1. Introduction

Welcome to the series of E-learning modules on Estimator. In this module we are going to cover the basic concepts of Estimator, Role of estimators, properties, role of Normal distribution, bias and variability and standard error.

By the end of this session, you will be able to:

- Explain about an estimator
- Explain the role and types of estimators
- Describe the conditions for the best estimators
- Explain the methods of obtaining estimators
- Explain the role of central limit theorem and Normal distribution on the estimators
- Explain Bias, Variability and Standard Error

One of the objectives of sampling theory is to obtain the best possible estimates of the population parameters with limited resources.

An estimator is a sample statistic. A statistic is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population.

For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn.

An estimator is a rule, usually expressed as a formula that tells us how to calculate an estimate based on information in the sample.

An estimator is a sample characteristic and is a function of the sample observations. Any sample statistic that is used to estimate a population parameter is called an estimator.

The sample mean x bar can be an estimator of the population mean mue, the sample proportion p can be an estimator of the population proportion P. We can also use a sample range as an estimator of the population Range.

Suppose y_1, y_2,……..,y_n are the n observations drawn from a population.
Then a function of sample observations t is equal to t of (yone, ytwo, etc ,y_n) is known as an estimator of the parameter of the population.

In Statistics, an estimator is a rule for calculating an estimate of a given quantity based on observed data. Thus, the rule and its result (the estimate) are distinguished. The two key concepts of Statistical Inference are Parameters and Estimators.

Types of Estimators

There are two types of estimators.

Point Estimators and Interval Estimators.

The Point Estimator yield single-valued results, which include the possibility of single vector-valued results and results that can be expressed as a single function.

In Interval Estimators the result would be a range of possible values (or vectors or functions).

Table illustrates several populations, population parameters and estimators

Table 1

Population in which we are interested	Population parameters we wish to estimate	Sample statistic we will use as an estimator
Employees in a furniture factory	Mean turn over per year	Mean turn over for a period of one month
Applicants for town manager of chapel hill	Mean formal education (years)	Mean formal education of every fifth applicant
Teenagers in a given community	Proportion who have criminal records	Proportion of a sample of fifty teenagers who have criminal records

2. Criteria of a Good Estimator

Criteria of a good estimator
Statistical theory concerned with the properties of estimators; that is, with defining properties that can be used to compare different estimators for the same quantity, based on the same data. Such properties can be used to determine the best rules to use under given circumstances.

However, in **Robust Statistics, s**tatistical theory goes on to consider the balance between having good properties, if rightly defined assumptions hold, and having less good properties that hold under wider conditions.

Some estimators are better estimators than others. Fortunately we can evaluate the quality of a statistic as an estimator by using four criteria:

1) Unbiasedness: This is a desirable property of a good estimator to have. The term unbiasedness refers to the fact that a sample mean is an unbiased estimator of a population mean because the mean of the sampling distribution of sample mean taken from the sample population is equal to the population mean itself. We can say that a statistic is an unbiased estimator if on average it tends to assume values above the population parameter being estimated as frequently and to the same extent as it tends to assume values that are below the population parameter being estimated.

Examples:

1. The sample mean, y bar is an unbiased estimator of the population mean, mue.

2. The sample variance s square is equal to su**mmation** (y_i minus, y bar whole square divided by n minus one is an unbiased estimator of the population variance, sigma square.
3. The sample proportion, **p** is an unbiased estimator of the population proportion, P.

Unbiased estimators determine the tendency, on the average, for the statistics to assume values close to the parameter of interest.

2) Efficiency: Another desirable property of a good estimator is that it should be efficient. Efficiency refers to the size of the Standard error of the statistic. If we compare two statistics from a sample of same size and try to decide which one is more efficient estimator, we would pick a statistic that has a smaller standard error or standard deviation of sampling distribution. It makes sense that a statistic of smaller standard error (with less variation) will have more chance of producing an estimate nearer to the population parameter under consideration.

Example:
The median **is** an unbiased estimator of **mue when** the sample distribution is normally distributed; but its standard error is one point two five greater than that of the sample mean, so the sample mean is a more efficient estimator than the median.

3) Consistency: A statistic is a consistent estimator of a population parameter. As the sample size increases it becomes almost certain that the value of statistic comes very close to the value of the population parameter. If an estimator is consistent it becomes more reliable with large samples. Thus, if you are wondering whether to increase the sample size to get more information about a population parameter, first, find out whether your statistic is a consistent estimator. If it is not you will waste your time and money by taking large sample.

The sample mean and sample proportions are consistent estimators, since from their formulas as n get big, the standard errors gets small.

4) Sufficiency: An estimator is sufficient if it makes so much use of information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

The care that the statisticians must take in picking an estimator
A given sample statistic is not always the best estimator of its analogous population parameter. Consider a symmetrically distributed population in which the values of the median and then mean coincide.

In this instance, a sample mean would be an unbiased estimator of a population median also a consistent estimator of a population median. Because as the sample size increase the value of the sample mean would tend to come very close to the population median. And sample mean would be more efficient estimator of the population median than the sample median itself. It is because in large samples the sample mean has a smaller standard error than the sample median. At the same time the sample median in a symmetrically distributed population would be an unbiased and consistent estimator of the population mean but not the most efficient estimator because in large samples its standard error is larger than that of the sample mean.

3. Central Limit Theorem

Central Limit Theorem
When sampling is not from a Normal population the sample size plays a critical role. When n is small the shape of the distribution will depend largely on the shape of the parent population but when n gets large (n>30) the shape of the sampling distribution become more and more like a Normal distribution irrespective of the shape of the parent population.

The theorem which explains this sort of relationship between the shape of the population distribution and the sampling distribution of the mean is known as the Central Limit Theorem. It ensures that the sampling distribution of the mean approaches Normal distribution as the sample size increases.

The significance of the central limit theorem lies in the fact that it permits us to use sample estimators to make inference about the population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

The samples in surveys are often large enough so that estimates made from them are approximately Normally distributed. Further more in Probability Sampling we have formulae that give the mean and variance of the estimators.
Suppose that we have taken a sample by a procedure known to give an unbiased estimator and have computed the sample estimate mue cap and its standard deviation S.D (mue cap) (Often called alternatively its standard error).

How good is the estimator?
We cannot know the exact value of the error of the estimate (mue cap minus mue) but from the properties of the Normal curve, the chances are
Zero point three two that the absolute error (mue cap minus mue) exceeds S.D (mue cap).
Zero point zero five that the absolute error (mue cap minus mue) exceeds one point nine six S.D (mue cap) = twice S.D (mue cap).
Zero point zero one that the absolute error (mue cap minus mue) exceeds two point five eight S.D (mue cap).

4. Bias in Estimators

Bias and its effects

In sample survey theory it is necessary to consider biased estimators for two reasons:

1. In some of the most common problems, particularly in the estimation of ratios, estimators that are otherwise convenient and suitable are found to be biased.
2. Even with estimators that are unbiased in probability sampling, errors of measurement and nonresponsive may produce biases in the numbers that we are able to compute from the data.

Bias is a term which refers to how far the average statistic lies from the parameter it is estimating, that is, the error which arises when estimating a quantity. Errors from chance will cancel each other out in the long run, whereas those from bias will not.

These are the Systematic errors produced by your sampling procedure.
For example, if you sample people and ask them whether they watch cricket match or not.
The percentage always comes out too high (maybe because you are interviewing your friends and your whole group who really likes cricket).

The amount of bias is B equals to Estimated value minus true value of the parameter.
As a working rule the effect of bias on the accuracy of the value of an estimator is negligible if the Bias is less than one tenth of the Standard deviation of the estimate.
That is B is less than zero point one by sigma where B is the absolute value of the bias.
Anyhow if the sample is large enough we can be confident that B by sigma will not exceed zero point one.

The Mean Square Error

In order to compare a biased estimator with an unbiased estimator or two estimators with different amounts of bias, a useful criterion is the Mean Square Error (M.S.E) of the estimator, measured from the population value that is being estimated which gives an accuracy of an estimator.

M.S.E of mue cap = E (mue cap minus mue) whole square

Observe the figure which illustrates bias and precision, where the target value is the bulls' eye.

Because of the difficulty of ensuring that no unsuspected bias enters into estimates we will usually speak of the precision of an estimator instead of the accuracy. Accuracy refers to the size of the deviation from the true mean mue whereas precision refers to the size of the deviations from the (estimated) mean obtained by repeated applications of the sampling procedure.

A character that describes the population is called a parameter. It is often difficult or impossible to measure the entire population, parameters are most often estimated. Suppose we don't have an idea what a population value is and we want to estimate it using an appropriate sample estimator. The first, easiest way to do this is to obtain a sample estimator whose value itself is a point estimate of the parameter value.

For example:
1. I could randomly select one hundred of campus students and ask how long it takes them to get to the campus. I could then average those times to get an estimate of the average commuting time for all of campus students.
2. You want to know the mean income of the subscribers to a particular magazine. You draw a random sample of one hundred subscribers and determine that their mean income is twenty-seven thousand dollars (a statistic). You conclude that the populations mean income mue is likely to be close to twenty-seven thousand dollars as well.

The estimators used to estimate the population parameters from a sample are:
Estimator of Population Mean Capital Y bar equals to y bar sample mean, equals to summation yi by n, i runs from 1 to n
Estimator of Population total Y = N y bar = capital N by n into summation yi, i runs from 1 to n.
runs from 1 to
Estimator Population Ratio R = y bar by x bar = summation yi by summation xi
n.

5. Sampling Distribution and Standard Error

Sampling Distribution:It is possible to draw more than one sample from the same population and the value of an estimator will in general vary from sample to sample. For example, the average value in a sample is a statistic. The average values in more than one sample, drawn from the same population, will not necessarily be equal.
The sampling distribution describes probabilities associated with an estimator when a random sample is drawn from a population.

The sampling distribution is the probability distribution or probability density function of an estimator.
Sample Means and Deviations: In general, we use estimators as a means of characterizing the nature of some sample on the basis of a few key indicators.

The first indicator is known as The Sample Mean:

- The mean quantity in some sample represents the average value or the most probable value in the sample.
- A sample mean is calculated by summing up the individual measurements and dividing by the number of measurements, usually denoted as n

All samples can be characterized by a mean value regardless of the shape of the distribution.

The second indicator we use is known as the Sample Variance, also known as the Sample Dispersion. This is usually denoted by the letter s.

In general, the dispersion is a more important quantity than the sample mean. The dispersion represents the range of the data about the mean value. Understanding the role of dispersion is the most critical aspect of understanding and interpreting statistical sampling data.The calculation of dispersion in a distribution is very important because it represents a uniform way to determine probabilities and to determine if some event in the data is expected (i.e. probable) or is significantly different than expected (i.e. improbable).Now that we have an understanding of means and dispersions we have a simple way for determining if two distributions are fundamentally different.

The **variability** of a statistic is determined by the spread of its sampling distribution.

- In general, larger samples will have smaller variability. This is because as the sample size increases, the chance of observing extreme values decreases and the observed values for the statistic will group more closely around the mean of the sampling distribution.

- Furthermore, if the population size is significantly larger than the sample size, then the size of the population will not affect the variability of the sampling distribution (i.e., a sample of size one hundred from a population of size one lakh will have the same variability as a sample of size one hundred from a population of size ten lakh).

Standard Error

- The standard error of an estimator is the standard deviation of a statistic
- The variability of a statistic is measured by the Standard Error of Sample Estimators
- The values of population parameters are often unknown, making it impossible to compute the standard deviation of a statistic. When this occurs, we use the standard error
- The standard error is computed from known sample statistics, and it provides an unbiased estimate of the standard deviation of the estimator

The table below shows how to compute the standard error.

Table 2

Statistic	Standard Error
Sample mean, x bar	SE_x equals s by sqrt(n)
Sample proportion, p	SE_p equals sqrt [p(1 - p) by n]
Difference between means, x_1 bar minus x_2 bar	SE_{x1-x2} equals sqrt [s one square by n one plus s two square by ntwo]
Difference between proportions, p_1 minus p_2	SE_{p1-p2} equals sqrt [p one(one minus p one) by n one plus p two(one minus ptwo) by ntwo]

Two characteristics are valuable in an estimator:
Observe the figure

1. The sampling distribution of the point estimator should be centered over the true value of the parameter to be estimated.
2. The spread (as measured by the variance) of the sampling distribution should be as small as possible.

Here’s a summary of our learning in this session:

- Statistical definition of an estimator
- Types of estimators
- Properties of estimators
- Role of Central Limit Theorem and Normal Distribution
- Bias in estimators
- Sampling Distribution and Standard error