we know the distribution of $\boldsymbol{X}$. There are mainly three techniques to obtain the distribution of $\boldsymbol{Y} = g(\boldsymbol{X})$.

1.1.1 Technique 1

In Technique 1, we try to find the joint cumulative distribution function (JCDF) of $\boldsymbol{Y} = g(\boldsymbol{X})$ given the distribution of $\boldsymbol{X}$. Recall that, for a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, the JCDF at $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$. As the JCDF exists for any random vector, this technique, in principle, can be used for any random vector. This technique is best understood using examples.

Example 1.1. Let the random variable (RV) X has the following probability mass function (PMF):

$$f(x) = \begin{cases} \frac{1}{7} & \text{if } x = -2, -1, 0, 1 \\ \frac{3}{14} & \text{if } x = 2, 3 \\ 0 & \text{otherwise.} \end{cases}$$

Consider $Y = X^2$. Let us denote the cumulative distribution function (CDF) of a random variable X by $F_X(\cdot)$. Clearly, for $y < 0$, $F_Y(y) = P(X^2 \leq y) = P(X \in \emptyset) = 0$. For $y \geq 0$,

$$F_Y(y) = P\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right).$$

Now, for $0 \leq y < 1$,

$$F_Y(y) = P(X = 0) = \frac{1}{7}.$$

For $1 \leq y < 4$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1) = \frac{3}{7}.$$

For $4 \leq y < 9$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1 \text{ or } 2 \text{ or } -2) = \frac{11}{14}.$$

For $y \geq 9$,

$$F_Y(y) = P(X = 0 \text{ or } 1 \text{ or } -1 \text{ or } 2 \text{ or } -2 \text{ or } 3) = 1.$$

Hence, the CDF of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{7} & \text{if } 0 \leq y < 1 \\ \frac{3}{7} & \text{if } 1 \leq y < 4 \\ \frac{11}{14} & \text{if } 4 \leq y < 9 \\ 1 & \text{if } y \geq 9. \end{cases} \quad ||$$

Example 1.2. Let the RV X has the following probability density function (PDF):

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

Again consider the RV $Y = X^2$. For $y < 0$, $F_Y(y) = 0$. Like the previous example, for $y \geq 0$, $F_Y(y) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right)$. Now, for $0 \leq y < 1$,

$$F_Y(y) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{|x|}{2} dx = \frac{y}{2}.$$

For $1 \leq y < 4$,

$$F_Y(y) = \int_{-1}^{1} \frac{|x|}{2} dx + \int_{1}^{\sqrt{y}} \frac{x}{3} dx = \frac{1}{6}\left(2 + y\right).$$

For $y \geq 4$,

$$F_Y(y) = \int_{-1}^{2} f(x) = 1. \quad ||$$

Example 1.3. Let the RV X has the following PDF:

$$f(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we want to find the distribution of $Y = [X]$. Here, $[x]$ denotes the largest integer not exceeding x. First notice that $F_Y(y) = P\left(Y \leq y\right) = P\left([X] \leq y\right) = 0$ for all $y < 0$. For $0 \leq y < 1$,

$$F_Y(y) = P\left([X] \leq y\right) = P\left(X < 1\right) = \int_{-\infty}^{1} f(x) dx = \int_{0}^{1} e^{-x} dx = 1 - e^{-1}.$$

For $1 \leq y < 2$,

$$F_Y(y) = P([X] \leq y) = P(X < 2) = \int_{-\infty}^{2} f(x)dx = \int_{0}^{2} e^{-x}dx = 1 - e^{-2}.$$

In general, for $i \leq y < i+1$, where $i = 0, 1, 2, \ldots$,

$$F_Y(y) = P([X] \leq y) = P(X < i+1) = \int_{-\infty}^{i+1} f(x)dx = \int_{0}^{i+1} e^{-x}dx = 1 - e^{-(i+1)}.$$

Thus, the CDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - e^{-(i+1)} & \text{if } i \leq y < i+1, i = 0, 1, 2, \ldots. \end{cases}$$ ||

Example 1.4. Let X_1 and X_2 be independent and identically distributed (*i.i.d.*) $U(0, 1)$ random variables. Suppose we want to find the CDF of $Y = X_1 + X_2$. Now,

$$F_Y(y) = P(Y \leq y) = P(X_1 + X_2 \leq y) = \int\int_{x_1+x_2 \leq y} f_{X_1, X_2}(x_1, x_2)dx_1dx_2. \tag{1.1}$$

As $X_1 \sim U(0, 1)$, $X_2 \sim U(0, 1)$ are independent RVs, the joint probability density function (JPDF) of (X_1, X_2) is given by

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} 1 & \text{if } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the JPDF of (X_1, X_2) is positive only on the unit square $(0, 1) \times (0, 1)$, which is indicated by gray shade in Figure 1.1. Now, to compute the integration in (1.1), we need to consider the following cases.

For $y < 0$, consider the Figure 1.1a. As the integrand in (1.1) is zero over the region $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ for $y < 0$,

$$F_Y(y) = 0.$$

For $0 \leq y < 1$, consider the Figure 1.1b. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

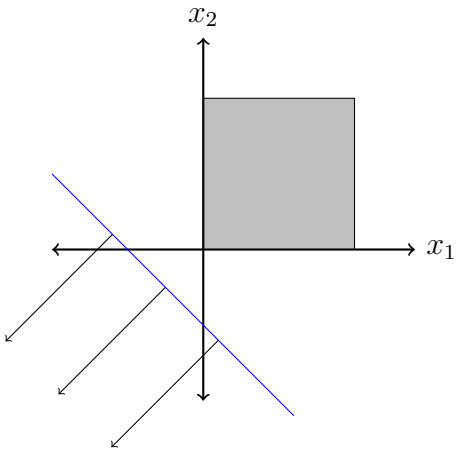
$$F_Y(y) = \int_{0}^{y} \int_{0}^{y-x_2} dx_1dx_2 = \frac{1}{2}y^2.$$

For $1 \leq y < 2$, consider the Figure 1.1c. The integrand is positive only on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

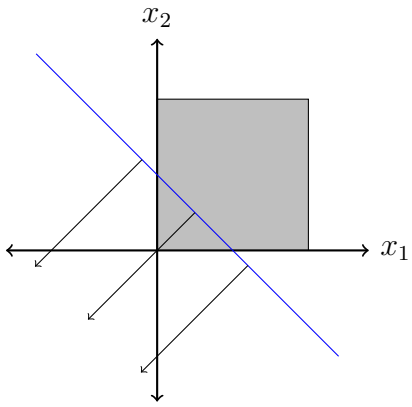
$$F_Y(y) = 1 - \int_{y-1}^{1} \int_{y-x_2}^{1} dx_1dx_2 = 1 - \frac{1}{2}(2 - y)^2.$$

For $y \geq 2$, consider the Figure 1.1d. The integrand is positive on the shaded region in the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$ and the square $(0, 1) \times (0, 1)$ is completely inside the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 \leq y\}$. Therefore,

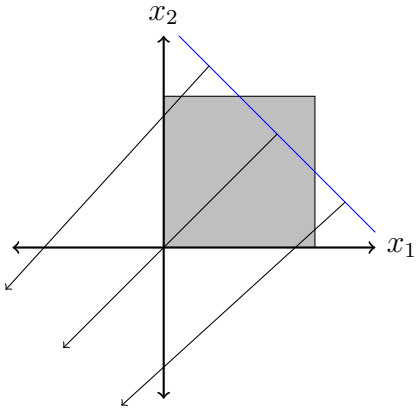
$$F_Y(y) = 1.$$



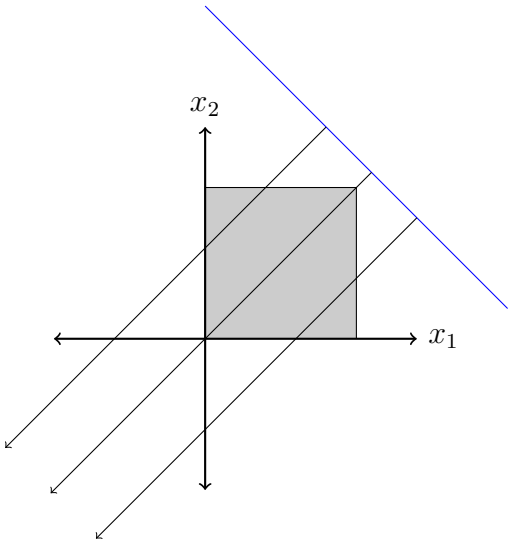
(a) $y < 0$



(b) $0 \leq y < 1$



(c) $1 \leq y < 2$



(d) $y \geq 2$

Figure 1.1: Plot for Example 1.4

Thus, the CDF of $Y = X_1 + X_2$ is given by

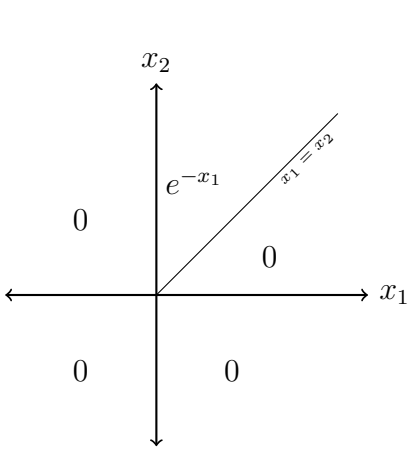
$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2}y^2 & \text{if } 0 \leq y < 1 \\ 1 - \frac{1}{2}(2-y)^2 & \text{if } 1 \leq y < 2 \\ 1 & \text{if } y \geq 2. \end{cases} \qquad ||$$

Example 1.5. Let the JPDF of (X_1, X_2) be given by

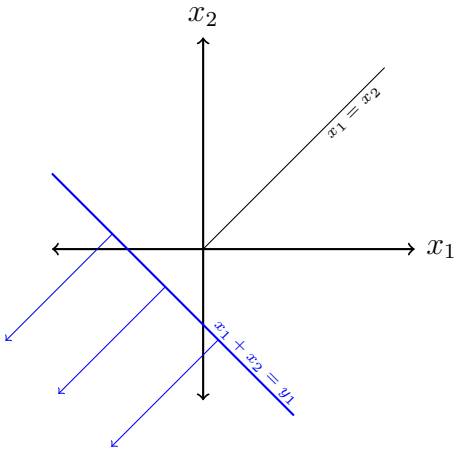
$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} e^{-x_1} & \text{if } 0 < x_1 < x_2 < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that we want to find the JCDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_2 - X_1$. Note that the JPDF of (X_1, X_2) is positive only on the set $S_{X_1, X_2} = \{(x_1, x_2) \in \mathbb{R}^2 : 0 < x_1 < x_2 < \infty\}$. See Figure 1.2a. Now, let $A_{y_1, y_2} = \{(x_1, x_2) \in \mathbb{R} : x_1 + x_2 \leq y_1, x_2 - x_1 \leq y_2\}$. Then

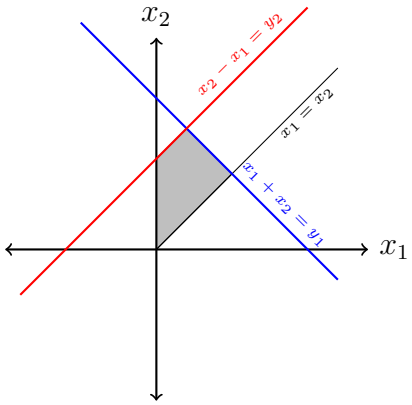
$$F_{Y_1, Y_2}(y_1, y_2) = P(X_1 + X_2 \leq y_1, X_2 - X_1 \leq y_2) = \int\int_{A_{y_1, y_2}} f_{X_1, X_2}(x_1, x_2) dx_2 dx_1. \quad (1.2)$$



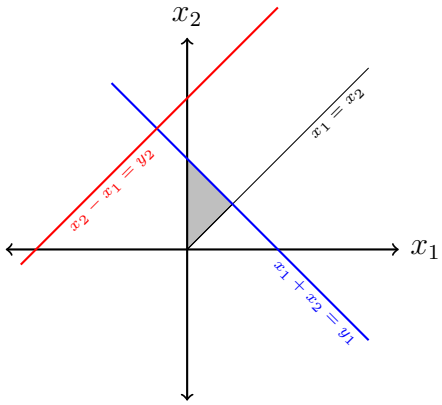
(a) JPDF of (X_1, X_2)



(b) $y_1 < 0$



(c) $0 < y_2 < y_1 < \infty$



(d) $0 < y_1 < y_2 < \infty$

Figure 1.2: Plot for Example 1.5

Suppose that $y_1 < 0$. Then $F_{Y_1}(y_1) = 0$. See the Figure 1.2b. As $F_{Y_1,Y_2}(y_1, y_2) \leq \min\{F_{Y_1}(y_1), F_{Y_2}(y_2)\}$, $F_{Y_1,Y_2}(y_1, y_2) = 0$ for $y_1 < 0$. Similarly, $F_{Y_1,Y_2}(y_1, y_2) = 0$ for $y_2 < 0$. For $0 < y_2 < y_1 < \infty$, $A_{y_1,y_2} \cap S_{X_1,X_2}$ is the shaded region of the Figure 1.2c. Therefore,

$$\begin{aligned} F_{Y_1,Y_2}(y_1, y_2) &= \int_0^{\frac{y_1-y_2}{2}} \int_{x_1}^{x_1+y_1} e^{-x_1} dx_2 dx_1 + \int_{\frac{y_1-y_2}{2}}^{\frac{y_1}{2}} \int_{x_1}^{y_1-x_1} e^{-x_1} dx_2 dx_1 \\ &= y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2)e^{-\frac{y_1-y_2}{2}}. \end{aligned}$$

For $0 < y_1 < y_2 < \infty$, $A_{y_1,y_2} \cap S_{X_1,X_2}$ is indicated by the shaded region in the Figure 1.2d. Therefore,

$$F_{Y_1,Y_2}(y_1, y_2) = \int_0^{\frac{y_1}{2}} \int_{x_1}^{y_1-x_1} e^{-x_1} dx_2 dx_1 = y_1 + 2e^{-\frac{y_1}{2}} - 2.$$

Thus, the JCDF of $(Y_1, Y_2) = (X_1 + X_2, X_2 - X_1)$ is given by

$$F_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 < 0 \text{ or } y_2 < 0 \\ y_1 + e^{-\frac{y_1}{2}} - (y_1 - y_2 + 2)e^{-\frac{y_1-y_2}{2}} & \text{if } 0 < y_2 \leq y_1 < \infty \\ y_1 + 2e^{-\frac{y_1}{2}} - 2 & \text{if } 0 < y_1 < y_2 < \infty. \end{cases} \quad ||$$

The basic idea here is to write the event $\boldsymbol{Y} \leq \boldsymbol{y}$ as $\boldsymbol{X} \in A_{\boldsymbol{y}}$ for appropriate set $A_{\boldsymbol{y}}$. In Example 1.3, we have written $Y \leq y$ as $X \in (-\infty, i+1)$ for $y \in [i,\, i+1)$. Then using the distribution of X, one needs to find the probability of the event $X \in A_{\boldsymbol{y}}$.

1.1.2 Technique 2

In Technique 2, we try to find joint probability mass function (JPMF) (if $\boldsymbol{Y}$ is a discrete random vector) or JPDF (if $\boldsymbol{Y}$ is a continuous random vector) of Y directly without finding its' CDF. Obviously, first we need to understand whether $\boldsymbol{Y}$ is discrete random vector or continuous random vector. This technique is mainly based on two theorems. The first theorem consider the case when $\boldsymbol{X}$ is discrete random vector. We will see that if $\boldsymbol{X}$ is discrete random vector, then $\boldsymbol{Y}$ is also a discrete random vector. The second theorem addresses the case when $\boldsymbol{X}$ is continuous random vector. We will see the under some conditions, $\boldsymbol{Y}$ is a continuous random vector if $\boldsymbol{X}$ is continuous random vector. With examples, we will illustrate that if the conditions do not hold, then $\boldsymbol{Y}$ can be discrete random vector as well as continuous random vector. Hence, those conditions are important.

Theorem 1.1. *Let $\boldsymbol{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ be a discrete random vector with joint probability mass function (JPMF) $f_{\boldsymbol{X}}$ and support $S_{\boldsymbol{X}}$. Let $g_i : \mathbb{R}^n \to \mathbb{R}$ for all $i = 1,\, 2,\, \ldots,\, k$. Let $Y_i = g_i(\boldsymbol{X})$ for $i = 1,\, 2,\, \ldots,\, k$. Then $\boldsymbol{Y} = (Y_1, \ldots,\, Y_k)$ is a discrete random vector with JPMF*

$$f_{\boldsymbol{Y}}(y_1,\, \ldots,\, y_k) = \begin{cases} \displaystyle\sum_{\boldsymbol{x} \in A_{\boldsymbol{y}}} f_{\boldsymbol{X}}(\boldsymbol{x}) & \text{if } (y_1,\, \ldots,\, y_k) \in S_{\boldsymbol{Y}} \\ 0 & \text{otherwise,} \end{cases}$$

where $A_{\boldsymbol{y}} = \{\boldsymbol{x} \in S_{\boldsymbol{X}} : g_i(\boldsymbol{x}) = y_i,\, i = 1,\, \ldots,\, k\}$ and $S_{\boldsymbol{Y}} = \{(g_1(\boldsymbol{x}),\, \ldots,\, g_k(\boldsymbol{x})) : \boldsymbol{x} \in S_{\boldsymbol{X}}\}$.

Proof: The proof of the theorem is skipped. □

Example 1.6. Let the RV X has the following PMF:

$$f(x) = \begin{cases} \frac{1}{7} & \text{if } x = -2,\, -1,\, 0,\, 1 \\ \frac{3}{14} & \text{if } x = 2,\, 3 \\ 0 & \text{otherwise.} \end{cases}$$

Consider $Y = X^2$ and suppose that we want to find PMF or PDF, whatever applicable, of Y. Note that the support of X is $S_X = \{-2,\, -1,\, 0,\, 1,\, 2,\, 3\}$. Intuition says that Y should takes value from the set $D = \{0,\, 1,\, 4,\, 9\}$ with positive probabilities. Based on this intuition, we will try to find $P\,(Y = y)$ for all $y \in D$ and then check if $\sum_{y \in D} P\,(Y = y)$ equal one or not.

$$P\,(Y = 0) = P\,(X = 0) = \frac{1}{7}.$$
$$P\,(Y = 1) = P\,(X = 1 \text{ or } -1) = \frac{2}{7}.$$
$$P\,(Y = 4) = P\,(X = 2 \text{ or } -2) = \frac{5}{14}.$$
$$P\,(Y = 9) = P\,(X = 3 \text{ or } -3) = \frac{3}{14}.$$

Note that again to compute $P(Y=y)$, we first find the inverse image of $Y=y$ as $X \in A_y$ and then used the distribution of X. Thus, $A_y = \{x \in \mathbb{R} : x^2 = y\}$. In the last case, $P(Y=9)$, suggests that even we do not need to consider all the elements x such that $x^2 = 9$. We need to only consider those x, which are in S_X and $x^2 = y$. Thus, we can take $A_y = \{x \in S_X : x^2 = y\}$. It is clear that $\sum_{y \in D} P(Y=y) = 1$. Hence, Y is a discrete random variable (DRV) with support D and PMF

$$f(y) = \begin{cases} \frac{1}{7} & \text{if } y = 0 \\ \frac{2}{7} & \text{if } y = 1 \\ \frac{5}{14} & \text{if } y = 4 \\ \frac{4}{14} & \text{if } y = 9 \\ 0 & \text{otherwise.} \end{cases} \qquad ||$$

Example 1.7. Let $X \sim Bin(n, p)$. Suppose that we are interested to find the distribution of $Y = n - X$. As X is a DRV, using the above theorem, Y is also DRV. Here, $S_X = \{0, 1, \ldots, n\} = S_Y$. For any $y \in S_Y$, $A_y = \{n - y\}$. Hence, the PMF of Y is

$$\begin{aligned} f_Y(y) &= \begin{cases} f_X(n-y) & \text{if } y = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \binom{n}{n-y} p^{n-y}(1-p)^{n-n+y} & \text{if } y = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \binom{n}{y} (1-p)^y p^{n-y} & \text{if } y = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, $Y \sim Bin(n, 1-p)$. Note that $Y = n - X$ is the number of failures out of n trials. Therefore, this result is well justified. ||

Example 1.8. Let $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$. Also, assume that X_1 and X_2 are independent. Then $Y = X_1 + X_2 \sim Poi(\lambda_1 + \lambda_2)$. To see it, we can apply Theorem 1.1. First note that the JPMF of (X_1, X_2) is given by

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!} & \text{if } x_1 = 0, 1, \ldots; x_2 = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1, X_2} = \{0, 1, 2, \ldots\} \times \{0, 1, 2, \ldots\}$, which implies that $S_Y = \{0, 1, 2, \ldots\}$. For $y \in S_Y$, $A_y = \{(x, y-x) : x = 0, 1, \ldots, y\}$. Hence, using the Theorem 1.1, for $y \in S_Y$,

$$f_Y(y) = \sum_{(x_1, x_2) \in A_y} \frac{e^{-(\lambda_1+\lambda_2)} \lambda_1^{x_1} \lambda_2^{x_2}}{x_1! x_2!} = \frac{e^{-(\lambda_1+\lambda_2)}}{y!} \sum_{x=0}^{y} \binom{y}{x} \lambda_1^x \lambda_2^{y-x} = \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^y .$$

Thus, the PMF of $Y = X_1 + X_2$ is

$$f_Y(y) = \begin{cases} \frac{1}{y!} e^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^y & \text{if } y = 0, 1, \ldots \\ 0 & \text{otherwise,} \end{cases}$$

which is PMF of a $P(\lambda_1 + \lambda_2)$. Hence, $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. ||

Example 1.9. Let $X_1 \sim Bin(n_1, p)$ and $X_2 \sim Bin(n_2, p)$. We also assume that X_1 and X_2 are independent. Suppose that we want to find the PMF of $Y = X_1 + X_2$. Note that X_1 and X_2 are the numbers of successes out of n_1 and n_2 independent Bernoulli trials, respectively. In both the cases the probability of success is p. Therefore, Y is the number of successes out of $n_1 + n_2$ Bernoulli trials with success probability p. As X_1 and X_2 are independent, these $n_1 + n_2$ Bernoulli trials can be assumed to be independent. Hence, the distribution of Y must be $Bin(n_1 + n_2, p)$. Let us now check if we get the same distribution using the Theorem 1.1. The JPMF of X_1 and X_2 is

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{n_1+n_2-x_1-x_2} & \text{if } x_1 = 0, 1, \ldots, n_1;\ x_2 = 0, 1, \ldots, n_2 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $S_{X_1, X_2} = \{0, 1, \ldots, n_1\} \times \{0, 1, \ldots, n_2\}$. Without loss of generality, we assume that $n_1 \leq n_2$. If not, exchange the roles of X_1 and X_2. Now, $S_Y = \{0, 1, \ldots, n_1 + n_2\}$. For $y \in S_Y$,

$$\begin{aligned} A_y &= \{(x_1, x_2) \in S_{X_1, X_2} : x_1 + x_2 = y\} \\ &= \begin{cases} \{(x, y-x) : x = 0, 1, \ldots, y\} & \text{if } 0 \leq y \leq n_1 \\ \{(x, y-x) : x = 0, 1, \ldots, n_1\} & \text{if } n_1 < y \leq n_2 \\ \{(x, y-x) : x = y - n_2, \ldots, n_1\} & \text{if } n_2 < y \leq n_1 + n_2. \end{cases} \end{aligned}$$

Hence, for $y \in S_Y$ and $y \leq n_1$,

$$f_Y(y) = \sum_{x=0}^{y} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

The last equality can be proved by collecting the coefficient of x^y from both sides of the following expression:

$$(1+x)^{n_1}(1+x)^{n_2} = \left\{\sum_{i=0}^{n_1} \binom{n_1}{i} x^i\right\} \times \left\{\sum_{i=0}^{n_2} \binom{n_2}{i} x^i\right\}.$$

For $y \in S_Y$ and $n_1 < y \leq n_2$,

$$f_Y(y) = \sum_{x=0}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

For $y \in S_Y$ and $n_2 < y \leq n_1 + n_2$,

$$f_Y(y) = \sum_{x=y-n_2}^{n_1} \binom{n_1}{x}\binom{n_2}{y-x} p^y (1-p)^{n_1+n_2-y} = \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y}.$$

Thus, $X_1 + X_2 \sim Bin\,(n_1 + n_2, p)$. Note that independence of X_1 and X_2 and same value of probability of success are important for the result. ||

Theorem 1.2. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a continuous random vector with JPDF $f_{\boldsymbol{X}}$.*

1. *Let $y_i = g_i(\boldsymbol{x})$, $i = 1, 2, \ldots, n$ be $\mathbb{R}^n \to \mathbb{R}$ functions such that*

$$\boldsymbol{y} = g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x}))$$

is one-to-one. That means that there exists the inverse transformation $x_i = h_i(\boldsymbol{y})$, $i = 1, 2, \ldots, n$ defined on the range of the transformation.

2. *Assume that both the mapping and its' inverse are continuous.*

3. *Assume that partial derivatives* $\frac{\partial x_i}{\partial y_j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, *exist and are continuous.*

4. *Assume that the Jacobian of the inverse transformation*

$$J \doteq \det\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1,2,\ldots,n} \neq 0$$

on the range of the transformation.

Then $\boldsymbol{Y} = (g_1(\boldsymbol{X}), \ldots, g_n(\boldsymbol{X}))$ *is a continuous random vector with JPDF*

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h_1(\boldsymbol{y}), \ldots, h_n(\boldsymbol{y}))|J|.$$

Proof: The proof of this theorem can be done using transformation of variable technique for multiple integration. However, the proof is skipped here. □

Remark 1.1. Note that g is a vector valued function. As g should be one-to-one, the dimension of g should be same as dimension of the argument of g. Though we have written that $g_i : \mathbb{R}^n \to \mathbb{R}$ in the previous theorem, the conclusion of the theorem is valid if we replace $g_i : \mathbb{R}^n \to \mathbb{R}$ by $g_i : S_{\boldsymbol{X}} \to \mathbb{R}$. Moreover, the theorem gives us sufficient conditions for $g(\boldsymbol{X})$ to be a continuous random vector, when $\boldsymbol{X}$ is continuous random vector. Thus, $g(\boldsymbol{X})$ can be a continuous random vector even if the conditions of the previous theorem do not hold true. †

Example 1.10. Let $X \sim U(0, 1)$. Suppose that $g(x) = -\ln x$ for $x \in (0, 1)$. Also, the support of X is $S_X = (0, 1)$, which is an interval. Clearly, $g'(x) < 0$ for all $x \in (0, 1)$. The inverse of $g(\cdot)$ is $g^{-1}(y) = e^{-y}$ for all $y \in g(S_X) = (0, \infty)$. Hence, $Y = -\ln X$ is a continuous random variable (CRV) with PDF

$$\begin{aligned} f_Y(y) &= \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} e^{-y} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Therefore, $Y = -\ln X \sim Exp(1)$. ||

Example 1.11. Let $X \sim Exp(1)$. Suppose we are interested to find the distribution of $Y = X^2$. Here, $g(x) = x^2$ for $x \in S_X = (0, \infty)$. Also, $g'(x) = 2x > 0$ for all $x > 0$. Hence, $Y = X^2$ is a CRV. Note that $g^{-1}(y) = \sqrt{y}$. Thus, the PDF of Y is

$$f_Y(y) = \begin{cases} e^{-\sqrt{y}} \times \frac{1}{2\sqrt{y}} & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that in this case the function $g(x) = x^2$ defined on $\mathbb{R}$ is not strictly monotone. However, we need to check only on the support of X and $g(\cdot)$ is strictly monotone on $(0, \infty)$. ||

Example 1.12. Let $X \sim N(0, 1)$. Suppose that we want to find the distribution of $Y = X^2$. Note that the support of X is $\mathbb{R}$ and $g'(x) = 2x$ does not take only positive or negative values on $\mathbb{R}$. Hence, we cannot use Theorem 1.2. However, we can use technique 1 to obtain the CDF of Y and then check the type of the RV Y. The CDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 2\Phi\left(\sqrt{y}\right) - 1 & \text{if } y \geq 0. \end{cases}$$

It is easy to see that $F_Y(y) = \int_{-\infty}^{y} f_Y(t)dt$, where

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{y}}\phi\left(\sqrt{y}\right) & \text{if } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, Y is a CRV. This example shows that even if some of the conditions of the Theorem 1.2 do not hold true, the RV Y could be CRV. Thus, the conditions in the Theorem 1.2 are important and they are sufficient conditions, but not necessary. ||

Example 1.13. Let X_1 and X_2 be *i.i.d.* $U(0, 1)$ random variables. We want to find the JPDF of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. Clearly,

$$g_1(x_1, x_2) = x_1 + x_2 \quad \text{and} \quad g_2(x_1, x_2) = x_1 - x_2.$$

Thus, $\boldsymbol{y} = (y_1, y_2) = g(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2)) = (x_1 + x_2, x_1 - x_2)$. Now, if $(x_1, x_2) \neq (\tilde{x}_1, \tilde{x}_2)$, then $g(x_1, x_2) \neq g(\tilde{x}_1, \tilde{x}_2)$. If not, then $x_1 + x_2 = \tilde{x}_1 + \tilde{x}_2$ and $x_1 - x_2 = \tilde{x}_1 - \tilde{x}_2$, which implies $x_1 = \tilde{x}_1$ and $x_2 = \tilde{x}_2$. This is a contradiction. Hence, the function $g(\cdot, \cdot)$ is one-to-one. The inverse function is given by $h(y_1, y_2) = (h_1(y_1, y_2), h_2(y_1, y_2))$, where $x_1 = h_1(y_1, y_2) = \frac{1}{2}(y_1 + y_2)$ and $x_2 = h_2(y_1, y_2) = \frac{1}{2}(y_1 - y_2)$. Clearly, both the mapping and inverse mapping are continuous. Now,

$$\frac{\partial x_1}{\partial y_1} = \frac{1}{2}, \frac{\partial x_1}{\partial y_2} = \frac{1}{2}, \frac{\partial x_2}{\partial y_1} = \frac{1}{2}, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = -\frac{1}{2}.$$

All the partial derivatives are continuous. The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2} \neq 0.$$

Thus, all the four conditions of the Theorem 1.2 hold, and hence, $\boldsymbol{Y} = (Y_1, Y_2)$ is a continuous random vector with JPDF

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= f_{X_1, X_2}\left(\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right)\left|-\frac{1}{2}\right| \\ &= \begin{cases} \frac{1}{2} & \text{if } 0 < y_1 + y_2 < 2,\ 0 < y_1 - y_2 < 2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Note that in Example 1.4, we have found the distribution of $X_1 + X_2$. You may find the marginal distribution of $X_1 + X_2$ from JPDF above and check if you are getting same marginal distribution. ||

Example 1.14. Let X_1 and X_2 be *i.i.d.* $N(0,\,1)$ random variables. We want to find the PDF of $Y_1 = X_1/X_2$. Note that we cannot use Theorem 1.2 directly here as we have a single function $g_1(x_1,\,x_2) = \frac{x_1}{x_2}$. Thus, we need to bring an auxiliary new function $g_2(x_1,\,x_2)$ such that $g(x_1,\,x_2) = (g_1(x_1,\,x_2),\,g_2(x_1,\,x_2))$ satisfies all the conditions of Theorem 1.2. Let us take $g_2(x_1,\,x_2) = x_2$. Clearly, $g(x_1,\,x_2)$ is a one-to-one function. Here, the inverse function is $h(y_1,\,y_2) = (h_1(y_1,\,y_2),\,h_2(y_1,\,y_2))$, where $x_1 = h_1(y_1,\,y_2) = y_1y_2$ and $x_2 = h_2(y_1,\,y_2) = y_2$. It is easy to see that mapping g and its' inverse are continuous. Also,

$$\frac{\partial x_1}{\partial y_1} = y_2,\ \frac{\partial x_1}{\partial y_2} = y_1,\ \frac{\partial x_2}{\partial y_1} = 0, \quad \text{and} \quad \frac{\partial x_2}{\partial y_2} = 1.$$

All the partial derivatives are continuous. Hence, the Jacobian is

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2.$$

Thus, all the four conditions of the Theorem 1.2 hold, and hence, $\boldsymbol{Y} = \left(\frac{X_1}{X_2},\,X_2\right)$ is a continuous random vector with JPDF

$$f_{Y_1,Y_2}(y_1,\,y_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2}\,|y_2| \quad \text{for } (y_1,\,y_2) \in \mathbb{R}^2.$$

Now, we can find the marginal PDF of Y_1 from the JPDF of $(Y_1,\,Y_2)$. The marginal PDF of Y_1 is given by

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \frac{|y_2|}{2\pi} e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi}\int_0^{\infty} y_2 e^{-\frac{1}{2}(1+y_1^2)y_2^2} dy_2 = \frac{1}{\pi(1+y_1^2)}$$

for all $y_1 \in \mathbb{R}$. Thus, $Y_1 \sim Cauchy(0,\,1)$. ||

1.1.3 Technique 3

The Technique 3 depends on the moment generating function (MGF). Hence, first we need to define the MGF of a random vector.

Definition 1.1 (Moment Generating Function)**.** *Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector. The MGF of $\boldsymbol{X}$ at $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$ is defined by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right)$$

provided the expectation exists in a neighborhood of origin $\mathbf{0} = (0,\,0,\,\ldots,\,0)$.

Definition 1.2. *Two n-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to have the same distribution, denoted by $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$, if $F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{Y}}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Theorem 1.3. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two n-dimensional random vectors. Let $M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t}$ in a neighborhood around $\mathbf{0}$, then $\boldsymbol{X} \overset{d}{=} \boldsymbol{Y}$.*

Proof: The proof is out of scope of the course. □

Example 1.15. Let $X \sim N(\mu, \sigma^2)$. Suppose we are interested to find the distribution of $Y = a + bX$, which is a linear combination of X. Assume that $b \neq 0$. Otherwise $Y = a$ with probability one. First, let us try to find the MGF of Y. Note that

$$E\left(e^{tY}\right) = E\left(e^{t(a+bX)}\right) = e^{ta}E\left(e^{tbX}\right) = e^{ta}M_X(tb) \quad \text{for all } t \in \mathbb{R}.$$

Hence,

$$E(e^{tY}) = e^{ta}e^{\mu bt + \frac{1}{2}b^2t^2\sigma^2} = e^{(a+b\mu)t + \frac{1}{2}(b\sigma)^2t^2}$$

for all $t \in \mathbb{R}$. Suppose that $Z \sim N(a + b\mu, b^2\sigma^2)$. Then the MGF of Z is

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

for all $t \in \mathbb{R}$. Thus, the MGFs of Y and Z are same for all $t \in \mathbb{R}$. Thus, $Y \stackrel{d}{=} Z \sim N(a + b\mu, b^2\sigma^2)$. Note that to use the technique 3, we need to identify the MGF of Y. ||

Example 1.16. Let X_i, $i = 1, 2, \ldots, k$ be independent $Bin(n_i, p)$ RVs. Let us try to find the distribution of $Y = \sum_{i=1}^{k} X_i$. Now, the MGF of Y is

$$M_Y(t) = E\left(e^{tY}\right) = E\left(\exp\left(t\sum_{i=1}^{k} X_i\right)\right) = E\left(\prod_{i=1}^{k} e^{tX_i}\right) = \prod_{i=1}^{k} E(e^{tX_i}) = \prod_{i=1}^{k} M_{X_i}(t).$$

The fourth equality is true as the RVs X_1, X_2, $\ldots$, X_k are independent. Note that the MGF of $X \sim Bin(n, p)$ is $M_X(t) = (1 - p + pe^t)^n$ for all $t \in \mathbb{R}$. Thus, the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k}(1 - p + pe^t)^{n_i} = (1 - p + pe^t)^{\sum_{i=1}^{k} n_i}$$

for $t \in \mathbb{R}$. Let $Z \sim Bin\left(\sum_{i=1}^{k} n_i, p\right)$, then $M_Z(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Thus, $Y \stackrel{d}{=} Z \sim Bin\left(\sum_{i=1}^{k} n_i, p\right)$. Note that this example is an extension of Example 1.9. ||

Example 1.17. Let $X_1, X_2, \ldots, X_k \stackrel{i.i.d.}{\sim} Exp(\lambda)$ and $Y = \sum_{i=1}^{k} X_i$. Then the MGF of Y is

$$M_Y(t) = \prod_{i=1}^{k} M_{X_i}(t) = [M_{X_1}(t)]^k = \left(1 - \frac{t}{\lambda}\right)^{-k}$$

for all $t < \lambda$. The second equality is due to the fact that X_i has same distribution for all $i = 1, 2, \ldots, k$. Let $Z \sim Gamma(k, \lambda)$. Then $M_Z(t) = M_Y(t)$ for all $t < \lambda$. Hence, $Y \sim Gamma(k, \lambda)$. ||

Example 1.18. Let X_i, $i = 1, 2, \ldots, k$ be independent $N(\mu_i, \sigma_i^2)$ RVs. Then $\sum_{i=1}^{k} X_i \sim N\left(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \sigma_i^2\right)$. This can be proved following the same technique as the last example. I am leaving it as an exercise. ||

1.2 Bivariate Normal Distribution

Definition 1.3 (Expectation of a Random Vector)**.** *Expectation of a random vector is given by*

$$E(\boldsymbol{X}) = (EX_1, EX_2, \ldots, EX_n)' = \boldsymbol{\mu}.$$

Definition 1.4 (Variance-Covariance Matrix of a Random Vector)**.** *The variance-covariance matrix of a n-dimensional random vector, denoted by Σ, is defined by*

$$\Sigma = [Cov(X_i, X_j)]_{i,j=1}^{n} = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'.$$

Definition 1.5 (Univariate Normal Distibution)**.** *A CRV X is said to have a univariate normal distribution if the PDF of X is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for all } x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. In this case, $X \sim N(\mu, \sigma^2)$ is used to denote the RV X follows a normal distribution with parameters μ and σ^2.

Remark 1.2. Note that if $X \sim N(\mu, \sigma^2)$, then all moments of X exist. In particular, $E(X)$ and $Var(X)$ exist, and they are given by $E(X) = \mu$ and $Var(X) = \sigma^2$. This means that a normal distribution is completely specified by its mean and variance. †

Theorem 1.4 (MGF of Univariate Normal Distribution)**.** *If $X \sim N(\mu, \sigma^2)$, then the MGF of X is $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ for all $t \in \mathbb{R}$.*

Proof: The proof is straight forward from the definition of MGF. □

Definition 1.6 (Bivariate Normal)**.** *A two dimensional random vector $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is said to have a bivariate normal distribution if $aX_1 + bX_2$ is a univariate normal for all $(a, b) \in \mathbb{R}^2 \setminus (0, 0)$.*

Theorem 1.5. *If $\boldsymbol{X}$ has bivariate normal distribution, then each of X_1 and X_2 is univariate normal. Hence, $E(X_1)$, $E(X_2)$, $Var(X_1)$, $Var(X_2)$, and $Cov(X_1, X_2)$ exist.*

Proof: Taking $a = 1$ and $b = 0$, $aX_1 + bX_2 = X_1$ follows normal distribution. Similarly, X_2 follows normal distribution. As all moments of a normal RV exist, $E(X_1)$, $E(X_2)$, $Var(X_1)$, and $Var(X_2)$ exist. As $|Cov(X_1, X_2)| \leq \sqrt{Var(X_1)Var(X_2)}$, $Cov(X_1, X_2)$ exists. □

Let us denote $\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = Var(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, where $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$, $\sigma_{11} = Var(X_1)$, $\sigma_{22} = Var(X_2)$, and $\sigma_{12} = \sigma_{21} = Cov(X_1, X_2)$.

Theorem 1.6. *Let $\boldsymbol{X}$ be a bivariate normal random vector. If $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = Var(\boldsymbol{X})$, then for any fixed $\boldsymbol{u} = (a, b) \in \mathbb{R}^2 \setminus (0, 0)$,*

$$\boldsymbol{u}'\boldsymbol{X} \sim N(\boldsymbol{u}'\boldsymbol{\mu}, \boldsymbol{u}'\Sigma\boldsymbol{u}).$$

Proof: As $\boldsymbol{u}'\boldsymbol{X} = aX_1 + bX_2$, $\boldsymbol{u}'\boldsymbol{X}$ follows a univariate normal distribution. Now,

$$E(\boldsymbol{u}'\boldsymbol{X}) = a\mu_1 + b\mu_2 = \boldsymbol{u}'\boldsymbol{\mu}.$$

and

$$Var(\boldsymbol{u}'\boldsymbol{X}) = a^2\sigma_{11} + b^2\sigma_{22} + 2ab\sigma_{12} = \boldsymbol{u}'\Sigma\boldsymbol{u}.$$

Thus, $\boldsymbol{u}'\boldsymbol{X} \sim N(\boldsymbol{u}'\boldsymbol{\mu}, \boldsymbol{u}'\Sigma\boldsymbol{u})$. □

Theorem 1.7 (MGF of Bivariate Normal Distribution)**.** *Let $\boldsymbol{X}$ be a bivariate normal random vector with $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = Var(\boldsymbol{X})$, then the MGF of $\boldsymbol{X}$ is given by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}}$$

for all $\boldsymbol{t} \in \mathbb{R}^2$.

Proof: The JMGF of $\boldsymbol{X}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left(e^{\boldsymbol{t}'\boldsymbol{X}}\right) = M_{\boldsymbol{t}'\boldsymbol{X}}(1). \tag{1.3}$$

As $\boldsymbol{X}$ has a bivariate normal distribution, $\boldsymbol{t}'\boldsymbol{X} \sim N(\boldsymbol{t}'\boldsymbol{\mu}, \boldsymbol{t}'\Sigma\boldsymbol{t})$. Now, using Theorem 1.4, the proof is immediate. □

The Theorem 1.7 shows that the bivariate normal distribution is completely specified by the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix Σ. We will use the notation $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ to denote that the random vector $\boldsymbol{X}$ follows a bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ.

Theorem 1.8 (Marginal Distribution)**.** *If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, then $X_1 \sim N(\mu_1, \sigma_{11})$ and $X_2 \sim N(\mu_2, \sigma_{22})$.*

Proof: The proof of the theorem is immediate from Theorem 1.6. □

The converse of the Theorem 1.8 is not true in general. Consider the following example in this regard.

Example 1.19. Let $X \sim N(0, 1)$. Let Z be a DRV, which is independent of X and

$$P(Z = 1) = 0.5 = P(Z = -1).$$

Then $Y = ZX \sim N(0, 1)$. To see it, notice that for all $y \in \mathbb{R}$,

$$\begin{aligned} P(Y \leq y) &= P(ZX \leq y) \\ &= P(ZX \leq y | Z = 1)\, P(Z = 1) + P(ZX \leq y | Z = -1)\, P(Z = -1) \\ &= \frac{1}{2}P(X \leq y) + \frac{1}{2}P(X \geq -y) \\ &= \Phi(y). \end{aligned}$$

Thus, $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. However, (X, Y) is not a bivariate normal random vector. To see it, observe that

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

That means that $X + Y$ does not follow a univariate normal distribution, and hence, (X, Y) is not a bivariate normal random vector, though $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. ||

Theorem 1.9. *If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ and $Cov(X_1, X_2) = 0$, then X_1 and X_2 are independent.*

Proof: In this case, $\Sigma = diag(\sigma_{11}, \sigma_{22})$. Hence, the JMGF of (X_1, X_2) is

$$\begin{aligned} M_{X_1, X_2}(t_1, t_2) &= e^{t_1\mu_1 + \frac{1}{2}\sigma_{11}t_1^2} \times e^{t_2\mu_2 + \frac{1}{2}\sigma_{22}t_2^2} \\ &= M_{X_1}(t_1) M_{X_2}(t_2), \end{aligned}$$

where $M_{X_i}(\cdot)$ is the MGF of X_i, $i = 1, 2$. This shows that X_1 and X_2 are independent. □

Note that two random variable can be dependent even if covariance between them zero. The bivariate normal random vector is special in this respect.

Theorem 1.10 (Probability Density Function)**.** *Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then, for all $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{X}$ has a joint PDF given by*

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}} \end{aligned}$$

where $\sigma_1 = \sqrt{\sigma_{11}}$, $\sigma_2 = \sqrt{\sigma_{22}}$, ρ is correlation coefficient between X_1 and X_2.

Proof: The proof of this theorem is out of scope. □

Theorem 1.11 (Conditional Probability Density Function)**.** *Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then for all $x_2 \in \mathbb{R}$, the conditional PDF of X_1 given $X_2 = x_2$ is given by*

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sigma_{1|2}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1 - \mu_{1|2}}{\sigma_{1|2}}\right)^2\right] \quad \text{for } x_1 \in \mathbb{R},$$

where $\mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and $\sigma_{1|2}^2 = \sigma_1^2(1-\rho^2)$. Thus, $X_1|X_2 = x_2 \sim N\left(\mu_{1|2}, \sigma_{1|2}^2\right)$.

Proof: Easy to see from the fact that

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Of course, you need to perform some algebra. □

Corollary 1.1. *Under the condition of the Theorem 1.11, $E(X_1|X_2 = x_2) = \mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and $Var(X_1|X_2 = x_2) = \sigma_{1|2}^2 = \sigma_1^2(1-\rho^2)$ for all $x_2 \in \mathbb{R}$. Hence, the conditional variance does not depend on x_2.*

Proof: Straight forward from the previous theorem. □

1.3 Some Results on Independent and Identically Distributed Normal RVs

Theorem 1.12. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Then*

$$\sum_{i=1}^{n} X_i^2 \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right) \equiv \chi_n^2.$$

Proof: The MGF of X_1^2 is given by

$$M_{X_1^2}(t) = E\left(e^{tX_1^2}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(\frac{1}{2}-t)x^2} dx = (1-2t)^{-\frac{1}{2}},$$

for $t < \frac{1}{2}$. Hence, the MGF of $T = \sum_{i=1}^{n} X_i^2$

$$M_T(t) = \prod_{i=1}^{n} M_{X_i^2}(t) = (1-2t)^{-\frac{n}{2}},$$

where $t < \frac{1}{2}$. Thus, $T = \sum_{i=1}^{n} X_i^2 \sim Gamma(\frac{n}{2}, \frac{1}{2})$. This distribution is also known as χ^2 distribution with degrees of freedom n. Thus, the sum of squares of n *i.i.d.* $N(0, 1)$ RVs has a χ^2 distribution with degrees of freedom n. □

Theorem 1.13. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then $\overline{X}$ and S^2 are independently distributed and*

$$\overline{X} \sim N(\mu, \sigma^2/n) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Proof: Let A be an $n \times n$ orthogonal matrix, whose first row is

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

Note that such a matrix exists as we can start with the row and construct a basis of $\mathbb{R}^n$. Then Gram-Schmidt orthogonalization will give us the required matrix. As A is orthogonal, its' inverse exists and $A^{-1} = A^T$, the transpose of A. Now, consider the transformation of random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ given by

$$\boldsymbol{Y} = A\boldsymbol{X}.$$

First, we shall try to find the distribution of $\boldsymbol{Y}$. Note that the transformation $g(\boldsymbol{x}) = A\boldsymbol{x}$ is a one-to-one transformation as A is invertible. The inverse transformation is given by $\boldsymbol{x} = A'\boldsymbol{y}$. Hence, the Jacobian of the inverse transformation is $J = det(A)$. As A is orthogonal, absolute value of $det(A)$ is one. Now, as $X_1, X_2, \ldots, X_n$ are *i.i.d.* $N(\mu, \sigma^2)$ RVs, the JPDF of $\boldsymbol{X}$, for $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)' \in \mathbb{R}^n$, is

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})\right], \end{aligned}$$

where $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)'$ is a n component vector. Thus, the JPDF of $\boldsymbol{Y}$, for $\boldsymbol{y} \in \mathbb{R}^n$, is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{X}}(A'\boldsymbol{y}) \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(A'\boldsymbol{y} - \boldsymbol{\mu})'(A'\boldsymbol{y} - \boldsymbol{\mu})\right] \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{\eta})'(\boldsymbol{y} - \boldsymbol{\eta})\right], \end{aligned}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)' = A\boldsymbol{\mu}$. Note that $\eta_1 = \sqrt{n}\mu$. Moreover,

$$\boldsymbol{\eta}'\boldsymbol{\eta} = \boldsymbol{\mu}'\boldsymbol{\mu} \implies \sum_{i=1}^{n} \eta_i^2 = n\mu^2 \implies \sum_{i=2}^{n} \eta_i^2 = n\mu^2 - \eta_1^2 = 0.$$

Thus, $\eta_i = 0$ for $i = 2, 3, \ldots, n$. Hence, the JPDF of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2} \left\{ \prod_{i=2}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{y_i^2}{2\sigma^2}} \right\} \quad \text{for } \boldsymbol{y} = (y_1, y_2, \ldots, y_n)' \in \mathbb{R}^n.$$

Therefore, $Y_1, Y_2, \ldots, Y_n$ are independent RVs and $Y_1 \sim N(\sqrt{n}\mu, \sigma^2)$ and $Y_i \sim N(0, \sigma^2)$ for $i = 2, 3, \ldots, n$, where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$. Now,

$$Y_1 = \sqrt{n}\,\overline{X} \implies \sqrt{n}\,\overline{X} \sim N(\sqrt{n}\mu, \sigma^2) \implies \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Again,

$$\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X} \implies \sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = (n-1)S^2.$$

For $i = 2, 3, \ldots, n$, $\frac{Y_i}{\sigma}$ are *i.i.d.* $N(0, 1)$ RVs. Thus, using the previous theorem

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} \left(\frac{Y_i}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

Notice that $\overline{X}$ is a function of Y_1 only, and S^2 is a function of $Y_2, Y_3, \ldots, Y_n$. As Y_i's are independent, $\overline{X}$ and S^2 are independent. □

Definition 1.7 (t-distribution)**.** *A CRV X is said to have a Student's t-distribution (or simply, t-distribution) with n degrees of freedom if the PDF of X is given by*

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{for } t \in \mathbb{R}.$$

We will use the notation $X \sim t_n$ to denote that the RV X has a t-distribution with n degrees of freedom.

Theorem 1.14. *Let $X \sim N(0, 1)$ and $Y \sim \chi^2_n$ be two independent RVs. Then the RV $T = \dfrac{X}{\sqrt{Y/n}} \sim t_n$.*

Proof: This theorem can be proved using the transformation technique 2. Note that the JPDF of X and Y is

$$f_{X,Y}(x,\, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \times \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}} \quad \text{for } x \in \mathbb{R},\, y > 0.$$

Take $V = \sqrt{\frac{Y}{n}}$. Then the inverse mapping is $x = tv$ and $y = nv^2$. The Jacobian of the transformation is

$$J = \begin{vmatrix} v & t \\ 0 & 2nv \end{vmatrix} = 2nv^2 > 0.$$

Thus, the JPDF of T and V is

$$f_{T,V}(t,\, v) = \frac{n^{\frac{n}{2}}}{2^{\frac{n-1}{2}}\sqrt{\pi}\Gamma(\frac{n}{2})} v^n e^{-\frac{1}{2}nv^2\left(1+\frac{t^2}{n}\right)} \quad \text{for } t \in \mathbb{R},\, v > 0.$$

Therefore, for $t \in \mathbb{R}$, the marginal PDF of T is

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,V}(t,\, v)dv \\ &= \frac{n^{\frac{n}{2}}}{2^{\frac{n-1}{2}}\sqrt{n}\Gamma(\frac{n}{2})} \int_0^\infty v^n e^{-\frac{1}{2}nv^2\left(1+\frac{t^2}{n}\right)} dv \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \end{aligned}$$

□

Corollary 1.2. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Then*

$$\sqrt{n}\,\frac{\overline{X}-\mu}{S} \sim t_{n-1},$$

where S is the positive square root of S^2.

Proof: From Theorem 1.13, it is clear that $\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma} \sim N(0,\, 1)$. Therefore,

$$\frac{\sqrt{n}\,\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \sqrt{n}\,\frac{\overline{X}-\mu}{S} \sim t_{n-1}.$$

□

Definition 1.8 (F-distribution)**.** *A CRV X is said to have a F-distribution with n and m degrees of freedom if the PDF of X is given by*

$$f(x) = \frac{1}{B\left(\frac{n}{2},\, \frac{m}{2}\right)} \left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1+\frac{n}{m}x\right)^{-\frac{n+m}{2}} \quad \textit{for } x > 0.$$

We will use the notation $X \sim F_{n,m}$ to denote that the RV X has a F-distribution with n and m degrees of freedom.

Theorem 1.15. *Let $X \sim \chi^2_n$ and $Y \sim \chi^2_m$ are two independent RVs. Then*

$$F = \frac{X/n}{Y/m} = \frac{mX}{nY} \sim F_{n,\,m}.$$

Proof: The JPDF of X and Y is

$$f_{X,\,Y}(x,\,y) = \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} e^{-\frac{1}{2}(x+y)} \quad \text{for } x > 0,\, y > 0.$$

Taking $V = Y$, the inverse transformation is $x = \frac{n}{m} fv$ and $y = v$. The Jacobian of the inverse transformation is

$$J = \begin{vmatrix} \frac{n}{m} v & \frac{n}{m} f \\ 0 & 1 \end{vmatrix} = \frac{n}{m} v > 0.$$

Thus, the JPDF of F and V is

$$f_{F,\,V}(f,\,v) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} f^{\frac{n}{2}-1} v^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(1+\frac{n}{m}f)v} \quad \text{for } f > 0,\, v > 0.$$

Therefore, for $f > 0$, the marginal PDF of F is

$$\begin{aligned} f_F(f) &= \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{m+n}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} f^{\frac{n}{2}-1} \int_0^\infty v^{\frac{m+n}{2}-1} e^{-\frac{1}{2}(1+\frac{n}{m}f)v} dv \\ &= \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{B\left(\frac{n}{2},\,\frac{m}{2}\right)} f^{\frac{n}{2}-1} \left(1 + \frac{n}{m} f\right)^{\frac{n+m}{2}}. \end{aligned}$$

□

Corollary 1.3. *Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$. Also, assume that X_i's and Y_j's are independent. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, $\overline{Y} = \frac{1}{m}\sum_{i=1}^m Y_i$, and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \overline{Y})^2$. Then*

$$\frac{\sigma_2^2 S_X^2}{\sigma_1^2 S_Y^2} \sim F_{n-1,m-1}.$$

Proof: The proof is straight forward from the Theorems 1.15 and 1.13. □

1.4 Modes of Convergence

This section will deal with convergence properties of a sequence of RVs. There are several modes of convergence of sequence of RVs. Here, we will discuss four modes of convergence for a sequence of RVs $\{X_n\}$. These are quite useful concepts in probability. They have applications in different other fields including Statistics.

Definition 1.9 (Almost Sure Convergence)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges almost surely or with probability (w.p.) 1 to a random variable X if*

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = 1.$$

Example 1.20. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability (for any interval $I \subseteq \mathcal{S}$, $P(I) =$ length of I). Define the sequence of RVs by

$$X_n(\omega) = 1_{[0,\frac{1}{n}]}(\omega) \quad \text{for all } n = 1,\, 2,\, 3,\, \ldots.$$

Then X_n converges almost surely to the zero RV. Here, the zero RV means a RV, say X, defined on the same probability space $(\mathcal{S},\, \mathcal{F},\, P)$ such that $X(\omega) = 0$ for all $\omega \in \mathcal{S}$. To see it, notice that for any fixed $\omega \in (0,\, 1]$, we can find an n_0 such that $\frac{1}{n} < \omega$ for all $n \geq n_0$. Thus, $X_n(\omega) \to 0 = X(\omega)$ as $n \to \infty$. Therefore, $\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\} = (0,\, 1]$ and hence,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P\left((0,\, 1]\right) = 1.$$

Thus, $X_n \to 0$ almost surely. ||

Definition 1.10 (Convergence in Probability)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. We say that X_n converges in probability to a random variable X if for any $\epsilon > 0$,*

$$P(|X_n - X| > \epsilon) \to 0 \quad \textit{as } n \to \infty.$$

Example 1.21. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs using $X_n = 1_{[0,\frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same probability space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\, \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

Example 1.22. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs using $X_n = n1_{[0,\frac{1}{n}]}$. Then X_n converges in probability to the zero random variable. Let X denote the zero RV defined on the same probability space. To see it, notice that for any fixed $\epsilon > 0$, $|X_n - X| > \epsilon$ only on the interval $[0,\, \frac{1}{n}]$. Thus,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \implies \lim_{n\to\infty} P\left(|X_n - X| > \epsilon\right) = 0.$$

Therefore, $X_n \to X$ in probability. ||

It may seem that convergence almost surely and convergence in probability are equivalent. However, this is not true, as the following example shows.

Example 1.23. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n},\, \frac{m}{2^n}]} \quad \text{for } m = 1,\, 2,\, \ldots,\, 2^n;\; n = 1,\, 2,\, 3,\, \ldots.$$

Note that $X_{1,1} = 1_{[0,1/2]}, X_{2,1} = 1_{[1/2,1]}, X_{1,2} = 1_{[0,1/4]}, X_{2,2} = 1_{[1/4,1/2]}, X_{3,2} = 1_{[1/2,3/4]}, X_{4,2} = 1_{[3/4,1]}$ and so on. This sequence of RVs $\{X_{m,n}\}$ can be visualized as follows (see Figure 1.3). We start with the interval $[0,\, 1]$. First, we divide the interval into two equal parts, $[0,\, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. The first RV $X_{1,1}$ is 1 on the first part and 0 on the second part. The second RV $X_{2,1}$ is 1 on the second part and 0 on the first part. Then, we divide the interval into 2^2 equal

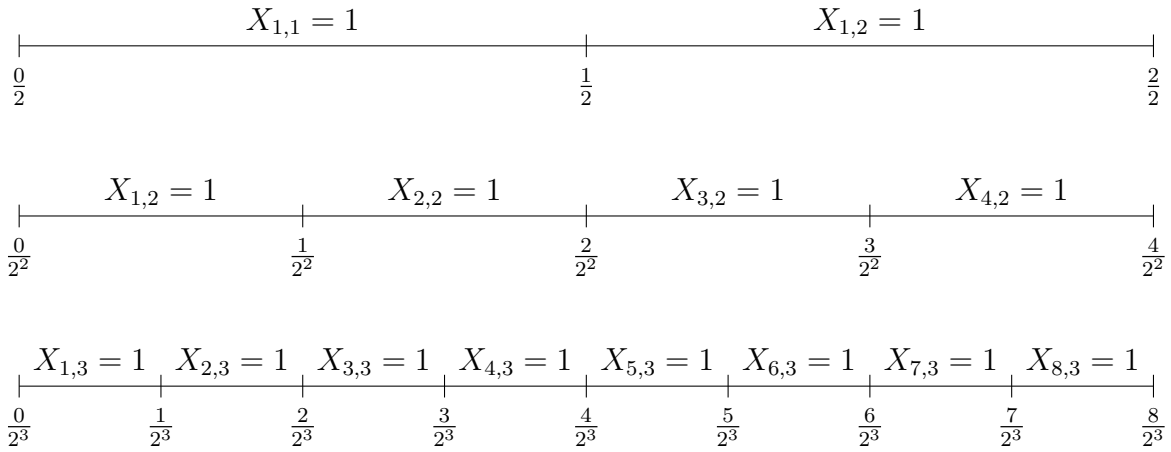


Figure 1.3: Figure for Example 1.23

parts, *viz.*, $[0, \frac{1}{2^2}]$, $[\frac{1}{2^2}, \frac{2}{2^2}]$, $[\frac{2}{2^2}, \frac{3}{2^2}]$, and $[\frac{3}{2^2}, 1]$. Now, the third RV $X_{1,2}$ is 1 on the first part $[1, \frac{1}{4}]$ and 0 otherwise. The fourth RV $X_{2,2}$ is 1 on the second part $[\frac{1}{4}, \frac{1}{2}]$ and 0 otherwise. The fifth RV $X_{3,2}$ equals 1 on the third part $[\frac{1}{2}, \frac{3}{4}]$ and 0 otherwise. Finally, the sixth RV $X_{4,2}$ is 1 on the fourth part $[\frac{3}{4}, 1]$ and 0 otherwise. Next, we divide the interval $[0, 1]$ into 2^3 equal parts and define the next 8 RVs in the similar manner. This procedure continues.

Let us assume that X be a RV defined on the same probability space and $X = 0$. Then, for any $\epsilon > 0$,

$$P\left(|X_{m,n} - X| > \epsilon\right) = \frac{1}{2^n} \implies \lim_{n\to\infty} P\left(|X_{m,n} - X| > \epsilon\right) = 0.$$

Therefore, $X_{m,n} \to X$ in probability. However, for any fixed $\omega \in \mathcal{S}$, there exists a subsequence of the sequence of real numbers $\{X_{m,n}(\omega)\}$ that converges to one and another subsequence that converges to zero. Therefore, $\{X_{m,n}(\omega)\}$ does not converge for all $\omega \in \mathcal{S}$. Thus,

$$P\left(\{\omega \in \mathcal{S} : X_{m,n} \text{ converges}\}\right) = P\left(\emptyset\right) = 0.$$

This shows that $X_{m,n}$ do not converge to any RV almost surely. This example shows that a sequence of RVs, which converges in probability, may not converge almost surely. ||

Definition 1.11 (Convergence in r^{th} Mean)**.** *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. For $r = 1, 2, 3, \ldots$, we say that X_n converges in r^{th} mean to a random variable X if*

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

Example 1.24. Let $\mathcal{S} = [0, 1], \mathcal{F} = \mathcal{B}([0, 1])$, and P be a uniform measure. Define $X_n = 1_{[0,\frac{1}{n}]}$. Then X_n converges in 1st mean to the zero random variable. To see it, notice that

$$E|X_n - X| = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

where X is a zero RV defined on the same probability space. ||

Definition 1.12 (Convergence in Distribution)**.** *Let $\{X_n\}$ be a sequence of RVs and X be a RV. Let $F_n(\cdot)$ and $F(\cdot)$ denote the CDF of X_n and X, respectively. We say that X_n converges in distribution to a random variable X if*

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

for all x where F is continuous.

Unlike the first three modes of convergence, here X_n's can be defined on different probability spaces. We are only interested if the sequence of CDFs converges to a CDF. This flexibility makes this mode of convergence very useful.

Example 1.25. Suppose X_n's are random variables such that $P(X_n = \frac{1}{n}) = 1$. Then, the CDF of X_n is

$$F_n(x) = \begin{cases} 0 & \text{if } x < \frac{1}{n} \\ 1 & \text{if } x \geq \frac{1}{n}, \end{cases}$$

which converges pointwise to the function

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

for all $x \neq 0$, which is the point of discontinuity of the function $F(\cdot)$. Now, $F(\cdot)$ is the CDF of the RV X, which takes value 0 with probability one. Therefore, X_n converges in distribution to the zero RV. ||

The following theorems states the relation between different modes of convergence.

Theorem 1.16. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ almost surely.*

Proof: This prove is skipped here. □

Theorem 1.17. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in probability if $X_n \to X$ in rth mean for any $r = 1, 2, 3, \ldots$.*

Proof: Let $X_n \to X$ in rth mean. Then, using Markov inequality, for any $\epsilon > 0$,

$$P\left(|X_n - X| > \epsilon\right) \leq \frac{E|X_n - X|^r}{\epsilon^r} \to 0 \quad \text{as } n \to \infty.$$

As probability of an event is always non-negative,

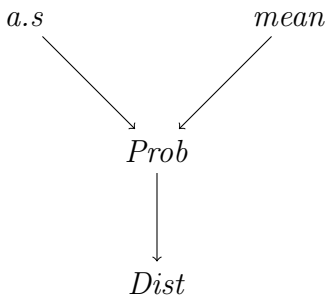
$$P\left(|X_n - X| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_n \to X$ in probability. □

Theorem 1.18. *Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Let X be a random variable defined on the same probability space $(\mathcal{S}, \mathcal{F}, P)$. Then $X_n \to X$ in distribution if $X_n \to X$ in probability.*

Proof: The proof is skipped here. □

The following figure depicts the relationship between several modes of convergence pictorially. Note that the arrows are one-sided. What about other sides? Moreover, there is no arrows between almost sure convergence and rth mean convergence. The following examples show that in general one mode of convergence does not imply other, whenever there is no directed arrows in the above figure. The Example 1.23 shows that probability convergence does not imply almost sure convergence.



Example 1.26. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$, and P be a uniform probability. Define the sequence of RVs by

$$X_{m,n} = 1_{[\frac{m-1}{2^n}, \frac{m}{2^n}]} \quad \text{for } m = 1, 2, \ldots, 2^n; n = 1, 2, 3, \ldots.$$

Then

$$E|X_{m,n}| = \frac{1}{2^n} \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_{m,n} \to X = 0$ in 1st mean. However, in Example 1.23, we have seen that $X_{m,n}$ does not convergence almost surely. This example shows that rth mean convergence does not imply almost sure convergence. ||

Example 1.27. Let $\mathcal{S} = [0,1], \mathcal{F} = \mathcal{B}([0,1])$ and P be a uniform probability. Define $X_n = n1_{[0,\frac{1}{n}]}$. Now, taking $X = 0$,

$$P\left(|X_n - X| > \epsilon\right) = \frac{1}{n} \to 0 \quad \text{as } n \to \infty,$$

for any $\epsilon > 0$. Thus, $X_n \to X$ in probability. Using the logic used in Example 1.20,

$$P\left(\{\omega \in \mathcal{S} : X_n(\omega) \to X(\omega)\}\right) = P((0,\,1]) = 1.$$

Thus, $X_n \to X$ almost surely. However, X_n does not converge to X in rth mean. To see it, notice that

$$E|X_n - X|^r = n^{r-1} \to \begin{cases} 1 & \text{if } r = 1 \\ \infty & \text{if } r > 1. \end{cases}$$

This example shows that probability convergence or almost sure convergence do not imply rth mean convergence. ||

Example 1.28. Let X be a $N(0,1)$ RV defined on some probability space $(\mathcal{S}, \mathcal{F}, P)$. Define $X_n = X$ for all n. Notice that the CDFs of X_n are same for all $n = 1, 2, \ldots$ and is given by $\Phi(\cdot)$. Moreover, the CDFs of X and $-X$ are also $\Phi(\cdot)$. Thus, X_n converges in distribution to $-X$. However, X_n does not converge to $-X$ in probability. To see it, we can proceed as follows: for $\epsilon > 0$,

$$P\left(|X_n + X| \le \epsilon\right) = P\left(2|X| \le \epsilon\right) = 2\Phi\left(\frac{\epsilon}{2}\right) - 1 \neq 1.$$

This example shows that distribution convergence does not imply probability convergence, even if the random variables are defined on the same probability space. ||

Theorem 1.19. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space and X_n converges in distribution to some constant c, then X_n also converges in probability to c.*

Proof: As X_n converges to a constant c,

$$F_n(x) \to F(x) = \begin{cases} 0 & \text{if } x < c \\ 1 & \text{if } x \geq c \end{cases}$$

as $n \to \infty$. Now, fix $\varepsilon > 0$. Then,

$$\begin{aligned} 0 \leq P\left(|X_n - c| > \varepsilon\right) &= P\left(X_n > c + \varepsilon\right) + P\left(X_n < c - \varepsilon\right) \\ &\leq 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon) \to 1 - 1 + 0 = 0 \end{aligned}$$

as $n \to \infty$. Note that as $c + \varepsilon > c$ and $c - \varepsilon < c$, $F_n(c + \varepsilon) \to 1$ and $F_n(c - \varepsilon) \to 0$. Thus, $X_n \to c$ in probability. □

Corollary 1.4. *Suppose $\{X_n\}$ is a sequence of RVs defined on a probability space. Then, $X_n \to c$ in distribution if and only if $X_n \to c$ in probability, where c is a constant.*

Proof: The proof of the corollary is straight forward by combining the previous theorem and Theorem 1.18. □

The following theorems provide several properties of different modes of convergence. The proof of the theorems are skipped here.

Theorem 1.20. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ w. p. 1 and $Y_n \to Y$ w. p. 1. Then*

- *$X_n + Y_n \to X + Y$ w. p. 1.*
- *$X_n Y_n \to XY$ w. p. 1.*
- *$f(X_n) \to f(X)$ w. p. 1, for any f continuous.*

Theorem 1.21. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in probability and $Y_n \to Y$ in probability. Then*

- *$X_n + Y_n \to X + Y$ in probability.*
- *$X_n Y_n \to XY$ in probability.*
- *$f(X_n) \to f(X)$ in probability, for any f continuous.*

Theorem 1.22. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$.*

- *If $X_n \to X$ in r^{th} mean and $Y_n \to Y$ in r^{th} mean, then $X_n + Y_n \to X + Y$ in r^{th} mean.*
- *If $X_n \to X$ in r^{th} mean then $f(X_n) \to f(X)$ in r^{th} mean, for any f bounded continuous.*

Theorem 1.23. *Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Suppose $X_n \to X$ in distribution and $Y_n \to c$ in probability for some constant c. Then*

- $X_n + Y_n \to X + c$ *in distribution.*
- $X_n Y_n \to cX$ *in distribution.*
- $f(X_n) \to f(X)$ *in distribution, for any* f *continuous.*

Example 1.29. Let $X, Y \sim N(0, 1)$ and X and Y be independent RVs. Take $X_n = X$ and $Y_n = Y$ for all $n = 1, 2, 3, \ldots$. Then, $X_n \to X$ in distribution and $Y_n \to X$ in distribution. Now, $X_n + Y_n = X + Y \sim N(0, 2)$ and $2X \sim N(0, 4)$. Thus, $X_n + Y_n$ does not converges to $2X$ in distribution. This example shows that $X_n + Y_n$ may not converge to $X + Y$ in distribution if $X_n \to X$ in distribution and $Y_n \to Y$ in distribution. You can easily check that the same conclusion is also true for product. ||

Theorem 1.24. *Let X_n be a RV with MGF $M_n(t)$ for $n = 1, 2, 3, \ldots$. Let X be a RV with MGF $M(t)$. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, as $n \to \infty$, then $X_n \to X$ in distribution.*

Theorem 1.25. *Let X_n be a DRV with PMF $f_n(\cdot)$ for $n = 1, 2, 3, \ldots$. Let X be a DRV with PMF $f(\cdot)$. If, for all $x \in \mathbb{R}$, $f_n(x) \to f(x)$ as $n \to \infty$, then $X_n \to X$ in distribution.*

Theorem 1.26. *Let X_n be a CRV with PDF $f_n(\cdot)$ for $n = 1, 2, 3, \ldots$. Let X be a CRV with PDF $f(\cdot)$. If, for all $x \in \mathbb{R}$, $f_n(x) \to f(x)$ as $n \to \infty$, then $X_n \to X$ in distribution.*

Example 1.30. Let $X_n \sim Bin(n, p_n)$, where $p_n \to 0$ and $np_n = \lambda\,(> 0)$. Then, for $n = 1, 2, 3, \ldots$, the MGF of X_n is

$$M_n(t) = \left(1 - p_n + p_n e^t\right)^n = \left(1 + \frac{\lambda}{n}\left(e^t - 1\right)\right)^n \to e^{\lambda\left(e^t - 1\right)}$$

for all $t \in \mathbb{R}$. Note that if $X \sim Poi(\lambda)$, then the MGF of X is

$$M(t) = e^{\lambda\left(e^t - 1\right)} \quad \text{for } t \in \mathbb{R}.$$

Thus, $X_n \to X$ in distribution.

This example tells us the motivation behind the Poisson distribution. We can use Poisson distribution to approximate the probability of a Binomial distribution when probability of success is very small and number of trials is very large. ||

Example 1.31. Under the conditions of the previous example, we can prove that $X_n \to X$ in distribution using Theorem 1.25. To see it, we can proceed as follows.

$$\begin{aligned} P\left(X_n = k\right) &= \binom{n}{k} p_n^k \left(1 - p_n\right)^{n-k} \\ &= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \times \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &\to \frac{e^{-\lambda}\lambda^k}{k!}. \end{aligned}$$

Notice that the support of X_n is the set $\{0, 1, 2, \ldots, n\}$. When $n \to \infty$, the support becomes $\{0, 1, 2, \ldots\}$. ||

Example 1.32. Let $X_n \sim U(0, 1 + 1/n)$ for $n = 1, 2, 3, \ldots$. Then the PDF of X_n is

$$f_n(x) = \begin{cases} \frac{1}{1+\frac{1}{n}} & \text{if } 0 < x < 1 + \frac{1}{n} \\ 0 & \text{otherwise} \end{cases} \longrightarrow f(x) = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

which is the PDF of a RV $X \sim U(0, 1)$. Thus, $X_n \to X$ in distribution. ||

1.5 Limit Theorems

In this section, we will discuss two very famous and useful theorems, *viz.*, strong law of large numbers (SLLN) and central limit theorem (CLT). We will skip the proofs, but we will see some applications.

Theorem 1.27 (Strong Law of Large Numbers)**.** *Let $\{X_n\}$ be a sequence of i.i.d. RVs with finite mean μ. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\{\overline{X}_n\}$ converges to μ almost surely.*

Proof: The proof is skipped. □

Let us loosely discuss the intuitive idea of the previous theorem. Suppose that we want to find the average height of all Indians. Ideally, we need to go to each and every Indian and record their height. Finally, the average should be calculated based on the observations on height. This average is called population average or population mean. It is a very costly (in terms of money and time) process. Alternatively, we can take a representative sample of the Indian population. Here, sample represents a subset of original population. Then, we can collect the height data for each and every person in the sample and then calculate the mean of those sample observations. This mean is called sample mean. If the number of persons in the sample is very small (say, 5 or 10), the calculated sample mean may not be close to the original population mean. However, if we keep on increasing the sample size (the number of persons in the sample), the sample mean should get closer to population mean. The above theorem provided theoretical justification of this intuitive idea. Note that μ and $\overline{X}$ are population and sample means, respectively. Thus, loosely speaking, the SLLN states that sample mean converges to population mean almost surely as we increase the sample size.

Example 1.33 (Bernoulli proportion converges to success probability)**.** Suppose that a sequence of independent trials is performed. Let E be a fixed event. Letting

$$X_i = \begin{cases} 1 & \text{if } E \text{ occurs on the } i\text{th trial} \\ 0 & \text{if } E \text{ does not occur on the } i\text{th trial,} \end{cases}$$

we have by the SLLN that, with probability one,

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \to \mu = E(X_1) = P(E).$$

Since, $X_1 + X_2 + \ldots + X_n$ represents the number of times that the event E occurs in the first n trials, we may interpret it as stating that, with probability one, the limiting proportion of time that the event E occurs is $P(E)$. ||

Example 1.34 (Monte Carlo Integration)**.** Suppose that we want to integrate

$$I = \int_a^b h(x)dx.$$

If we cannot do it explicitly, we can use numerical technique like Simpson's $\frac{1}{3}$rd rule. Here, we will see another technique based on the SLLN. Suppose that a and b are finite real numbers. Note that the above integration can be rewritten as

$$I = (b-a)\int_a^b h(x)\frac{1}{b-a}dx = (b-a)E\left(Y\right),$$

where $Y = h(X)$ and $X \sim U(a,\, b)$. Let $\{X_n\}$ be a sequence of *i.i.d.* RVs with common distribution $U(a,\, b)$ and assume that $Y_n = h(X_n)$ for $n = 1,\, 2,\, 3,\, \ldots$. Now, SLLN says that, with probability one,

$$\overline{Y}_n = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \to E(Y) = \frac{I}{b-a} \implies \frac{b-a}{n}\sum_{i=1}^{n} h(X_i) \to I.$$

Thus, we can generate N random numbers from $U(a,\, b)$ and then, the integration I can be approximated by $\frac{b-a}{N}\sum_{i=1}^{N} h(X_i)$. Here, N is a large integer (the popular choices are 5000 or 10000). The generation from $U(a,\, b)$ can be done using any standard software like R, MATLAB, python etc. ||

Theorem 1.28 (Central Limit Theorem)**.** *Let $\{X_n\}$ be a sequence of i.i.d. RVs with mean μ and variance $\sigma^2 < \infty$. Then, as $n \to \infty$,*

$$P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le a\right) \to \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

for all $a \in \mathbb{R}$.

Proof: The proof is skipped. □

The central limit theorem (CLT) says that

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \to Z \sim N(0,\, 1) \quad \text{in distribution.}$$

Thus, the CDF of standardized sample mean can be approximated (for large sample size) using the CDF of a standard normal distribution, whenever X_n's are i.i.d. RVs with finite mean μ and finite variance σ^2. In other words, the CDF of sample mean can be approximated using the CDF of a $N(\mu,\, \frac{\sigma^2}{n})$ distribution. Note that CLT holds true for any distribution of X_n as long as the variance is finite.

Example 1.35 (Normal Approximation to the Binomial)**.** Let $X_n \sim Bin(n,\, p)$. Then

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

We will use CLT to prove this statement. Let $\{Y_n\}$ be a sequence of i.i.d. RVs where $Y_1 \sim Bernoulli(p)$. Then, we know that

$$\sum_{i=1}^{n} Y_i \stackrel{d}{=} X_n \implies \overline{Y}_n \stackrel{d}{=} \frac{X_n}{n}.$$

Now, $E(Y_n) = p$ and $Var(Y_n) = p(1-p)$ for all $n = 1, 2, 3, \ldots$. Thus,

$$P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq a\right) = P\left(\sqrt{n}\frac{\overline{Y}_n - p}{\sqrt{p(1-p)}} \leq a\right) \to \Phi(a) \quad \text{as } n \to \infty.$$

The equality in the above line is due to the fact that $\overline{Y}_n$ and $\frac{X_n}{n}$ have same distribution. The convergence is due to the CLT. ||

Example 1.36. The lifetimes of a special type of battery is a RV with mean 40 hours and standard deviation 20 hours. A battery is used until it fails, at which point it is replaced by a new one. Assume a stockpile of 25 such batteries, the lifetimes of which are independent, we want to approximate the probability that over 1100 hours of use can be obtained. Let X_i denote the lifetime of the ith battery to be put in use. Then, we are interested in

$$p = P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right),$$

which can be approximated as follows:

$$\begin{aligned}
p &= P\left(X_1 + X_2 + \ldots + X_{25} > 1100\right) \\
&= P\left(\overline{X}_{25} > 44\right) \\
&= P\left(\sqrt{25}\,\frac{\overline{X}_{25} - 40}{20} > \sqrt{25}\,\frac{44 - 40}{20}\right) \\
&\approx P\left(Z > 1\right), \text{ where } Z \sim N(0,\, 1). \text{ This is due to CLT.} \\
&= 1 - \Phi(1) \approx 0.1587,
\end{aligned}$$

as $\Phi(1) \approx 0.8413$. This values can be found from the normal table. ||

Chapter 2

Point Estimation

2.1 Introduction to Statistical Inference

Statistical tools are very popular and useful in almost all fields of study. Whenever we need to analyze data, we can use statistical tools. Now-a-days, statistical tools are used in news, exit poll of an election, sports, science and technology, social sciences, to mention a few. In this course, we will try to learn some basic statistical tools.

In a typical statistical problem, our aim is to find information regarding numerical characteristic(s) of a collection of items/persons/products. This collection is called *population*. For example, I may want to know the average height of Indian citizens. Ideally, I should reach each and every citizen and measure their heights. However, it is a very costly (in terms of money and time) procedure. Likewise, it is not possible to enumerate each and every individual in the population due to cost constrain in most of the situations, though it is possible in principle. In some other cases, it is not possible, even in principle, to enumerate each and every item in the population. For example, suppose that a company wants to find the average lifetime of an electronic item manufactured by the company. To calculate average lifetime, we need lifetime of each and every item. The lifetime of an item is only known if the item fails. Therefore, to have the lifetime of each item, we need to put all the items on a life test and wait for their failure. This will be complete disaster for the company as they do not have any item to sell after the experiment.

One approach to address these issues is to take a subset of the population based on which we try to find out the value of the numerical characteristic. Obviously, it will not be exact, and hence, it is an estimate. This subset is called a *sample*. We should choose the sample such that it will be a good representative of the population. Otherwise the estimate may not be close (in some sense) to the original value, which we do not know. There are different ways of selecting sample from a population. We will not discuss this issue here. We will consider one such sample which is called *random sample* (definition will be given).

As different elements of a population may have different values of the numerical characteristic under study, we will model it with a random variable and the uncertainty using a probability distribution. Let X be a random variable (either discrete or continuous random variable), which denotes the numerical characteristic under consideration. Our job is to find the probability distribution. Note that once the probability distribution is determined, the numerical summary of the distribution can be found. The numerical summary includes mean or expectation, variance, median, etc. Now, there are two possibilities:

1. X has a CDF F with known functional form except perhaps some parameters. Here

our aim is to (educated) guess value of the parameters. For example, in some case we may have $X \sim N(\mu, \sigma^2)$, where the functional form of the PDF is known, but the parameters μ and/or σ^2 may be unknown. In this case, we need to find value of the unknown parameters based on a sample. This is known as *parametric inference*. In this course, we will mainly consider parametric inference.

2. X has a CDF F who's functional form is unknown. This is known as *nonparametric inference*. We will not discuss nonparametric inference in this course.

2.2 Parametric Inference

In the standard framework of parametric inference, we start with a data, say $(x_1, x_2, \ldots, x_n)$. Each x_i is an observation on the numerical characteristic under study. There are n observations and n is fixed, pre-assigned, and known positive integer. Our job is to identify (based on a data) the CDF (or equivalently PMF/PDF) of the RV X, which denote the numerical characteristic in the population.

Definition 2.1 (Random Sample)**.** *The random variables $X_1, X_2, \ldots, X_n$ is said to be a random sample (RS) of size n from the population F if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with marginal CDF F. If F has a PMF/PDF f, we will write that $X_1, \ldots, X_n$ is a RS from a PMF/PDF f.*

In practice, we have a data. A natural question is: How to model a data using RS? Notice that the first observation in the sample can be one of the member of the population. For example, if we take a sample of size 200 from the population of Indian citizen, the first height in the sample corresponds to of one of the citizen of India. Thus, a particular observation is one of the realizations from the whole population. Therefore, it can be seen as a realization of a random variable. Let X_i denote the ith observation for $i = 1, 2, \ldots, n$, where n is the sample size. Then, a meaningful assumption is that each X_i has same CDF F, as X_i is a copy of X. Now, if we can ensure that the observation are taken such a way that the value of one does not effect the others, then we can assume that $X_1, X_2, \ldots, X_n$ are independent. Thus, a RS can be used to model the situation.

Note that JCDF of a RS $X_1, \ldots, X_n$ is

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

Similarly, JPMF/JPDF of a RS $X_1, \ldots, X_n$ from PMF/PDF f is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

In a typical problem of parametric inference, we further assume that the functional form of the CDF/PMF/PDF of RV X is known, but the CDF/PMF/PDF involves unknown but fixed real or vector valued parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. Thus, if the value of $\boldsymbol{\theta}$ is known, the stochastic properties of the numerical characteristic is completely known. Therefore, our aim is to find the value of $\boldsymbol{\theta}$ or a function of $\boldsymbol{\theta}$. We also assume that the possible values of $\boldsymbol{\theta}$ belong to a set Θ, which is called *parametric space*. Here, $\boldsymbol{\theta}$ is an indexing or a labelling parameter. We say that $\boldsymbol{\theta}$ is an *indexing or a labelling parameter* if the CDF/PMF/PDF is

uniquely specified by $\boldsymbol{\theta}$. That means that $F(x, \boldsymbol{\theta}_1) = F(x, \boldsymbol{\theta}_2)$ for all $x \in \mathbb{R}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, where $F(\cdot, \boldsymbol{\theta})$ is the CDF of X.

As discussed, our main aim is to identify the CDF/PMF/PDF of the RV X based on a RS. In other words, we want to identify which member of the family $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ can represent the CDF of X, which is equivalent to decide the value of $\boldsymbol{\theta}$ in Θ based on a realization of a RS. Note that, as we know the functional form of the CDF of X, the value of $\boldsymbol{\theta} \in \Theta$ completely specifies the member in $\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. Here, it is assumed implicitly that the data has information regarding the unknown parameter. Though we have not quantify the information yet, we will see it in the Section 2.5.

Example 2.1. Suppose that 100 seeds of a brand were planted one in each pot and let X_i equals one or zero according as the seed in the ith pot germinates or not. The data consists of $(x_1, x_2, \ldots, x_{100})$, where each x_i is either one or zero. The data is regarded as a realization of $(X_1, X_2, \ldots, X_{100})$, where the RVs are *i.i.d.* with $P(X_i = 1) = \theta = 1 - P(X_i = 0)$. Here, θ is the probability that a seed germinates, therefore the natural parametric space is $\Theta = [0, 1]$. The objective is to estimate the value of θ or a function $\psi(\theta)$. For example, consider $\psi_1(\theta) = \binom{10}{8}\theta^8(1-\theta)^2$, which is the probability that in a batch of 10 seeds, exactly 8 seeds will germinate. Another function of interest could be

$$\psi_2(\theta) = \begin{cases} 1 & \text{if } \theta \geq 0.90 \\ 0 & \text{otherwise.} \end{cases}$$

This corresponds to the situation when the brand would be recommended to the farmer provided the probability of germination of a seed is at least 0.90. Thus, if the estimate of $\psi_2(\theta)$ is one, the brand is recommended.

It is easy to see that θ is an indexing parameter. Suppose that $F(x, \theta_1) = F(x, \theta_2)$ for all $x \in \mathbb{R}$. In particular, take $x = \frac{1}{2}$. Then

$$F\left(\frac{1}{2}, \theta_1\right) = F\left(\frac{1}{2}, \theta_2\right) \implies 1 - \theta_1 = 1 - \theta_2 \implies \theta_1 = \theta_2. \qquad ||$$

Example 2.2. Consider determination of gravitational constant g. A standard way to estimate g is to use the pendulum experiment and use the formula

$$g = \frac{2\pi^2 l}{T^2},$$

where l is the length of the pendulum and T is the time required for a fixed number of oscillations. However, due to various reasons including the skill of the experimenter, calibration of the measuring instruments, a variation is observed in the calculated values of g using the previous formula. Let the repeated experiments are performed and the calculated values of g are $X_1, X_2, \ldots, X_n$. In this case, we can use the model $X_i = g + \epsilon_i$, where ϵ_i is the random error. Assuming the errors are normally distributed with mean zero and variance σ^2, $X_i \sim N(g, \sigma^2)$, and the parameter is $\boldsymbol{\theta} = (g, \sigma^2)$ with parametric space $\Theta = \mathbb{R} \times \mathbb{R}^+$. Here, we may be interested to estimate g or σ^2. Note that σ^2 represents the skill of the experimenter. If the experimenter's skill is not up to the mark, variation will be high, and hence, σ^2 will be large.

In this case, it can be shown that $\boldsymbol{\theta}$ is an indexing parameter. Let $\Phi(\cdot, \mu, \sigma^2)$ denote the CDF of $N(\mu, \sigma^2)$ distribution and $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1^2)$ and $\boldsymbol{\theta}_2 = (\mu_1, \sigma_2^2)$. Now, consider

$$\Phi(x, \mu_1, \sigma_1^2) = \Phi(x, \mu_2, \sigma_2^2) \quad \text{for all } x \in \mathbb{R}.$$

Then the corresponding MGFs will be same for all $t \in \mathbb{R}$. Thus,

$$\begin{aligned}
&\exp\left(\mu_1 t + \frac{1}{2}\sigma_1^2 t^2\right) = \exp\left(\mu_2 t + \frac{1}{2}\sigma_2^2 t^2\right) \quad \text{for all } t \in \mathbb{R} \\
&\implies \mu_1 t + \frac{1}{2}\sigma_1^2 t^2 = \mu_2 t + \frac{1}{2}\sigma_2^2 t^2 \quad \text{for all } t \in \mathbb{R} \\
&\implies \mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2.
\end{aligned}$$

||

Example 2.3. Suppose that we are interested to estimate the average height of a large community of people. Let we assume that $N(\mu, \sigma^2)$ distribution is a plausible distribution. We know that the natural parametric space for a normal distribution is $\Theta = \mathbb{R} \times \mathbb{R}^+$. However, as the average of heights of persons is always a positive real number, it is realistic to assume that $\mu > 0$. Hence, a better choice of Θ is $\mathbb{R}^+ \times \mathbb{R}^+$ in the current situation. Thus, we may need to choose the parametric space based on the background of the problem. ||

Example 2.4. Consider a series system with two components. A series system works if all its components work. Thus, in this case, the system works if both the components work. Let Z and Y denote the lifetimes of the first and second components, respectively. Also, assume that Z and Y are independent exponential RVs with rate θ and λ, respectively. However, we cannot observe both Z and Y, but we can observe $X = \min\{Z, Y\}$. It is easy to see that X follows an exponential distribution with rate $\theta + \lambda$. Clearly, $\alpha = \theta + \lambda$ is an indexing parameter. However, (θ, λ) is not an indexing parameter as, for example, $\theta = 1$, $\lambda = 3$ and $\theta = 2$, $\lambda = 2$ would give rise to the same distribution of X.

This example shows that there are practical situations, where the way data arises leads to a parameter that is not an indexing parameter. This issue is referred as the problem of non-identifiability. We will not consider the problem of non-identifiability in the course and mainly concentrate on the cases in which the data arises from a probability distribution with real or vector valued indexing parameter. ||

Definition 2.2 (Statistic)**.** *Let $X_1, \ldots, X_n$ be a RS. Let $T(x_1, \ldots, x_n)$ be a real-valued function having domain that includes the sample space, χ^n, of $X_1, X_2, \ldots, X_n$. Then the RV $\boldsymbol{Y} = T(X_1, \ldots, X_n)$ is called a statistic if it is not a function of unknown parameters.*

Note that our aim is to find a guess value of unknown parameters based on a RS. Hence, we are considering a function of RS. If the function involve any unknown parameters, we will not be able to compute the value of the function given a realization of a RS. Hence, the function that involves unknown parameters is of no use in this respect. Therefore, we define a statistic as a function of RS, but statistic should not involve an unknown parameter. Note that the distribution of a statistic may depend on unknown parameters.

Example 2.5. Let $X_1, \ldots, X_n$ be a RS from a $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ are examples of statistic. However, $\frac{\overline{X}-\mu}{\sigma}$ is not a statistic. Note that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$. Clearly, the distribution of $\overline{X}$ depends on the unknown parameters. ||

Definition 2.3 (Point Estimator and Estimate)**.** *In the context of estimation, a statistic is called a point estimator (or simply estimator). A realization of a point estimator is called an estimate.*

In the above definition of an estimator, we do not mention about the parameter that is to be estimated and its parametric space. However, in practice, we need to take care of the parameter to be estimated and its parametric space. For example, to estimate population variance, we should not use an estimator that can be negative.

There are several methods to find an estimator. We will consider three of them in this course: 1) method of moment estimator (MME), 2) maximum likelihood estimator (MLE) and 3) least square estimator (LSE). We will study the first two methods in this chapter and the third method will be discussed when we will study regression.

Before discussing the methods of estimation, we will study sufficiency, information, ancillary, and completeness. These are useful concepts for the theory of estimation.

2.3 Sufficient Statistic

Recall that our aim is to estimate unknown parameter $\boldsymbol{\theta}$ based on a realization of a RS using a suitable statistic or estimator. Of course, the RS $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ has all the "information" regarding unknown parameter $\boldsymbol{\theta}$. One should use a statistic that has same amount of "information" that the data have regarding $\boldsymbol{\theta}$. We can take $\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{X}$. However, it is not interesting in most of the situations as one should take a summary of the data that capture all the "information". Therefore, in most of the cases, we will consider a function $\boldsymbol{T} : \chi^n \to \mathbb{R}^m$, where $m < n$. In most of the times, the value of m is much smaller than that of n. Such summary or statistic is as good as the whole data and is called sufficient for $\boldsymbol{\theta}$.

If a quantity vary with the change in another quantity, then there is some information in the first quantity regarding the second. On the other hand, if the first quantity do not change with the second quantity, then the first does not have any information regarding the second. Similarly, if the distribution of a statistic does not involve the unknown parameter $\boldsymbol{\theta}$, then the statistic does not have any information regarding $\boldsymbol{\theta}$. Motivated by this understanding, a sufficient statistic for $\boldsymbol{\theta}$ can be defined as follows.

Definition 2.4 (Sufficient Statistic)**.** *A statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ is called a sufficient statistic for unknown parameter $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{T} = \boldsymbol{t}$ does not include $\boldsymbol{\theta}$ for all $\boldsymbol{t}$ in the support of $\boldsymbol{T}$.*

Thus, given the value $\boldsymbol{t}$ of a sufficient statistic $\boldsymbol{T}$, conditionally there is no information left in $\boldsymbol{X}$ regarding $\boldsymbol{\theta}$. In other words, $\boldsymbol{X}$ is trying to tell us a story regarding $\boldsymbol{\theta}$ and any statistic is a gist of the story. If we have the gist $\boldsymbol{T}$, a sufficient statistic, the original story is redundant as the gist has all the information that the original story has regarding $\boldsymbol{\theta}$. Note that $\boldsymbol{X}$ is a sufficient statistic. However, we are interested in a summary statistic in most of the situations.

Example 2.6. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$, $p \in (0, 1)$. Take $T = \sum_{i=1}^{n} X_i$. We know that $T \sim Bin(n, p)$. Now, for $t = 0, 1, \ldots, n$,

$$
\begin{aligned}
& P(X_1 = x_1, \ldots, X_n = x_n | T = t) \\
&= \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)} \\
&= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

$$= \begin{cases} \frac{\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^{n} x_i = t \\ 0 & \text{otherwise,} \end{cases}$$

which does not include p. Hence, $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for p. ||

We can verify if a statistic is sufficient or not using the definition of sufficient statistic. That means that we first need to guess a correct statistic and then we can use the definition to show that it is actually a sufficient statistic for the unknown parameters. However, the next theorem gives necessary and sufficient conditions, which can be used to find a sufficient statistic. Therefore, the next theorem is very useful.

Theorem 2.1 (Neyman-Fisher Factorization Theorem)**.** *Let $X_1, \ldots, X_n$ be RS with JPMF or JPDF $f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Then $\boldsymbol{T} = \boldsymbol{T}(X_1, \ldots, X_n)$ is sufficient for $\boldsymbol{\theta}$ if and only if*

$$f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta}) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}\left(\boldsymbol{T}(\boldsymbol{x})\right),$$

where $h(\boldsymbol{x})$ does not involve $\boldsymbol{\theta}$, $g_{\boldsymbol{\theta}}(\cdot)$ depends $\boldsymbol{x}$ only through $\boldsymbol{T}(\boldsymbol{x})$.

Proof: We will proof the theorem only for the discrete case.
Only if part: Let us notice that $\{\boldsymbol{X} = \boldsymbol{x}\} \subseteq \{\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x})\}$. Now,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta}) &= P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) \\ &= P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x})) \\ &= P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x}))\, P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x})). \end{aligned}$$

As $\boldsymbol{T}$ is a sufficient statistic, $P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x}))$ does not involve $\boldsymbol{\theta}$. Therefore, we can take $h(\boldsymbol{x}) = P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x}))$. On the other hand, $P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{T}(\boldsymbol{x}))$ is a function of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ only through $\boldsymbol{T}(\boldsymbol{x})$. Thus, $g_{\boldsymbol{\theta}}(t) = P_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{X}) = \boldsymbol{t})$.
If Part: For $\boldsymbol{t}$ in the support of $\boldsymbol{T}$, the conditional PMF of $\boldsymbol{X}$ given $\boldsymbol{T} = \boldsymbol{t}$ is

$$f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}|\boldsymbol{t}) = \frac{P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t})}{P_{\boldsymbol{\theta}}(\boldsymbol{T} = \boldsymbol{t})}.$$

Now, notice that if $\boldsymbol{T}(\boldsymbol{x}) \neq \boldsymbol{t}$, then $P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t}) = 0$, Thus, $f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}|\boldsymbol{t}) = 0$ for $\boldsymbol{T}(\boldsymbol{x}) \neq \boldsymbol{t}$. For $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}$,

$$\begin{aligned} f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}|\boldsymbol{t}) &= \frac{P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{T} = \boldsymbol{t})}{P_{\boldsymbol{\theta}}(\boldsymbol{T} = \boldsymbol{t})} \\ &= \frac{P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})}{\sum_{\{x: \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}\}} P_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})} \\ &= \frac{h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{x}))}{\sum_{\{\boldsymbol{x}: \boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{t}\}} h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{x}))} \end{aligned}$$

$$= \frac{h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{t})}{\sum\limits_{\{\boldsymbol{x}:\boldsymbol{T}(\boldsymbol{x})=\boldsymbol{t}\}} h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{t})}$$
$$= \frac{h(\boldsymbol{x})}{\sum\limits_{\{\boldsymbol{x}:\boldsymbol{T}(\boldsymbol{x})=\boldsymbol{t}\}} h(\boldsymbol{x})},$$

which does not involve the parameter $\boldsymbol{\theta}$. Therefore, $\boldsymbol{T}$ is a sufficient statistic. □

Example 2.7. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Poi(\lambda)$, $\lambda > 0$. Here the JPMF is

$$f(\boldsymbol{x}, \lambda) = \frac{e^{-n\lambda}\lambda^{n\overline{x}}}{\prod_{i=1}^{n}(x_i!)} = h(\boldsymbol{x}) g_{\lambda}\left(T(\boldsymbol{x})\right),$$

where $h(\boldsymbol{x}) = \left[\prod_{i=1}^{n}(x_i!)\right]^{-1}$, $g_{\lambda}(t) = e^{-n\lambda}\lambda^{nt}$, and $T(\boldsymbol{x}) = \bar{x}$. This shows that $T = \overline{X}$ is a sufficient statistic for λ. ||

Example 2.8. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma > 0$. Denoting $\boldsymbol{\theta} = (\mu, \sigma^2)$, the JPDF, for $\boldsymbol{x} \in \mathbb{R}^n$, is

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right] \\ &= h(\boldsymbol{x}) g_{\boldsymbol{\theta}}\left(\boldsymbol{T}(\boldsymbol{x})\right), \end{aligned}$$

where

$$\begin{aligned} h(\boldsymbol{x}) &= \frac{1}{(2\pi)^{\frac{n}{2}}}, \\ \boldsymbol{T}(\boldsymbol{x}) &= \left(\sum_{i=1}^{n}x_i^2, \sum_{i=1}^{n}x_i\right), \\ g_{\boldsymbol{\theta}}(t_1, t_2) &= \frac{1}{\sigma^n}\exp\left[-\frac{1}{2\sigma^2}\left(t_1 - 2\mu t_2 + n\mu^2\right)\right]. \end{aligned}$$

Therefore, using Theorem 2.1, a sufficient statistic for (μ, σ^2) is $(\sum_{i=1}^{n}X_i, \sum_{i=1}^{n}X_i^2)$. ||

Example 2.9. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. Here, the JPDF is

$$f(\boldsymbol{x}, \theta) = \frac{1}{\theta^n} I_{(0,\infty)}(x_{(1)}) I_{(0,\theta)}(x_{(n)}) = h(\boldsymbol{x}) g_{\theta}\left(T(\boldsymbol{x})\right),$$

where $h(\boldsymbol{x}) = I_{(0,\infty)}(x_{(1)})$, $g_{\theta}(t) = \frac{1}{\theta^n} I_{(0,\theta)}(t)$, and $T(\boldsymbol{x}) = x_{(n)}$. Hence, $T = X_{(n)}$ is a sufficient statistic for θ, where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$. ||

Example 2.10. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathbb{R}$. Here, the JPDF is

$$f(\boldsymbol{x}, \theta) = h(\boldsymbol{x}) g_{\theta}(\boldsymbol{T}(\boldsymbol{x})),$$

where $h(x) = 1$, $g_{\theta}(\boldsymbol{t}) = I_{(\theta-1/2,\theta+1/2)}(x_{(1)}) I_{(\theta-1/2,\theta+1/2)}(x_{(n)})$, and $\boldsymbol{T}(\boldsymbol{x}) = (x_{(1)}, x_{(n)})$. Hence, $\boldsymbol{T} = \left(X_{(1)}, X_{(n)}\right)$ is a sufficient for θ, where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. ||

Theorem 2.2. *If $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$, then for any one-to-one function of $\boldsymbol{T}$ is also sufficient for $\boldsymbol{\theta}$.*

Proof: Let $\boldsymbol{S} = \widetilde{g}(\boldsymbol{T})$ be a one-to-one function. Then inverse of $\tilde{g}$ exists and $\boldsymbol{T} = \widetilde{g}^{-1}(\boldsymbol{S})$. Now, using Theorem 2.1,

$$f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta}) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(\boldsymbol{T}(\boldsymbol{x})) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}\left(\widetilde{g}^{-1}(\boldsymbol{S}(\boldsymbol{x}))\right).$$

Here, $h(\boldsymbol{x})$ does not involve $\boldsymbol{\theta}$ and $g_{\boldsymbol{\theta}}\left(\widetilde{g}^{-1}(\boldsymbol{S}(\boldsymbol{x}))\right)$ depends on $\boldsymbol{\theta}$ and $\boldsymbol{x}$ only through $\boldsymbol{S}(\boldsymbol{x})$. Thus, $\boldsymbol{S}$ is a sufficient statistic. □

Example 2.11 (Continuation of Example 2.8)**.** Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma > 0$. We have seen in Example 2.8 that $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is a sufficient statistic for (μ, σ^2). As the mapping $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) \to \left(\overline{X}, S^2\right)$ is one-to-one, using the previous theorem $\left(\overline{X}, S^2\right)$ is sufficient for (μ, σ^2), where $S^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$. ||

Example 2.12. Let $X_1, X_2 \overset{i.i.d.}{\sim} N(\mu, 1)$. Using Theorem 2.1, it is easy to show that $X_1 + X_2$ is a sufficient statistics for μ. Is $T = X_1 + 2X_2$ a sufficient statistics for μ? Answer to the question is negative, T is not a sufficient statistic for μ. Note that it is difficult to use Theorem 2.1 to show that a statistic is not a sufficient.

If a statistic $\boldsymbol{T}$ is sufficient statistic for $\boldsymbol{\theta}$, then the conditional distribution of any other statistic given $\boldsymbol{T} = \boldsymbol{t}$ must be independent of $\boldsymbol{\theta}$. On the other hand, if the conditional distribution of a statistic given $\boldsymbol{T} = \boldsymbol{t}$ involves $\boldsymbol{\theta}$, then the conditional distribution $X_1, X_2, \ldots, X_n$ must depend on $\boldsymbol{\theta}$, and hence, $\boldsymbol{T}$ is not a sufficient statistic for the unknown parameter. Here, we can use this argument to show that $T = X_1 + 2X_2$ is not a sufficient statistic for μ. In fact, we will show that the conditional distribution of X_1 given $X_1 + 2X_2 = t$ involves μ.

Note that $(X_1, X_1 + 2X_2)$ is a bivariate normal random vector with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix Σ, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu \\ 3\mu \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} 1 & 1 \\ 0 & 5 \end{pmatrix}.$$

Therefore, the conditional distribution of X_1 given $X_1 + 2X_2 = t$ is univariate normal with mean $\frac{1}{5}(t + 2\mu)$ and variance $\frac{4}{5}$. Thus, the conditional distribution of X_1 given $X_1 + 2X_2 = t$ involves μ, and hence, $T = X_1 + 2X_2$ is not a sufficient statistic. This example also shows that any function of sufficient statistic is not sufficient as the random sample is itself a sufficient statistic. ||

Remark 2.1. One-dimensional parameter may have multidimensional sufficient statistic. Please revisit the Example 2.10. Moreover, $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ do not imply that the jth component of $\boldsymbol{T}$ is sufficient for the jth component of $\boldsymbol{\theta}$ even if $\boldsymbol{T}$ and $\boldsymbol{\theta}$ are of same dimension. It only tells that $\boldsymbol{T}$ is jointly sufficient for $\boldsymbol{\theta}$. †

2.4 Minimal Sufficiency

In the previous section, we have seen that the Theorem 2.1 can be used to find a sufficient statistic. We have also seen that the RS itself is a sufficient statistic. However, we want to reduce the data by considering an appropriate summary statistic. Of course, we should take a summary which has all "information" that present in the original data. Thus, we want a

shortest summary statistic that has all "information" regarding the parameter $\boldsymbol{\theta}$. Now, a natural question arises: How to define a "shortest" sufficient statistic?

Let $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ be two sufficient statistics. Then we say that $\boldsymbol{T}_2$ represents a further reduction if $\boldsymbol{T}_2$ is a function of $\boldsymbol{T}_1$. Note that any statistic, being a function defined on sample space, say χ^n, of the RS of size n, induces a partition over χ^n. Thus, $\boldsymbol{T}_1$ induces a finer partition over χ^n than that induced by $\boldsymbol{T}_2$. Keeping the above discussion in mind, we have the following definition of minimal sufficient statistic.

Definition 2.5 (Minimal Sufficiency Statistic)**.** *A sufficient statistic $\boldsymbol{T}$ is called minimal sufficient statistic if $\boldsymbol{T}$ is a function of any other sufficient statistic.*

Let a two-dimensional statistic $\boldsymbol{T} = (T_1, T_2)$ be a minimal sufficient statistic for $\boldsymbol{\theta}$. Is it possible to reduce it further? Yes, of course. For example, we may take $S_1 = T_1$ or $S_2 = T_2$, or $S_3 = \frac{1}{2}(T_1 + T_2)$, etc. Now, the next question is: Can the statistic S_1 or S_2 or S_3 individually be sufficient for $\boldsymbol{\theta}$? The answer to the question is no, none of them are sufficient statistic for $\boldsymbol{\theta}$. For example, consider S_1. If possible, assume that S_1 is a sufficient statistic for $\boldsymbol{\theta}$. Then $\boldsymbol{T}$, being a minimal sufficient statistic, must be a function of S_1. However, this is a contradiction, as $\boldsymbol{T}$ cannot be uniquely specified from the value of S_1 alone. Thus, S_1 cannot be a sufficient statistic. A minimal sufficient statistic $\boldsymbol{T}$ cannot be reduced any further to another sufficient statistic. In this sense, minimal sufficient statistic is the shortest and best sufficient statistic. The next theorem provides us a way to find minimal sufficient statistic.

Theorem 2.3. *Let $X_1, X_2, \ldots, X_n$ be a RS from a population with PMF/PDF $f(\cdot, \boldsymbol{\theta})$. Consider*

$$h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})}{\prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})} \quad \text{for } \boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{y} = (y_1, \ldots, y_n) \in \chi^n.$$

Suppose that there is a statistic $\boldsymbol{T}$ such that for any two points $\boldsymbol{x}, \boldsymbol{y} \in \chi^n$, the expression $h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$ does not involve $\boldsymbol{\theta}$ if and only if $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{y})$. Then $\boldsymbol{T}$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.

Proof: The proof is little involved and skipped here. □

Example 2.13 (Contitution of Example 2.6)**.** Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$. We have seen that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic. Is this minimal sufficient statistic? We can answer the question using the previous theorem. Note that here

$$\chi^n = \{(x_1, x_2, \ldots, x_n) : x_i = 0 \text{ or } 1,\ i = 1, 2 \ldots, n\}.$$

Let $\boldsymbol{x}, \boldsymbol{y} \in \chi^n$. Then

$$h(\boldsymbol{x}, \boldsymbol{y}, p) = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}.$$

Hence, $h(\boldsymbol{x}, \boldsymbol{y}, p)$ would become free form p if and only if $\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \implies \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. Thus, using Theorem 2.3, $T = \sum_{i=1}^{n} X_i$ is minimal sufficient statistic for p. ||

Example 2.14 (Continuation of Example 2.8)**.** Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Here, $\chi^n = \mathbb{R}^n$ and $\boldsymbol{\theta} = (\mu, \sigma^2)$. Then a simple calculation shows that

$$h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \exp\left[-\frac{1}{2\sigma^2}\left\{\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) - 2\mu\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right\}\right].$$

Clearly, $h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$ does not involve $\boldsymbol{\theta}$ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$. Therefore, $\boldsymbol{T} = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is a minimal sufficient statistic. ||

Example 2.15 (Continuation of Example 2.9)**.** Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, where $\theta > 0$. Then

$$h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{I_{(0, x_{(n)})}(x_{(1)})}{I_{(0, y_{(n)})}(y_{(1)})} \times \frac{I_{(0,\theta)}(x_{(n)})}{I_{(0,\theta)}(y_{(n)})}.$$

Clearly the first part on the right hand side does not involve θ. We will show that the second part on the right hand side does not involve θ if and only if $x_{(n)} = y_{(n)}$. It is easy to see that the second part does not involve θ if $x_{(n)} = y_{(n)}$. Now, assume that the second part does not involve θ. We claim that $x_{(n)} = y_{(n)}$. If possible, suppose that our claim is not correct. Therefore either $x_{(n)} < y_{(n)}$ or $x_{(n)} > y_{(n)}$. If $x_{(n)} < y_{(n)}$, then for $\theta > y_{(n)}$, the second part is 1, but for $x_{(n)} < \theta < y_{(n)}$, the second part is ∞. Therefore, the second part is not free of θ. This is a contradiction, and hence, $x_{(n)} \not< y_{(n)}$. Similarly, we can work with the case $x_{(n)} > y_{(n)}$ to show $y_{(n)} \not< x_{(n)}$. Thus, $h(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})$ does not involve θ if and only if $x_{(n)} = y_{(n)}$. Therefore, $T = X_{(n)}$ is a minimal sufficient statistic. ||

Example 2.16 (Continuation of Example 2.13)**.** This example shows that the Theorem 2.3 can be used to show that a statistic is not a sufficient statistic. Let us take $n = 3$ in Example 2.13. We have seen that $T = X_1 + X_2 + X_3$ is minimal sufficient statistic for p. Let us consider the statistic $U = X_1X_2 + X_3$. Is U sufficient for p? If possible, assume that U is a sufficient statistic for p. Then T, being a minimal sufficient statistic, must be a function of U. That means that given any observed value of U, the observed value of T can be obtained uniquely. Now, consider the event $\{U = 0\}$. Note that the even $\{U = 0\}$ is union of the following three events.

$$\{X_1 = 0, X_2 = 0, X_3 = 0\},$$
$$\{X_1 = 0, X_2 = 1, X_3 = 0\},$$
$$\{X_1 = 1, X_2 = 0, X_3 = 0\}.$$

It is clear that if we observe $U = 0$, then the observed value of T is either 0 or 1. However, we cannot tell a unique value for T. Thus, T is not a function of U, and hence, U is not a sufficient statistic. ||

2.5 Information

In the previous sections, we have mentioned that we would work with sufficient or minimal sufficient statistic, as they provide reduction of dimension and preserve all "information" that are present in the RS. However, we have not quantify information. We will quantify it in the current section.

Let X be a RV with PMF or PDF $f(\cdot, \theta)$, which depends on a real valued parameter $\theta \in \Theta$. As mentioned, the variation in the PMF or PDF $f(x, \theta)$ with respect to $\theta \in \Theta$ for fixed value of x provides us information about θ. For example, suppose that X has a binomial distribution with PMF

$$f(X = x, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

and let $n = 10$ and $x = 2$. Then we have $f(x, \theta)$ varies with θ as given in Table 2.1. Note that $P(X = 2)$ at $\theta = 0.8$ and 0.9 are given as 0.000 in the above table. However, they are not exactly zero. These probabilities are rounded off to three decimal places. It is this variation that provides some information about θ. If the variation is large, then we have more information about θ. On the other hand if the variation is less, we have less information.

Table 2.1: Variation in PMF with respect to parameter

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$f(2, \theta)$	0.194	0.302	0.233	0.121	0.044	0.011	0.001	0.000	0.000

You may find some resembles with the following situations. Suppose that our job is to identify the place based on a landscape picture given to us. It is known that the place is either Kolkata or Shimla. Now, we can see the picture and look for mountains. If there are mountains, then it is a picture of Shimla, otherwise it is a picture of Kolkata. It is so easy as there is a huge variation in landscape of these two places. You can now think the places are the values of the unknown parameters and pictures are the PMF or PDF of X. On the other hand, if we are asked to identify between Shimla and Kausani based on a landscape picture, it would be very difficult as there is less variation in the landscapes of these two places.

Note that we measure the change in a function with respect to a variable using derivative of the function with respect to the variable. Following it, here we consider $\frac{\partial}{\partial \theta} \ln f(x, \theta)$. However, this partial derivative, in general, depend on x. As we are interested to measure the variation (with respect to x) in change, we can consider the variance of the partial derivative, $Var\left(\frac{\partial}{\partial \theta} \ln f(X, \theta)\right)$. Now, to define information, we need following assumptions, which are called regularity conditions.

1. Let $S_\theta = \{x \in \mathbb{R} : f(x, \theta) > 0\}$ denote the support of the PMF or PDF $f(\cdot, \theta)$ and $S = \cup_{\theta \in \Theta} S_\theta$. Here, we assume that S_θ does not depend on θ, *i.e.*, $S_\theta = S$ for all $\theta \in \Theta$.

2. We also assume that the PDF (or PMF) $f(\cdot, \theta)$ is such that differentiation (with respect to θ) and integration (or sum) (with respect to x) are interchangeable.

Now, assume that X is a CRV. Then

$$E_\theta\left[\frac{\partial}{\partial \theta} \ln f(X, \theta)\right] = \int_S \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) dx = \int_S \frac{\partial f(x, \theta)}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int_S f(x, \theta) dx = 0,$$

as $\int_S f(x, \theta) dx = 1$. Thus,

$$Var\left(\frac{\partial \ln f(X, \theta)}{\partial \theta}\right) = E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial \theta}\right)^2\right].$$

The DRV case can be handled in a similar manner by replacing the integration by a summation sign. This discussion give us the following quantification of information.

Definition 2.6 (Fisher Information)**.** *The Fisher information (or simply information) about parameter θ contained in X is defined by*

$$\mathcal{I}_X(\theta) = E_\theta\left[\left(\frac{\partial \ln f(X,\,\theta)}{\partial\theta}\right)^2\right].$$

Note that $\mathcal{I}_X(\theta) = 0$ if and only if $\frac{\partial}{\partial\theta}\ln f(x,\,\theta) = 0$ with probability one, which means that the PMF or PDF of X does not involve θ. An alternative form of Fisher information can be obtained as follows.

$$\begin{aligned}
&\frac{\partial^2}{\partial\theta^2}\int_S f(x,\,\theta) = 0\\
&\implies \frac{\partial}{\partial\theta}\int_S \frac{\partial \ln f(x,\,\theta)}{\partial\theta} f(x,\,\theta)dx = 0\\
&\implies \int_S \frac{\partial^2 \ln f(x,\,\theta)}{\partial\theta^2} f(x,\,\theta)dx + \int_S \left[\frac{\partial \ln f(x,\,\theta)}{\partial\theta}\right]^2 f(x,\,\theta)dx = 0\\
&\implies \int_S \left[\frac{\partial \ln f(x,\,\theta)}{\partial\theta}\right]^2 f(x,\,\theta)dx = -\int_S \frac{\partial^2 \ln f(x,\,\theta)}{\partial\theta^2} f(x,\,\theta)dx\\
&\implies \mathcal{I}_X(\theta) = -E_\theta\left(\frac{\partial^2 \ln f(X,\,\theta)}{\partial\theta^2}\right).
\end{aligned}$$

Example 2.17. Let $X \sim Poi(\lambda)$, where $\lambda > 0$. The PMF of X is

$$f(x,\,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

for $x = 0,\,1,\,2,\,\ldots$. Therefore,

$$\begin{aligned}
&\ln f(x,\,\lambda) = -\lambda + x\ln\lambda - \ln(x!)\\
&\implies \frac{\partial}{\partial\lambda}\ln f(x,\,\lambda) = -1 + \frac{x}{\lambda}\\
&\implies \mathcal{I}_X(\lambda) = E_\lambda\left[\left(\frac{\partial}{\partial\lambda}\ln f(X,\,\lambda)\right)^2\right] = E_\lambda\left[\left(\frac{X-\lambda}{\lambda}\right)^2\right] = \frac{1}{\lambda}.
\end{aligned}$$

Recall that $E(X) = Var(X) = \lambda$. That means that as λ increases, the variability of X increases. Therefore, information about λ (mean) go down as λ increases. ||

Example 2.18. Let $X \sim N(\mu,\,\sigma^2)$, where σ is known and $\mu \in \mathbb{R}$ is unknown parameters. The PDF of X is

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

for all $x \in \mathbb{R}$. Therefore,

$$\ln f(x,\,\mu) = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\begin{aligned}
&\implies \frac{\partial}{\partial \mu} \ln f(x,\, \mu) = \frac{x-\mu}{\sigma^2} \\
&\implies \frac{\partial^2}{\partial \mu^2} \ln f(x,\, \mu) = -\frac{1}{\sigma^2} \\
&\implies \mathcal{I}_X(\mu) = \frac{1}{\sigma^2}.
\end{aligned}$$ ||

Definition 2.7 (Fisher Information)**.** *The Fisher information contained in a collection of RVs, say* $\boldsymbol{X}$*, is defined by*

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial \theta} \ln f_{\boldsymbol{X}}(\boldsymbol{X},\, \theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial \theta^2} \ln f_{\boldsymbol{X}}(\boldsymbol{X},\, \theta)\right],$$

where $f_{\boldsymbol{X}}(\cdot,\, \theta)$ *is the JPDF of* $\boldsymbol{X}$ *under* θ*.*

Theorem 2.4. *Let* $X_1,\, X_2,\, \ldots,\, X_n$ *be a RS from a population with PMF or PDF* $f(\cdot,\, \theta)$*, where* $\theta \in \Theta$*. Let* $\mathcal{I}_{\boldsymbol{X}}(\theta)$ *denote the Fisher information contained in the RS, then*

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = n\mathcal{I}_{X_1}(\theta) \quad \text{for all } \theta \in \Theta.$$

Proof: For a RS, the JPMF or JPDF is

$$f_{\boldsymbol{X}}(\boldsymbol{x},\, \theta) = \prod_{i=1}^{n} f(x_i,\, \theta),$$

where $\boldsymbol{x} = (x_1,\, x_2,\, \ldots,\, x_n)$. Thus,

$$\begin{aligned}
\ln f_{\boldsymbol{X}}(\boldsymbol{x},\, \theta) &= \sum_{i=1}^{n} \ln f(x_i,\, \theta) \\
\implies \frac{\partial^2}{\partial \theta^2} f_{\boldsymbol{X}}(\boldsymbol{x},\, \theta) &= \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(x_i,\, \theta) \\
\implies \mathcal{I}_{\boldsymbol{X}}(\theta) &= -E_\theta\left[\frac{\partial^2}{\partial \theta^2} \ln f_{\boldsymbol{X}}(\boldsymbol{X},\, \theta)\right] \\
&= \sum_{i=1}^{n} -E_\theta\left[\frac{\partial^2}{\partial \theta^2} \ln f(X_i,\, \theta)\right] \\
&= \sum_{i=1}^{n} \mathcal{I}_{X_i}(\theta) \\
&= n\mathcal{I}_{X_1}(\theta).
\end{aligned}$$ □

Theorem 2.5. *Let* $\boldsymbol{X}$ *be a RS and* $\boldsymbol{T}$ *be a statistic. Then* $\mathcal{I}_{\boldsymbol{X}}(\theta) \geq \mathcal{I}_{\boldsymbol{T}}(\theta)$ *for all* $\theta \in \Theta$*. The equality holds for all* $\theta \in \Theta$ *if and only if* $\boldsymbol{T}$ *is a sufficient statistic for* θ*.*

Proof: Here we provide an outline of the proof for the continuous case only. Let $\boldsymbol{X} = (X_1,\, X_2,\, \ldots,\, X_n)$ and $\boldsymbol{T} = (T_1,\, T_2,\, \ldots,\, T_k)$ be continuous random vector with $n > k$. Now, the Fisher information contained in $\boldsymbol{T}$ can be calculated using the JPDF of $\boldsymbol{T}$. Notice that

$$f_{\boldsymbol{X}}(\boldsymbol{x},\, \theta) = f_{\boldsymbol{T}}(\boldsymbol{t},\, \theta)\, f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x},\, \theta|\boldsymbol{t}) \implies \ln f_{\boldsymbol{X}}(\boldsymbol{x},\, \theta) = \ln f_{\boldsymbol{T}}(\boldsymbol{t},\, \theta) + \ln f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x},\, \theta|\boldsymbol{t}).$$

Thus,

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = \mathcal{I}_{\boldsymbol{T}}(\theta) + E_\theta\left[\mathcal{I}_{\boldsymbol{X}|\boldsymbol{T}}(\theta)\right],$$

where $\mathcal{I}_{\boldsymbol{X}|\boldsymbol{T}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{X}, \theta|\boldsymbol{t})\right)^2 \middle| \boldsymbol{T}\right]$ is called conditional Fisher information. Clearly, $\mathcal{I}_{\boldsymbol{X}|\boldsymbol{T}}(\theta) \geq 0$. Therefore, $\mathcal{I}_{\boldsymbol{X}}(\theta) \geq \mathcal{I}_{\boldsymbol{T}}(\theta)$.

Now, the equality holds if and only if

$$E_\theta\left[\mathcal{I}_{\boldsymbol{X}|\boldsymbol{T}}(\theta)\right] = 0 \iff \mathcal{I}_{\boldsymbol{X}|\boldsymbol{T}}(\theta) = 0 \iff f_{\boldsymbol{X}|\boldsymbol{T}}(\boldsymbol{x}|\boldsymbol{t}) \text{ does not involve } \theta.$$

Thus, $\boldsymbol{T}$ is sufficient statistic for θ. □

Example 2.19 (Continuation of Example 2.17)**.** Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Poi(\lambda)$ with $\lambda > 0$. We have seen that $\mathcal{I}_{X_1}(\lambda) = \frac{1}{\lambda}$. Thus, $\mathcal{I}_{\boldsymbol{X}}(\lambda) = \frac{n}{\lambda}$.

Now, assume that $T = \sum_{i=1}^n X_i$. We know that $T \sim Poi(n\lambda)$. Thus, the logarithm of the PMF of T is

$$\begin{aligned}
&\ln f_T(t, \lambda) = -n\lambda + t\ln(n\lambda) - \ln(t!) \quad \text{for } t = 0, 1, \ldots \\
&\implies \frac{\partial^2}{\partial\lambda^2}\ln f_T(t, \lambda) = -\frac{t}{\lambda^2} \\
&\implies \mathcal{I}_T(\lambda) = \frac{n}{\lambda}.
\end{aligned}$$

Hence, Fisher information contained in the RS is same as that contained in T. Therefore, T is a sufficient statistic for λ. ||

2.6 Ancillary Statistic

The concept of ancillary statistic is apparently opposite to that of sufficiency. Unlike, a sufficiency statistic, which contains all information about $\boldsymbol{\theta}$, a ancillary statistic does not contain any information about $\boldsymbol{\theta}$. It does not mean that the ancillary statistic are not useful in statistical analysis. For example, the fixed sample size n seldom has any information about $\boldsymbol{\theta}$, but it is a very important quantity in any statistical analysis.

Definition 2.8 (Ancillary Statistic)**.** *A statistic $\boldsymbol{T}$ is called an ancillary statistic for $\boldsymbol{\theta}$ if the distribution of $\boldsymbol{T}$ does not involve $\boldsymbol{\theta}$.*

Example 2.20. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, where $\mu \in \mathbb{R}$ is unknown parameter. Then, $T_1 = X_1 - X_2$ is an ancillary statistic for μ as $T_1 \sim N(0, 2)$, which does not involve μ. Similarly, we can check that $T_2 = X_1 + X_2 + \ldots + X_{n-1} - (n-1)X_n$ and S^2 are ancillary statistics for μ.

Let us now consider $\boldsymbol{T} = (T_1, T_2)$. It is easy to check that $\boldsymbol{T} \sim N_2(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & n(n-1) \end{pmatrix}.$$

Thus, the distribution of $\boldsymbol{T}$ does not involve μ, and hence, $\boldsymbol{T}$ is ancillary for μ. ||

Example 2.21. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ is unknown parameter vector. In this case, $T_1 = X_1 - X_2$ or $T_2 = X_1 + X_2 + \ldots + X_{n-1} - (n-1)X_n$ are not ancillary. Now, consider the statistic $T_3 = \frac{X_1 - X_2}{S}$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. We will show that T_3 is an ancillary statistic for $\boldsymbol{\theta}$. Let $Y_i = \frac{X_i - \mu}{\sigma}$. Then, $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} N(0, 1)$. Moreover,

$$T_3 = \frac{X_1 - X_2}{S} = \frac{\sqrt{n-1}\,(Y_1 - Y_2)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}.$$

As the distributions of $Y_1, Y_2, \ldots, Y_n$ do not involve $\boldsymbol{\theta}$, the distribution of T_3 does not involve $\boldsymbol{\theta}$. Therefore, T_3 is an ancillary statistic for $\boldsymbol{\theta}$. Note that we do not need to find the PDF of T_3 explicitly to show that the distribution of T_3 does not depend on $\boldsymbol{\theta}$. ||

Example 2.22. Let $X_1, X_2 \overset{i.i.d.}{\sim} N(\mu, 1)$. Take, $T_1 = X_1 - X_2$ and $T_2 = X_1$. We have seen in Example 2.20 that T_1 is ancillary for μ. Also notice that $\mathcal{I}_{T_2}(\mu) = 1$ and $\mathcal{I}_{X_1, X_2}(\mu) = 2$. Therefore, T_2 is not sufficient statistic for μ. However, $\boldsymbol{T} = (T_1, T_2)$ is a sufficient statistic for μ. To see it, notice that there exists a one-to-one function between $\boldsymbol{T}$ and (X_1, X_2).

This example shows that it is possible to have two statistics T_1 and T_2 such that T_1 is ancillary for a parameter, T_2 has some information about the same parameter, but not sufficient for the parameter, and yet (T_1, T_2) is jointly sufficient for the parameter. ||

Example 2.23. Let (X, Y) is a bivariate normal random vector with $E(X) = E(Y) = 0$, $Var(X) = Var(Y) = 1$ and correlation coefficient ρ, where $\rho \in (-1, 1)$ is an unknown parameter. Consider two statistics $T_1 = X$ and $T_2 = Y$. Then, $T_1, T_2 \sim N(0, 1)$, and hence, ancillary statistics for ρ. However, (T_1, T_2), being equivalent to (X, Y), is minimal sufficient statistics for ρ.

This example shows that it is possible to have two statistics T_1 and T_2 such that both of T_1 and T_2 has no information about the parameter, and yet, (T_1, T_2) has all information. ||

2.7 Completeness

Suppose that $\boldsymbol{X}$ is a RS with common PMF or PDF $f(\cdot, \boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{T}$ be a statistic with its PMF or PDF $g(\boldsymbol{t}, \boldsymbol{\theta})$ for $t \in \mathcal{T}$, the support of $\boldsymbol{T}$, and $\boldsymbol{\theta} \in \Theta$.

Definition 2.9. *The family $\{g(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is called the family of distributions induced by the statistic T.*

Example 2.24. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$ is the unknown parameter. Consider the statistic $T = \overline{X}$. We know that $T \sim N(\mu, n^{-1})$. Thus, the PDF of T is

$$\phi(t, \mu) = \frac{\sqrt{n}}{\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2n}\right] \quad \text{for } t \in \mathbb{R}.$$

Therefore, the family of distributions induced by T is $\{\phi(\cdot, \mu) : \mu \in \mathbb{R}\}$. ||

Definition 2.10 (Complete Family)**.** *A family $\{g(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is called complete if for any real valued function $h(\boldsymbol{t})$ defined for all $\boldsymbol{t} \in \mathcal{T}$,*

$$E_{\boldsymbol{\theta}}(h(\boldsymbol{T})) = 0 \text{ for all } \boldsymbol{\theta} \in \Theta \text{ implies } h(\boldsymbol{T}) = 0 \text{ with probability } 1.$$

Definition 2.11 (Complete Statistic)**.** *A statistic* $\boldsymbol{T}$ *is called complete if the family induced by the statistic* $\boldsymbol{T}$ *is complete.*

Example 2.25. A statistic T is distributed as $Bernoulli(p)$, $0 < p < 1$. The family induced by T is $\{g(\cdot, p) : 0 < p < 1\}$, where

$$g(t, p) = \begin{cases} p^t(1-p)^{1-t} & \text{if } t = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

Consider any real valued function $h(t)$ such that $E_p[h(T)] = 0$ for all $0 < p < 1$. That means that

$$E_p[h(T)] = (1-p)h(0) + ph(1) = p\{h(1) - h(0)\} + h(0) = 0 \text{ for all } p \in (0, 1).$$

Note that $p\{h(1) - h(0)\} + h(0) = 0$ is a linear equation in p. This can have at most one solution. However, we are demanding that $p\{h(1) - h(0)\} + h(0) = 0$ for all $p \in (0, 1)$. Thus, the expression $p\{h(1) - h(0)\} + h(0)$ must be identically zero, and hence, the coefficients must be zero. Therefore, $h(0) = 0$ and $h(1) - h(0) = 0$, which implies that $h(0) = h(1) = 0$. This shows that $h(T) = 0$ with probability one. Hence, T is complete. ||

Example 2.26. Consider $g(t, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{t^2}{2\sigma^2}\right]$ for $t \in \mathbb{R}$ and $\sigma > 0$. In this case the corresponding family is not complete. To see it, take $h(t) = t$. Clearly, $E_\sigma(h(T)) = 0$ for all $\sigma > 0$. However, $P_\sigma(h(T) = 0) = P_\sigma(T = 0) = 0$. Therefore, the family is not complete. ||

2.8 Complete Sufficient Statistic

Definition 2.12 (Complete Sufficient Statistic)**.** *A statistic* $\boldsymbol{T}$ *is called complete sufficient statistic for* $\boldsymbol{\theta}$ *if* $\boldsymbol{T}$ *is sufficient for* $\boldsymbol{\theta}$ *and* $\boldsymbol{T}$ *is complete.*

Example 2.27. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$, where $p \in (0, 1)$. We have seen that $T = \sum_{i=1}^n X_i$ is sufficient for p. Now, we will verify that T is complete sufficient statistic by showing that T is complete statistic. Consider $h(\cdot)$ be a real valued function such that $E_p[h(T)] = 0$ for all $p \in (0, 1)$. First, notice that $T \sim Bin(n, p)$. Now, for $\nu = \frac{p}{1-p}$, we have

$$E_p[h(T)] = \sum_{t=0}^{n} h(t)\binom{n}{t}p^t(1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} \binom{n}{t} h(t)\nu^t = 0.$$

As we assume that $E_p[h(T)] = 0$ for all $p \in (0, 1)$, $\sum_{t=0}^n \binom{n}{t}h(t)\nu^t = 0$ for all $\nu > 0$. As $\sum_{t=0}^n \binom{n}{t}h(t)\nu^t$ is a nth degree polynomial in ν, the polynomial equation has at most n solutions in $(0, 1)$. As we are demanding the polynomial is zero for all values of $\nu > 0$, the coefficients are zero. Thus, $\binom{n}{t}h(t) = 0$ for all $t = 0, 1, 2, \ldots, n$, which implies that $h(t) = 0$ for all $t = 0, 1, \ldots, n$. Hence, $E_p[h(T)] = 0$ for all $p \in (0, 1)$ implies that $h(T) = 0$ with probability one. Therefore, T is a complete statistic. ||

Theorem 2.6. *Suppose that a statistic* $\boldsymbol{T}$ *is complete. Let* $\boldsymbol{U}$ *be another statistic with* $\boldsymbol{U} = g(\boldsymbol{T})$*, where* g *is a one-to-one function. Then* $\boldsymbol{U}$ *is complete.*

Proof: The proof is skipped here. □

Theorem 2.7. *A complete sufficient statistic is minimal sufficient.*

Proof: The proof is skipped. □

Example 2.28. Suppose that $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \theta^2)$ with an unknown parameter $\theta > 0$. Then, $\boldsymbol{T} = \left(\overline{X}, S^2\right)$ is a minimal sufficient statistic for θ, where sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Now, it is easy to check that $E_\theta\left[\frac{n}{n+1}\overline{X}^2\right] = \theta^2 = E_\theta(S^2)$ for all $\theta > 0$. Thus, $E_\theta\left[\frac{n}{n+1}\overline{X}^2 - S^2\right] = 0$ for all $\theta > 0$. Now, if we take $h(\boldsymbol{t}) = \frac{n}{n+1}\overline{x}^2 - s^2$, where $\boldsymbol{t} = (\overline{x}, s^2)$, then $E_\theta[h(T)] = 0$ for all $\theta > 0$. However, $P_\theta(h(\boldsymbol{T}) = 0) = 0$ (why?). Therefore, $\boldsymbol{T}$ is minimal sufficient but not complete. This example shows that the converse of the previous theorem is not true. ||

2.9 Families of Distributions

In this section, we will briefly discuss several families of distributions that are commonly encountered in Statistics.

2.9.1 Location Family

Definition 2.13 (Location Family)**.** *Let $g(\cdot)$ be a PDF. Then the family of distributions*

$$\mathcal{F} = \{g(x - \theta) : \theta \in \mathbb{R},\, x \in \mathbb{R}\}$$

is called location family of distributions. Here, the parameter θ is called location parameter.

Note that $f(x, \theta) = g(x - \theta)$ is a PDF for all $\theta \in \mathbb{R}$. Thus, the previous definition is meaningful.

Example 2.29. A $N(\mu, 1)$ distribution, with $\mu \in \mathbb{R}$, forms a location family, where μ is the location parameter. To see it, start with $g(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ for all $x \in \mathbb{R}$. Then it is easy to see that $f(x, \mu) = g(x - \mu)$ is the PDF of a $N(\mu, 1)$ distribution. ||

Example 2.30. Let us start with the $Exp(1)$ distribution as the base distribution. That means that the form of the PDF $g(\cdot)$ is given by

$$g(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefor, for any $\theta \in \mathbb{R}$, $f(\cdot, \theta)$ is given by

$$f(x, \theta) = \begin{cases} e^{-(x-\theta)} & \text{if } x > \theta \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the family of distributions $\mathcal{F} = \{f(x, \theta) : \theta \in \mathbb{R},\, x \in \mathbb{R}\}$ is a location family, where θ is location parameter. ||

In a location family, the PDF moves along the horizontal axis as the value of the location parameter changes. For example, the PDF of $N(0,\,1)$ distribution is centered at zero, but the PDF of $N(5,\,1)$ is centered at five. Similarly, if we take $\theta = 0$ in Example 2.30, the PDF is positive on the positive part of the real line. If we take $\theta = -5$, the PDF is positive for all $x > -5$. However, the shapes of the PDFs remain same in both examples. Thus, the shape of the PDF does not change, only the position of the PDF changes with location parameter.

Theorem 2.8. *Let $X_1,\, X_2,\, \ldots,\, X_n$ be a RS from a PDF, which belongs to a location family. Then the statistic $\boldsymbol{T} = (X_1 - X_n,\, X_2 - X_n,\, \ldots,\, X_{n-1} - X_n)$ is ancillary for the location parameter.*

Proof: Let the common PDF of $X_1,\, X_2,\, \ldots,\, X_n$ be of the form $g(x-\theta)$. Let us define $Y_i = X_i - \theta$ for $i = 1,\, 2,\, \ldots,\, n$. Then $Y_1,\, Y_2,\, \ldots,\, Y_n$ are *i.i.d.* RVs with common PDF $g(x)$, which does not involve θ. Thus, the JPDF of $Y_1,\, Y_2,\, \ldots,\, Y_n$ does not involve θ. Therefore, the distribution of

$$\boldsymbol{T} = (X_1 - X_n,\, X_2 - X_n,\, \ldots,\, X_{n-1} - X_n) = (Y_1 - Y_n,\, Y_2 - Y_n,\, \ldots,\, Y_{n-1} - Y_n)$$

does not involve θ. □

Remark 2.2. Note that we have written $\boldsymbol{T}$ as a function of $X_i - X_n$ for $i = 1,\, 2,\, \ldots,\, n-1$ in the previous theorem. However, the previous theorem holds true if $\boldsymbol{T}$ is taken as a function of $X_i - X_j$ for $i \neq j$. †

2.9.2 Scale Family

Definition 2.14 (Scale Family)**.** *Let $g(\cdot)$ be a PDF. Then the family of distributions*

$$\mathcal{F} = \left\{ \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right) : \sigma > 0,\, x \in \mathbb{R} \right\}$$

is called scale family of distributions. Here, the parameter σ is called scale parameter.

Note that $f(x,\,\sigma) = \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right)$ is a PDF for all $\sigma > 0$. Thus, the previous definition is meaningful.

Example 2.31. A $N(0,\,\sigma^2)$ distribution, with $\sigma > 0$, forms a scale family, where σ is scale parameter. To see it, start with $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ for all $x \in \mathbb{R}$. Then it is easy to see that $f(x,\,\sigma) = \frac{1}{\sigma} g\left(\frac{x}{\sigma}\right)$ is the PDF of a $N(0,\,\sigma^2)$ distribution. ||

Example 2.32. Let us start with the $Exp(1)$ distribution as the base distribution. Now, for any $\theta > 0$, $f(\cdot,\,\theta)$ is given by

$$f(x,\,\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the family of distributions $\mathcal{F} = \{f(x,\,\theta) : \theta > 0,\, x \in \mathbb{R}\}$ is a scale family, where θ is scale parameter. ||

In a scale family, the PDF squeezes or expands as the value of the scale parameter changes. For example, the PDF of $N(0,\,1)$ and $N(0,\,5)$ distributions are centered at $x = 0$, but the PDF of $N(0,\,5)$ is more flat compare to that of $N(0,\,1)$. On the other hand, the PDF of $N(0,\,0.5)$ is more peaked compare to the PDF of a standard normal distribution. Similar effect can be seen in Example 2.32.

Theorem 2.9. *Let $X_1, X_2, \ldots, X_n$ be a RS from a PDF, which belongs to a scale family. Then the statistic $\boldsymbol{T} = \left(\frac{X_1}{X_n}, \frac{X_2}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right)$ is ancillary for the scale parameter.*

Proof: Let the common PDF of $X_1, X_2, \ldots, X_n$ be of the form $\frac{1}{\sigma} g\left(\frac{x}{\sigma}\right)$. Let us define $Y_i = \frac{X_i}{\sigma}$ for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n$ are *i.i.d.* RVs with common PDF $g(x)$, which does not involve σ. Thus, the JPDF of $Y_1, Y_2, \ldots, Y_n$ does not involve σ. Therefore, the distribution of

$$\boldsymbol{T} = \left(\frac{X_1}{X_n}, \frac{X_2}{X_n}, \ldots, \frac{X_{n-1}}{X_n}\right) = \left(\frac{Y_1}{Y_n}, \frac{Y_2}{Y_n}, \ldots, \frac{Y_{n-1}}{Y_n}\right)$$

does not involve σ. □

Remark 2.3. Note that we have written $\boldsymbol{T}$ as a function of $\frac{X_i}{X_n}$ for $i = 1, 2, \ldots, n-1$ in the previous theorem. However, the previous theorem holds true if $\boldsymbol{T}$ is taken as a function of $\frac{X_i}{X_j}$ for $i \neq j$. †

2.9.3 Location-Scale Family

Definition 2.15 (Location-Scale Family)**.** *Let $g(\cdot)$ be a PDF. Then the family of distributions*

$$\mathcal{F} = \left\{\frac{1}{\sigma} g\left(\frac{x-\theta}{\sigma}\right) : \theta \in \mathbb{R},\, \sigma > 0,\, x \in \mathbb{R}\right\}$$

is called location-scale family of distributions. Here, the parameters θ and σ are called location parameter and scale parameter, respectively.

Note that $f(x, \theta, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\theta}{\sigma}\right)$ is a PDF for all $\theta \in \mathbb{R}$ and $\sigma > 0$. Thus, the previous definition is meaningful.

Example 2.33. A $N(\mu, \sigma^2)$ distribution, with $\mu \in \mathbb{R}$ and $\sigma > 0$, forms a location-scale family. To see it, start with $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ for all $x \in \mathbb{R}$. Then it is easy to see that $f(x, \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right)$ is the PDF of a $N(\mu, \sigma^2)$ distribution. In this case, μ and σ are location and scale parameters, respectively. ||

Example 2.34. Let us start with the $Exp(1)$ distribution as the base distribution. Now, for any $\mu \in \mathbb{R}$ and $\theta > 0$, $f(\cdot, \mu, \theta)$ is given by

$$f(x, \mu, \theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x-\mu}{\theta}} & \text{if } x > \mu \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the family of distributions $\mathcal{F} = \{f(x, \mu, \theta) : \mu \in \mathbb{R},\, \theta > 0,\, x \in \mathbb{R}\}$ is a location-scale family, where μ and θ are location and scale parameters, respectively. ||

Movement of the PDF along the horizontal axis as well as squeezing and expanding effects are noticed in a location-scale family. For example, suppose that the data consists of daily maximum temperatures in Celsius of a city. Now, if we postulate a normal distribution for temperatures, changing the unit to Fahrenheit would amount to shift in location and scale. Recall the relationship between Celsius and Fahrenheit: $C = \frac{5}{8}(F - 32)$.

Theorem 2.10. *Let $X_1, X_2, \ldots, X_n$ be a RS from a PDF, which belongs to a location-scale family. Then the statistic $\boldsymbol{T} = \left(\frac{X_1 - X_n}{S}, \frac{X_2 - X_n}{S}, \ldots, \frac{X_{n-1} - X_n}{S}\right)$ is ancillary for the location and scale parameters, where S is the positive square root of sample variance.*

Proof: Let the common PDF of $X_1, X_2, \ldots, X_n$ be of the form $\frac{1}{\sigma} g\left(\frac{x-\theta}{\sigma}\right)$. Let us define $Y_i = \frac{X_i - \theta}{\sigma}$ for $i = 1, 2, \ldots, n$. Then $Y_1, Y_2, \ldots, Y_n$ are *i.i.d.* RVs with common PDF $g(x)$, which does not involve θ. Thus, the JPDF of $Y_1, Y_2, \ldots, Y_n$ does not involve θ. Now, notice that

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} (\theta + \sigma Y_i) = \theta + \sigma \overline{Y}$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\theta + \sigma Y_i - \theta - \sigma\overline{Y}\right)^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sigma^2 S_Y^2.$$

Therefore, the distribution of

$$\boldsymbol{T} = \left(\frac{X_1 - X_n}{S}, \frac{X_2 - X_n}{S}, \ldots, \frac{X_{n-1} - X_n}{S}\right) = \left(\frac{Y_1 - Y_n}{S_Y}, \frac{Y_2 - Y_n}{S_Y}, \ldots, \frac{Y_{n-1} - Y_n}{S_Y}\right)$$

does not involve θ and σ. □

Remark 2.4. Note that we have written $\boldsymbol{T}$ as a function of $\frac{X_i - X_n}{S}$ for $i = 1, 2, \ldots, n-1$ in the previous theorem. However, the previous theorem holds true if $\boldsymbol{T}$ is taken as a function of $\frac{X_i - X_j}{S}$ for $i \neq j$. †

2.9.4 Exponential Family

Definition 2.16 (Exponential Family)**.** *Let a RV X have its PMF or PDF given by $f(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$. We say that $f(\cdot, \boldsymbol{\theta})$ belongs to a k-parameter exponential family if*

$$f(x, \boldsymbol{\theta}) = a(\boldsymbol{\theta}) g(x) \exp\left[\sum_{i=1}^{k} b_i(\boldsymbol{\theta}) R_i(x)\right], \tag{2.1}$$

for all $x \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$, with some appropriate forms for $g(x) \geq 0$, $a(\boldsymbol{\theta}) \geq 0$, $b_i(\boldsymbol{\theta})$, and $R_i(x)$, $i = 1, 2, \ldots, k$.

Remark 2.5. It is crucial to note that $a(\boldsymbol{\theta})$, $b_1(\boldsymbol{\theta})$, $b_2(\boldsymbol{\theta})$, $\ldots$, $b_k(\boldsymbol{\theta})$ cannot involve x and $R_1(x)$, $R_2(x)$, $\ldots$, $R_k(x)$, $g(x)$ cannot involve $\boldsymbol{\theta}$. †

Remark 2.6. To have statistically meaningful parameterization, we will assume that

1. Neither b_i's nor R_i's satisfy linear constrains.
2. Θ contained a k-dimensional rectangle. Such an exponential family is said to have full rank. †

Example 2.35. Let $X \sim Bin(n, p)$, where n is known, but $p \in (0, 1)$ is unknown parameter. Then the PMF of X belongs to a one-parameter exponential family. To see it, take $k = 1$, $\theta = p$, $a(\theta) = (1-p)^n$, $b_1(\theta) = \ln\left(\frac{p}{1-p}\right)$, $g(x) = \binom{n}{x}$, and $R_1(x) = x$ in (2.1). ||

Example 2.36. Let $X \sim N(\mu, \sigma^2)$. It is easy to see that the PDF of X can be expressed in the form of (2.1) with $k = 2$, $\boldsymbol{\theta} = (\mu, \sigma)$, $R_1(x) = x$, $R_2(x) = x^2$, $b_1(\boldsymbol{\theta}) = \frac{\mu}{\sigma^2}$, $b_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}$, $a(\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}}$, $g(x) = 1$. Thus, the PDF of a normal distribution with mean μ and variance σ^2 belongs to a two-parameter exponential family. ||

Example 2.37. Let the PDF of a RV X be

$$f(x, \theta) = \frac{1}{\theta} \exp\left[-\frac{x-\theta}{\theta}\right] I_{(\theta,\infty)}(x),$$

where $\theta > 0$. Here, the term $I_{(\theta,\infty)}(x)$ cannot be absorbed within $a(\theta)$, $b(\theta)$, $g(x)$, or $R(x)$, and so this distribution does not belong to a one-parameter exponential family. ||

Example 2.38. Suppose that X has a $N(\theta, \theta^2)$ distribution with $\theta > 0$. The PDF can be expressed as

$$f(x, \theta) = \frac{1}{\theta\sqrt{2\pi}} \exp\left[-\frac{(x-\theta)^2}{2\theta^2}\right] = \frac{1}{\theta\sqrt{2\pi e}} \exp\left[-\frac{x^2}{2\theta^2} + \frac{x}{\theta}\right],$$

which does not have the same form as in (2.1) and it does not belong to a one-parameter exponential family. ||

Theorem 2.11. *Let $X_1, X_2, \ldots, X_n$ be a RS from a common PMF or PDF*

$$f(x, \boldsymbol{\theta}) = a(\boldsymbol{\theta}) g(x) \exp\left[\sum_{j=1}^{k} b_j(\boldsymbol{\theta}) R_j(x)\right]$$

belongs to a k-parameter exponential family with full rank. Let us denote the statistic $T_j = \sum_{i=1}^{n} R_j(X_i)$ for $j = 1, 2, \ldots, k$. Then, $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$ is jointly minimal sufficient for $\boldsymbol{\theta}$.

Proof: The proof is not very easy and skipped here. □

Theorem 2.12. *Let $X_1, X_2, \ldots, X_n$ be a RS from a common PMF or PDF*

$$f(x, \boldsymbol{\theta}) = a(\boldsymbol{\theta}) g(x) \exp\left[\sum_{j=1}^{k} b_j(\boldsymbol{\theta}) R_j(x)\right]$$

belongs to a k-parameter exponential family with full rank. Let us denote the statistic $T_j = \sum_{i=1}^{n} R_j(X_i)$ for $j = 1, 2, \ldots, k$. Then, minimal sufficient statistic $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$ for $\boldsymbol{\theta}$ is complete.

Proof: The proof is quite involved and therefore skipped here. □

Example 2.39. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$ is unknown parameter. We have seen in Example 2.36 that $N(\mu, \sigma^2)$ belongs to a two-parameter exponential family with $R_1(x) = x$ and $R_2(x) = x^2$. As the parametric space $\mathbb{R} \times \mathbb{R}^+$ contains a two-dimensional rectangle, it is of full rank. Therefore, using the previous theorem, $\boldsymbol{T} = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is complete sufficient statistic for (μ, σ^2). ||

2.10 Basu's Theorem

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector with JPMF or JPDF $f(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is unknown parameter. Now, assume that $\boldsymbol{U}$ and $\boldsymbol{W}$ be two statistic based on $\boldsymbol{X}$. Then Basu's theorem provide an elegant method to show independence of two appropriate statistics $\boldsymbol{U}$ and $\boldsymbol{W}$, which is otherwise a tedious job. Note that for Basu's theorem it is not necessary to have independent and identically distributed RVs. Of course, we are going to use Basu's theorem mainly for RS in this course.

Theorem 2.13 (Basu's Theorem)**.** *Suppose that $\boldsymbol{U}$ is a complete sufficient statistic for $\boldsymbol{\theta}$ and $\boldsymbol{W}$ is an ancillary statistic for $\boldsymbol{\theta}$. Then $\boldsymbol{U}$ and $\boldsymbol{W}$ are independent.*

Proof: For simplicity, we will prove the theorem for discrete case only. Let the supports of $\boldsymbol{U}$ and $\boldsymbol{W}$ be denoted by $\mathcal{U}$ and $\mathcal{W}$, respectively. Now, we need to show that

$$P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{U} = \boldsymbol{u}) = P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}) \text{ for all } \boldsymbol{w} \in \mathcal{W}, \boldsymbol{u} \in \mathcal{U} \text{ and } \boldsymbol{\theta} \in \Theta.$$

For all $\boldsymbol{w} \in \mathcal{W}$, notice that $P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w})$ does not involve $\boldsymbol{\theta}$ as $\boldsymbol{W}$ is ancillary for $\boldsymbol{\theta}$. Denote $h(\boldsymbol{w}) = P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w})$ for all $\boldsymbol{w} \in \mathcal{W}$. Also, $P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{U} = \boldsymbol{u})$ is free of $\boldsymbol{\theta}$ as $\boldsymbol{U}$ is sufficient statistic for $\boldsymbol{\theta}$. Now, for fixed $\boldsymbol{w} \in \mathcal{W}$, let us denote $g_{\boldsymbol{w}}(\boldsymbol{u}) = P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{U} = \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{U}$. Then

$$h(\boldsymbol{w}) = P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}) = \sum_{\boldsymbol{u} \in \mathcal{U}} P_{\boldsymbol{\theta}}(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{U} = \boldsymbol{u}) P_{\boldsymbol{\theta}}(\boldsymbol{U} = \boldsymbol{u}) = E_{\boldsymbol{\theta}}(g_{\boldsymbol{w}}(\boldsymbol{U})).$$

Thus, $E_{\boldsymbol{\theta}}[g_{\boldsymbol{w}}(\boldsymbol{U}) - h(\boldsymbol{w})] = 0$ for all $\boldsymbol{\theta} \in \Theta$. As $\boldsymbol{U}$ is complete and $g_{\boldsymbol{w}}(\boldsymbol{U}) - h(\boldsymbol{w})$ is a statistic, we have $g_{\boldsymbol{w}}(\boldsymbol{U}) = h(\boldsymbol{w})$ with probability one for each fixed $\boldsymbol{w} \in \mathcal{W}$. Therefore, $\boldsymbol{U}$ and $\boldsymbol{W}$ are independent. □

Example 2.40. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with $n \geq 2$. Further assume that $\mu \in \mathbb{R}$ is unknown, but $\sigma > 0$ is known. In this case, $\overline{X}$ is complete sufficient statistic for μ. It can be seen from the fact that $N(\mu, \sigma^2)$ belongs to one-parameter exponential family. On the other hand, S^2 is ancillary statistic. Thus, using Basu's theorem, $\overline{X}$ and S^2 are independent. Of course, we have seen a stronger result in Theorem 1.13.

The sample range is defined by $V = X_{(n)} - X_{(1)}$. As $N(\mu, \sigma^2)$ belongs to location family of distributions, it is easy to see that V is ancillary. Then, $\overline{X}$ and $\frac{S}{V}$ are independent. Similarly, $\overline{X}$ and $X_{(n)} - \overline{X}$ are independent. In the same spirit, $\overline{X}$ and $\left(X_{(1)} - \overline{X}\right)^2$ are independent. ||

2.11 Method of Finding Estimator

2.11.1 Method of Moment Estimator

MME was first introduced by Karl Pearson in the year 1902. The basic method can be summarized in following algorithm:

1. Suppose that we have a RS of size n form a population with PMF/PDF $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is the unknown parameter vector. We want to find estimator of $\boldsymbol{\theta}$.

2. Calculate first k (number of unknown parameters) moments $\mu'_1, \ldots, \mu'_k$ of $f(x; \boldsymbol{\theta})$, where $\mu'_r = E_{\boldsymbol{\theta}}(X^r)$.

3. Calculate first k sample moments $m_1', \ldots, m_k'$, where $m_r' = \frac{1}{n}\sum_{i=1}^n X_i^r$.

4. Equate $\mu_r' = m_r'$ for $r = 1, 2, \ldots, k$.

5. Solve the system of k equations (if they are consistent) for θ_i's. The solutions are the MMEs of the unknown parameters.

Example 2.41. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$, $\theta \in (0, 1) = \Theta$. Here, we have one parameter θ. Thus, $k = 1$. $E(X_1) = \theta$. Hence, we get the MME of θ is $\widehat{\theta} = \overline{X}$. ||

Example 2.42. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Theta$. Here $k = 2$, $E(X) = \mu$, and $E(X^2) = \sigma^2 + \mu^2$. Hence, we get the MMEs of μ and σ^2 are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, respectively. ||

Example 2.43. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\sigma > 0$. Here $k = 1$. However, as $E(X) = 0$, equating $E(X) = \overline{X}$ does not provide any solution (inconsistent). Alternatively, we can find $E(X^2) = \sigma^2$ and equate to m_2' to obtain MME of σ^2 as $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$. ||

Example 2.44. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \theta^2)$, $\theta > 0$. Here $k = 1$. $E(X) = \theta$. Equating $E(X) = \overline{X}$, we get MME of θ is $\widehat{\theta} = \overline{X}$. However, this may not be a meaningful estimator as $\overline{X}$ can be negative with positive probability, while $\theta > 0$. ||

Remark 2.7. Previous two examples show that there are some degrees of arbitrariness in this method. †

2.11.2 Maximum Likelihood Estimator

The MLE was first proposed by R. A. Fisher in 1912. This is one of the most popular methods of estimation. Let us start with an example.

Example 2.45. Let a box has some red balls and some black balls. It is known that number of black balls to red balls is in 1:1 or 1:2 ratio. We want to find whether it is 1:1 or 1:2. To perform the task, suppose that two balls are drawn randomly and with replacement from the box. Let X be the number of black balls out of two drawn balls. Then $X \sim Bin(2, p)$, where p is the probability that a drawn ball is black. In this case, as the ratio of the black to red balls is either 1:1 or 1:2, p can tale values $\frac{1}{2}$ or $\frac{1}{3}$. Thus, the parametric space is $\Theta = \left\{\frac{1}{2}, \frac{1}{3}\right\}$. Now, the problem of deciding whether the ratio is 1:1 or 1:2 boils down to estimate the value of p.

Let us consider the following table, where the entries are $P_p(X = x)$ for each possible values of x and p. From first column, we see that $P(X = 0)$ is maximum if $p = \frac{1}{3}$. Thus, it

	$x = 0$	$x = 1$	$x = 2$
$p = \frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$
$p = \frac{1}{3}$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

is more likely to occur $X = 0$ (that is no black balls in the sample) under $p = \frac{1}{3}$ than under $p = \frac{1}{2}$. Therefore, if we observer $X = 0$, it is plausible to take $p = \frac{1}{3}$ and the maximum likelihood estimate (MLE) of p is $\frac{1}{3}$. Similarly, the second column shows it is more likely to occur $X = 1$ under $p = \frac{1}{2}$ than under $p = \frac{1}{3}$. Therefore, the MLE of p is $\frac{1}{2}$. Similarly, from

third column, we observe that $P(X = 2)$ is maximum if $p = \frac{1}{2}$, and hence, MLE of $p = \frac{1}{2}$. Therefore, the MLE of p is

$$\widehat{p} = \begin{cases} \frac{1}{3} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1,\, 2. \end{cases}$$

Note that if $X = 0$ occur, it is more likely that there are lesser number of black balls, and hence, the estimate turns out to be 1:2. For other values of X, it is 1:1. ||

Motivated by the previous example, we have following definitions.

Definition 2.17 (Likelihood Function)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a RS from a population with PMF/PDF $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$. The function*

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

considered as a function of $\boldsymbol{\theta} \in \Theta$ for any fixed $\boldsymbol{x} \in \mathscr{X}$ ($\mathscr{X}$ is support of the RS, which is also called sample space of the RS), is called the likelihood function.

Definition 2.18 (Maximum Likelihood Estimator)**.** *For a sample point $\boldsymbol{x} \in \mathscr{X}$, let $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a value in Θ at which $L(\boldsymbol{\theta}, \boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. Then MLE of the parameter $\boldsymbol{\theta}$ based on a RS $\boldsymbol{X}$ is $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$.*

Unlike MME, by definition, MLE always lies in the parametric space. Moreover, the problem of finding MLE boils down to finding maxima of likelihood function. For finding maxima, we can use any method that is applicable for a particular problem. For example, if $L(\boldsymbol{\theta}, \boldsymbol{x})$ is twice differentiable, then one can find $\widehat{\boldsymbol{\theta}}$ using simple calculus. In regular cases, we can equivalently maximize the log-likelihood function $l(\boldsymbol{\theta}, \boldsymbol{x}) = \log L(\boldsymbol{\theta}, \boldsymbol{x})$, as $\log(\cdot)$ is an strictly increasing function. In such cases, we can find MLE by solving

$$\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}, \boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}, \boldsymbol{x}) = 0 \tag{2.2}$$

simultaneously. The equation (2.2) is called likelihood equation.

Now onwards, for brevity, we will write $L(\boldsymbol{\theta})$ instead of $L(\boldsymbol{\theta}, \boldsymbol{x})$ if not special emphasis is needed on $\boldsymbol{x}$. Similarly, we will use $l(\boldsymbol{\theta})$.

Example 2.46. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P(\lambda)$, where $\lambda > 0$. For $\lambda > 0$, the likelihood function for λ is

$$L(\lambda, \boldsymbol{x}) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} (x_i!)}.$$

Therefore, the log-likelihood function is

$$l(\lambda, \boldsymbol{x}) = \ln L(\lambda, \boldsymbol{x}) = -n\lambda + n\overline{x} \ln \lambda - \sum_{i=1}^{n} \ln(x!).$$

$\frac{dl}{d\lambda} = 0 \implies \lambda = \overline{x}$. Also $\frac{d^2 l}{d\lambda^2} < 0$ for all $\lambda > 0$. Hence $l(\lambda, \boldsymbol{x})$ maximizes at $\lambda = \overline{x}$ and the MLE of λ is $\widehat{\lambda} = \overline{X}$. ||

Example 2.47. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$. The likelihood function is

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right].$$

Now, the maximization of $L(\mu)$ is equivalent to minimization of $\sum_{i=1}^{n}(x_i - \mu)^2$ over $\mu \in \mathbb{R}$. It is known that $\sum_{i=1}^{n}(x_i - \mu)^2$ attains it's minimum at $\mu = \bar{x}$. Therefore, the MLE of μ is $\widehat{\mu} = \overline{X}$. ||

Example 2.48. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. The log-likelihood function is

$$l(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Now, we need to simultaneously solve $\dfrac{\partial}{\partial\mu} l(\mu, \sigma) = 0$ and $\dfrac{\partial}{\partial\sigma^2} l\left(\mu, \sigma^2\right) = 0$ for μ and σ^2

$$\frac{\partial l}{\partial \mu} = 0 \implies \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \implies \mu = \bar{x},$$
$$\frac{\partial l}{\partial \sigma^2} = 0 \implies -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \implies \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We need to find the Hessian matrix evaluated at $(\widehat{\mu}, \widehat{\sigma}^2) = \left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ to check if the likelihood function attains its maximum at $(\mu, \sigma^2) = (\widehat{\mu}, \widehat{\sigma}^2)$. It is easy to see that

$$\left.\frac{\partial^2}{\partial\mu^2} l\left(\mu, \sigma^2\right)\right|_{(\widehat{\mu}, \widehat{\sigma}^2)} = -\frac{n}{2\widehat{\sigma}^2},$$
$$\left.\frac{\partial^2}{\partial(\sigma^2)^2} l\left(\mu, \sigma^2\right)\right|_{(\widehat{\mu}, \widehat{\sigma}^2)} = -\frac{n}{2\widehat{\sigma}^4},$$
$$\left.\frac{\partial^2}{\partial\mu\partial\sigma^2} l\left(\mu, \sigma^2\right)\right|_{(\widehat{\mu}, \widehat{\sigma}^2)} = 0.$$

Thus, the Hessian matrix evaluated at $(\widehat{\mu}, \widehat{\sigma}^2)$ is

$$H = \begin{pmatrix} -\frac{n}{2\widehat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\widehat{\sigma}^4} \end{pmatrix}.$$

If the Hessian matrix evaluated at $(\widehat{\mu}, \widehat{\sigma}^2)$ is negative definite, then the likelihood function attains its' maximum at $(\widehat{\mu}, \widehat{\sigma}^2)$. As the first diagonal is negative and determinant is positive, H is negative definite and likelihood function attains its maximum at $(\widehat{\mu}, \widehat{\sigma}^2)$. Thus, the MLE of μ and σ^2 are $\overline{X}$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, respectively. ||

Example 2.49. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$, where $\sigma > 0$. The log-likelihood is

$$l\left(\sigma^2\right) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2.$$

Thus,

$$\frac{\partial}{\partial\sigma^2}l\left(\sigma^2\right)=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}x_i^2 \qquad \text{and} \qquad \frac{\partial^2}{\partial(\sigma^2)^2}l\left(\sigma^2\right)=\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}x_i^2.$$

Now, $\frac{\partial}{\partial\sigma^2}l\left(\sigma^2\right)=0$ implies that $\sigma^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2=\widehat{\sigma}^2$, say. Moreover,

$$\frac{\partial^2}{\partial\left(\sigma^2\right)^2}l\left(\sigma^2\right)\Big|_{\widehat{\sigma}^2}=-\frac{n}{2\widehat{\sigma}^2}<0.$$

Therefore, the MLEs of σ^2 is $\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}X_i^2$. ||

Remark 2.8. Note that the estimator of σ^2 are different in the last two examples. It shows that the MLE may update itself based on any available information on the parameters. †

Example 2.50. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, $\mu \leq 0$. Thus, the parametric space is $\Theta=(-\infty, 0]$. For $\mu \leq 0$, the log-likelihood function is

$$l(\mu)=C-\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2,$$

where C is a constant, which does not depend on μ. Now, we need to find the point in Θ at which the log-likelihood function attains its maximum. Note that $\frac{d}{d\mu}l\left(\mu\right)=n(\overline{x}-\mu)$. Clearly, for $\overline{x}>0$, $\frac{d}{d\mu}l\left(\mu\right)=0$ does not possess a solution in the parametric space. However, if $\overline{x}>0$, $\frac{d}{d\mu}l\left(\mu\right)>0$ for all $\mu \leq 0$. Hence, for $\overline{x}>0$, $l(\mu)$ is an increasing function and it takes it maximum value at $\mu=0$. On the other hand, if $\overline{x}\leq 0$, $\frac{d}{d\mu}l\left(\mu\right)=0$ possesses a solution and it is $\mu=\overline{x}$. Moreover, $\frac{d^2}{d\mu^2}l\left(\mu\right)=-n$, which is negative for all values of $\mu \leq 0$. Hence, the MLE of μ is

$$\widehat{\mu}=\begin{cases}\overline{X} & \text{if } \overline{X}\leq 0\\ 0 & \text{otherwise.}\end{cases}$$ ||

Example 2.51. Let X_1 be a sample of size one from $Bernoulli(\frac{1}{1+e^{\theta}})$, where $\theta \geq 0$. In this case $L(\theta, 0)=\frac{e^{\theta}}{1+e^{\theta}}$ and $L(\theta, 1)=\frac{1}{1+e^{\theta}}$. Clearly, MLE does not exist for $x=0$ as $L(\theta, 0)$ is a increasing function of θ. On the other hand, MLE exist for $x=1$, the likelihood function is an decreasing function. Therefore, MLE exists for $x=1$ and it is $\widehat{\theta}=0$. This example shows that there are situations when MLE does not exist. ||

Example 2.52. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta>0$. The likelihood function is

$$\begin{aligned}L(\theta)&=\frac{1}{\theta^n} \quad \text{if } 0<x_1, \ldots, x_n\leq\theta\\ &=\frac{1}{\theta^n} \quad \text{if } \theta\geq x_{(n)}=\max\left\{x_1, x_2, \ldots, x_n\right\}.\end{aligned}$$

Clearly, $L(\theta)$ is a decreasing function on $\theta \geq x_{(n)}$ and it takes it maximum value at $\theta=x_{(n)}$. Hence, the MLE of θ is $\widehat{\theta}=X_{(n)}$. ||

Example 2.53. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, $\theta \in \mathbb{R}$. The likelihood function is

$$\begin{aligned} L(\theta) &= 1 \text{ if } \theta - \frac{1}{2} \leq x_1, \ldots, x_n \leq \theta + \frac{1}{2} \\ &= 1 \text{ if } x_{(n)} - \frac{1}{2} \leq \theta \leq x_{(1)} + \frac{1}{2}, \end{aligned}$$

where $x_{(n)} = \max\{x_1, \ldots, x_n\}$ and $x_{(1)} = \min\{x_1, \ldots, x_n\}$. As $X_{(n)} - X_{(1)} \leq 1$ with probability one, $\left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$ is a non-empty interval. Also $L(\theta)$ maximizes at any point in the interval. Hence, any point in the interval

$$\left[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}\right]$$

is a MLE of θ. In particular, a MLE of θ is $\widehat{\theta} = \alpha\left(X_{(n)} - \frac{1}{2}\right) + (1-\alpha)\left(X_{(1)} + \frac{1}{2}\right)$ for any value of $\alpha \in [0, 1]$. This example shows that MLE may not be unique. ||

Theorem 2.14 (Invariance Property of MLE)**.** *If $\widehat{\boldsymbol{\theta}}$ is MLE of $\boldsymbol{\theta}$, then for any function $\tau(\cdot)$ defined on Θ, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\widehat{\boldsymbol{\theta}})$.*

Proof: The proof of above theorem is straight forward for a strictly monotone function $\tau(\cdot)$. However, the proof is little involved for a general function and, therefore, skipped here. □

Example 2.54. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P(\lambda)$, $\lambda > 0$. To find the MLE of $P(X_1 = 0)$, we can proceed as follows. Note that $P(X_1 = 0) = e^{-\lambda}$ and we know that the MLE of λ is $\overline{X}$. Hence, the MLE of $P(X_1 = 0)$ is $e^{-\overline{X}}$. ||

Theorem 2.15. *Let $\boldsymbol{T}$ be a sufficient statistics for $\boldsymbol{\theta}$. If a unique MLE exist for $\boldsymbol{\theta}$, it is a function of $\boldsymbol{T}$. If MLE of $\boldsymbol{\theta}$ exist but is not unique, then one can find a MLE that is a function of $\boldsymbol{T}$ only.*

Proof: Using Theorem 2.1,

$$L(\boldsymbol{\theta}) = h(x)\, g_{\boldsymbol{\theta}}(\boldsymbol{T}).$$

This shows that maximization of $L(\boldsymbol{\theta})$ boils down to maximization of the function $g_{\boldsymbol{\theta}}(\boldsymbol{T})$. If a unique MLE $\widehat{\boldsymbol{\theta}}$ exists that maximizes $L(\boldsymbol{\theta})$, it also maximizes $g_{\boldsymbol{\theta}}(\boldsymbol{T})$ and hence, $\widehat{\boldsymbol{\theta}}$ is a function of $\boldsymbol{T}$. If MLE of $\boldsymbol{\theta}$ is not unique, we can choose a particular MLE $\widehat{\boldsymbol{\theta}}$ form the set of all MLEs, which is a function of $\boldsymbol{T}$ only. □

Example 2.55. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We know that the MLE is unique and $X_{(n)}$, which is also sufficient. Thus, the unique MLE is a function of sufficient statistic in this case. ||

Example 2.56. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathbb{R}$. Using Example 2.10 a sufficient statistic for θ is $\boldsymbol{T} = (X_{(1)}, X_{(n)})$. Also, we have seen in Example 2.53 that MLE exists but is not unique. Any point in the interval $\left[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}\right]$ is a MLE of θ. Hence, $\frac{1}{2}\left(X_{(1)} + X_{(n)}\right)$ is a MLE and it is also a function of $\boldsymbol{T}$. On the other hand, $Q = \left(\sin^2 X_1\right)\left(X_{(n)} - \frac{1}{2}\right) + \left(1 - \sin^2 X_1\right)\left(X_{(1)} - \frac{1}{2}\right)$ is also a MLE but not a function of $\boldsymbol{T}$ only. ||

2.12 Criteria to Compare Estimators

We have considered two different methods of estimation. Now, a natural question is to ask: Which method provide a better estimator in a particular situation? Or in other words, if we have multiple estimator for an unknown parameter, then which one is "best"? To find the best estimator, we need to consider error that we may commit if we use an estimator to estimate a parameter. We should choose an estimator with least error. As an estimator is a is a function of a RS, the error will vary with realization of the RS. Therefore, to have a meaningful measure of an error, we should consider average of error over all possible realizations of RS. There are different measures of error. We will discuss some of them here along with some desirable properties of an estimator based on different measures of the error. In this section, we will assume that $\tau : \Theta \to \mathbb{R}$ and we are interested to estimate $\tau(\boldsymbol{\theta})$.

2.12.1 Unbiasedness, Variance, and Mean Squared Error

Definition 2.19 (Unbiased Estimator)**.** *A real valued estimator T is said to be an unbiased estimator (UE) of a parametric function $\tau(\boldsymbol{\theta})$ if $E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$. Here it is assumed that $E_{\boldsymbol{\theta}}(T)$ exists. An estimator is called biased if it is not unbiased.*

Note that $E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta})$ implies $E(T - \tau(\boldsymbol{\theta})) = 0$. Thus, unbiasedness tells us that on an average, there is no error. The average is taken over all possible realizations of the RS.

Definition 2.20 (Bias)**.** *Bias of a real valued statistic T as an estimator of $\tau(\boldsymbol{\theta})$ is defined by*

$$B_T(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(T) - \tau(\boldsymbol{\theta}) \quad \text{for all} \quad \theta \in \Theta.$$

Example 2.57. Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$. Then $\overline{X}$ is an unbiased estimator for μ. To see it, notice that, for all $\mu \in \mathbb{R}$,

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu. \qquad ||$$

Example 2.58. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We saw that the MLE of θ is $X_{(n)}$. Now, we want to check if $X_{(n)}$ is unbiased or not. First, we will find the CDF of $X_{(n)}$. Note that $F(x) = P\left(X_{(n)} \leq x\right) = 0$ for all $x \leq 0$. Similarly, $F(x) = 1$ for all $x \geq \theta$. Now, for $0 < x < \theta$,

$$F(x) = P\left(X_{(n)} \leq x\right) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = \left(\frac{x}{\theta}\right)^n.$$

Thus, the CDF of $X_{(n)}$ is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 \leq x < \theta \\ 1 & \text{otherwise.} \end{cases}$$

and the PDF of $X_{(n)}$ is

$$f(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$E_\theta(X_{(n)}) = \int_0^\theta \frac{nx^n}{\theta} dx = \frac{n}{n+1}\theta$$

for all $\theta > 0$. Hence, $X_{(n)}$ is a biased estimator for θ. The bias of $X_{(n)}$ is $B_{X_{(n)}}(\theta) = -\frac{1}{n+1}\theta$. As bias tends to zero as $n \to \infty$, we can make bias as small as we wish by taking sufficiently large sample size. It is very easy to see that $T = \frac{n+1}{n} X_{(n)}$ is an unbiased estimator of θ. ||

Example 2.59. Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$. Define $T_1 = X_1$, $T_2 = \frac{1}{2}(X_1 + X_2) \ldots, T_n = \overline{X}$. It is easy to verify that $E(T_i) = \mu$ for all $\mu \in \mathbb{R}$ and for all $i = 1, 2, \ldots, n$. Thus, T_i is an unbiased estimator of μ for all $i = 1, 2, \ldots, n$. This example shows that there may be more than one unbiased estimator for a parametric function. Which one should we prefer? We will discuss the answer to the question after the next example. ||

Example 2.60. Let X be distributed as $Bin(2, p)$, where $p \in (0, 1)$. Suppose that $\tau(p) = \frac{1}{p}$. We want to check if there is an UE for $\frac{1}{p}$. Here, we will show that the UE of $\frac{1}{p}$ does not exist. If possible, assume that there exists an UE, say $\delta(X)$ for $\frac{1}{p}$. Then $\delta(X)$ satisfies

$$E_p(\delta(X)) = \frac{1}{p} \implies \delta(0)\binom{2}{0}q^2 + \delta(1)\binom{2}{1}pq + \delta(2)\binom{2}{2}p^2 = \frac{1}{p},$$

for all $p \in (0, 1)$, where $q = 1 - p$. Now, for $p \to 0$, the left side tends to $\delta(0)$ and the right side tends to ∞. Hence, the equality cannot be true for all $p \in (0, 1)$ and UE for $1/p$ does not exist in this case. This example shows that UE may not exist for a parametric function. ||

Definition 2.21 (U-estimable Function)**.** *A parametric function $\tau(\boldsymbol{\theta})$ is called U-estimable, if there exits an UE T of $\tau(\boldsymbol{\theta})$.*

Definition 2.22 (Mean Square Error)**.** *Mean square error (MSE) of a real valued statistic T as an estimator of $\tau(\boldsymbol{\theta})$ is defined by*

$$MSE_T(\boldsymbol{\theta}) = E\left[(T - \tau(\boldsymbol{\theta}))^2\right],$$

provided the expectation exists.

Note that MSE gives us average square distance between the estimator and the true value of the parametric function. Hence, an estimator with smaller value of MSE is preferred.

Theorem 2.16. $MSE_T(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}(T) + (B_T(\boldsymbol{\theta}))^2$.

Proof:

$$\begin{aligned} MSE(T) &= E(T - \theta)^2 \\ &= E(T - E(T) + E(T) - \theta)^2 \\ &= E(T - E(T))^2 + E(E(T) - \theta)^2 + 2E((T - E(T))(E(T) - \theta)) \\ &= Var(T) + (Bias(T))^2. \end{aligned}$$

□

Corollary 2.1. *If T is an UE for θ, then $MSE_T(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}(T)$.*

Proof: The proof is straight forward from the previous theorem, as the bias of an UE is zero. □

Example 2.61 (Continuation of Example 2.59)**.** Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$ and finite variance σ^2. Let $T_1 = X_1$ and $T_i = \frac{1}{i}\sum_{j=1}^{i} X_j$ for $i = 2, 3, \ldots, n$. Then T_i is an UE for μ for all $i = 1, 2, \ldots, n$. Which one should we prefer? Note that

$$MSE(T_i) = Var(T_i) = \frac{\sigma^2}{i}$$

for $i = 1, 2, \ldots, n$. Hence, T_n has smallest MSE among these estimator and we will prefer T_n over other estimators. Note that only T_n is based on all observations. ||

Example 2.62. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $n > 1$. Then, using Example 2.48, the MLE of μ and σ^2 are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, respectively. Using the Example 2.57, $\widehat{\mu}$ is an UE for μ.

Now, note that $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}$. Hence,

$$E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = n - 1 \implies E(\widehat{\sigma^2}) = \frac{n-1}{n}\sigma^2 \quad \text{for all } \sigma > 0.$$

Thus, $\widehat{\sigma}^2$ is a biased estimator of σ^2. However, $S^2 = \frac{n}{n-1}\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is UE for σ^2. Now,

$$Var\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = 2(n-1) \implies Var(\widehat{\sigma}^2) = \frac{2n-2}{n^2}\sigma^4 \text{ and } Var(S^2) = \frac{2}{n-1}\sigma^4.$$

Hence,

$$MSE(\widehat{\sigma}^2) = Var(\widehat{\sigma}^2) + \left(Bias(\widehat{\sigma}^2)\right)^2 = \frac{2n-1}{n^2}\sigma^4 \text{ and } MSE(S^2) = Var(S^2) = \frac{2}{n-1}\sigma^4.$$

Now,

$$\frac{2}{n-1} - \frac{2n-1}{n^2} = \frac{3n-1}{n^2(n-1)} > 0 \implies MSE\left(\widehat{\sigma}^2\right) < MSE\left(S^2\right).$$

This example shows that biased estimator may have lower MSE and hence, may be preferred over an UE. ||

2.12.2 Best Unbiased Estimator

We are interested to find the "best" estimator among all UEs of a parametric function. Recall that there are situations where a parametric function does not have a UE. In such situations, looking for best unbiased estimator makes no sense. Therefore, in this subsection, we will only consider U-estimable parametric functions. How should we compare the performance of two UEs? We will use MSE to compare them. Recall that MSE of an UE is same as the variance of the UE. Thus, we have following definition.

Definition 2.23 (Uniformly Minimum Variance Unbiased Estimator)**.** *Let the set of all UEs of a parametric function $\tau(\boldsymbol{\theta})$ be denoted by $\mathcal{C}$, which is assumed to be non-empty. An estimator $T \in \mathcal{C}$ is called a uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\boldsymbol{\theta})$ if for all estimator $T^* \in \mathcal{C}$,*

$$Var_{\boldsymbol{\theta}}(T) \leq Var_{\boldsymbol{\theta}}(T^*) \quad \textit{for all } \boldsymbol{\theta} \in \Theta.$$

Theorem 2.17. *Let $X_1, X_2, \ldots, X_n$ be a RS from common PMF/PDF $f(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$. Let T be a real valued estimator with $Var_{\boldsymbol{\theta}}(T) < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Also assume that $\mathcal{U}$ be the set of all unbiased estimators of zero such that $Var_{\boldsymbol{\theta}}(U) < \infty$ for all $U \in \mathcal{U}$ and all $\boldsymbol{\theta} \in \Theta$. Then, a necessary and sufficient condition for T to be a UMVUE of its expectation $\tau(\boldsymbol{\theta})$ is that*

$$Cov_{\boldsymbol{\theta}}(T, U) = E_{\boldsymbol{\theta}}(TU) = 0 \quad \text{for all } U \in \mathcal{U} \text{ and for all } \boldsymbol{\theta} \in \Theta.$$

Proof: Necessity: Let T be a UMVUE of its expectation $\tau(\boldsymbol{\theta})$. We want to prove that $E(TU) = 0$ for all $U \in \mathcal{U}$ and $\boldsymbol{\theta} \in \Theta$. Fix $U \in \mathcal{U}$ and $\boldsymbol{\theta} \in \Theta$. Then, for arbitrary real constant λ, $T^* = T + \lambda U$ is an UE of $\tau(\boldsymbol{\theta})$, as $E(T^*) = E(T) + \lambda E(U) = \tau(\boldsymbol{\theta})$. Now, as T is a UMVUE of $\tau(\boldsymbol{\theta})$, for all $\lambda \in \mathbb{R}$,

$$Var_{\boldsymbol{\theta}}(T^*) \geq Var_{\boldsymbol{\theta}}(T) \implies \lambda^2 Var_{\boldsymbol{\theta}}(U) + 2\lambda Cov_{\boldsymbol{\theta}}(T, U) \geq 0.$$

That means that the discriminant of the quadratic equation $\lambda^2 Var_{\boldsymbol{\theta}}(U) + 2\lambda Cov_{\boldsymbol{\theta}}(T, U) = 0$ is zero. Here, the discriminant is $4\left(Cov_{\boldsymbol{\theta}}(T, U)\right)^2$, and hence, $Cov_{\boldsymbol{\theta}}(U, T) = 0$

Sufficiency: Assume that $E(TU) = 0$ for all $U \in \mathcal{U}$ and $\boldsymbol{\theta} \in \Theta$. We want to show that T is an UMVUE of its expectation $\tau(\boldsymbol{\theta})$. Let T^* be any UE of $\tau(\boldsymbol{\theta})$. If $Var_{\boldsymbol{\theta}}(T^*) = \infty$, there is nothing to prove. Hence assume that $Var_{\boldsymbol{\theta}}(T^*) < \infty$ for all $\boldsymbol{\theta} \in \Theta$. It is clear that $T - T^*$ is an UE of zero. Moreover, $Var_{\boldsymbol{\theta}}(T - T^*) = Var_{\boldsymbol{\theta}}(T) + Var_{\boldsymbol{\theta}}(T^*) - 2Cov_{\boldsymbol{\theta}}(T, T^*) < \infty$, as $Var_{\boldsymbol{\theta}}(T) < \infty$, $Var_{\boldsymbol{\theta}}(T^*) < \infty$ and $Cov_{\boldsymbol{\theta}}(T, T^*) \leq \sqrt{Var_{\boldsymbol{\theta}}(T) Var_{\boldsymbol{\theta}}(T^*)} < \infty$. Thus, $T - T^* \in \mathcal{U}$ so that

$$E_{\boldsymbol{\theta}}\left[T\left(T - T^*\right)\right] = 0 \implies E_{\boldsymbol{\theta}}\left(T^2\right) = E_{\boldsymbol{\theta}}\left(TT^*\right) \implies Var_{\boldsymbol{\theta}}(T) = Cov_{\boldsymbol{\theta}}(T, T^*),$$

as $E_{\boldsymbol{\theta}}(T) = E_{\boldsymbol{\theta}}(T^*) = \tau(\boldsymbol{\theta})$. Now,

$$Var_{\boldsymbol{\theta}}(T) = Cov_{\boldsymbol{\theta}}(T, T^*) \leq \sqrt{Var_{\boldsymbol{\theta}}(T) Var_{\boldsymbol{\theta}}(T^*)} \implies Var_{\boldsymbol{\theta}}(T) \leq Var_{\boldsymbol{\theta}}(T^*). \qquad \square$$

Remark 2.9. Note that any constant statistic is an UMVUE of its expectation, as the variance of a constant statistics is zero, which is minimum possible value of variance of any RV. Leaving this constant case, there are three cases. Case 1: No non-constant U-estimable function has a UMVUE. Case 2: Some, but not all, non-constant U-estimable function have UMVUE. Case 3: Every U-estimable function has a UMVUE. †

Theorem 2.18. *If T is UMVUE of $\tau(\boldsymbol{\theta})$, then it is the unique UMVUE of $\tau(\boldsymbol{\theta})$. Note that here unique means unique with probability one.*

Proof: We will prove this theorem by contradiction. If possible, let us assume that there exist another UMVUE T^* of $\tau(\boldsymbol{\theta})$ such that $P(T \neq T^*) > 0$. Then, $T - T^* \in \mathcal{U}$, and hence

$$Cov_{\boldsymbol{\theta}}(T, T^*) = Var_{\boldsymbol{\theta}}(T) \qquad \text{and} \qquad Cov_{\boldsymbol{\theta}}(T, T^*) = Var_{\boldsymbol{\theta}}(T^*).$$

Thus, we have $\left[Cov_{\boldsymbol{\theta}}(T, T^*)\right]^2 = Var_{\boldsymbol{\theta}}(T) Var_{\boldsymbol{\theta}}(T^*)$, which implies that $T = a + bT^*$ with probability one for some real constant a and b. Therefore,

$$Var_{\boldsymbol{\theta}}(T) = Var_{\boldsymbol{\theta}}(a + bT^*) = b^2 Var_{\boldsymbol{\theta}}(T^*) \implies b^2 = 1.$$

For $b = 1$, $T = a + T^*$. Taking expectation,

$$E_{\boldsymbol{\theta}}(T) = a + E_{\boldsymbol{\theta}}(T^*) \implies a = 0.$$

Thus, $T = T^*$ with probability one. For $b = -1$, $T = a - T^*$. Taking expectation, $a = 2\tau(\boldsymbol{\theta})$, which cannot happen as T and T^* are not function of $\boldsymbol{\theta}$. Therefore, we have $T = T^*$ with probability one. $\square$

It is, in general, quite difficult to enumerate the whole set of UEs of a parametric function in search of UMVUE. In Section 2.12.3, we will discuss Rao-Blackwell theorem, which provides a way to improve an UE based on a sufficient statistic. By improvement, we mean reduction of variance. Then, in Section 2.12.4, we will discuss several methods of finding the UMVUE of a U-estimable parametric function.

2.12.3 Rao-Blackwell Theorem

Theorem 2.19 (Rao-Blackwell Theorem)**.** *Suppose that T is an UE of a real valued parametric function $\tau(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Also assume that $\boldsymbol{U}$ is sufficient statistic for $\boldsymbol{\theta}$. Then*

(a) $W = E(T|\boldsymbol{U})$ is an UE of $\tau(\boldsymbol{\theta})$.

(b) $Var_{\boldsymbol{\theta}}(W) \leq Var_{\boldsymbol{\theta}}(T)$ for all $\boldsymbol{\theta} \in \Theta$. The equality holds if an only if $T = W$ with probability one.

Proof: (a) As $\boldsymbol{U}$ is a sufficient statistic for $\boldsymbol{\theta}$, the conditional distribution of T given $\boldsymbol{U}$ does not involve $\boldsymbol{\theta}$. Therefore, $W = E(T|U)$ is a function of RS and does not involve $\boldsymbol{\theta}$, and hence, is a statistic. Now,

$$E_{\boldsymbol{\theta}}(W) = E_{\boldsymbol{\theta}}[E(T|\boldsymbol{U})] = E_{\boldsymbol{\theta}}(T) = \tau(\boldsymbol{\theta}).$$

for all $\boldsymbol{\theta} \in \Theta$. This shows that W is an UE for $\tau(\boldsymbol{\theta})$.

(b) Note that

$$Var_{\boldsymbol{\theta}}(T) = Var_{\boldsymbol{\theta}}[E(T|\boldsymbol{U})] + E_{\boldsymbol{\theta}}[Var(T|\boldsymbol{U})] \geq Var_{\boldsymbol{\theta}}(W),$$

as $E_{\boldsymbol{\theta}}[Var(T|\boldsymbol{U})] \geq 0$. Equality holds if and only if $E_{\boldsymbol{\theta}}[Var(T|\boldsymbol{U})] = 0$, which implies and is implied by $Var(T|\boldsymbol{U}) = 0$. Thus, given $\boldsymbol{U}$, T is constant, and hence, $W = E(T|\boldsymbol{U}) = T$ with probability one. □

Example 2.63. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoilli(p)$, where $0 < p < 1$. We want to estimate $\tau(p) = p$. Note that $T = X_1$ is an UE of p. Also $U = \sum_{i=1}^{n} X_i$ is a sufficient statistic for p. Now, the support of U is $\{0, 1, \ldots, n\}$. For $u = 0$, $P(X_1 = 0|U = 0) = 1$. Thus, $E(X_1|U = 0) = 0 = \bar{x}$. For $u \in \{1, 2, \ldots, n\}$,

$$\begin{aligned}
E(T|U = u) &= 1 \times P(T = 1|U = u) + 0 \times P(T = 0|U = u) \\
&= P(T = 1|U = u) \\
&= \frac{P(X_1 = 1, \sum_{i=1}^{n} X_i = u)}{P(\sum_{i=1}^{n} X_i = u)} \\
&= \frac{P(X_1 = 1, \sum_{i=2}^{n} X_i = u - 1)}{P(\sum_{i=1}^{n} X_i = u)} \\
&= \frac{P(X_1 = 1)\, P(\sum_{i=2}^{n} X_i = u)}{P(\sum_{i=1}^{n} X_i = u)} \\
&= \frac{p \times \binom{n-1}{u-1} p^{u-1}(1-p)^{n-u}}{\binom{n}{u} p^{u}(1-p)^{n-u}} \\
&= \frac{\binom{n-1}{u-1}}{\binom{n}{u}}
\end{aligned}$$

$$= \frac{u}{n}$$
$$= \overline{x}.$$

Here, the fifth equality is obtained using independence of X_1 and $\sum_{i=2}^{n} X_i$. For sixth equality, note that $\sum_{i=1}^{n} X_i \sim Bin(n,\, p)$ and $\sum_{i=2}^{n} X_i \sim Bin(n-1,\, p)$. Thus, the Rao-Blackwellized version of an initial UE $T = X_1$ is $\overline{X}$. Note that the initial UE X_1 is naive and practically useless estimator of p. ||

Example 2.64. Let $X_1,\, X_2,\, \ldots,\, X_n \overset{i.i.d.}{\sim} Bernoilli(p)$, where $0 < p < 1$. Suppose that we want to estimate $\tau(p) = p(1-p)$. Note that $\tau(p) = P\left(X_1 = 1,\, X_2 = 0\right)$. Therefore, an UE of $p(1-p)$ is

$$T = \begin{cases} 1 & \text{if } X_1 = 1,\, X_2 = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Now, to obtain Rao-Blackwellized version of T, we need to find $E(T|U)$, where $U = \sum_{i=1}^{n} X_i$. As $P\left(T = 0|U = 0\right) = 1$, $E\left(T|U = 0\right) = 0$. For $u \in \{1,\, 2,\, \ldots,\, n\}$,

$$\begin{aligned} E\left(T|U = u\right) &= P\left(T = 1|U = u\right) \\ &= \frac{P\left(X_1 = 1,\, X_2 = 0,\, \sum_{i=1}^{n} X_i = u\right)}{P\left(\sum_{i=1}^{n} X_i = u\right)} \\ &= \frac{P\left(X_1 = 1\right) P\left(X_2 = 0\right) P\left(\sum_{i=3}^{n} X_i = u - 1\right)}{P\left(\sum_{i=1}^{n} X_i\right)} \\ &= \frac{p(1-p)\binom{n-2}{u-1} p^{u-1}(1-p)^{n-u-1}}{\binom{n}{u} p^{u}(1-p)^{n-u}} \\ &= \frac{\binom{n-2}{u-1}}{\binom{n}{u}} \\ &= \frac{u(n-u)}{n(n-1)} \\ &= \frac{n\overline{x}(1-\overline{x})}{n-1}. \end{aligned}$$

Thus, the Rao-Blackwellized version of initial UE T is $\frac{n}{n-1}\overline{X}\left(1-\overline{X}\right)$. ||

Example 2.65. Let $X_1,\, X_2,\, \ldots,\, X_n \overset{i.i.d.}{\sim} N(\mu,\, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown, but $\sigma > 0$ is known. Suppose that we consider unbiased estimation of $\tau(\mu) = \mu$. Consider $T = X_1$, which is an UE of μ. Also, take $U = \overline{X}$, which is a sufficient statistic for μ. Note that $\left(X_1,\, \overline{X}\right)$ has a bivariate normal distribution with mean vector $(\mu,\, \mu)$ and variance-covariance matrix

$$\begin{pmatrix} \sigma^2 & \frac{\sigma^2}{n} \\ \frac{\sigma^2}{n} & \frac{\sigma^2}{n} \end{pmatrix}.$$

Therefore, the conditional distribution of T given $U = u \in \mathbb{R}$ is $N\left(u,\, \frac{n-1}{n}\sigma^2\right)$. Thus, $E\left(T|U = u\right) = u$ for all $u \in \mathbb{R}$ and Rao-Blackwellized version of the initial UE T is $\overline{X}$. ||

Example 2.66. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown, but $\sigma > 0$ is known. Suppose that we want to estimate $\tau(\mu) = \mu^2$ unbiasedly. Note that $T = X_1^2 - \sigma^2$ is an UE of μ^2. Let us take $U = \overline{X}$, which is a sufficient statistic for μ. Now, for $U = u \in \mathbb{R}$,

$$\begin{aligned} E(T|U=u) &= E\left(X_1^2 - \sigma^2 | U = u\right) \\ &= E\left(X_1^2 | U = u\right) - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 + u^2 - \sigma^2 \\ &= u^2 - \frac{1}{n}\sigma^2. \end{aligned}$$

Hence, the Rao-Blackwellized version of initial UE T is $\left(\overline{X}^2 - \frac{\sigma^2}{n}\right)$. Now, note that

$$P\left(\overline{X}^2 - \frac{\sigma^2}{n} < 0\right) > 0$$

for all values of μ and σ, but μ^2 is always non-negative. Therefore, unbiasedness criteria may create an awkward estimator. ||

2.12.4 Uniformly Minimum Variance Unbiased Estimator

In this section, we will discuss the methods of finding UMVUE of a parametric function. We will discuss mainly two methods. First method is based on Cramer-Rao inequality and the second one is based on Lahmann-Scheffe theorem.

Cramer-Rao Inequality

In this subsection, we will assume that $X_1, X_2, \ldots, X_n$ is a RS from a common PMF/PDF $f(\cdot, \theta)$, where $\theta \in \Theta \subset \mathbb{R}$.

Theorem 2.20 (Cramer-Rao Inequality)**.** *Suppose that T is an unbiased estimator of a real valued parametric function $\tau(\theta)$. Assume that $\frac{d}{d\theta}\tau(\theta)$, denoted by $\tau'(\theta)$, is finite for all $\theta \in \Theta$. Then, for all $\theta \in \Theta$, under the assumptions 1 and 2 of Section 2.5, we have*

$$Var_\theta(T) \geq \frac{(\tau'(\theta))^2}{n\mathcal{I}_{X_1}(\theta)}.$$

The expression on the right hand side of the inequality is call Cramer-Rao lower bound (CRLB).

Proof: As $E_\theta(T) = \tau(\theta)$,

$$\begin{aligned} \tau'(\theta) &= \frac{d}{d\theta}\int\int\ldots\int T(x_1, x_2, \ldots, x_n)\prod_{i=1}^n f(x_i, \theta)dx_1dx_2\ldots dx_n \\ &= \int\int\ldots\int T(x_1, x_2, \ldots, x_n)\left[\frac{d}{d\theta}\prod_{i=1}^n f(x_i, \theta)\right]dx_1dx_2\ldots dx_n. \end{aligned}$$

Now, using

$$\prod_{i=1}^n f(x_i, \theta) = \exp\left[\sum_{i=1}^n \ln f(x_i, \theta)\right] \text{ for all } x_i \in \mathcal{X},$$

we obtain

$$\frac{d}{d\theta}\prod_{i=1}^{n} f(x_i,\,\theta) = \exp\left[\sum_{i=1}^{n}\ln f(x_i,\,\theta)\right]\sum_{i=1}^{n}\frac{d}{d\theta}\ln f(x_i,\,\theta)$$
$$= \left[\prod_{i=1}^{n} f(x_i,\,\theta)\right]\left[\sum_{i=1}^{n}\frac{d}{d\theta}\ln f(x_i,\,\theta)\right] \text{ for all } x_i \in \mathcal{X}.$$

Hence, $\tau'(\theta)$ can be rewritten as

$$\tau'(\theta) = \int\int\ldots\int T(x_1,\,x_2,\,\ldots,\,x_n)\left[\sum_{i=1}^{n}\frac{d}{d\theta}\ln f(x_i,\,\theta)\right]\left[\prod_{i=1}^{n} f(x_i,\,\theta)\right]dx_1dx_2\ldots dx_n$$
$$= E_\theta\left(TY\right),$$

where $Y = \sum_{i=1}^{n}\frac{d}{d\theta}\ln f(X_i,\,\theta)$. As $E_\theta(Y) = 0$, we have

$$\tau'(\theta) = E_\theta\left(TY\right) = Cov_\theta\left(T,\,Y\right).$$

Thus,

$$\left[\tau'(\theta)\right]^2 = \left[Cov_\theta\left(T,\,Y\right)\right]^2 \leq Var_\theta(T)Var_\theta(Y) \implies Var_\theta(T) \geq \frac{\left[\tau'(\theta)\right]^2}{Var_\theta(Y)} = \frac{\left[\tau'(\theta)\right]^2}{n\mathcal{I}_{X_1}(\theta)}. \quad \square$$

Remark 2.10. The Cramer-Rao inequality provides a lower bound for variance of an UE of the parametric function $\tau(\theta)$. Thus, if one can find an UE T of $\tau(\theta)$ such that $Var(T)$ equals CRLB for all $\theta \in \Theta$, then T is the UMVUE of $\tau(\theta)$. However, note that if there is an UE T of $\tau(\theta)$ such that the variance of T is greater than CRLB, we cannot decide if T is UMVUE of $\tau(\theta)$. In fact, we will discuss example, where the variance of the UMVUE is strictly greater than CRLB. †

Example 2.67. Let $X_1,\,X_2,\,\ldots,\,X_n \overset{i.i.d.}{\sim} Poi(\lambda)$, where $\lambda > 0$ is unknown parameter. Let us consider $\tau(\lambda) = \lambda$. The Fisher information is $\mathcal{I}_{X_1}(\lambda) = \frac{1}{\lambda}$. Thus, CRLB is

$$\frac{\left(\tau'(\theta)\right)^2}{n\,\mathcal{I}_{X_1}(\theta)} = \frac{\lambda}{n}.$$

On the other hand, $\overline{X}$ is an UE of λ. Note that $Var(\overline{X}) = \frac{\lambda}{n}$. Thus, variance of $\overline{X}$ coincide with CRLB. Therefore, $\overline{X}$ is UMVUE of λ. ||

Example 2.68. Let $X_1,\,X_2,\,\ldots,\,X_n \overset{i.i.d.}{\sim} N(\mu,\,\sigma^2)$, where $\mu \in \mathbb{R}$ is unknown parameter and $\sigma > 0$ is known. Consider $\tau(\mu) = \mu$. Then, $\overline{X}$ is an UE for μ. In this case Fisher information is $\mathcal{I}_{X_1}(\mu) = \frac{1}{\sigma^2}$. Therefore, CRLB is $\frac{\sigma^2}{n}$, which is same as variance of $\overline{X}$. Thus, $\overline{X}$ is the MUVUE of μ. ||

Example 2.69. Let $X_1,\,X_2,\,\ldots,\,X_n \overset{i.i.d.}{\sim} Poi(\lambda)$, where $\lambda > 0$ is unknown parameter. Take $\tau(\lambda) = e^{-\lambda}$. Then, a naive UE is

$$T = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise.} \end{cases}$$

We know that $U = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ. Following the technique used in Example 2.64, it can be shown that a improved UE is

$$W = \left(1 - \frac{1}{n}\right)^U.$$

Is W the UMVUE of $\tau(\lambda)$? Note that $U \sim Poi(n\lambda)$, and hence, the MGF of U is

$$E\left(e^{tU}\right) = \exp\left[n\lambda\left(e^t - 1\right)\right] \quad \text{for all } t \in \mathbb{R}.$$

Now,

$$\begin{aligned}
E\left(W^2\right) &= E\left[\left(1 - \frac{1}{n}\right)^{2U}\right] \\
&= E\left[\exp\left\{\ln\left(1 - \frac{1}{n}\right)^{2U}\right\}\right] \\
&= E\left[\exp\left\{2U \ln\left(1 - \frac{1}{n}\right)\right\}\right] \\
&= \exp\left[n\lambda\left(\left(1 - \frac{1}{n}\right)^2 - 1\right)\right] \\
&= e^{-\lambda\left(2 - \frac{1}{n}\right)}.
\end{aligned}$$

Thus, $Var(W) = E(W^2) - E^2(W) = e^{-\lambda\left(2-\frac{1}{n}\right)} - e^{-2\lambda} = e^{-2\lambda}\left(e^{\frac{\lambda}{n}} - 1\right)$. On the other hand, Fisher information $\mathcal{I}_{X_1}(\lambda) = \frac{1}{\lambda}$. Therefore, CRLB is $\frac{\lambda}{n}e^{-2\lambda}$. As $e^{\frac{\lambda}{n}} > 1 + \frac{\lambda}{n}$, $Var(W)$ is greater than CRLB. Thus, we cannot decide weather W is UMVUE of $\tau(\lambda)$ using CRLB. ||

Lehmann-Scheffee Theorems

Theorem 2.21 (Lehmann-Scheffe Theorem I)**.** *Suppose that T is an UE of a real valued parametric function $\tau(\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{U}$ be a complete sufficient statistic for $\boldsymbol{\theta}$. Define $g(\boldsymbol{u}) = E_{\boldsymbol{\theta}}[T|\boldsymbol{U} = \boldsymbol{u}]$ for all $\boldsymbol{u} \in \mathcal{U}$. Then, the statistic $W = g(\boldsymbol{U})$ is the unique (with probability one) UMVUE of $\tau(\boldsymbol{\theta})$.*

Proof: Let W^* be an UE of $\tau(\boldsymbol{\theta})$ and an function of $\boldsymbol{U}$ only. As W and W^* are UEs of $\tau(\boldsymbol{\theta})$, $E_{\boldsymbol{\theta}}(W - W^*) = 0$ for all $\boldsymbol{\theta} \in \Theta$. As $\boldsymbol{U}$ is complete, we have $W - W^* = 0$ with probability one, *i.e.*, $W^* = W$ with probability one. Thus, UE based on $\boldsymbol{U}$ is unique.

Now, assume that T^* is an UE of $\tau(\boldsymbol{\theta})$ (but not necessarily a function of $\boldsymbol{U}$ only) and $V = E(T^*|\boldsymbol{U})$. Then, using Rao-Blackwell theorem, V is an UE of $\tau(\boldsymbol{\theta})$ and $Var_{\boldsymbol{\theta}}(V) \leq Var_{\boldsymbol{\theta}}(T^*)$ for all $\boldsymbol{\theta} \in \Theta$. Noting that V is a function of $\boldsymbol{U}$ only, $V = W$ with probability one. Thus, $Var_{\boldsymbol{\theta}}(W) \leq Var_{\boldsymbol{\theta}}(T^*)$. This shows that W is unique UMVUE of $\tau(\boldsymbol{\theta})$. □

Theorem 2.22 (Lehmann-Scheffe Theorem II)**.** *Suppose that U is a complete sufficient statistic for $\boldsymbol{\theta} \in \Theta$. Also, suppose that a statistic $W = g(\boldsymbol{U})$ is an UE of a real valued parametric function $\tau(\boldsymbol{\theta})$. Then, W is the unique (with probability one) UMVUE of $\tau(\boldsymbol{\theta})$.*

Proof: Let T be any UE of $\tau(\boldsymbol{\theta})$. As there exists unique UE of $\tau(\boldsymbol{\theta})$ based on $\boldsymbol{U}$, $E(T|\boldsymbol{U})$ is same as W with probability one. Thus, using Rao-Blackwell theorem, $Var_{\boldsymbol{\theta}}(W) \leq Var_{\boldsymbol{\theta}}(T)$ for all $\boldsymbol{\theta} \in \Theta$. Hence, W is the UMVUE of $\tau(\boldsymbol{\theta})$. □

Example 2.70. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Poi(\lambda)$, where $\lambda > 0$ is unknown parameter. We wish to estimate $\tau(\lambda) = e^{-\lambda}$. In Example 2.69, we have seen that an UE of $\tau(\lambda)$ is

$$W = \left(1 - \frac{1}{n}\right)^U,$$

where $U = \sum_{i=1}^n X_i$ is complete sufficient statistic. Now, W is a function of U only. Therefore, using Lahmann-Scheffe theorems, W is unique UMVUE of $\tau(\lambda)$. Note that using CRLB, we cannot decide if W is UMVUE of $\tau(\lambda)$ or not, as variance of W is greater than CRLB. This is an example where variance of UMVUE is greater than CRLB. ||

Example 2.71. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown parameter, but $\sigma > 0$ is known. Consider $\tau(\mu) = \mu^2$. In Example 2.66, we found that an UE of μ^2 is

$$W = \overline{X}^2 - \frac{\sigma^2}{n}.$$

Clearly, W is a function of complete sufficient statistic $\overline{X}$ of μ. Thus, using Lehmann-Scheffe theorems, W is UMVUE of μ^2. ||

Example 2.72. Suppose that $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, where $\theta > 0$ is unknown parameter. The estimand is $\tau(\theta) = \theta$. Clearly, in this case we cannot use CRLB technique, as the distribution does not belong to regular family. However, we know that $X_{(n)}$ is complete sufficient statistic of θ. It is easy to see that the PDF of $X_{(n)}$ is

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

Therefore, $E\left[X_{(n)}\right] = \frac{n}{n+1}\theta$, which implies that $\frac{n+1}{n} X_{(n)}$ is UMVUE of θ. ||

Example 2.73. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown parameters. Suppose that we wish to estimate $\tau(\mu, \sigma) = \mu + \sigma$. We know that $\left(\overline{X}, S^2\right)$ is complete sufficient statistic for (μ, σ^2). Now, if we can find a statistic, which is a function of $\left(\overline{X}, S^2\right)$ only and an UE of $\mu + \sigma$, we are done. Note that $\overline{X}$ is an UE of μ. We will, now, try to find an UE of σ based on S. Recall that

$$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Thus, the PDF of T is

$$f(t) = \frac{1}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right)} t^{\frac{n-3}{2}} e^{-\frac{t}{2}} \quad \text{if } t > 0.$$

Therefore,

$$E(S) = E\left(\frac{\sigma\sqrt{T}}{\sqrt{n-1}}\right) = \frac{\sigma}{2^{\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right) \sqrt{n-1}} \int_0^\infty t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt = \frac{\sigma \Gamma\left(\frac{n}{2}\right) \sqrt{2}}{\Gamma\left(\frac{n-1}{2}\right) \sqrt{n-1}}.$$

This shows that $a_n S$ is an UE of σ, where

$$a_n = \frac{\sqrt{n-1}\,\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{2}\,\Gamma\left(\frac{n}{2}\right)}.$$

Thus, $W = \overline{X} + a_n S$ is an UE of $\mu + \sigma$. As W is a function of complete sufficient statistic $\left(\overline{X},\, S^2\right)$ only, using Lehmann-Scheffe theorems, W is the UMVUE of $\mu + \sigma$. ||

2.12.5 Large Sample Properties

Note that an estimator is a function of the sample size also, though we do not emphasis it earlier. In this section, we will study the effect of large sample size on an estimator. If we keep on increasing sample size, it is expected that the sample will cover almost all the population, and hence, an estimator of a parametric function should be closer to the true value of the parametric function.

To emphasis the sample size, a real values estimator calculated based on a RS of size n will be denoted by T_n in the current section. For example, if we consider the sample mean, then $T_1 = X_1$ is the estimator calculated based on a RS of size one, the estimator $T_2 = \frac{1}{2}(X_1 + X_2)$ is calculated based on a RS of size 2, so on. Therefore, we have a sequence of estimators (RVs) $\{T_n\}_{n\geq 1}$. Here, we want to study if T_n is very close to the true value of the parametric function or not when n is very large, *i.e.*, $n \to \infty$. Note that $\{T_n\}_{n\geq 1}$ is a sequence of RVs. We will consider convergence in probability in this course and we have the following definition.

Definition 2.24 (Consistent Estimator)**.** *Let T_n be an estimator based on a RS of size n. The estimator T_n is said to be consistent for θ if the sequence of RVs $\{T_n : n \geq 1\}$ converges to θ in probability for all $\theta \in \Theta$, i.e., if for all $\varepsilon > 0$ and all $\theta \in \Theta$,*

$$\lim_{n\to\infty} P\left(|T_n - \theta| \leq \varepsilon\right) = 1.$$

Remark 2.11. Consistency says us that for a sample with reasonably large size, T_n is close to the true value of parameter with high probability. †

Example 2.74. Let $X_1, X_2 \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$. Then, using WLLN, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator for μ. ||

Example 2.75. $X_1, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We saw that the MLE of θ is $X_{(n)}$. Let us see if the MLE is consistent estimator of θ. Note that the CDF of $X_{(n)}$ is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 \leq x < \theta \\ 1 & \text{if } x \geq \theta. \end{cases}$$

Thus, for $\epsilon > 0$,

$$\begin{aligned} P\left(\left|X_{(n)} - \theta\right| \leq \epsilon\right) &= P\left(\theta - \epsilon \leq X_{(n)} \leq \theta + \epsilon\right) \\ &= F\left(\theta + \epsilon\right) - F\left(\theta - \epsilon\right) \\ &= \begin{cases} 1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n & \text{if } 0 < \epsilon < \theta \\ 1 - 0 & \text{if } \epsilon \geq \theta, \end{cases} \end{aligned}$$

which converse to one for all values of θ. Therefore, $X_{(n)} \to \theta$ in probability and $X_{(n)}$ is a consistent estimator of θ. ||

Theorem 2.23 (Consistency of MLE). *Let $X_1, X_2, \ldots, X_n$ be a RS from the population having PMF/PDF $f(x;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider the following assumptions.*

1. *$\frac{\partial}{\partial\theta}\ln f(x;\theta)$, $\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)$, $\frac{\partial^3}{\partial\theta^3}\ln f(x;\theta)$ are finite for all $x \in \mathbb{R}$ and for all $\theta \in \Theta$.*
2. *$\int_{-\infty}^{+\infty}\frac{\partial}{\partial\theta}f(x;\theta)dx = 0$, $\int_{-\infty}^{+\infty}\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = 0$, and $\int_{-\infty}^{+\infty}\left\{\frac{\partial}{\partial\theta}f(x;\theta)\right\}^2 dx > 0$ for all $\theta \in \Theta$.*
3. *For all $\theta \in \Theta$, $\left|\frac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < a(x)$, where $E\left(a(X_1)\right) < b$ for a constant b which is independent of θ.*

Under these three assumptions, the likelihood equation has solution denoted by $\widehat{\theta}_n(\boldsymbol{x})$, such that $\widehat{\theta}_n(\boldsymbol{X})$ is consistent estimator of θ.

Proof: The proof is skipped here. □

Theorem 2.24 (Asymptotic Normality of MLE). *Under the three assumptions of the Theorem 2.23,*

$$\sqrt{n\mathcal{I}_{X_1}(\theta)}\left(\widehat{\theta}_n(\boldsymbol{X}) - \theta\right) \to Z$$

in distribution, where $Z \sim N(0,1)$, and $\mathcal{I}_{X_1}(\theta)$ is Fisher information based on a RS of size one.

Proof: The proof is skipped here. □

Example 2.76. Let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} Bernoulli(p)$. The MLE of p based on a sample of size n is $\widehat{p}_n = \overline{X}_n$ and $\mathcal{I}_{X_1}(p) = \frac{1}{p(1-p)}$. Using above theorems, $\widehat{p}_n$ is consistent for p and $\sqrt{n}(\widehat{p}_n - p) \to N(0, p(1-p))$ in distribution. ||

Example 2.77. Let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} P(\lambda)$. The MLE of λ based on a sample of size n is $\widehat{\lambda}_n = \overline{X}_n$ and $\mathcal{I}_{X_1}(\lambda) = \frac{1}{\lambda}$. Using above theorems, $\widehat{\lambda}_n$ is consistent for λ and $\sqrt{n}(\widehat{\lambda}_n - \lambda) \to N(0, \lambda)$ in distribution. ||

Example 2.78. Let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} U(0,\theta)$. The MLE of θ based on a sample of size n is $\widehat{\theta}_n = X_{(n)}$. Note that the first condition of assumption 2 does not hold. Hence, we cannot use previous theorems here. However, we have already discussed that $X_{(n)}$ is consistent for θ. On the other hand, one can show that $n\left(\theta - X_{(n)}\right) \to Z$ in distribution, where Z has an exponential distribution with mean θ. To see it, note that the CDF of $T_n = n\left(\theta - X_{(n)}\right)$ is

$$\begin{aligned}
F_{T_n}(t) &= P\left(n\left(\theta - X_{(n)}\right) \le t\right) \\
&= P\left(X_{(n)} \ge \theta - \frac{t}{n}\right) \\
&= 1 - F_{X_{(n)}}\left(\theta - \frac{t}{n}\right) \\
&= \begin{cases} 1-0 & \text{if } \theta - \frac{t}{n} < 0 \\ 1 - \left(\frac{\theta - \frac{t}{n}}{\theta}\right)^n & \text{if } 0 \le \theta - \frac{t}{n} < \theta \\ 1-1 & \text{if } \theta - \frac{t}{n} \ge \theta \end{cases}
\end{aligned}$$

$$= \begin{cases} 0 & \text{if } t \leq 0 \\ 1 - \left(1 - \frac{t}{n\theta}\right)^n & \text{if } 0 < t \leq n\theta \\ 1 & \text{if } t > n\theta. \end{cases}$$

Now, for $t \leq 0$, $F_{T_n}(t)$ converges to zero. For $t > 0$, we can find n large enough so that $t \leq n\theta$. Therefore, for $t > 0$, $F_{T_n}(t) \to 1 - e^{-\frac{t}{\theta}}$. Hence, as $n \to \infty$

$$F_{T_n}(t) \to F(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 - e^{-\frac{t}{\theta}} & \text{if } t > 0, \end{cases}$$

where $F(\cdot)$ is the CDF of an exponential RV with expectation θ. ||

Chapter 3

Tests of Hypotheses

3.1 Introduction

In the previous section, we have discussed point estimation, where we try to find meaningful guesses for unknown parameters or parametric functions. In testing of hypothesis, we do not guess the value of the parametric function. We try to check if a given statement about parameters is true or not. Let us start with examples.

Example 3.1. The Cherry Blossom Run is a 10 mile race that takes place every year in D.C. In 2009, there were 14974 participants and average running time of all the participants was 103.5 minutes. Now, in the year 2010, one may want to ask the question: Were runners faster in 2010 compared to 2009? Of course, if we have running times of all the participant, we can find the average running time on the year 2010 and compare it with average running time of the year 2009. However, assume that it is not possible to have the running times of all the participants in 2010. In this case, how should we proceed and answer the question? Again, we will rely on a RS, say of size n, drawn from the 2010 runners, and denote the running time by $X_1, X_2, \ldots, X_n$. Based on historic data, it my be plausible to assume that the distribution of the running time is a normal with variance 373. Hence, we are given a RS $X_1, X_2, \ldots, X_n$ and we want to know if $X_1 \sim N(103.5, 373)$. This is a hypothesis about the distribution of running time and we want to test the hypothesis. There are many ways this hypothesis could be false:

- $E(X_1) \neq 103.5$
- $Var(X_1) \neq 373$
- X_1 is not normal.

From the analysis of the past data, it is found that the last two assumptions are reasonable and hence we put them as model assumptions. Thus, the only way in which the hypothesis $X_1 \sim N(103.5, 373)$ could be false is $\mu = E(X_1) \neq 103, 5$. Using modeling assumptions, we have reduced the number of ways the hypothesis $X_1 \sim N(103.5, 373)$ may be false. Now, we want to test: Is $\mu = 103.5$ or $\mu < 103.5$? Note the we have written $\mu < 103.5$, as our initial question was: Is the runner faster in the year 2010? Thus, μ must be less than or equal to 103.5. Now, the only way the hypothesis can be false is if $X_1 \sim N(\mu, 373)$ for some $\mu < 103.5$. This is an example, where we compare an expected value to a fixed reference number (here 103.5).

Simple heuristic would be: If $\overline{X} < 103.5$, then $\mu < 103.5$, where $\overline{X}$ is the sample mean based on a RS of size n drawn form runners of the year 2010. It is easy to understand that it can go wrong if we select, by chance, the fast runners in the sample. Better heuristic could be: If $\overline{X} < 103.5 - a$ then $\mu < 103.5$ for some a. We will try to make this intuitions more precise as we proceed. ||

Example 3.2. Pharmaceutical companies use hypothesis testing to test if a new drug is efficient. To do so, a group of patients are randomly divided into two groups. One of the groups is administered with the drug and the other is administered with placebo. The first and the second groups are called test group and control group, respectively. Assume that the drug is a cough syrup. Let μ_1 denote the expected number of expectorations per hour after a patient has used placebo and μ_2 denote the expected number of expectorations per hour after a patient has used the syrup. We want to know if $\mu_2 < \mu_1$. In this case, two expectations are compared. One of the method to attack this problem is to draw RS from both the groups. Let $X_1, X_2, \ldots, X_{n_1}$ denote a RS of size n_1 from the control group. Let $Y_1, \ldots, Y_{n_2}$ denote a RS of size n_2 drawn from test group. We want to test if $\mu_2 = \mu_1$ or $\mu_2 < \mu_1$. If $\mu_1 = \mu_2$, then the new drug is not efficient. If $\mu_2 < \mu_1$, then the new drug has some effect. Note that we have taken $\mu_2 < \mu_1$ and we have not considered if $\mu_2 > \mu_1$. The reason for the same is as follows: As we are trying to check the efficiency of the new drug, we implicitly assume that $\mu_2 \leq \mu_1$. Heuristically, we should compare $\overline{X}$ and $\overline{Y}$. ||

Example 3.3. Let a coin is tossed 80 times, and head are obtained 55 times. Can we conclude that the coin is fair based on this data? Let, for $i = 1, 2, \ldots, 80$, X_i be an indicator RV, which takes value 1 if ith toss is a head and takes value zero if the ith toss is a tail. Then, we have a RS $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$, where p is the probability of getting a head in a toss. We want to test $p = 0.5$ or $p \neq 0.5$. Intuitively, it makes sense to use $\overline{X}$ to check if $p = 0.5$ or not. Here, the observed value of $\overline{X}$ is $\overline{x} = 55/80 = 0.6875$. If p is actually equal to 0.5, then, using CLT, we have

$$T_n = \frac{\sqrt{n}\left(\overline{X}_n - 0.5\right)}{\sqrt{0.5 \times (1 - 0.5)}} \approx N(0, 1).$$

Now, if the number of heads is too small or too large (*i.e.*, the value of $\overline{x}$ is not close to 0.5), we should go for biased coin. If the number of head is moderate (*i.e.*, the value of $\overline{x}$ is close to 0.5), we should choose that the coin is fair. In the first case ($\overline{x}$ not close to 0.5), the absolute observed value of T_n will be large. The absolute observed value of T_n will be close to zero in the second case. This discussion suggests that we should reject the fact that the coin is fair if $|T_n| > C$ for some appropriate real constant C.

Here, the observed value of T_n is 3.3541, which is too extreme with respect to a standard normal distribution as $P(|Z| > 3.35) \approx 0.0008$, where $Z \sim N(0, 1)$. Therefore, it is quite reasonable to reject the hypothesis $p = 0.5$ based on the data. ||

Example 3.4. A coin is tossed 80 times, and head are obtained 35 times. Can we conclude that the coin is significantly fair? Here, the observed value of T_n is -1.1180. Data do not suggest to reject the fact that the coin is fair, as the observed value of T_n is not extreme with respect to a standard normal distribution. Note that $P(|Z| > 1.11) = 0.267$. ||

In the last two examples, we have talked about extreme or not extreme. The question is: Which values are considered as extreme and which are not? More precisely, we are rejecting $p = 0.5$ if the observation belong to the set

$$\{\boldsymbol{x} : |T_n(\boldsymbol{x})| > C\}.$$

What value of C should we choose so that we can make correct decision? This issues will be discussed as we proceed in the current chapter.

Definition 3.1 (Statistical Hypothesis)**.** *A statistical hypothesis or simply hypothesis is a statement about the unknown parameters.*

Definition 3.2 (Null and Alternative Hypotheses)**.** *Suppose that one wants to choose between two reasonable hypotheses $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ and $\Theta_1 \cap \Theta_2 = \emptyset$. We call $H_0 : \boldsymbol{\theta} \in \Theta_0$ and $H_1 : \boldsymbol{\theta} \in \Theta_1$ as null hypothesis and alternative hypothesis, respectively.*

Definition 3.3 (Simple and Composite Hypothesis)**.** *A hypothesis is called simple if it specifies the underlying distribution. Otherwise, it is called composite hypothesis.*

Our aim is to choice one hypothesis among null and alternative hypotheses. As we will see that the roles of these two hypotheses are asymmetric, we need to be careful while labeling these two hypotheses.

3.2 Errors and Errors Probabilities

As illustrated in Examples 3.3 and 3.4, the decision to accept or reject null hypothesis will be taken based on a RS. We will consider a reasonable statistic and make the choice based on the statistic. If the observed value of the statistic belongs to an appropriate set, we reject null hypothesis and if the value of the statistic does not belong to the set, we accept null hypothesis. Now, the statistic belongs to a set can be alternatively written as the sample observation belongs to a subset of $\mathbb{R}^n$. In Examples 3.3 and 3.4, suppose that we reject the null hypothesis if and only if $|T_n| > 1.96$. The set $\{|T_n| > 1.96\}$ is the set of all sample points $\boldsymbol{x}$ such that $T_n(\boldsymbol{x}) > 1.96$ or $T_n(\boldsymbol{x}) < -1.96$. Thus, the condition $|T_n| > 1.96$ actually divides $\mathbb{R}^n$ into two parts, *viz.*, $\{\boldsymbol{x} \in \mathbb{R}^n : |T_n(\boldsymbol{x})| > 1.96\}$ and $\{\boldsymbol{x} \in \mathbb{R}^n : |T_n(\boldsymbol{x})| \leq 1.96\}$. If the sample observation $\boldsymbol{x}$ belongs to the first partition, we reject the null hypothesis. If the sample observation $\boldsymbol{x}$ belongs the second partition, we accept the null hypothesis. Of course, we need to find a "meaningful" partition so that we can make correct decision.

Definition 3.4 (Critical Region and Acceptance Region)**.** *Suppose that we want to test $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$ based on a RS of size n drawn from a population having PMF/PDF $f(\cdot, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1$. Let R be a subset of $\mathbb{R}^n$ such that we reject H_0 if and only if $\boldsymbol{x} \in R$, where $\boldsymbol{x}$ denotes a realization of the RS. Then, R is called rejection region or critical region, and R^c is called acceptance region.*

In this process, there are four possibility as described in the following table. There are two cases, where we do not commit any error. These cases are accepting null hypothesis when it is actually true and rejecting null hypothesis when it is actually false. The green ticks in the following table signify that there are no errors. We commit errors in other two cases, *viz.*, accepting null hypothesis when it is actually false or rejecting null hypothesis when it is actually true.

	H_0 true	H_1 true
Accept H_0	✓	Type-II Error
Reject H_0	Type-I Error	✓

Definition 3.5 (Type-I and Type-II Errors)**.** *The error committed by rejecting H_0 when it is actually true is called Type-I Error. The error committed by accepting H_0 when it is actually false is called Type-II Error.*

Example 3.5. Let $X_1, X_2, \ldots, X_9 \overset{i.i.d.}{\sim} N(\theta, 1)$. Suppose that we are want to test $H_0 : \theta = 5.5$ against $H_1 : \theta = 7.5$. For comparison purpose, let us consider four critical regions:

$$R_1 = \emptyset, \quad R_2 = \left\{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 7\right\}, \quad R_3 = \left\{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 6\right\}, \quad R_4 = \mathbb{R}^9.$$

As we are trying to test hypotheses regarding population mean, it is intuitively make sense to use sample mean. That is the reason to take R_2 and R_3 in terms of sample mean $\bar{x}$. On the other hand, R_1 and R_4 are two extremes. The critical regions R_1 accepts null hypothesis irrespective of the realization of the RS. Similarly, we always reject the null hypothesis if the critical region is R_4.

Note that Type-I or Type-II errors are events. Thus, we may talk about the probabilities of these errors. For critical region R_3,

$$\begin{aligned} P\left(\text{Type-I Error}\right) &= P_{\theta=5.5}\left(\overline{X} > 6\right) = 1 - \Phi\left(3(6 - 5.5)\right) = 0.06681 \\ P\left(\text{Type-II Error}\right) &= P_{\theta=7.5}\left(\overline{X} \leq 6\right) = \Phi\left(3(6 - 7.5)\right) \sim 0. \end{aligned}$$

Similarly, the probabilities of errors for other critical regions can be calculated and given in following table.

	R_1	R_2	R_3	R_4
P(Type-I)	0	~ 0	0.06681	1
P(Type-II)	1	0.06681	~ 0	0

In this example, $R_1 \subset R_2 \subset R_3 \subset R_4$. Notice that as we increase the size of the critical region, probability of Type-I error increases and that of Type-II error decreases. In other words, if we try to reduce probability of one error, probability of the other one increases. ||

Definition 3.6 (Power Function)**.** *The power function of a critical region, denoted by $\beta : \Theta_1 \cup \Theta_0 \to [0, 1]$, is the probability of rejecting the null hypothesis H_0 when $\boldsymbol{\theta}$ is the true value of the parameter, i.e.,*

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\boldsymbol{X} \in R).$$

It is clear that $\beta(\cdot)$ is the probability of Type-I error if $\boldsymbol{\theta} \in \Theta_0$. For $\boldsymbol{\theta} \in \Theta_1$, $\beta(\cdot)$ is one minus probability of Type-II error.

Example 3.6. Let $X_1, X_2, \ldots, X_9 \overset{i.i.d.}{\sim} N(\theta, 1)$. Suppose that we are want to test $H_0 : \theta = 5.5$ against $H_1 : \theta = 7.5$. The power function of the critical region $R_2 = \{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 7\}$ is

$$\beta(\theta) = P_\theta\left(\boldsymbol{X} \in R\right) = P_\theta\left(\overline{X} > 7\right) = P_\theta\left(3\left(\overline{X} - \theta\right) > 21 - 3\theta\right) = 1 - \Phi(21 - 3\theta).$$

for $\theta = 5.5$ and 7.5. ||

3.3 Best Test

Note that we want to find a "meaningful" partition. As mentioned, "meaningful" means that we take correct decisions by rejecting (accepting) the null hypothesis when it is actually false (true). That means that we want $\boldsymbol{x}$ to be in R ($\boldsymbol{x} \notin R$) when null hypothesis is false (true). A critical region R which minimizes the probabilities of both the errors could be a "meaningful" choice. Unfortunately, as shown in the previous example, the reduction of probability of one type of error forces to increase the probability of other type of error, in general. Optimization in such a situation can be done in several ways. For tests of hypotheses, the method is as follows: Put an upper bound on the probability of Type-I error and try to minimize the probability of Type-II error subject to the upper bound of the probability of Type-I error.

Definition 3.7 (Size of a Test)**.** *Let $\alpha \in (0, 1)$ be a fixed real number. A test for $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$ with power function $\beta(\cdot)$ is called a size α test if*

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \beta(\boldsymbol{\theta}) = \alpha.$$

Definition 3.8 (Level of a Test)**.** *A test is called level α if $\beta(\boldsymbol{\theta}) \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$.*

Size of a test can be considered as worst possible probability of Type-I error. If a test is of size α, then it is of level α.

Example 3.7. Let $X_1, X_2, \ldots, X_9 \overset{i.i.d.}{\sim} N(\theta, 1)$. Suppose that we are want to test $H_0 : \theta = 5.5$ against $H_1 : \theta = 7.5$. We have seen that the power function of the critical region $R_2 = \{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 7\}$ is

$$\beta(\theta) = 1 - \Phi(21 - 3\theta).$$

for $\theta = 5.5$ and 7.5. In this case, $\Theta_0 = \{5.5\}$ is single-tone. Therefore, the size of the test is

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(5.5) = 1 - \Phi(4.5) \simeq 3.4 \times 10^{-6}.$$

This test is of level α for any $\alpha \in [1 - \Phi(4.5), 1]$. ||

Definition 3.9 (Critical or Test Function)**.** *A function $\psi : \mathscr{X}^n \to [0, 1]$ is called a critical function or test function, where $\psi(\boldsymbol{x})$ stands for the probability of rejecting H_0 when $\boldsymbol{X} = \boldsymbol{x}$ is observed. Here, $\mathscr{X}^n$ is the sample space of the random sample of size n.*

Example 3.8. Let $X_1, X_2, \ldots, X_9 \overset{i.i.d.}{\sim} N(\theta, 1)$. Suppose that we are want to test $H_0 : \theta = 5.5$ against $H_1 : \theta = 7.5$. Let us consider two critical regions $R_1 = \{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 6\}$ and $R_2 = \{\boldsymbol{x} \in \mathbb{R}^9 : \bar{x} > 7\}$. The critical regions R_1 and R_2, respectively, can be expressed as the test functions

$$\psi_1(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \bar{x} > 6 \\ 0 & \text{if } \bar{x} \leq 6, \end{cases} \qquad \text{and} \qquad \psi_2(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \bar{x} > 7 \\ 0 & \text{if } \bar{x} \leq 7. \end{cases}$$

Note that the power function for R_1 is

$$\beta_1(\theta) = P_\theta(\overline{X} > 6) = 1 - \Phi(18 - 3\theta) \text{ for } \theta = 5.5,\ 7.5,$$

which can be expressed as $E_\theta(\psi_1(\boldsymbol{X}))$. ||

Thus, the last example shows that the test function is an alternative way to write a critical region. Then, what does we gain by defining test function? First note that $\psi(\boldsymbol{x})$ being a probability, can take any value between zero and one (not only values 0 and 1). This is the gain. We will illustrate it in Example 3.9. However, before going into the example, let us discuss couple of definitions in terms of test function.

Definition 3.10 (Power Function)**.** *The power function of a test function is defined by*

$$\beta(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\psi(\boldsymbol{X})\right) \text{ for all } \boldsymbol{\theta} \in \Theta_0 \cup \Theta_1.$$

Once the power function is defined, we can now define size or level of a test using the power function as given in Definitions 3.7 and 3.8.

Definition 3.11 (Randomized and Non-randomized Tests)**.** *A test is called randomized test if $\psi(\boldsymbol{x}) \in (0, 1)$ for some $\boldsymbol{x}$. Otherwise, it is called a non-randomized test.*

Example 3.9. Let X be a sample of size one form a $Bin(3, p)$ distribution. We want to test $H_0 : p = \frac{1}{4}$ against $H_1 : p = \frac{3}{4}$. The probabilities of $X = x$ under H_0 is given in the table below:

x	Prob. under H_0
0	27/64
1	27/64
2	9/64
3	1/64

Do we have a critical region of size $\alpha_1 = \frac{5}{32}$? The answer is yes, and the critical region is given by $\{2, 3\}$ as $P(X = 2 \text{ or } 3) = \frac{5}{32}$ under H_0.

Does a critical region of size $\alpha_2 = \frac{1}{32}$ exist? The answer is no, there is no critical region of size $\frac{1}{32}$. However, we have a randomized test of of size $\frac{1}{32}$, and it is given by

$$\psi(x) = \begin{cases} 1 & \text{if } x = 3 \\ \frac{1}{9} & \text{if } x = 2 \\ 0 & \text{otherwise,} \end{cases}$$

as $E_{p=\frac{1}{4}}\left(\psi(X)\right) = 1 \times \frac{1}{64} + \frac{1}{9} \times \frac{9}{64} = \frac{1}{32}$. Hence, in this case though a critical region of size 1/32 does not exist, a randomized test function of the same size exists. This is the gain of defining a test function over critical region. ||

Remark 3.1. Test functions are more general in the sense that all critical regions can be represented as a test function, but the converse is not true. †

Remark 3.2. Let for a fixed $\boldsymbol{x}_0$, $\psi(\boldsymbol{x}_0) = 0.6$. If $\boldsymbol{X} = \boldsymbol{x}_0$ is observed, how should we accept or reject H_0? We will perform a random experiment with two outcomes (toss of a coin), with one (say head) has probability 0.4, and other (say tail) has probability 0.6. If tail occur, we reject H_0, otherwise we accept it. †

Definition 3.12 (Uniformly Most Powerful Test)**.** *Consider the collection $\mathscr{C}_\alpha$ of all level α tests for $H_0 : \boldsymbol{\theta} \in \Theta_0$ against $H_1 : \boldsymbol{\theta} \in \Theta_1$. A test belonging to $\mathscr{C}_\alpha$ with power function $\beta(\cdot)$ is called uniformly most powerful (UMP) level α test if $\beta(\boldsymbol{\theta}) \geq \beta^*(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta_1$, where $\beta^*(\cdot)$ is the power function of any other test in $\mathscr{C}_\alpha$. If the alternative hypothesis is simple (that means that Θ_1 is singleton), the test is called most powerful (MP) level α test.*

Remark 3.3. Note that here we are putting a bound on probability of type one error. The bound is α. Among all the tests whose probability of Type-I error is bounded by α, we are trying to find one for which probability of Type-II error is minimum. A test satisfies this criterion is called a UMP level α test. †

Remark 3.4. When $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ for some fixed $\boldsymbol{\theta}_1$, *i.e.*, H_1 is simple, it boils down to check if $\beta(\boldsymbol{\theta}_1) \geq \beta^*(\boldsymbol{\theta}_1)$. Hence, the word 'uniformly' is removed. †

3.4 Simple Null Vs. Simple Alternative

Theorem 3.1 (Neyman-Pearson Lemma)**.** *Let $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$ be two fixed numbers in Θ. The MP level α test for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ is given by*

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } L(\boldsymbol{\theta}_1) > kL(\boldsymbol{\theta}_0) \\ \gamma & \text{if } L(\boldsymbol{\theta}_1) = kL(\boldsymbol{\theta}_0) \\ 0 & \text{if } L(\boldsymbol{\theta}_1) < kL(\boldsymbol{\theta}_0), \end{cases}$$

where $k \geq 0$ and $\gamma \in [0, 1]$ such that $\beta(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}(\psi(\boldsymbol{X})) = \alpha$. Here, $L(\cdot)$ is the likelihood function.

Proof: Let $\psi^*(\cdot) \in \mathscr{C}_\alpha$ and the power function of $\psi^*(\cdot)$ be $\beta^*(\cdot)$. We need to show that $\beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1)$. Let us consider the quantity

$$Q(\boldsymbol{x}) = (\psi(\boldsymbol{x}) - \psi^*(\boldsymbol{x}))(L(\boldsymbol{\theta}_1) - kL(\boldsymbol{\theta}_0)).$$

First, we will prove that $Q(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathscr{X}^n$. Let us define

$$\begin{aligned} \mathscr{X}_1^n &= \{\boldsymbol{x} \in \mathscr{X}^n : L(\boldsymbol{\theta}_1) - kL(\boldsymbol{\theta}_0) > 0\}, \\ \mathscr{X}_2^n &= \{\boldsymbol{x} \in \mathscr{X}^n : L(\boldsymbol{\theta}_1) - kL(\boldsymbol{\theta}_0) = 0\}, \\ \mathscr{X}_3^n &= \{\boldsymbol{x} \in \mathscr{X}^n : L(\boldsymbol{\theta}_1) - kL(\boldsymbol{\theta}_0) < 0\}. \end{aligned}$$

First notice that $\mathscr{X}^n = \mathscr{X}_1^n \cup \mathscr{X}_2^n \cup \mathscr{X}_3^n$. Now, if $\boldsymbol{x} \in \mathscr{X}_1^n$, $\psi(\boldsymbol{x}) \geq \psi^*(\boldsymbol{x})$ as $\psi(\boldsymbol{x}) = 1$ and $0 \leq \psi^*(\boldsymbol{x}) \leq 1$. Therefore, $Q(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathscr{X}_1^n$. For $\boldsymbol{x} \in \mathscr{X}_2^n$, $Q(\boldsymbol{x}) = 0$. For $\boldsymbol{x} \in \mathscr{X}_3^n$, $\psi(\boldsymbol{x}) = 0 \leq \psi^*(\boldsymbol{x})$. Therefore, $Q(\boldsymbol{x}) \geq 0$ for $\boldsymbol{x} \in \mathscr{X}_3^n$. Thus,

$$\begin{aligned} &\int_{\mathscr{X}^n} Q(\boldsymbol{x})\, d\boldsymbol{x} \geq 0 \\ &\implies \int_{\mathscr{X}^n} (\psi(\boldsymbol{x}) - \psi^*(\boldsymbol{x}))\, L(\boldsymbol{\theta}_1)\, d\boldsymbol{x} + k \int_{\mathscr{X}^n} (\psi(\boldsymbol{x}) - \psi^*(\boldsymbol{x}))\, L(\boldsymbol{\theta}_0)\, d\boldsymbol{x} \geq 0 \\ &\implies \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1) - k(\beta(\boldsymbol{\theta}_0) - \beta^*(\boldsymbol{\theta}_0)) \geq 0 \\ &\implies \beta(\boldsymbol{\theta}_1) - \beta^*(\boldsymbol{\theta}_1) \geq k(\beta(\boldsymbol{\theta}_0) - \beta^*(\boldsymbol{\theta}_0)) \geq 0, \end{aligned}$$

as $\beta(\boldsymbol{\theta}_0) = \alpha$ and $\beta^*(\boldsymbol{\theta}_0) \leq \alpha$. This completes the proof. □

Note that $L(\theta_1) > kL(\theta_0)$ can be expressed as $\dfrac{L(\theta_1)}{L(\theta_0)} > k$ if $L(\theta_0) > 0$. Hence, loosely speaking, the MP test rejects the null hypothesis for large values of the ratio.

Example 3.10. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where σ is known. Let $\mu_0 < \mu_1$ be two fixed real numbers. We are interested to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$. Here, the likelihood function is

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

Therefore,

$$\frac{L(\mu_1)}{L(\mu_0)} = \exp\left[\frac{1}{2\sigma^2}\left\{2n\bar{x}(\mu_1-\mu_0) + n\left(\mu_0^2-\mu_1^2\right)\right\}\right].$$

Now.

$$\begin{aligned} \frac{L(\mu_1)}{L(\mu_0)} > k &\iff \exp\left[\frac{1}{2\sigma^2}\left\{2n\bar{x}(\mu_1-\mu_0) + n\left(\mu_0^2-\mu_1^2\right)\right\}\right] > k \\ &\iff \frac{1}{2\sigma^2}\left\{2n\bar{x}(\mu_1-\mu_0) + n\left(\mu_0^2-\mu_1^2\right)\right\} > \ln k \\ &\iff 2n(\mu_1-\mu_0)\bar{x} > 2\sigma^2\ln k - n\left(\mu_0^2-\mu_1^2\right) \\ &\iff \bar{x} > \frac{2\sigma^2\ln k - n\left(\mu_0^2-\mu_1^2\right)}{2n(\mu_1-\mu_0)} = \kappa \text{ (say), as } \mu_1 > \mu_0. \end{aligned}$$

Thus, $\dfrac{L(\mu_1)}{L(\mu_0)} > k$ if and only if $\bar{x} > \kappa$ for some constant κ. Hence, using Neymann-Pearson lemma, the test function of the MP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \bar{x} > \kappa \\ \gamma & \text{if } \bar{x} = \kappa \\ 0 & \text{if } \bar{x} < \kappa, \end{cases}$$

where κ and γ are such that

$$\begin{aligned} E_{\mu_0}(\psi(\boldsymbol{X})) = \alpha &\iff P_{\mu_0}\left(\overline{X} > \kappa\right) + \gamma P_{\mu_0}\left(\overline{X} = \kappa\right) = \alpha \\ &\iff P_{\mu_0}\left(\overline{X} > \kappa\right) = \alpha \\ &\iff P_{\mu_0}\left(\sqrt{n}\,\frac{\overline{X}-\mu_0}{\sigma} > \sqrt{n}\,\frac{\kappa-\mu_0}{\sigma}\right) = \alpha \\ &\iff \sqrt{n}\,\frac{\kappa-\mu_0}{\sigma} = z_\alpha, \end{aligned}$$

where z_α is a real number such that $P(Z > z_\alpha) = \alpha$ for a RV $Z \sim N(0, 1)$ (please see the Figure 3.1). Here, we can take any value of $\gamma \in [0, 1]$ as $P_{\mu_0}(\overline{X} = \kappa) = 0$. In this situation γ is taken to be zero and that makes it a non-randomized test. Hence, the MP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}}{\sigma}(\bar{x}-\mu_0) > z_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

||

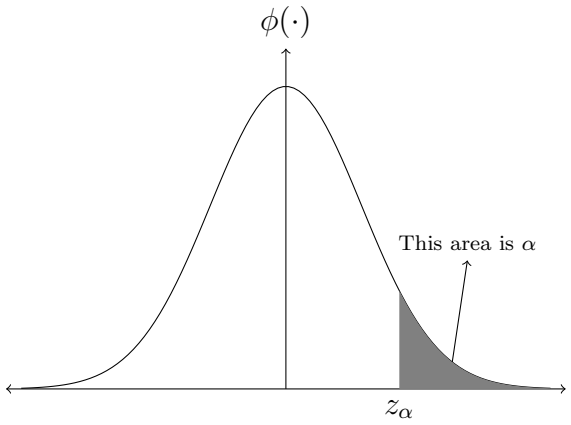


Figure 3.1: Upper α-point for a standard normal distribution

Remark 3.5. Note the way of solving the problem. We try to simplify $\dfrac{L(\mu_1)}{L(\mu_0)} > k$ so that we can write an equivalent condition on a statistic whose distribution under H_0 is known or can be found. If this statistic is a continuous random variable, we will have a non-randomized test. Otherwise we may need to consider $\gamma \in (0,\, 1)$ making the test a randomized one. †

Example 3.11. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$. Let $0 < \theta_1 < \theta_0 < 1$ be two real numbers. We are interested to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Let $T = \sum_{i=1}^{n} X_i \sim Bin(n,\, \theta_0)$ under H_0. Now,

$$\frac{L(\theta_1)}{L(\theta_0)} = \left(\frac{\theta_1}{\theta_0} \times \frac{1-\theta_0}{1-\theta_1}\right)^t > k \iff t < k_1,$$

for some constant k_1. It see it, notice that

$$g(t) = \left(\frac{\theta_1}{\theta_0} \times \frac{1-\theta_0}{1-\theta_1}\right)^t$$

is a decreasing function of t for $\theta_0 > \theta_1$. Hence, the MP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } t < k_1 \\ \gamma & \text{if } t = k_1 \\ 0 & \text{if } t > k_1, \end{cases}$$

where $E_{\theta_0}(\psi(\boldsymbol{X})) = P_{\theta_0}(T < k_1) + \gamma P_{\theta_0}(T = k_1) = \alpha$. Let $\widetilde{K} \in \{1,\, 2,\, \ldots,\, n\}$ be such that

$$P_{\theta_0}\left(T < \widetilde{K}\right) \le \alpha < P_{\theta_0}\left(T \le \widetilde{K}\right).$$

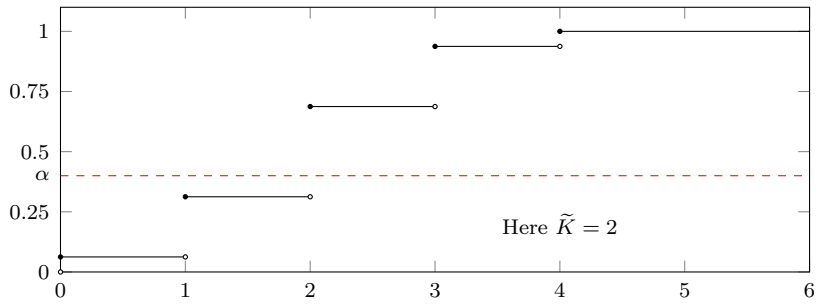


Figure 3.2: CDF of Bin(4,0.5), $\alpha = 0.4$

Take $k_1 = \widetilde{K}$ and $\gamma = \dfrac{\alpha - P_{\theta_0}\left(T < \widetilde{K}\right)}{P_{\theta_0}\left(T = \widetilde{K}\right)}$. The MP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } t < \widetilde{K} \\ \frac{\alpha - P_{\theta_0}\left(T < \widetilde{K}\right)}{P_{\theta_0}\left(T = \widetilde{K}\right)} & \text{if } t = \widetilde{K} \\ 0 & \text{if } t > \widetilde{K}. \end{cases} \qquad ||$$

Example 3.12. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$. Let $\theta_0 > \theta_1 > 0$ be two fixed real numbers. We are interested to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Now,

$$\frac{L(\theta_1)}{L(\theta_0)} = \begin{cases} \left(\frac{\theta_0}{\theta_1}\right)^n & \text{if } x_{(n)} < \theta_1 \\ 0 & \text{if } \theta_1 \leq x_{(n)} < \theta_0. \end{cases}$$

It shows that the ratio $\dfrac{L(\theta_1)}{L(\theta_0)}$ is a non-increasing function of $x_{(n)}$. Hence, $\dfrac{L(\theta_1)}{L(\theta_0)} > k$ if and only if $x_{(n)} < k_1$, for some constant k_1. Therefore, the MP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } x_{(n)} < k_1 \\ 0 & \text{otherwise,} \end{cases}$$

where k_1 is such that

$$E_{\theta_0}\left(\psi(\boldsymbol{X})\right) = \alpha \iff P\left(X_{(n)} > k_1\right) = \alpha \iff k_1 = \theta_0 \alpha^{\frac{1}{n}}. \qquad ||$$

3.5 One-sided Composite Alternative

Suppose that the null hypothesis $H_0 : \theta = \theta_0$ is simple, but the alternative hypothesis is of the form $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$. In such cases, the alternative hypothesis is called one-sided. The alternative hypothesis $H_1 : \theta > \theta_0$ is upper-sided composite hypothesis, and $H_1 : \theta < \theta_0$ is lower-sided composite hypothesis. In this section, we will discuss mainly two methods to construct UMP test for such cases.

3.5.1 UMP Test via Neyman-Pearson Lemma

Though the Neyman-Pearson lemma is for simple null vs. simple alternative, it can be used to find the UMP test for simple null vs. one-sided alternative in some situations. We will explain this method with the help of the following example.

Example 3.13. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where σ is known. Let μ_0 be a real number. We are interested to test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. Note that here the alternative is not a simple. However, we can use NP lemma to find UMP level α test. Let $\mu_1 > \mu_0$ be a real number. Then, we know that the MP level α test for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}}{\sigma}\left(\overline{x} - \mu_0\right) > z_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

Please revisit Example 3.10. Note that this test does not depend on μ_1 as long as $\mu_1 > \mu_0$. Hence, same test is UMP level α test for $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. ||

3.5.2 UMP Test via Monotone Likelihood Ratio Property

Definition 3.13 (Monotone Likelihood Ratio)**.** *A family of distributions* $\{f(x,\,\theta) : \theta \in \Theta\}$ *with* $\Theta \subseteq \mathbb{R}$ *is said to have a monotone likelihood ratio (MLR) property in a real valued statistic* $T(\boldsymbol{X})$ *if for any* $\theta < \theta^*$*, the ratio* $\dfrac{L(\theta^*,\,\boldsymbol{x})}{L(\theta,\,\boldsymbol{x})}$ *is a nondecreasing function of* $T(\boldsymbol{x})$*.*

Remark 3.6. In the previous definition, $\dfrac{L(\theta^*,\,\boldsymbol{x})}{L(\theta,\,\boldsymbol{x})}$ should be nondecreasing or nonincreasing. If the likelihood ratio $\dfrac{L(\theta^*,\,\boldsymbol{x})}{L(\theta,\,\boldsymbol{x})}$ is nonincreasing instead of nondecreasing, the effect would be felt in the placement of rejection region. †

Example 3.14. Let $X_1,\, X_2,\, \ldots,\, X_n \overset{i.i.d.}{\sim} N(\mu,\,\sigma^2)$, where $\mu \in \mathbb{R}$ is unknown, but $\sigma > 0$ is known. Then for arbitrary real numbers $\mu^* > \mu$,

$$\frac{L(\mu^*,\,\boldsymbol{x})}{L(\mu,\,\boldsymbol{x})} = \exp\left[(\mu^* - \mu)\,\frac{T(\boldsymbol{x})}{\sigma^2} + \frac{n}{2\sigma^2}\left(\mu^2 - \mu^{*2}\right)\right],$$

where $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. Now, it is easy to see that the likelihood ratio is an increasing function in T. Therefore, we have a MLR (increasing) property in T. ||

Example 3.15. Let $X_1,\, X_2,\, \ldots,\, X_n$ be a RS from a exponential distribution with mean $\frac{1}{\lambda}$. Then, for arbitrary positive constants $\lambda^* > \lambda$,

$$\frac{L(\lambda^*,\,\boldsymbol{x})}{L(\lambda,\,\boldsymbol{x})} = \left(\frac{\lambda^*}{\lambda}\right)^n \exp\left[-\left(\lambda^* - \lambda\right) T(\boldsymbol{x})\right],$$

where $T(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. Now, as the likelihood ratio is a decreasing function in T, we have MLR (decreasing) property in T. ||

Theorem 3.2. *Let* $X_1,\, X_2,\, \ldots,\, X_n$ *be a RS from a PMF/PDF* $f(x,\,\theta)$ *with MLR (nondecreasing) property in* $T(\boldsymbol{x})$*, where* $\theta \in \Theta \subseteq \mathbb{R}$*. For testing* $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$*, there exists a UMP level* α *test, which is given by*

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) > k \\ \gamma & \text{if } T(\boldsymbol{x}) = k \\ 0 & \text{if } T(\boldsymbol{x}) < k, \end{cases}$$

where k *and* γ *are such that* $E_{\theta_0}\left(\psi(\boldsymbol{X})\right) = \alpha$*.*

Proof: The proof is skipped. □

Example 3.16. Let $X_1,\, X_2,\, \ldots,\, X_n \overset{i.i.d.}{\sim} N(\mu,\,\sigma^2)$, where μ is unknown, but $\sigma > 0$ is known. Let us consider the testing of $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. In this case, we have MLR (nondecreasing) property in $T = \sum_{i=1}^{n} X_i$. The UMP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) > k \\ 0 & \text{otherwise,} \end{cases}$$

where k is such that $P_{\mu_0}(T(\boldsymbol{X}) > k) = \alpha$. Solving for k, we have

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}}{\sigma}\left(\overline{x} - \mu_0\right) > z_\alpha \\ 0 & \text{otherwise.} \end{cases}$$ ||

Notice that we have following four different combinations of MLR (nondecreasing or nonincreasing) property and null hypothesis ($H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$).

	$H_0 : \theta \leq \theta_0$	$H_0 : \theta \geq \theta_0$
MLR (nondecreasing)	Case I	Case II
MLR (nonincreasing)	Case III	Case IV

Of course, for each null hypothesis, the alternative is $\theta \in \mathbb{R} - \Theta_0$. The previous theorem provides the UMP test for Case I. For other three combinations, the UMP tests can be obtained in an almost similar fashion and are mentioned below.

Theorem 3.3. *Let $X_1, X_2, \ldots, X_n$ be a RS from a PMF/PDF $f(x, \theta)$ with MLR (nondecreasing) property in $T(\boldsymbol{x})$, where $\theta \in \Theta \subseteq \mathbb{R}$. For testing $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$, there exists a UMP level α test, which is given by*

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) < k \\ \gamma & \text{if } T(\boldsymbol{x}) = k \\ 0 & \text{if } T(\boldsymbol{x}) > k, \end{cases}$$

where k and γ are such that $E_{\theta_0}(\psi(\boldsymbol{X})) = \alpha$.

Theorem 3.4. *Let $X_1, X_2, \ldots, X_n$ be a RS from a PMF/PDF $f(x, \theta)$ with MLR (nonincreasing) property in $T(\boldsymbol{x})$, where $\theta \in \Theta \subseteq \mathbb{R}$. For testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, there exists a UMP level α test, which is given by*

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) < k \\ \gamma & \text{if } T(\boldsymbol{x}) = k \\ 0 & \text{if } T(\boldsymbol{x}) > k, \end{cases}$$

where k and γ are such that $E_{\theta_0}(\psi(\boldsymbol{X})) = \alpha$.

Theorem 3.5. *Let $X_1, X_2, \ldots, X_n$ be a RS from a PMF/PDF $f(x, \theta)$ with MLR (nonincreasing) property in $T(\boldsymbol{x})$, where $\theta \in \Theta \subseteq \mathbb{R}$. For testing $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$, there exists a UMP level α test, which is given by*

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } T(\boldsymbol{x}) > k \\ \gamma & \text{if } T(\boldsymbol{x}) = k \\ 0 & \text{if } T(\boldsymbol{x}) < k, \end{cases}$$

where k and γ are such that $E_{\theta_0}(\psi(\boldsymbol{X})) = \alpha$.

3.6 Simple Null Vs. Two-sided Alternative

In this section, we will discuss testing of a simple null hypothesis $H_0 : \theta = \theta_0$ against a two-sided alternative $H_1 : \theta \neq \theta_0$, where $\theta_0 \in \Theta$. A natural question is: Does there exist a UMP level α test? The answer is "yes" in some situations and "no" in some other situations. We will not discuss it in full detail. We will discuss two examples and conclude this section.

Example 3.17. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, where $\theta > 0$. We wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. In this case, it can be show that the test function of UMP level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } X_{(n)} \geq \theta_0 \text{ and } X_{(n)} \leq \theta_0 \alpha^{\frac{1}{n}} \\ 0 & \text{otherwise.} \end{cases}$$

The proof can be break into two parts. In the first part, we will show that any test function $\psi^*(\boldsymbol{x})$ such that

$$E_{\theta_0}\left[\psi^*(\boldsymbol{X})\right] = \alpha,\ E_{\theta}\left[\psi^*(\boldsymbol{X})\right] \leq \alpha \text{ for } \theta \leq \theta_0, \text{ and } \psi^*(\boldsymbol{x}) = 1 \text{ when } X_{(n)} > \theta_0$$

corresponds to a UMP level α test for $H_0' : \theta \leq \theta_0$ against $H_1' : \theta > \theta_0$. Then in the second part, we will prove that $\psi(\boldsymbol{x})$ is UMP level α test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ using the first part.

To see the first part, we can proceed as follows. Note that using the MLR property, one can obtain the test function of the UMP level α test for H_0' against H_1' as

$$\psi_0(\boldsymbol{x}) = \begin{cases} 1 & \text{if } X_{(n)} > \theta_0(1-\alpha)^{\frac{1}{n}} \\ 0 & \text{if } X_{(n)} \leq \theta_0(1-\alpha)^{\frac{1}{n}}. \end{cases}$$

The power function associated with the test function $\psi_0(\cdot)$ is given by

$$\begin{aligned} E_{\theta}\left[\psi_0(\boldsymbol{X})\right] &= P_{\theta}\left(X_{(n)} > \theta_0(1-\alpha)^{\frac{1}{n}}\right) \\ &= \int_{\theta_0(1-\alpha)^{\frac{1}{n}}}^{\theta} \frac{nt^{n-1}}{\theta^n} dt \\ &= 1 - (1-\alpha)\left(\frac{\theta_0}{\theta}\right)^n. \end{aligned}$$

Notice that ψ^* is a level α test. Moreover,

$$\alpha = E_{\theta_0}\left[\psi^*(\boldsymbol{X})\right] = E_{\theta_0}\left[E\left(\psi^*(\boldsymbol{X})|X_{(n)}\right)\right] = E_{\theta_0}\left[g(X_{(n)})\right],$$

where $g(x) = E\left[\psi^*(\boldsymbol{X})|X_{(n)} = x\right]$, which does not depend on θ as $X_{(n)}$ is sufficient statistic. Also, as $\psi^*(\boldsymbol{x}) = 1$ for $X_{(n)} > \theta_0$, $g(t) = 1$ for all $t > \theta_0$. Now, for $\theta > \theta_0$, the power function of the test ψ^* is

$$\begin{aligned} E_{\theta}\left[\psi^*(\boldsymbol{X})\right] &= E_{\theta}\left[g(X_{(n)})\right] \\ &= \int_0^{\theta_0} g(t)\frac{nt^{n-1}}{\theta^n} dt + \int_{\theta_0}^{\theta} g(t)\frac{nt^{n-1}}{\theta^n} dt \\ &= \left(\frac{\theta_0}{\theta}\right)^n \int_0^{\theta_0} g(t)\frac{nt^{n-1}}{\theta_0^n} dt + \int_{\theta_0}^{\theta} \frac{nt^{n-1}}{\theta^n} dt \\ &= \alpha\left(\frac{\theta_0}{\theta}\right)^n + 1 - \left(\frac{\theta_0}{\theta}\right)^n \\ &= 1 - (1-\alpha)\left(\frac{\theta_0}{\theta}\right)^n. \end{aligned}$$

Thus, the power functions of ψ^* and ψ_0 coincide on $\theta > \theta_0$. Therefore, ψ^* is UMP level α test for testing H_0' against H_1'. This completes the proof of the first part.

To see the second part, let us first calculate level of the test ψ.

$$\begin{aligned} E_{\theta_0}[\psi(\boldsymbol{X})] &= P_{\theta_0}\left(X_{(n)} \geq \theta_0\right) + P_{\theta_0}\left(X_{(n)} \leq \theta_0 \alpha^{\frac{1}{n}}\right) \\ &= \int_0^{\theta_0 \alpha^{\frac{1}{n}}} \frac{n t^{n-1}}{\theta_0^n} dt \\ &= \alpha. \end{aligned}$$

Thus, ψ is a level α test. Using the first part, ψ is UMP level α test for testing $H_0 : \theta = \theta_0$ against $H_1' : \theta > \theta_0$. Now, using MLR property, the UMP level α test for testing $H_0 : \theta = \theta_0$ against $H_1'' : \theta < \theta_0$ is given by

$$\psi_{00}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } X_{(n)} \leq \theta_0 \alpha^{\frac{1}{n}} \\ 0 & \text{if } X_{(n)} > \theta_0 \alpha^{\frac{1}{n}}. \end{cases}$$

For $\theta < \theta_0$, the power function of ψ is

$$\begin{aligned} E_{\theta}[\psi(\boldsymbol{X})] &= P_{\theta}\left(X_{(n)} \geq \theta_0\right) + P_{\theta}\left(X_{(n)} \leq \theta_0 \alpha^{\frac{1}{n}}\right) \\ &= P_{\theta}\left(X_{(n)} \leq \theta_0 \alpha^{\frac{1}{n}}\right) \\ &= E_{\theta}\left[\psi_{00}(\boldsymbol{X})\right]. \end{aligned}$$

Therefore, ψ is UMP level α test for testing $H_0 : \theta = \theta_0$ against $H_1'' : \theta < \theta_0$. Hence, ψ is UMP level α test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. ||

Example 3.18. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where σ is known. Let μ_0 be a real number. We are interested to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. We will show here that the UMP test of any level does not exist. To see it, recall that the MP level α test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1 (> \mu_0)$ is given by

$$\psi_1(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}}{\sigma}(\overline{x} - \mu_0) > z_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, it can be shown (left as an exercises) that the MP level α test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1 (< \mu_0)$ is given by

$$\psi_2(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}}{\sigma}(\overline{x} - \mu_0) < -z_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

For $\mu > \mu_0$, $\psi_1(\cdot)$ has maximum power among all level α tests. For $\mu < \mu_0$, $\psi_2(\cdot)$ has maximum power among all level α tests. As these two tests are different, UMP level α test for $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ does not exist for all $\alpha \in (0, 1)$. ||

The main take away form the previous two examples is that there are situations where UMP test exits and there are some other situations where the UMP test does not exist. However, the problem of testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is practically quite meaningful. Hence, we need some alternative. One of such alternatives is likelihood ratio test, which depends on the concept of MLE and is discussed in the next section.

3.7 Likelihood Ratio Tests

In the previous section, we have seen that the UMP level α test does not exist for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ for all values of $\alpha \in (0, 1)$. Nontheless, these hypotheses are quite meaningful in practice. In the current section, we will discuss an alternative method to find a meaningful test. This test is called likelihood ratio test (LRT). Note that LRT may not be best, but they are quite intuitive and useful.

The general procedure to obtain a LRT can be described as follows. Suppose that we want to test $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$ at level $\alpha \in (0, 1)$. Consider the ratio

$$\Lambda(\boldsymbol{x}) = \frac{\sup\limits_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta}, \boldsymbol{x})}{\sup\limits_{\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1} L(\boldsymbol{\theta}, \boldsymbol{x})}.$$

This ratio $\Lambda(\boldsymbol{x})$ is called likelihood ratio test statistic. The test function of a LRT is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \Lambda(\boldsymbol{x}) < k \\ \gamma & \text{if } \Lambda(\boldsymbol{x}) = k \\ 0 & \text{if } \Lambda(\boldsymbol{x}) > k, \end{cases}$$

where $\gamma \in (0, 1)$ and $k > 0$ are such that $\sup\limits_{\boldsymbol{\theta} \in \Theta_0} E_\theta \left(\psi(\boldsymbol{X})\right) = \alpha$.

Note that $\sup\limits_{\boldsymbol{\theta} \in \Theta_0} L(\theta, \boldsymbol{x})$ can be interpreted as the best evidence for observing $\boldsymbol{X} = \boldsymbol{x}$, when the parameter $\boldsymbol{\theta}$ is restricted in Θ_0. Similarly, $\sup\limits_{\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1} L(\boldsymbol{\theta}, \boldsymbol{x})$ can be interpreted as the overall best evidence for observing $\boldsymbol{X} = \boldsymbol{x}$. Also, note that LRT rejects the null hypothesis if the value of $\Lambda(\boldsymbol{x})$ is "small". The rationale is that if the best evidence of observing $\boldsymbol{X} = \boldsymbol{x}$ under the null hypothesis is weak compared with overall evidence, then H_0 ought to be rejected.

It is clear from the definition of LRT statistic $\Lambda(\boldsymbol{x})$ that $0 \leq \Lambda(\boldsymbol{x}) \leq 1$. To see it, notice that the supremum in the numerator is over a smaller set than that in the denominator. Therefore, we need to find the cutoff point $k \in (0, 1)$ and $\gamma \in (0, 1)$ such that the test satisfies level condition, *i.e.*,

$$\sup_{\boldsymbol{\theta} \in \Theta_0} E_{\boldsymbol{\theta}} \left(\psi\left(\boldsymbol{X}\right)\right) = \alpha.$$

Example 3.19. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where σ is known. Let μ_0 be a real number. We are interested to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Here, $\Theta_0 = \{\mu_0\}$, $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$, and $\Theta_0 \cup \Theta_1 = \mathbb{R}$. Now,

$$\sup_{\mu \in \Theta_0} L(\mu) = L(\mu_0) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right],$$

and

$$\sup_{\mu \in \Theta_0 \cup \Theta_1} L(\mu) = L(\overline{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right].$$

The first equality of the last equation is true as $\overline{X}$ is MLE for $\mu \in \mathbb{R}$. As,

$$\sum_{i=1}^{n} (x_i - \mu_0)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n (\bar{x} - \mu_0)^2 ,$$

the LRT statistic is

$$\Lambda(\boldsymbol{x}) = \exp\left[-\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2\right].$$

We reject the null hypothesis if and only if

$$\Lambda(\boldsymbol{x}) < k \iff |\bar{x} - \mu_0| > k_1.$$

Thus, likelihood ratio level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } |\bar{x} - \mu_0| > k_1 \\ 0 & \text{otherwise,} \end{cases}$$

where k_1 is such that

$$\begin{aligned} E_{\mu_0}(\psi(\boldsymbol{X})) = \alpha &\iff P_{\mu_0}\left(\frac{\sqrt{n}}{\sigma}|\overline{X} - \mu_0| > \frac{\sqrt{n}}{\sigma}k_1\right) = \alpha \\ &\iff P_{\mu_0}\left(\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu_0) > \frac{\sqrt{n}}{\sigma}k_1\right) = \frac{\alpha}{2} \\ &\iff \frac{\sqrt{n}}{\sigma}k_1 = z_{\alpha/2}. \end{aligned}$$

Hence, the likelihood ratio level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}\,|\bar{x} - \mu_0|}{\sigma} > z_{\alpha/2} \\ 0 & \text{otherwise.} \end{cases}$$

Note that the test reject the null hypothesis if the distance between the sample mean and μ_0 is sufficiently large. ||

Remark 3.7. $\dfrac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma}$ can be considered as standardized distance between $\bar{x}$ and μ_0. Hence, this test rejects H_0 if the distance is large, which is quite intuitive. †

Example 3.20. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Let μ_0 be a real number. We are interested to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. Here, $\Theta_0 = \{\mu_0\} \times \mathbb{R}^+$, $\Theta_1 = \mathbb{R} \backslash \{\mu_0\} \times \mathbb{R}^+$, and hence, $\Theta_0 \cup \Theta_1 = \mathbb{R} \times \mathbb{R}^+$. Now,

$$\sup_{\Theta_0} L(\mu, \sigma^2) = \sup_{\sigma > 0} L(\mu_0, \sigma^2) = \left(\frac{2\pi e}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2\right)^{-n/2},$$

and

$$\sup_{\Theta_0 \cup \Theta_1} L(\mu, \sigma^2) = \left(\frac{2\pi e}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{-n/2}.$$

Hence, the LRT statistic is

$$\Lambda = \left(\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \mu_0)^2} \right)^{n/2} = \left(1 + \frac{n (\overline{x} - \mu_0)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{-n/2}.$$

Note that for $x > 0$, $f(x) = (1+x)^{-n/2}$ is a decreasing function in x. Hence,

$$\Lambda(\boldsymbol{x}) < k \iff \frac{n (\overline{x} - \mu_0)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} > k_1 \iff \frac{\sqrt{n} |\overline{x} - \mu_0|}{s} > k_2,$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$. The likelihood ratio test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}|\overline{x}-\mu_0|}{s} > k_2 \\ \gamma & \text{if } \frac{\sqrt{n}|\overline{x}-\mu_0|}{s} = k_2 \\ 0 & \text{otherwise,} \end{cases}$$

where k_2 is such that $E_{\Theta_0} (\psi(\boldsymbol{X})) = \alpha$. Now, we know that

$$\frac{\sqrt{n} \left(\overline{X} - \mu_0 \right)}{S} \sim t_{n-1} \text{ under } H_0.$$

Hence, $E_{\Theta_0} (\psi(\boldsymbol{X})) = \alpha \implies P_{\Theta_0} \left(\frac{\sqrt{n}|\overline{X}-\mu_0|}{S} > k_2 \right) = \alpha \implies k_2 = t_{n-1;\alpha/2}$, where $t_{n;\alpha}$ is a real number such that $P(Z > t_{n;\alpha}) = \alpha$, where $Z \sim t_n$. Also, notice that the PDF of a t_n distribution is symmetric about zero. Thus, the likelihood ratio test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}|\overline{x}-\mu_0|}{s} > t_{n-1;\alpha/2} \\ 0 & \text{otherwise.} \end{cases} \qquad ||$$

Example 3.21. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. Let σ_0 be a positive real number. We are interested to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$. Here, $\Theta_0 = \mathbb{R} \times \{\sigma_0^2\}$, $\Theta_1 = \mathbb{R} \times \mathbb{R}^+ \setminus \{\sigma_0^2\}$, and hence $\Theta_0 \cup \Theta_1 = \mathbb{R} \times \mathbb{R}^+$. Now,

$$\sup_{\Theta_0} L(\mu, \sigma^2) = \sup_{\mu \in \mathbb{R}} L(\mu, \sigma_0^2) = \left(2\pi\sigma_0^2 \right)^{-n/2} \exp \left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right]$$

and

$$\sup_{\Theta_0 \cup \Theta_1} L(\mu, \sigma^2) = \left(\frac{2\pi e}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{-n/2}.$$

Therefore, the LRT statistic is

$$\Lambda(\boldsymbol{x}) = \left\{ \left(\frac{\widehat{\sigma}^2}{\sigma_0^2} \right) \exp \left[-\frac{\widehat{\sigma}^2}{\sigma_0^2} \right] \right\}^{\frac{n}{2}},$$

where $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$. LRT rejects the null hypothesis if and only if $\Lambda(\boldsymbol{x}) < k$. To simplify it, let us check the behavior of the function $\Lambda(\boldsymbol{x})$ as a function of $\frac{\widehat{\sigma}^2}{\sigma_0^2}$, as the distribution of $\frac{n\widehat{\sigma}^2}{\sigma_0^2}$ is known under the null hypothesis. For this, consider the function

$$g(x) = xe^{-x} \quad \text{for } x > 0.$$

It can be easily seen that the function $g(\cdot)$ attains its unique maximum at $x = 1$. Moreover, $g(0) = 0$ and $\lim_{x\to\infty} g(x) = 0$. Thus, it is a bell shaped function. Therefore,

$$\Lambda(\boldsymbol{x}) < k \iff \frac{\widehat{\sigma}^2}{\sigma_0^2} < k_1 \text{ or } \frac{\widehat{\sigma}^2}{\sigma_0^2} > k_2.$$

Thus, the test function of likelihood ratio level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{\widehat{\sigma}^2}{\sigma_0^2} < k_1 \text{ or } \frac{\widehat{\sigma}^2}{\sigma_0^2} > k_2 \\ \gamma & \text{if } \frac{\widehat{\sigma}^2}{\sigma_0^2} = k_1 \text{ or } \frac{\widehat{\sigma}^2}{\sigma_0^2} = k_2 \\ 0 & \text{otherwise,} \end{cases}$$

where k_1 and k_2 are such that $E_{\Theta_0}(\psi(X)) = \alpha$. Now, we know that under H_0

$$T = \frac{n\widehat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Hence,

$$E_{\Theta_0}(\psi(\boldsymbol{X})) = \alpha \implies P_{\Theta_0}(T < nk_1) + P_{\Theta_0}(T > nk_2) = \alpha.$$

There are infinite number of choices of (k_1, k_2), which satisfy the above condition. The most popular choice of (k_1, k_2) can be described as follows. Take

$$P_{\Theta_0}(T < nk_1) = \frac{\alpha}{2} = P(T > nk_2) \implies nk_1 = \chi^2_{n-1;1-\alpha/2} \text{ and } nk_2 = \chi^2_{n-1;\alpha/2}.$$

Thus, the likelihood ratio level α test is given by

$$\psi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{n\widehat{\sigma}^2}{\sigma_0^2} < \chi^2_{n-1;1-\alpha/2} \text{ or } \frac{n\widehat{\sigma}^2}{\sigma_0^2} > \chi^2_{n-1;\alpha/2} \\ 0 & \text{otherwise.} \end{cases} \quad ||$$

3.8 p-value

In the previous sections, we have discussed hypothesis testing at a fixed level α. Note that this approach is one of the two standard approaches to the evaluation of hypotheses. To explain the other, first we need to define nested test.

Definition 3.14 (Nested Test)**.** *For varying level α, assume that the test is a non-randomized test with critical region R_α. The test is called nested if*

$$R_\alpha \subset R_{\alpha'} \quad \textit{for all } \alpha < \alpha'.$$

When a test is nested, it is good practice to determine not only whether the null hypothesis is accepted or rejected at a given level α, but also to determine the smallest level at which the null hypothesis would be rejected for the given observation. This smallest level is called p-value.

Definition 3.15 (p-value)**.** *The p-value of a nested test is defined by*

$$\widehat{p} = \widehat{p}(\boldsymbol{X}) = \inf\{\alpha \in [0, 1] : \boldsymbol{X} \in R_\alpha\}.$$

The p-value provides an idea of how strong the data contradict the null hypothesis. It also enables other to reach a verdict based on the level of their choice. If p-value is smaller than α, we reject the null hypothesis. Otherwise, we accept the null hypothesis.

Example 3.22. Let $X_1, X_2, \ldots, X_n$ be a RS form a population having normal distribution with unknown mean μ and variance $\sigma^2 = 1$. Consider $H_0 : \mu = \mu^0$ against $H_1 : \mu \neq \mu_0$. The have seen that the critical region of likelihood ratio level α test is given by

$$R_\alpha = \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sqrt{n}\frac{|\overline{x} - \mu_0|}{\sigma} > z_{\frac{\alpha}{2}} \right\}.$$

As for $\alpha < \alpha'$, $z_{\frac{\alpha}{2}} > z_{\frac{\alpha'}{2}}$, $R_\alpha \subset R_{\alpha'}$. Therefore, the test is a nested test and we can talk about p-value. As $\Phi(\cdot)$ is a strictly increasing function, we can write

$$\begin{aligned} R_\alpha &= \left\{ \boldsymbol{x} \in \mathbb{R}^n : \sqrt{n}\frac{|\overline{x} - \mu_0|}{\sigma} > z_{\frac{\alpha}{2}} \right\} \\ &= \left\{ \boldsymbol{x} \in \mathbb{R}^n : \Phi\left(\sqrt{n}\frac{|\overline{x} - \mu_0|}{\sigma}\right) > 1 - \frac{\alpha}{2} \right\} \\ &= \left\{ \boldsymbol{x} \in \mathbb{R}^n : \alpha > 2\left(1 - \Phi\left(\sqrt{n}\frac{|\overline{x} - \mu_0|}{\sigma}\right)\right) \right\}. \end{aligned}$$

Therefore, the p-value is

$$\widehat{p}(\boldsymbol{X}) = 2\left[1 - \Phi\left(\sqrt{n}\frac{|\overline{X} - \mu_0|}{\sigma}\right)\right]. \qquad ||$$

Chapter 4

Interval Estimation

4.1 Confidence Interval

In this section, we assume that the parameter under consideration is a real valued parameter. We are interested to find an interval in $\Theta \subseteq \mathbb{R}$ such that the interval covers the unknown parameter with a specified probability. Of course, the interval will be based on a RS. Interval estimation is quite useful in practice. For example, one may be interested to find a upper limit of mean of toxic level of some drug or food.

Note that for a RV X and two real constants $a > 0$ and $b > 0$,

$$P\left(a < X < b\right) = P\left(X < b < \frac{bX}{a}\right).$$

Though, these two probabilities are same, there is a basic difference in these two probability statements. For the LHS, we are taking about probability that a random quantity X belongs to a fixed interval $(a,\, b)$. For the RHS, we are taking about probability that a random interval $(X,\, \frac{bX}{a})$ contains a fixed point b. For example, let $X \sim U(0,\, 1)$, $a = 0.5$, and $b = 1$. In this case, $P\left(X < 1 < 2X\right) = P(0.5 < X < 1) = 0.5$.

Definition 4.1. *An interval estimate of a real valued parameter θ is any pair of functions $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ of random sample only (do not involve any unknown parameters) that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the support of the RS. The random interval $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ is called an interval estimator of θ.*

Remark 4.1. Though in the definition, the closed interval $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ is written, the interval may be closed, open or semi-open based on the problem. †

If $L(\boldsymbol{x}) = -\infty$, then $U(\boldsymbol{x})$ provides an upper limit and $(-\infty,\, U(\boldsymbol{X}))$ is called upper interval estimator. Similarly, if $U(\boldsymbol{x}) = \infty$, then $L(\boldsymbol{x})$ provides a lower limit, and $(L(\boldsymbol{X}),\, \infty)$ is called lower interval estimator.

Example 4.1. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu,\, 1)$. Consider $L_1(\boldsymbol{x}) = x_1 - 1$, $U_1(\boldsymbol{x}) = x_1 + 1$, $L_2(\boldsymbol{x}) = \bar{x} - 1$, and $U_2(\boldsymbol{x}) = \bar{x} + 1$. Then both $[L_1(\boldsymbol{X}),\, U_1(\boldsymbol{X})]$ and $[L_2(\boldsymbol{X}),\, U_2(\boldsymbol{X})]$ are interval estimator of μ. Which one should we use? Note that here the lengths of both intervals are same, hence, one should use the interval estimator which has higher probability that the random interval includes μ.

$$P\left(X_1 - 1 \leq \mu \leq X_1 + 1\right) = P\left(-1 \leq X_1 - \mu \leq 1\right) = 2\Phi(1) - 1,$$

$$P\left(\overline{X}-1\leq\mu\leq\overline{X}+1\right)=P\left(-\sqrt{n}\leq\sqrt{n}\left(\overline{X}-\mu\right)\leq\sqrt{n}\right)=2\Phi(\sqrt{n})-1.$$

Now, as $\Phi(\cdot)$ is an increasing function, we should prefer $[L_2(\boldsymbol{X})=\overline{X}-1,\, U_2(\boldsymbol{X})=\overline{X}+1]$ over $[L_1(\boldsymbol{X})=X_1-1,\, U_1(\boldsymbol{X})=X_1+1]$. ||

Remark 4.2. In the previous example, as the length of the intervals are same, we prefer an interval for which the probability that the random interval covers the parameter μ is highest. In other cases, we may have interval estimators that have equal probability of covering the parameter. In such cases, we should prefer an interval which has minimum length. We will not study such optimality issues in this course. †

Definition 4.2 (Coverage Probability)**.** *Coverage probability associated with an interval estimator* $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ *for* θ *is measured by*

$$P_\theta\left(L(\boldsymbol{X})\leq\theta\leq U(\boldsymbol{X})\right).$$

Definition 4.3 (Confidence Coefficient)**.** *The confidence coefficient associated with an interval estimator* $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ *is defined by*

$$\inf_{\theta\in\Theta}P_\theta\left(L(\boldsymbol{X})\leq\theta\leq U(\boldsymbol{X})\right).$$

Definition 4.4 (Confidence Interval)**.** *Let* $\alpha\in(0,\,1)$*. An interval estimator* $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ *is said to be a confidence interval (CI) of level* $1-\alpha$ *(or a* $100(1-\alpha)\%$ *confidence interval) if*

$$P_\theta\left(L(\boldsymbol{X})\leq\theta\leq U(\boldsymbol{X})\right)\geq 1-\alpha \text{ for all } \theta\in\Theta.$$

Remark 4.3. Typical values of α are 0.1, 0.05, 0.01. †

Remark 4.4. Clearly, we are loosing precision in interval estimation compared to point estimation. Do we have any gain? Consider the previous example. A reasonable point estimator of μ is $\overline{X}$. However, $P\left(\overline{X}=\mu\right)=0$ as $\overline{X}$ is a CRV. On the other hand,

$$P\left(\overline{X}-1\leq\mu\leq\overline{X}+1\right)>0.$$

Hence, in interval estimation we have some confidence which we gain by loosing precision. †

4.1.1 Interpretation of Confidence Interval

Let us try to interpret a CI. Let $[L(\boldsymbol{X}),\, U(\boldsymbol{X})]$ be an interval estimator of the parameter θ. Once we observe $\boldsymbol{X}=\boldsymbol{x}$, an interval estimate $[L(\boldsymbol{x}),\, U(\boldsymbol{x})]$ is a fixed interval. Also, recall that the parameter θ is an unknown but fixed entity. Therefore, no probability is attached to these observed interval estimate. The interpretation of the phrase "$(1-\alpha)$ confidence" can be discussed as follows. Suppose that the RS is drawn repeatedly. For the first observation $\boldsymbol{X}=\boldsymbol{x}_1$, the interval estimate is $[L(\boldsymbol{x}_1),\, U(\boldsymbol{x}_1)]$. For the second observation $\boldsymbol{X}=\boldsymbol{x}_2$, the interval estimate is $[L(\boldsymbol{x}_2),\, U(\boldsymbol{x}_2)]$, and so on. If we keep on repeating this procedure, we will have interval estimates

$$[L(\boldsymbol{x}_1),\, U(\boldsymbol{x}_1)],\, [L(\boldsymbol{x}_2),\, U(\boldsymbol{x}_2)],\, [L(\boldsymbol{x}_3),\, U(\boldsymbol{x}_3)],\, [L(\boldsymbol{x}_4),\, U(\boldsymbol{x}_4)],\, \ldots.$$

In a long haul, out of these conceptual interval estimates found, approximately $100(1-\alpha)\%$ would include the unknown value of the parameter θ. This interpretation goes hand in hand with the relative frequency definition of probability.

4.2 Method of Finding CI

There are several ways of construction of CI. In this section, we will discuss the construction of CI based on pivot. The definition of pivot in given below.

Definition 4.5. *A random variable $T = T(\boldsymbol{X}, \theta)$ is called a pivot (or a pivotal quantity) if the distribution of T does not involve any unknown parameters.*

Remark 4.5. Pivot is a function of random sample and unknown parameters, but its' distribution is independent of all unknown parameters. Hence, pivot is not a statistic in general. †

Remark 4.6. In general, we want to find a pivot that is a function of minimal sufficient statistic. †

Example 4.2. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$. Then $\overline{X} - \mu$ is a pivot as $\overline{X} - \mu \sim N(0, 1/n)$. ||

Example 4.3. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and μ and σ both are unknown. Then $\overline{X} - \mu$ is not a pivot as $\overline{X} - \mu \sim N(0, \sigma^2/n)$. However, $\frac{\sqrt{n}}{\sigma}\left(\overline{X} - \mu\right) \sim N(0, 1)$ and $\frac{\sqrt{n}}{S}\left(\overline{X} - \mu\right) \sim t_{n-1}$. Therefore, these are pivots. ||

Example 4.4. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Exp(\lambda)$. Then $2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{2n}$ (why?), and hence, is a pivot. ||

Once an appropriate pivot is found, the CI for a parameter θ can be obtained as follows. Let T be a pivot. Find two real numbers a and b such that

$$P_\theta\left(a \leq T(\boldsymbol{X}, \theta) \leq b\right) \geq 1 - \alpha.$$

Note that a and b are independent of all unknown parameters as the distribution of T does not involve any unknown parameter. Let us denote the set

$$C(\boldsymbol{x}) = \left\{\theta \in \Theta : a \leq T(\boldsymbol{x}; \theta) \leq b\right\}.$$

Then, $C(\boldsymbol{X})$ is a $100(1-\alpha)\%$ CI for θ. Note that $C(\boldsymbol{x})$ does not involve any unknown parameters as a and b are independent of all unknown parameters. Also notice that if $T(\boldsymbol{x}; \theta)$ is monotone in $\theta \in \Theta$ for each $\boldsymbol{x}$, then $C(\boldsymbol{x})$ is an interval. Otherwise it could be a general set.

4.2.1 One-sample Problems

Example 4.5. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma > 0$ is known. We are interested in μ. A pivot based on minimal sufficient statistics is $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0, 1)$. Let z_α be the upper α-point of the standard normal distribution. We can take $a = z_{1-\alpha/2} = -z_{\alpha/2}$ (as $N(0, 1)$ distribution is symmetric about zero) and $b = z_{\alpha/2}$. Now,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha \implies P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

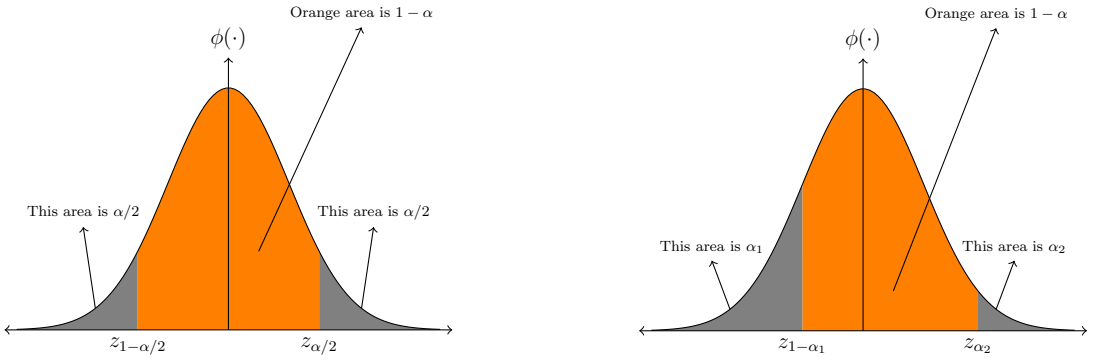


Figure 4.1: Symmetric and asymmetric CIs

Hence, a $100(1-\alpha)\%$ symmetric CI for μ is

$$C(\boldsymbol{X}) = \left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}\right]. \qquad ||$$

Note that the choice of $a = -z_{\frac{\alpha}{2}}$ and $b = z_{\frac{\alpha}{2}}$ corresponds to symmetric CI, as we leave $\frac{\alpha}{2}$ probability on both sides and take the middle part of the probability distribution. Of course, there are infinite number of choices for a and b. For example, let $\alpha_1 > 0$, $\alpha_2 > 0$ are to real numbers such that $\alpha_1 + \alpha_2 = \alpha$. Then, $a = z_{1-\alpha_1}$ and $b = z_{\alpha_2}$ can be considered (see the right panel of Figure 4.1).

Example 4.6. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is known and $\sigma > 0$ is unknown. We are interested in CI of σ^2. A pivot based on minimal sufficient statistics is $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2_n$. Let $\chi^2_{n,\alpha}$ be the upper α-point of a χ^2-distribution with degrees of freedom n. We can take $a = \chi^2_{n,1-\alpha/2}$ and $b = \chi^2_{n,\alpha/2}$. Hence, a $100(1-\alpha)\%$ symmetric CI for σ^2 is

$$C(\boldsymbol{X}) = \left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,1-\alpha/2}}\right]. \qquad ||$$

Example 4.7. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown. We are interested in CI of μ. A pivot based in minimal sufficient statistic is $\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Let $t_{n,\alpha}$ be the upper α-point of a t-distribution with degrees of freedom n. Then, we can take $a = t_{n-1,1-\alpha/2} = -t_{n-1,\alpha/2}$ (as t-distribution is symmetric about zero) and $b = t_{n-1,\alpha/2}$. Hence, a $100(1-\alpha)\%$ symmetric CI for μ is

$$C(\boldsymbol{X}) = \left[\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right]. \qquad ||$$

Example 4.8. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown. We are interested in CI for σ^2. A pivot based on minimal sufficient statistic is $\frac{1}{\sigma^2}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \sim \chi^2_{n-1}$. We can take $a = \chi^2_{n-1,1-\alpha/2}$ and $b = \chi^2_{n-1,\alpha/2}$. Hence, a $100(1-\alpha)\%$ symmetric CI for σ^2 is

$$C(\boldsymbol{X}) = \left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right]. \qquad ||$$

4.2.2 Two-sample Problems

Example 4.9. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$. Also, assume that X_i's and Y_j's are independent. Here, μ_1, μ_2, and σ are assumed to be unknown and we are interested to construct a $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$. Let us first try to construct a pivot based on minimal sufficient statistic $\left(\overline{X}, \overline{Y}, S^2\right)$, where the pooled sample variance S^2 is defined by

$$S^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2\right].$$

Now, notice that

$$\overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2,\, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right) \implies T = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,\, 1).$$

Of course, T is a pivot, but we cannot use it to construct the required confidence interval due to the presence of unknown σ in T. Also, note

$$\frac{(n+m-2)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2 \sim \chi^2_{n+m-2}.$$

Moreover, S^2 and $\left(\overline{X}, \overline{Y}\right)$ are independent. Therefore,

$$T_1 = \frac{\frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{\frac{(n+m-2)S^2}{(n+m-2)\sigma^2}}} = \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}.$$

Thus, T_1 is a pivot and we will use it to construct CI for $\mu_1 - \mu_2$.

$$P\left(-t_{n+m-2,\,\frac{\alpha}{2}} \le \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n} + \frac{1}{m}}} \le t_{n+m-2,\,\frac{\alpha}{2}}\right) = 1 - \alpha$$

Therefore, the symmetric $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$C(\boldsymbol{X}) = \left[\left(\overline{X} - \overline{Y}\right) - S\sqrt{\frac{1}{n} + \frac{1}{m}}t_{n+m-2,\,\frac{\alpha}{2}},\ \left(\overline{X} - \overline{Y}\right) + S\sqrt{\frac{1}{n} + \frac{1}{m}}t_{n+m-2,\,\frac{\alpha}{2}}\right]||$$

Example 4.10. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$. Also, assume that X_i's and Y_j's are independent. Here, μ_1, μ_2, σ_1, and σ_2 are assumed to be unknown and we are interested to construct a $100(1-\alpha)\%$ CI for $\dfrac{\sigma_2^2}{\sigma_1^2}$. In this case minimal sufficient statistic is $\left(\overline{X}, \overline{Y}, S_1^2, S_2^2\right)$, where S_1^2 and S_2^2 are sample variances based on the samples X_i's and Y_j's, respectively. In this case,

$$T = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} = \frac{\frac{(n-1)S_1^2}{(n-1)\sigma_1^2}}{\frac{(m-1)S_2^2}{(m-1)\sigma_2^2}} \sim F_{n-1,\,m-1}.$$

Thus,

$$P\left(F_{n-1,\,m-1,\,1-\frac{\alpha}{2}} \leq \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \leq F_{n-1,\,m-1,\,\frac{\alpha}{2}}\right) = 1-\alpha$$

so that a $100(1-\alpha)\%$ CI for $\dfrac{\sigma_2^2}{\sigma_1^2}$ is

$$C(\boldsymbol{X}) = \left[\frac{S_2^2}{S_1^2} F_{n-1,\,m-1,\,1-\frac{\alpha}{2}},\ \frac{S_2^2}{S_1^2} F_{n-1,\,m-1,\,\frac{\alpha}{2}}\right]. \qquad ||$$

4.3 Asymptotic CI

In many cases it is very difficult to find pivot for a small sample. For example, it is difficult to find a pivot to construct CI for successes probability of a Bernoulli distribution. However, we may able to find CI quite easily if the sample size is sufficiently large. This CI is called asymptotic confidence interval. For this purpose, convergence in distribution (mainly CLT or large sample distribution of MLE) and convergence in probability (consistent estimator) are handy tools.

4.3.1 Distribution Free Population Mean

Let X_1, X_2, ... be i.i.d. random variables with mean μ and finite variance σ^2. Then, using CLT

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \xrightarrow{\mathscr{L}} Z \sim N(0,\,1).$$

Thus, if we have a RS with large sample size n, we can approximate the distribution of $\frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{\sigma}$ using a standard normal distribution. Hence,

$$P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \leq z_{\alpha/2}\right) \approx 1-\alpha.$$

If σ is known and n is sufficiently large, we can use the last statement to find an asymptotic CI for μ and it is given by

$$\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\, \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right].$$

If σ is unknown, we can proceed as follows. Using WLLN, we have $\frac{S_n}{\sigma} \xrightarrow{P} 1$, and hence,

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{S_n} \xrightarrow{\mathscr{L}} Z \sim N(0,\,1).$$

Hence, $P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X}_n-\mu\right)}{S_n} \leq z_{\alpha/2}\right) \approx 1-\alpha$. An asymptotic CI for μ is given by

$$\left[\overline{X}_n - \frac{S_n}{\sqrt{n}} z_{\alpha/2},\, \overline{X}_n + \frac{S_n}{\sqrt{n}} z_{\alpha/2}\right].$$

Note that this method can be used for any distribution of X_1, X_2, ..., X_n, as long as the conditions of CLT hold true. Therefore, it is called distribution free.

4.3.2 Using MLE

Let $\widehat{\theta}_n$ be a consistent estimator of θ and $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathscr{L}} N(0, b^2(\theta))$, where $b(\theta) > 0$ for all $\theta \in \Theta$. Assume that $b(\cdot)$ is a continuous function. Then, $\frac{b(\widehat{\theta}_n)}{b(\theta)} \xrightarrow{P} 1$, and hence, $\frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{b(\widehat{\theta}_n)} \xrightarrow{\mathscr{L}} N(0, 1)$. A $100(1-\alpha)\%$ asymptotic CI for θ is given by

$$\left[\widehat{\theta}_n - \frac{b(\widehat{\theta}_n)}{\sqrt{n}} z_{\alpha/2},\, \widehat{\theta}_n + \frac{b(\widehat{\theta}_n)}{\sqrt{n}} z_{\alpha/2}\right].$$

Under some regularity conditions, we may use MLE of θ in place of $\widehat{\theta}_n$.

Example 4.11. Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernaoulli(p)$, where $p \in (0, 1)$. We are interested to construct asymptotic CI for p. We know that $\widehat{p}_n = \overline{X}_n \xrightarrow{P} p$ and $\frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{p(1-p)}} \xrightarrow{\mathscr{L}} N(0, 1)$. Here, $b(p) = \sqrt{p(1-p)}$, which is a continuous function in $p \in (0, 1)$. Hence, $\frac{\sqrt{n}\left(\overline{X}_n - p\right)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \xrightarrow{\mathscr{L}} N(0, 1)$. A $100(1-\alpha)\%$ asymptotic CI for p is

$$\left[\overline{X}_n - \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} z_{\alpha/2},\, \overline{X}_n + \sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}} z_{\alpha/2}\right]. \qquad ||$$