

UNIT 1

EXPONENTIAL FAMILY AND COMPLETENESS

Structure

1.1 Introduction	1.6 Properties of Complete Statistic
Expected Learning Outcomes	
1.2 Basic Terminology	1.7 Summary
1.3 Minimum Variance Unbiased Estimator	1.8 Terminal Questions
1.4 Exponential Family	1.9 Solutions /Answers
1.5 Completeness	

1.1 INTRODUCTION

In Block 2 of the course MST-016: Statistical Inference, you have studied various properties of an estimator (a function of sample observations which is used to estimate the unknown parameter), namely: unbiasedness, consistency, efficiency and sufficiency. Let us have a look at them.

- **Unbiasedness:** An estimator is said to be unbiased for a parameter θ if and only if the average/mean of the sampling distribution of the estimator is equal to the true value of the parameter. In other words, an estimator is said to be unbiased **if the expected value of the estimator is equal to the true value of the parameter being estimated**, that is, $E(T) = \theta$.
- **Consistency:** An estimator T_n is said to be a consistent estimator of a parameter θ if T_n converges to θ in probability, that is, $T_n \xrightarrow{P} \theta$.
- **Efficiency:** If T_1 and T_2 are two unbiased estimators of a parameter θ , then the estimator T_1 is said to be more efficient than the estimator T_2 if $\text{Var}(T_1) < \text{Var}(T_2)$ for all n .
- **Sufficiency:** An estimator is said to be a sufficient estimator/statistic if it captures all the information in the sample about the unknown parameter, leaving no additional useful information to be extracted.

In the continuation of the search for the best estimator, the practical question that may arise is **how we find the estimator which is best**. In this block, we

Tools You Will Need

The following terms are considered essential background material for this Unit. If you doubt your knowledge of any of these terms, you should review the appropriate Unit or section before proceeding:

- Sampling distributions (Units 2,3, 4 and 5 of MST-016: Statistical Inference).
- Probability distributions (MST-012: Probability and Probability Distributions).

will discuss two important methods for finding the best estimator of a parameter. These methods are as follows:

- Cramér-Rao inequality
- Lehmann-Scheffé theorem

We will discuss the Cramér-Rao inequality and the Lehmann-Scheffé theorem in the subsequent units of this block.

To apply the Cramér-Rao inequality and Lehmann-Scheffé theorem, you have to have some idea of a minimum variance unbiased estimator, Fisher information, sufficiency and completeness. Therefore, this unit is devoted to explaining some of the terms required for applying these.

This unit is divided into nine sections. Section 1.1 is introductory in nature. The basic terms used to understand and find UMVUE are defined in Section 1.2. Section 1.3 is devoted to explaining MVUE and UMVUE. Sections 1.4 and 1.5 explore the exponential family of distributions and the concept of completeness with examples. The properties of complete statistic are discussed in Section 1.6. The unit ends by providing a summary of what we have discussed in this unit in Section 1.7. The terminal questions and the solution of the SAQs/TQs are given in Sections 1.8 and 1.9, respectively.

In the next unit, we shall discuss the Cramér-Rao inequality, which is used to find UMVUE.

Expected Learning Outcomes

After studying this unit, you should be able to:

- ❖ describe the uniformly minimum variance unbiased estimator (UMVUE);
- ❖ describe the general structure of the exponential family and its importance;
- ❖ differentiate between sufficiency and completeness;
- ❖ define a complete statistic and explain its importance in finding UMVUE;
- ❖ apply the concept of the exponential family approach to check whether a statistic is complete or not; and
- ❖ describe various properties of a complete statistic.

1.2 BASIC TERMINOLOGY

Before discussing the Cramér-Rao inequality and the Lehmann-Scheffé theorem to find the UMVUE, we quickly review various distributions (discrete and continuous) and the concept of mathematical expectation, which you have studied in the course MST-012: Probability and Probability Distributions. These terms are very useful in finding UMVUE.

Discrete and Continuous Distributions

In Units 9 to 16 of MST-012, we have discussed standard discrete and continuous distributions as binomial, Poisson, normal, exponential, etc. We know that a population can be described with the help of a distribution, therefore, standard discrete and continuous distributions are also used in

statistical inference. Here, we discuss some standard discrete and continuous distributions in brief as in tabular form, and you have to learn at least the mean and variance of these distributions, which will help you to find UMVUE.

S. No.	Distribution	Parameter(s)	Mean	Variance
1	Bernoulli (discrete) $P[X = x] = p^x(1-p)^{1-x}; x = 0,1$	p	p	$p(1-p)$
2	Binomial (discrete) $P[X = x] = {}^n C_x p^x (1-p)^{n-x}; x = 0,1, \dots, n$	n & p	np	$np(1-p)$
3	Poisson (discrete) $P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; x = 0,1, \dots; \lambda > 0$	λ	λ	λ
4	Uniform (discrete) $P[X = x] = \frac{1}{n}; x = 1,2, \dots, n$	n	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
5	Hypergeometric (discrete) $P[X = x] = \frac{{}^M C_x {}^{N-M} C_{n-x}}{{}^N C_n}; x = 0,1, \dots, \min\{M,n\}$	N, M & n	$\frac{nM}{N}$	$\frac{NM(N-M)(N-n)}{N^2(N-1)}$
6	Geometric (discrete) $P[X = x] = p(1-p)^x; x = 0,1,2, \dots$	p	$\frac{1-p}{p}$	$\frac{(1-p)^2}{p}$
7	Negative Binomial (discrete) $P[X = x] = \binom{x+r-1}{r-1} p^r (1-p)^x; x = 0, 1, 2, \dots$	r & p	$\frac{r(1-p)}{p}$	$\frac{r(1-p)^2}{p}$
8	Normal (continuous) $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; -\infty < x < \infty;$ $\sigma > 0, -\infty < \mu < \infty$	μ & σ^2	μ	σ^2
9	Standard Normal (continuous) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}; -\infty < x < \infty$	--	0	1
10	Uniform (continuous) $f(x) = \frac{1}{b-a}; a < x < b, b > a$	a & b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
11	Exponential (continuous) $f(x) = \theta e^{-\theta x}; x \geq 0; \theta > 0$ Negative Exponential or simply exponential (continuous) $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}}; x \geq 0; \theta > 0$	θ	$\frac{1}{\theta}$	$\frac{1}{\theta^2}$
12	Gamma (continuous) $f(x) = \frac{b^a}{\Gamma(a)} e^{-bx} x^{a-1}; x > 0; a, b > 0$	a & b	$\frac{b}{a}$	$\frac{b}{a^2}$
.13	Beta First Kind (continuous) $f(x) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1}; 0 < x < 1$ $a > 0, b > 0$	a & b	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$

14	Beta Second Kind (continuous) $f(x) = \frac{1}{B(a,b)} \frac{x^{a-1}}{(1+x)^{a+b}}; x > 0; a, b > 0$	a & b	$\frac{a}{b-1}$	$\frac{a(a+b+1)}{(b-1)^2(b-2)}$
15	Standard Cauchy $f(x) = \frac{1}{\pi(1+x^2)}; -\infty < x < \infty$	---	Does not exist	Does not exist
16	Laplace $f(x) = \frac{1}{2b} e^{-\frac{ x-\mu }{b}}; -\infty < x < \infty$	μ & b	μ	$2b^2$

Mathematical Expectation

If X is a continuous random variable having the probability density function $f(x)$, then the expected value of X (mean) is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

and in general

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

If X is a discrete random variable having the probability mass function $p(x)$, then the expected value of X is defined as

$$E(X) = \sum_{i=1}^n x_i p(x_i)$$

and in general

$$E[g(x)] = \sum_{i=1}^n g(x_i) p(x_i)$$

Some properties of mathematical expectation are:

- $E(a) = a$ where 'a' is a constant
- $E(aX) = aE(X)$
- $E(aX \pm bY) = aE(X) \pm bE(Y)$

Variance

If X is a random variable, then the variance of X in terms of expectation is defined as

$$\text{Var}(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

Some properties of variance are:

- $\text{Var}(a) = 0$
- $\text{Var}(aX) = a^2 \text{Var}(X)$
- If random variables X and Y are independent, then
 $\text{Var}(aX \pm bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$

We now define the uniformly minimum variance unbiased estimator in the next section.

1.3 MINIMUM VARIANCE UNBIASED ESTIMATOR

In statistical estimation, the aim of the statistician is to estimate the unknown parameter (θ) on the basis of an estimator/statistic. A statistic which is used in estimation or drawing inferences about the population parameter is a function of sample observations. Therefore, there may be lots of estimators. For example, suppose you are interested in knowing the average age of Facebook users on the basis of a sample X_1, X_2, \dots, X_n of size n . Therefore, we may propose a number of estimators for estimating the unknown average age of Facebook users. Some possible estimators are:

- $T_1(X_1, X_2, \dots, X_n) = \text{Sample mean } (\bar{X}) = \frac{X_1 + X_2 + \dots + X_n}{n}$
- $T_2(X_1, X_2, \dots, X_n) = \text{Sample median } (\tilde{X}) = \text{median}(X_1, X_2, \dots, X_n)$
- $T_3(X_1, X_2, \dots, X_n) = \text{Sample mode } (X_0) = \text{mode}(X_1, X_2, \dots, X_n)$
- $T_4(X_1, X_2, \dots, X_n) = \frac{\max(X_1, X_2, \dots, X_n) + \min(X_1, X_2, \dots, X_n)}{2}$

Out of such estimators, an estimator is said to be the best estimator of a population parameter θ (average age of Facebook users) if it is close to the parameter θ and has the minimum variance. In other words, an estimator is best if the average of its sampling distribution is equal to the parameter and does not too spread out around the true value of the parameter, because if it is too spread out, then there will be a high probability that an estimate could be generated that will have a significant distance from the true value of the parameter. The foregoing considerations motivate to use of an unbiased estimator of the parameter θ that also has minimum variance among all unbiased estimators of θ . Such an estimator is called a **minimum variance unbiased estimator (MVUE)**. We can define it as follows:

An estimator T of the parameter θ is said to be a minimum variance unbiased estimator of θ if and only if

- (i) $E(T) = \theta$, that is, the estimator T is an unbiased estimator of the parameter θ ; and
- (ii) $\text{Var}(T) \leq \text{Var}(T')$ where T' is any other unbiased estimator of parameter θ .

The above definition implies that an estimator is a minimum variance unbiased estimator (MVUE) if and only if the estimator is unbiased and if there is no other unbiased estimator that has a smaller variance for any value of θ .

If an estimator is unbiased and has the smallest variance among the unbiased estimator for **all values of the parameter**, then:

- The estimator is efficient, and
- It is called a **uniformly minimum variance unbiased estimator (UMVUE)**.

Let us discuss the exponential family of distributions in the next section, which helps us to find the MVUE/UMVUE.

1.4 EXPONENTIAL FAMILY

We have discussed standard discrete and continuous distributions as binomial, Poisson, normal, exponential, etc. Out of these, many distributions are members of what is called the exponential class, or exponential family. You do not confuse exponential density or negative exponential density, which is a special case of the exponential family. The exponential family of distributions plays a crucial role in statistics and data science because it encompasses a wide range of commonly used probability distributions and has several important theoretical and practical properties.

The general form of the one-parameter exponential family is expressed in the following form:

$$f(x; \theta) = a(\theta)b(x)\exp[c(\theta)d(x)] \text{ for all values of the parameter } \theta.$$

where

- x : The observed data.
- $a(\theta)/c(\theta)$: A function of the parameter of the distribution.
- $b(x)/d(x)$: A non-negative function that does not depend on the parameter θ .

Similarly,

The general form of the k -parameter exponential family is expressed in the following form:

$$f(x; \theta_1, \theta_2, \dots, \theta_k) = a(\theta_1, \theta_2, \dots, \theta_k)b(x)\exp\sum_{j=1}^k c_j(\theta_1, \theta_2, \dots, \theta_k)d_j(x)$$

A probability density/mass function belongs to the exponential family or exponential class for a suitable choice of functions $a(\theta)$, $b(x)$, $c(\theta)$ and $d(x)$.

Let us take some examples to illustrate how we check whether a continuous or discrete distribution belongs to the exponential family or not.

Example 1: The lifetime X (in hours) of a battery follows an exponential distribution whose probability density function is given by

$$f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}; \quad x > 0, \theta > 0$$

where θ , represents the mean lifetime.

Check whether it belongs to the exponential family or not.

Solution: To check whether a distribution (continuous or discrete) belongs to the exponential family, we try to express the probability density or mass function of the given distribution in the general form of the exponential family. Therefore, we can express the probability density function of the given exponential distribution as

$$f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} = \frac{1}{\theta} \times 1 \times e^{-\frac{1}{\theta}x} = a(\theta)b(x)e^{c(\theta)d(x)}$$

where $a(\theta) = 1/\theta$, $b(x) = 1$, $c(\theta) = -1/\theta$ and $d(x) = x$

Since the probability density function of the exponential distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = 1/\theta$, $b(x) = 1$, $c(\theta) = -1/\theta$ and $d(x) = x$.

Some authors, use the following form of the one-parameter exponential family:

$$f(x; \theta) = \exp \left[\frac{\eta(\theta)T(x) + A(\theta) + B(x)}{\quad} \right]$$

Note: Most of the well-known distributions belong to the exponential family. Any probability density or mass function for which the range of the variable depends on the parameter does not belong to the exponential family. Therefore, a family of uniform densities does not belong to the exponential family.

Before moving to the next section, you can assess your understanding by answering the following Self-Assessment Question.

SAQ 1

The transport department of a country monitors the number of cars passing through a toll booth on a national highway. They observed, on average, 10 cars per minute passing through the highway toll booth. The number of cars per minute follows a Poisson distribution with a rate $\lambda = 10$ per minute. Check whether it belongs to the exponential family or not.

To find the uniformly minimum variance unbiased estimator of an unknown parameter using the Lehmann-Scheffé theorem, we use completeness. Let us understand what it is in the next section.

1.5 COMPLETENESS

The concept of completeness helps us to find the UMVUE. Completeness ensures that no other estimator can perform better by using additional information from the same sample. This concept is critical in statistical theory, especially when determining the **uniformly minimum variance unbiased estimator**.

Let us explore the concept of a **complete estimator** in more depth, including its formal definition, properties, examples, and practical significance.

An estimator is said to be complete if it incorporates all of the information from the data about the parameter being estimated, such that there is no unbiased estimator that could enhance it by including additional or different information.

Mathematically,

If X_1, X_2, \dots, X_n is a random sample of size n taken from a population whose probability density (or mass) function is $f(x; \theta)$ where θ is the population parameter, then an estimator $T = t(X_1, X_2, \dots, X_n)$ or family of $f(x; \theta)$ is said to be a complete estimator for the parameter θ or family of $f(x; \theta)$ if and only if for any function 'h' of the estimator T :

$$E[h(T)] = 0 \quad \text{for all values of the parameter } \theta$$

such that

$$h(T) = 0 \text{ almost surely (i.e., with probability 1).}$$

This means that no non-zero function of estimator T can have an expected value of zero for all possible values of the parameter θ .

Another way of saying that a statistic/estimator is a complete statistic as follows:

“An estimator T is said to be complete if and only if the only unbiased

The pdf of the uniform distribution $U[a, b]$ is given by

$$f(x) = \frac{1}{b-a}; \quad a < x < b, b > a$$

The range of the variable depends on the parameters 'a' and 'b'.

estimator of 0 that is a function of estimator T is the statistic that is identically 0 with probability 1.”

Completeness is particularly relevant when discussing **unbiased estimators**. In certain cases, a **complete and sufficient statistic** leads to an **unbiased estimator**. This result is part of the **Lehmann-Scheffé theorem**, which states that a complete and sufficient statistic is the **best unbiased estimator** (BUE) for a parameter, meaning it is unbiased and has the lowest variance among all unbiased estimators. Let us discuss the importance of a complete statistic in estimation.

Importance of Complete Statistic

- Completeness indicates that a statistic contains all relevant information about a parameter.
- The key difference between sufficiency and completeness lies in their purpose and how they relate to parameter estimation. Sufficiency is about capturing all the information in the sample relevant to a parameter. Once a sufficient statistic is known, the rest of the sample provides no additional information about that parameter. In contrast, completeness ensures uniqueness. A statistic is complete if no non-trivial function of it has an expected value of zero for all parameter values, unless that function is almost surely zero.
- A complete estimator ensures that the estimation process has fully utilised the data, leading to more reliable and efficient estimates.

To show an estimator is complete, you have to follow the following steps:

Step 1: First of all, we assume that the expected value of the function of the estimator T, say, h is zero, that is,

$$E[h(T)] = 0 \text{ for all values of the parameter } \theta$$

Step 2: After that, we find the value of $E[h(T)]$ using the probability density (mass) function of the estimator T, such as

$$E[h(T)] = \int_t h(t)g(t)dt \text{ (for continuous case)}$$

$$E[h(T)] = \sum_t h(t)P[T = t] \text{ (for discrete case)}$$

Step 3: Finally, we demonstrate that the only solution to $E[h(T)] = 0$ for all values of the parameter θ is $h(T) = 0$ (almost surely with respect to the distribution of T).

If the above steps hold, then the estimator T is a complete estimator/ statistic.

Let us consider the following examples, which help you to understand the process of verifying whether an estimator is complete or not.

Example 2: Suppose there is an experiment where you flip a coin. If the outcome of the flip is head, then you will win. Suppose the probability of getting a head is p, and you flip the coin n times. Also, suppose X_1, X_2, \dots, X_n represent the outcomes, then show that

- (i) $\sum_{i=1}^n X_i$ is a complete statistic.

(ii) $T = X_1 - X_2$ is not a complete statistic.

Solution: Here, X_1, X_2, \dots, X_n represent the outcomes of a coin which flipped independently n times and the probability of getting a head is constant (p), therefore, they follow the Bernoulli distribution with parameter p . To check whether $T = \sum_{i=1}^n X_i$ is a complete statistic or not, we consider

$$E[h(T)] = 0$$

To find the above expectation, we require the distribution of the estimator

$T = \sum_{i=1}^n X_i$. Since it is the sum of n independent Bernoulli variables, therefore,

the statistic/ estimator $T = \sum_{i=1}^n X_i$ follows a binomial distribution (n, p) whose

probability mass function is given by

$$P[T = t] = {}^n C_t p^t (1-p)^{n-t}; \quad t = 0, 1, \dots, n$$

Therefore, we can compute the expectation as follows:

$$E[h(T)] = \sum_{t=1}^n h(t) {}^n C_t p^t (1-p)^{n-t} = 0$$

We can write the above expression as

$$(1-p)^n \sum_{t=1}^n h(t) {}^n C_t \left(\frac{p}{1-p}\right)^t = 0$$

For all p , $0 < p < 1$, the factor $(1-p)^n$ is not 0 for any p in the range 0 to 1. Thus, it must be that

$$\sum_{t=1}^n h(t) {}^n C_t \left(\frac{p}{1-p}\right)^t = \sum_{t=1}^n h(t) {}^n C_t z^t = 0 \quad \left(\text{where } z = \frac{p}{1-p}\right)$$

It is a polynomial of degree n in z . For the polynomial to be 0 for all z , each coefficient must be 0. Since none of the ${}^n C_t$ terms is 0, this implies that $h(t) = 0$ for $t = 1, 2, \dots, n$. Since T takes values $0, 1, \dots, n$ with probability 1, this yields that $P[h(T) = 0] = 1$ for all p . Hence, $T = \sum_{i=1}^n X_i$ is a complete statistic.

We now check whether $T = X_1 - X_2$ is complete or not.

Here, $T = X_1 - X_2$ is defined as the difference between the outcomes of the first two coin flips. It can take values $-1, 0, 1$ based on the outcomes of X_1 and X_2 :

- $T = 1$ when $X_1 = 1, X_2 = 0$
- $T = 0$ when $X_1 = 0, X_2 = 0$ or $X_1 = 1, X_2 = 1$
- $T = -1$ when $X_1 = 0, X_2 = 1$

Also,

$$\begin{aligned} P[T = 1] &= P[X_1 = 1 \text{ and } X_2 = 0] \\ &= P[X_1 = 1] P[X_2 = 0] = p(1-p) \quad [\because X_1 \text{ and } X_2 \text{ are independent}] \end{aligned}$$

Similarly,

$$\begin{aligned}
P[T = 0] &= P[X_1 = 0 \text{ and } X_2 = 0 \text{ or } X_1 = 1 \text{ and } X_2 = 1] \\
&= P[X_1 = 0 \text{ and } X_2 = 0] + P[X_1 = 1 \text{ and } X_2 = 1] \\
&= P[X_1 = 0]P[X_2 = 0] + P[X_1 = 1]P[X_2 = 1] \\
&= (1-p)^2 + p^2
\end{aligned}$$

$$\begin{aligned}
P[T = -1] &= P[X_1 = 0 \text{ and } X_2 = 1] \\
&= P[X_1 = 0]P[X_2 = 1] = (1-p)p
\end{aligned}$$

Therefore, the probability mass function of T is

T	-1	0	1
P[T = t]	p(1-p)	(1-p) ² + p ²	p(1-p)

The probability mass function of T depends only on p but does not fully capture the information about p because it ignores the outcomes of the other n - 2 flips. Thus, T is not sufficient, and hence, it cannot be complete.

Let us take another example.

Example 3: The life of the lithium batteries of a brand used in cars follows a normal distribution with mean μ months and standard deviation of σ months. To estimate the average life of the batteries, a researcher measured the lifetimes of n randomly selected batteries as X_1, X_2, \dots, X_n , then show that the sample mean (\bar{X}) is a complete statistic for μ .

Solution: To check whether the sample mean \bar{X} for μ is a complete statistic or not, we consider

$$E[h(T)] = E[h(\bar{X})] = 0$$

To find the above expectation, we require the sampling distribution of the estimator \bar{X} . Since the life of the lithium batteries used in cars follows a normal distribution with mean μ months and a known standard deviation of σ months, therefore, the sampling distribution of the average life of the batteries follows a normal distribution with mean

$$E(\bar{X}) = \mu \text{ and variance } \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

That is

$\bar{X} \sim N(\mu, \sigma^2 / n)$ and its probability density function is given by

$$g(\bar{x}) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{1}{2} \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right)^2}; \quad -\infty < \bar{x} < \infty, -\infty < \mu < \infty$$

Therefore, we can compute the expectation as follows:

$$E[h(\bar{X})] = \int_{-\infty}^{\infty} h(\bar{x})g(\bar{x})d\bar{x} = \int_{-\infty}^{\infty} h(\bar{x}) \frac{1}{\sqrt{2\pi \times \sigma^2/n}} e^{-\frac{1}{2} \left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right)^2} d\bar{x} = 0$$

Putting $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = z$, we get $\bar{x} = \mu + z \frac{\sigma}{\sqrt{n}}$ and $d\bar{x} = \frac{\sigma}{\sqrt{n}} dz$

Therefore,

$$E[h(\bar{X})] = \int_{-\infty}^{\infty} h\left(\mu + z \frac{\sigma}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0$$

This integral

$$\int_{-\infty}^{\infty} h\left(\mu + z \frac{\sigma}{\sqrt{n}}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0$$

must hold for every real number μ .

This implies that the function 'h' must be such that its "shifted" version to zero under the standard normal distribution for any shift μ . The only function that satisfies this condition for all μ is the zero function almost everywhere. Hence, the sample mean \bar{X} is complete.

I think you understood how we check whether a statistic/estimator is complete or not. You may be curious to check it yourself. For that, you can try the following Self-Assessment Question.

SAQ 2

The magnitude of the earthquakes recorded in a region modelled as an exponential distribution with an unknown parameter θ whose pdf is given by

$$f(x; \theta) = \theta e^{-\theta x}; \quad x > 0, \theta > 0$$

If a seismologist measured the magnitude of the n random earthquakes in that region and denoted them as X_1, X_2, \dots, X_n , then show that $\sum_{i=1}^n X_i$ is complete.

In general, the method used to verify completeness must be derived on a case-by-case basis. To check whether a statistic is complete or not with the help of the definition, is cumbersome. In fact, most statisticians consider it **extremely difficult**. One slightly easier way is to use the “**Exponential Family**”.

Let us state a Theorem that gives us a minimal sufficient statistic (which you have studied in Unit 9 of MST-016) as well as the complete statistics using the exponential family.

Theorem 1: Let X_1, X_2, \dots, X_n be a random sample taken from a density/mass function $f(x; \theta)$. If the probability density/mass function $f(x; \theta)$ belongs to the exponential family, that is, expressed in the form

$$f(x; \theta) = a(\theta)b(x) \exp[c(\theta)d(x)]$$

then $\sum_{i=1}^n d(X_i)$ is a complete as well as minimal sufficient statistic.

Note: You can also use Theorem 1 to show that a statistic/estimator is a minimal sufficient statistic as discussed in Unit 9 of the course MST-016: Statistical Inference.

Let us take some examples to illustrate how we check whether a statistic/estimator is complete using the exponential family approach.

Example 4: The magnitude of the earthquakes recorded in a region modelled as an exponential distribution with an unknown parameter θ whose pdf is given by

$$f(x; \theta) = \theta e^{-\theta x}; \quad x > 0, \theta > 0$$

Check whether it belongs to the exponential family or not. If a seismologist measured the magnitude of the n random earthquakes in that region and denoted them as X_1, X_2, \dots, X_n , then show that $\sum_{i=1}^n X_i$ is a complete and sufficient statistic.

Solution: Here, we will show that $\sum_{i=1}^n X_i$ is a complete and sufficient statistic

using the exponential family approach. To check whether the given distribution belongs to the exponential family, we try to express the probability density function of the given distribution in the general form of the exponential family. Thus, we can express the probability density function of the given exponential distribution as

$$f(x; \theta) = \theta \times 1 \times e^{-\theta x} = a(\theta)b(x)e^{c(\theta)d(x)}$$

where $a(\theta) = \theta$, $b(x) = 1$, $c(\theta) = -\theta$ and $d(x) = x$

Since the density function of the exponential distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = \theta$, $b(x) = 1$, $c(\theta) = -\theta$ and $d(x) = x$.

Since the exponential distribution of the magnitude of the earthquakes belongs to the exponential family, therefore, according to Theorem 1, $\sum_{i=1}^n d(X_i) = \sum_{i=1}^n X_i$ is a complete as well as a sufficient statistic.

Let us consider another example.

Example 5: If the number of wrong calls received by a company follows a Poisson distribution with parameter λ whose pdf is given by

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots \quad \& \lambda > 0$$

Check whether it belongs to the exponential family or not. If a telecom company selects a random sample of the wrong calls in a week, say, $X_1, X_2,$

\dots, X_n , then show that $\sum_{i=1}^n X_i$ is a sufficient as well as a complete statistic.

Solution: To check whether the Poisson distribution belongs to the exponential family, we try to express the given distribution in the general form of the exponential family. Therefore, we can express the pmf of the given Poisson distribution as

$$\begin{aligned} P[X = x] &= \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \left(\frac{1}{x!} \right) \exp[(\log \lambda) x] = a(\theta)b(x)e^{c(\theta)d(x)} \left[\because \lambda^x = e^{\log(\lambda^x)} = e^{x \log(\lambda)} \right] \end{aligned}$$

where $a(\theta) = e^{-\lambda}$, $b(x) = \frac{1}{x!}$, $c(\theta) = \log \lambda$ and $d(x) = x$.

Since the probability mass function of the Poisson distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = e^{-\lambda}$, $b(x) = \frac{1}{x!}$, $c(\theta) = \log \lambda$ and $d(x) = x$.

Since the probability mass function of the Poisson distribution of the number of wrong calls belongs to the exponential family, therefore, according to Theorem

1, $\sum_{i=1}^n d(X_i) = \sum_{i=1}^n X_i$ is a complete as well as a sufficient statistic.

Before moving to the next section, you can assess your understanding by answering the following Self-Assessment Question.

SAQ 3

Consider Example 2 of flipping a coin. Show that $\sum_{i=1}^n X_i$ is a complete statistic using the exponential family approach.

Let us discuss the properties of the complete statistic in the next section.

1.6 PROPERTIES OF COMPLETE STATISTIC

After understanding the concept of complete statistic, we now describe some important properties of it as follows:

1. A complete estimator/statistic may be unbiased.
2. A complete estimator ensures that the estimator is using all information available in the data, leading to efficient and reliable estimates. A complete and sufficient estimator is the most efficient estimator if an efficient estimator exists.
3. If T is a complete statistic, then any unbiased estimator that is a function of T is unique. This ensures that no two different unbiased estimators depend on T .
4. If T is a complete and sufficient statistic, then any unbiased estimator of a parameter θ , that is, a function of T is the uniformly minimum variance unbiased estimator (UMVUE).
5. If T is a complete statistic for a parameter θ and $\varphi(T)$ is a one-to-one function of T then $\varphi(T)$ is also complete for θ . This property holds because a one-to-one function preserves all the information in T and do not lose any information about the parameter. For example, if $T = \sum X_i$ is a complete statistic for the parameter θ then $\bar{X} = \frac{1}{n} \sum X_i = \frac{T}{n}$ is also complete for θ because $\bar{X} = \frac{T}{n}$ is a one-to-one function of T .

Let us study the use of the properties of a complete statistic with the help of an example.

Example 6: Consider Example 1 of the life of the lithium batteries used in cars, which follows a normal distribution with mean μ months and standard

deviation σ months, then show that the sample mean \bar{X} is a complete statistic for μ .

Solution: To show that the sample mean is a complete statistic, first, we show that $\sum X_i$ is complete and then we use property of the complete estimator.

For that, we express the probability density function of the normal distribution with mean μ and variance σ^2 as follows:

$$\begin{aligned} f(x; \mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^2 + \mu^2 - 2\mu x)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2} e^{\frac{1}{\sigma^2}\mu x} = a(\theta)b(x)e^{c(\theta)d(x)} \end{aligned}$$

$$\text{where } a(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}, b(x) = e^{-\frac{1}{2\sigma^2}x^2}, c(\theta) = \frac{1}{\sigma^2}\mu \text{ and } d(x) = x$$

Since the probability density function of the normal distribution is expressed in the general form of the exponential family so it belongs to the exponential

family for $a(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\mu^2}{2\sigma^2}}, b(x) = e^{-\frac{1}{2\sigma^2}x^2}, c(\theta) = \frac{\mu}{\sigma^2}$ and $d(x) = x$.

Since the probability density function of the normal distribution of the life of the lithium batteries belongs to the exponential family, therefore, according to

Theorem 1, $\sum_{i=1}^n X_i$ is sufficient as well as complete statistic.

Since $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is one to one function of $\sum_{i=1}^n X_i$, therefore, by the property of completeness, \bar{X} is also complete for parameter μ .

Now, try the following Self-Assessment Questions.

SAQ 4

Write important properties of complete statistic.

We now end this unit by giving a summary of what we have covered in it.

1.7 SUMMARY

In this unit, we have covered the following points:

- An estimator is said to be the best estimator if the average of its sampling distribution is equal to the parameter and does not too spread out around the true value of the parameter.
- An estimator T of the parameter θ is said to be a minimum variance unbiased estimator of θ if and only if the estimator T is an unbiased estimator of the parameter θ and $\text{Var}(T) \leq \text{Var}(T')$ where T' is any other unbiased estimator of the parameter θ .
- A probability density/mass function belongs to the one-parameter exponential family if it can be expressed in the following general form:

$$f(x; \theta) = a(\theta)b(x)\exp[c(\theta)d(x)]$$

- An estimator is said to be complete if it incorporates all of the information from the data about the parameter being estimated.
- If a probability density/mass function $f(x;\theta)$ belongs to the exponential family, then it is a complete family.

1.8 TERMINAL QUESTIONS

1. Suppose a factory produces electronic items, and each item is inspected for defects. The probability of finding a defective item is $p = 0.2$. The number of inspections required to find the first defective item follows a geometric distribution whose pdf is given as follows:

$$P[X = x] = (1-p)p^{x-1}; \quad x = 1, 2, \dots$$

Suppose the quality control inspector inspected n independent items. Let

X_1, X_2, \dots, X_n represent the output of each inspection. Show that $T = \sum_{i=1}^n X_i$ is

a sufficient and complete statistic for p using both approaches and interpret the results.

1.9 SOLUTIONS / ANSWERS

Self-Assessment Questions (SAQs)

1. We know the probability mass function of the Poisson distribution is given as follows:

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots$$

We can express the pmf of the given Poisson distribution as

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \left(\frac{\lambda}{x!} \right) \exp[(\log \lambda)x] = a(\theta)b(x)e^{c(\theta)d(x)}$$

where $a(\theta) = e^{-\lambda}$, $b(x) = \frac{1}{x!}$, $c(\theta) = \log \lambda$ and $d(x) = x$

Since the probability mass function of the Poisson distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = e^{-\lambda}$, $b(x) = \frac{1}{x!}$, $c(\theta) = \log \lambda$ and $d(x) = x$.

2. To check whether the $T = \sum_{i=1}^n X_i$ is a complete statistic or not we consider

$$E[h(T)] = 0$$

To find the above expectation, we require the sampling distribution of the estimator $T = \sum_{i=1}^n X_i$. As we know, the sum of n independent exponentially distributed random variables with parameter θ follows the gamma distribution with parameters n and θ . Therefore, the probability density function of T is given by as follows:

$$g(t) = \frac{\theta^n}{\Gamma(n)} e^{-\theta t} t^{n-1}; \quad 0 < t < \infty$$

The pdf of gamma distribution with parameters a and b is given by

$$f(x) = \frac{b^a}{\Gamma(a)} e^{-bx} x^{a-1}; \quad x > 0; a, b > 0$$

Therefore, we can compute the expectation as follows:

$$E[h(T)] = \int_0^{\infty} h(t)g(t) dt = \int_0^{\infty} h(t) \frac{\theta^n}{\Gamma(n)} e^{-\theta t} t^{n-1} dt$$

$$E[h(T)] = \frac{\theta^n}{\Gamma(n)} \int_0^{\infty} h(t) t^{n-1} e^{-\theta t} dt = 0$$

The Laplace transform of a function $g(t)$ is defined by the following integral

$$L = \int_0^{\infty} g(t) e^{-st} dt$$

The term $e^{-\theta t}$ behaves like the Laplace transform of $h(t)t^{n-1}$, therefore,

$E[h(T)]$ is zero if $h(t)t^{n-1} = 0 \Rightarrow h(t) = 0$ (almost surely). Hence, $\sum_{i=1}^n X_i$ is a complete statistic.

3. The probability mass function of the Bernoulli distribution with parameter p is given as follows:

$$P[X = x] = p(x) = p^x (1-p)^{1-x}; \quad x = 0, 1$$

To show that $\sum_{i=1}^n X_i$ is a complete statistic using the exponential family

approach, first, we try to express the probability mass function of the Bernoulli distribution in the form of the exponential family:

Take natural logarithm on both sides of the above pmf, we get

$$\begin{aligned} P[X = x] &= e^{\log[p^x (1-p)^{1-x}]} = e^{x \log(p) + (1-x) \log(1-p)} \quad \left[\because e^{\log(x)} = x \right] \\ &= e^{x \log(p) + \log(1-p) - x \log(1-p)} = e^{x \log\left(\frac{p}{1-p}\right) + \log(1-p)} = e^{\log(1-p)} e^{\log\left(\frac{p}{1-p}\right)x} \end{aligned}$$

$$P[X = x] = (1-p) e^{\log\left(\frac{p}{1-p}\right)x} = a(\theta) b(x) e^{c(\theta)d(x)}$$

where $a(\theta) = (1-p)$, $b(x) = 1$, $c(\theta) = \log\left(\frac{p}{1-p}\right)$ and $d(x) = x$

Since the probability mass function of the Bernoulli distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = (1-p)$, $b(x) = 1$, $c(\theta) = \log\left(\frac{p}{1-p}\right)$ and $d(x) =$

x , therefore, according to Theorem 1, $\sum_{i=1}^n d_i = \sum_{i=1}^n X_i$ is a complete statistic

as well as a minimal sufficient statistic.

Terminal Questions (TQs)

1. Check Sufficiency (Factorization Theorem)

To find sufficient statistics for p , we can use the factorization theorem, which is discussed in Unit 9 of the course MST-016: Statistical Inference. To apply the factorization theorem, we have to find the joint probability density function of the sample values. The probability mass function of the geometric distribution with parameter p is given as follows:

$$P[X = x] = p(1-p)^x; \quad x = 0, 1, 2, \dots$$

We can obtain the joint probability mass function of X_1, X_2, \dots, X_n as

$$f(x_1, x_2, \dots, x_n; p) = P[X_1 = x_1] \cdot P[X_2 = x_2] \dots P[X_n = x_n]$$

$$= p(1-p)^{x_1} \cdot p(1-p)^{x_2} \dots p(1-p)^{x_n} = p^n (1-p)^{\sum_{i=1}^n x_i}$$

We now try to factor the above joint probability mass function as the product of two functions, one of which is a function of the parameter (p), and another is independent of the parameter (p). We can factor the joint probability mass function as

$$f(x_1, x_2, \dots, x_n; p) = p^n (1-p)^{\sum_{i=1}^n x_i} \cdot 1 = g[t(x), p] \cdot h(x_1, x_2, \dots, x_n)$$

where $g[t(x), p] = p^n (1-p)^{\sum_{i=1}^n x_i}$ is a function of the parameter p and the

observed sample values x_1, x_2, \dots, x_n only through $t(x) = \sum_{i=1}^n x_i$ and

$h(x_1, x_2, \dots, x_n) = 1$ is a function of observed sample values x_1, x_2, \dots, x_n and is independent of the parameter p .

Hence, by the factorization theorem of sufficiency, the $\sum_{i=1}^n X_i$ is a sufficient statistic for p .

Check Completeness

To check whether $T = \sum_{i=1}^n X_i$ is a complete statistic or not, we consider

$$E[h(T)] = 0$$

To find the above expectation, we require the distribution of the estimator

$T = \sum_{i=1}^n X_i$. Since it is the sum of n independent geometric variables with

the same parameter p , therefore, the statistic/ estimator $T = \sum_{i=1}^n X_i$ follows a negative binomial distribution (n, p) whose probability mass function is given by

$$P[T = t] = \binom{t+n-1}{n} p^n (1-p)^t; \quad t = 0, 1, 2, \dots$$

Therefore, we can compute the expectation as follows:

$$E[h(T)] = \sum_{t=0}^{\infty} h(t) \binom{t+n-1}{n} p^n (1-p)^t = 0$$

We can write the above expression as

$$p^n \sum_{t=1}^{\infty} h(t) \binom{t+n-1}{n} (1-p)^t = 0$$

For all p , $0 < p < 1$, the factor p^n is not 0 for any p in the range 0 to 1.

Thus, it must be that

$$\sum_{t=1}^{\infty} h(t) \binom{t+n-1}{n} (1-p)^t = 0$$

It is a polynomial in $(1 - p)$, and it will be 0 for all p if each coefficient is 0. Since none of the $\binom{t+n-1}{n}$ terms is 0, this implies that $h(t) = 0$ for $t = 0, 1, 2, \dots$. Since T takes values $0, 1, \dots$ with probability 1, this yields that $P[h(T) = 0] = 1$ for all p . Hence, $T = \sum_{i=1}^n X_i$ is a complete statistic.

Thus, using the first approach $\sum_{i=1}^n X_i$ is a sufficient as well as complete statistic. We now consider the second approach.

Exponential Family Approach

The pmf of the geometric distribution with parameter p is given as follows:

$$P[X = x] = p(1-p)^x; \quad x = 0, 1, 2, \dots$$

To show that $\sum_{i=1}^n X_i$ is a sufficient and complete statistic using the exponential family approach, first, we try to express the probability mass function of the geometric distribution in the form of the exponential family:

Taking the natural logarithm on both sides of the above pmf, we get

$$\log\{P[X = x]\} = \log(p(1-p)^x) = \log(p) + x \log(1-p)$$

Taking the anti-logarithm on both sides, we get

$$\begin{aligned} P[X = x] &= e^{[\log(p) + x \log(1-p)]} = e^{\log(p)} e^{x \log(1-p)} = p \times 1 \times e^{x \log(1-p)} \\ &= a(\theta) b(x) e^{c(\theta) d(x)} \end{aligned}$$

where $a(\theta) = p, b(x) = 1, c(\theta) = \log(1-p)$ and $d(x) = x$

Since the probability mass function of the geometric distribution is expressed in the general form of the exponential family so it belongs to the exponential family for $a(\theta) = p, b(x) = 1, c(\theta) = \log(1-p)$ and $d(x) = x$,

therefore, according to Theorem 1, $\sum_{i=1}^n d_i = \sum_{i=1}^n X_i$ is a sufficient and complete statistic.