**Eco-Hc-3036: Statistical Methods for Economics**

**Unit:-5 Sampling**

**Class:-3rd**

Semester:-III (H) Date:-03/09/2022

 Time: - 09:15am

**What is sampling theory?**

Sampling theory is a study of relationships existing between a population and samples drawn from the population. Sampling theory is applicable only to random samples. For this purpose the population or a universe may be defined as an aggregate of items possessing a common trait or traits. In other words, a universe is the complete group of items about which knowledge is sought. The universe may be finite or infinite. Infinite universe is one which has a definite and certain number of items, but when the number of items is uncertain and infinite, the universe is said to be an infinite universe. Similarly, the universe may be hypothetical or existent. In the former case the universe in fact does not exist and we can only imagine the items constituting it. Tossing of a coin or throwing a dice is examples of hypothetical universe. Existent universe is a universe of concrete objects i.e., the universe where the items constituting it really exist. On the other hand, the term sample refers to that part of the universe which is selected for the purpose of investigation. The theory of sampling studies the relationships that exist between the universe and the sample or samples drawn from it.

The main problem of sampling theory is the problem of relationship between a parameter and a statistic. The theory of sampling is concerned with estimating the properties of the population from those of the sample and also with gauging the precision of the estimate. This sort of movement from particular (sample) towards general (universe) is what is known as statistical induction or statistical inference. In clearer terms “from the sample we attempt to draw inference concerning the universe? In order to be able to follow this inductive method, we first follow a deductive argument which is that we imagine a population or universe (finite or infinite) and investigate the behavior of the samples drawn from this universe applying the laws of probability.” The methodology dealing with all this is known as sampling theory.

Sampling theory is designed to attain one or more of the following objectives:

* **Statistical estimation:** Sampling theory helps in estimating unknown population parameters from a knowledge of statistical measures based on sample studies. In other words, to obtain an estimate of parameter from statistic is the main objective of the sampling theory. The estimate can either be a point estimate or it may be an interval estimate. Point estimate is a single estimate expressed in the form of a single figure, but interval estimate has two limits viz., the upper limit and the lower limit within which the parameter value may lie. Interval estimates are often used in statistical induction.
* **Testing of hypotheses:** The second objective of sampling theory is to enable us to decide whether to accept or reject hypothesis; the sampling theory helps in determining whether observed differences are actually due to chance or whether they are really significant.
* **Statistical inference:** Sampling theory helps in making generalization about the population/ universe from the studies based on samples drawn from it. It also helps in determining the accuracy of such generalizations.

The theory of sampling can be studied under two heads viz., the sampling of attributes and the sampling of variables and that too in the context of large and small samples (By small sample is commonly understood any sample that includes 30 or fewer items, whereas alarge sample is one in which the number of items is more than 30). When we study some qualitative characteristic of the items in a population, we obtain statistics of attributes in the form of two classes; one class consisting of items wherein the attribute is present and the other class consisting of items wherein the attribute is absent. The presence of an attribute may be termed as a ‘success’ and its absence a ‘failure’. Thus, if out of 600 people selected randomly for the sample, 120 are found to possess a certain attribute and 480 are such people where the attribute is absent. In such a situation we would say that sample consists of 600 items (i.e., n = 600) out of which 120 are successes and 480 failures. The probability of success would be taken as 120/600 = 0.2 (i.e., p = 0.2) and the probability of failure or q = 480/600 = 0.8. With such data the sampling distribution generally takes the form of binomial probability distribution whose mean Formula would be equal to n × p and standard deviation s p d i would be equal to Formula. If n is large, the binomial distribution tends to become normal distribution which may be used for sampling analysis. We generally consider the following three types of problems in case of sampling of attributes:

1. The parameter value may be given and it is only to be tested if an observed ‘statistic’ is its estimate.
2. The parameter value is not known and we have to estimate it from the sample.
3. Examination of the reliability of the estimate i.e., the problem of finding out how far the estimate is expected to deviate from the true value for the population.

All the above stated problems are studied using the appropriate standard errors and the tests of significance which have been explained and illustrated in the pages that follow.
The theory of sampling can be applied in the context of statistics of variables (i.e., data relating to some characteristic concerning population which can be measured or enumerated with the help of some well defined statistical unit) in which case the objective happens to be :

1. to compare the observed and expected values and to find if the difference can be ascribed to the fluctuations of sampling;
2. to estimate population parameters from the sample, and
3. To find out the degree of reliability of the estimate.

The tests of significance used for dealing with problems relating to large samples are different from those used for small samples. This is so because the assumptions we make in case of large samples do not hold good for small samples. In case of large samples, we assume that the sampling distribution tends to be normal and the sample values are approximately close to the population values. As such we use the characteristics of normal distribution and apply what is known as z-test. When n is large, the probability of a sample value of the statistic deviating from the parameter by more than 3 times its standard error is very small (it is 0.0027 as per the table giving area under normal curve) and as such the z-test is applied to find out the degree of reliability of a statistic in case of large samples. Appropriate standard errors have to be worked out which will enable us to give the limits within which the parameter values would lie or would enable us to judge whether the difference happens to be significant or not at certain confidence levels. For instance, Formula would give us the range within which the parameter mean value is expected to vary with 99.73% confidence. Important standard errors generally used in case of large samples have been stated and applied in the context of real life problems in the pages that follow.

The sampling theory for large samples is not applicable in small samples because when samples are small, we cannot assume that the sampling distribution is approximately normal. As such we require a new technique for handling small samples, particularly when population parameters are unknown. Sir William S. Gosset (pen name Student) developed a significance test, known as Student’s t-test, based on t distribution and through it made significant contribution in the theory of sampling applicable in case of small samples. Student’s t-test is used when two conditions are fulfilled viz., the sample size is 30 or less and the population variance is not known. While using t-test we assume that the population from which sample has been taken is normal or approximately normal, sample is a random sample, observations are independent, there is no measurement error and that in the case of two samples when equality of the two populations means is to be tested, we assume that the population variances are equal. For applying t-test, we work out the value of test statistic (i.e., ‘t’) and then compare with the table value of t (based on ‘t’ distribution) at certain level of significance for given degrees of freedom. If the calculated value of ‘t’ is either equal to or exceeds the table value, we infer that the difference is significant, but if calculated value of t is less than the concerning table value of t, the difference is not treated as significant. The following formulae are commonly used to calculate the t value:

Population Parameters versus Sample Statistics

A descriptive measure for an entire population is a ''**parameter**.'' There are many population parameters. For example, the population size (N) is one parameter, and the mean diastolic blood pressure or the mean body weight of a population would be other parameters that relate to continuous variables. Other population parameters focus on discrete variables, such as the percentage of current smokers in the population or the percentage of people with type 2 diabetes mellitus. Health-related behaviors can also be thought of this way, such as the percentage of the population that gets vaccinated against the flu each year or the percentage who routinely wear a seatbelt when driving.

It is generally not feasible to directly measure parameters, since it requires collecting information from all members of the population. We therefore take samples from the population, and the descriptive measures for a sample are referred to as **''sample statistics''** or simply **''statistics.''** For example, the mean diastolic blood pressure, the mean body weight, and the percentage of smokers in a sample from the population would be sample statistics.