Statistical Inference and Multivariate Analysis (MA 324)

Class Notes
January – May, 2021

Instructor
Ayon Ganguly
Department of Mathematics
IIT Guwahati

Contents

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 will work 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 on $\boldsymbol{\theta}$ and $\boldsymbol{x}$ 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 $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a sufficient statistic for (μ, σ^2). As the mapping $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \to \left(\overline{X}, S^2\right)$ 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

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

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, $x_{(n)} = y_{(n)}$. Similarly, we can work with the case $x_{(n)} > y_{(n)}$. 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 Chennai 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 Chennai. 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,

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

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

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,

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

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 ease 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 ease 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 ease 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 $T^* \in \mathcal{C}$,*

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

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. ||