3

Theory of Sampling

1. INTRODUCTION

There are two methods of collection of data : (1) Census Method (2) Sampling Method. In census method, information relating all the units of population or universe is collected whereas in sampling method, information about some selected units is collected. In this chapter, we consider some basic concepts of these methods.

Universe or Population : A population or universe is the total set of elements of interest for a given problem. For example, if we want to get information about the attitudes of voters toward a metropolitan transit agency. The population here consist of all eligible voters in the city. The elements of the population are the individual voters.

A population can be of two types :

(*a*) Finite Population

(*b*) Infinite Population

A finite population consists of a finite number of elements. Many of the populations of interest in Economics, Business and Social Sciences are finite. The set of all persons living in India is an example of a large finite population. Other examples of finite population are number of students in a college, number of doctors in a city etc.

On the other hand, an infinite population consists of an indefinitely large number of elements. For example, number of stars in the sky. Information about a finite population can be obtained either by a census or by a sample. A census of a finite population is a study of every elements of the population. For example, a company with 50 employees is interested in employees' preferences for a new pension plan. Here the population is finite as well as small. It is easy to reach every employee, hence, a census is appropriate.

On the other hand, a sample is a part of the population under study. In other words, if we select some of elements of population with the intention of

finding out something about the population from which they are taken, we refer to that sub-group of elements as a sample. For example, an economist considers data on saving plans by households based on a sample of 10,000 households from the population of India.

How far would the information we get from the samples is likely to be same as the information that we would get, if the given 'population' or 'universe' as a whole was studied, would depend on the way sample is selected.

2. NEED FOR SAMPLING

There are a number of reasons why sampling is used so often for finite populations :

(1) Saving Time : In sampling, much time is saved than census as fewer data have to be collected.

(2) Less Cost : A sample can provide useful information at much lower cost than a census. It saves labour also.

(3) Accuracy : A sample provides as accurate as, or even more accurate information than a census because errors can be controlled more effectively in a small undertaking than in a larger one.

(4) Reliability : A sample represents totally the universe. The results derived from a good sample will be more reliable.

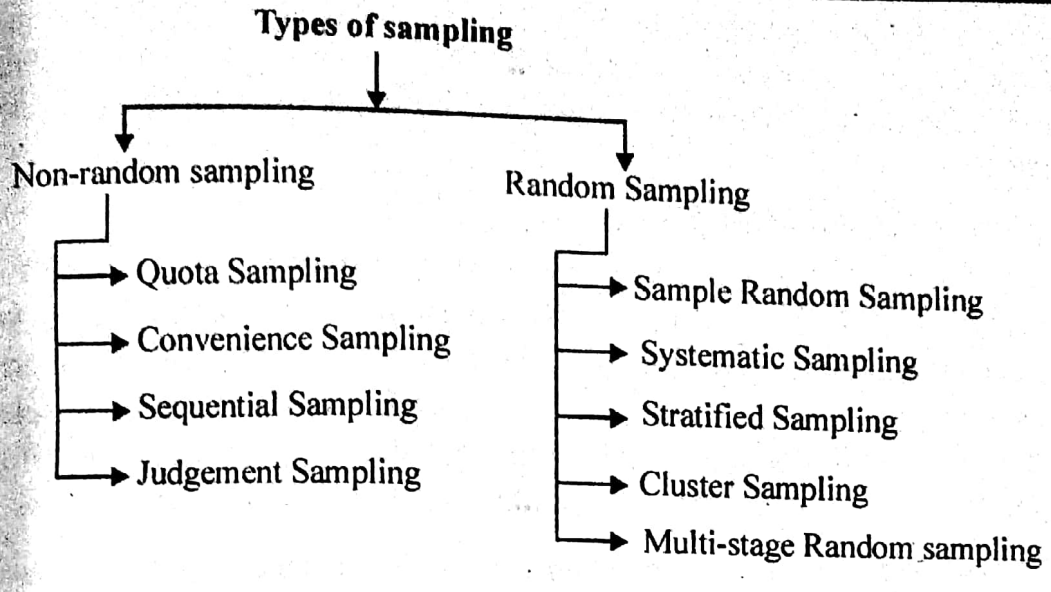
(5) Only Method : When size of population is very large, only sampling method is appropriate.

Sampling, however, is not without its limitations. Some of the limitations are discussed below :

(1) Sampling demands exercises of great care and caution. In the absence of it, the results obtained may be misleading.

(2) If population is heterogeneous, a very large sample is required to get reliable information about it. A small sample will not represent the population.

(3) It needs a specified knowledge about sampling to select a sample from the universe. Lacking this knowledge will lead to incorrect results.



(I) NON-RANDOM SAMPLING OR NON-PROBABILITY SAMPLING

In non-random sampling methods, every element has no specific chance for being included in the sample. Here, the investigator or the researcher himself purposively selects certain elements which he thinks are the best representatives of the population. For better selection of the elements, the researcher may adopt a certain criteria of selection.

There are mainly four types of non-random sampling :

(i) Quota Sampling

In quota sampling, the universe is first divided into certain parts or groups. Then the sample is allocated among these groups.

For instance, in a household survey interviewer are given certain 'quotas' with respect to certain characteristics such as sex, age and income. The actual seleciton of persons is left to the interviewer. The interviewers often choose persons readily available. Also, he can simply substitute another person from the same quota for a person who is not immediately available for interview. This method is very commonly used in market surveys.

If the interviewers are skilled and experienced, this method gives very reliable results. But the method is not free from the personal bias.

(ii) Convenience Sampling

Here the researcher or the interviewer selects the sample units purposively according to his convenience in the matter of location or contacts. This method is suitable for making pilot surveys in which questionnare is pre-tested. But this method is also not free from the personal bias of the interviewer.

(iii) Sequential Sampling

In this type of sampling, a number of samples are selected one after the

other in order of a sequence till a satisfactory sample is obtained. This method is suitable when the universe is infinite. For example, for testing, wheat or rice etc. Sequential sampling method is not suitable for a finite population or universe.

(Iv) Judgement Sampling

In this method of sampling, choice of the sampling units depends upon the judgement of the investigator. The investigator selects those units which he thinks are most typical of the population with regard to the characteristics under investigation. For example, if sample of ten students is to be selected from a class of fifty for analysing the reading habits of students, the investigator would select ten students who, he thinks, are representative of the class.

Judgement sampling method is used in solving many economic and business problems when only a small number of sampling units are in the universe. But this method is not scientific as the choice of the sampling units may be affected by the personal bias of the investigator. Thus the success of this method depends upon the excellence in judgement.

(II) RANDOM SAMPLING OR PROBABILITY SAMPLING

In probability sampling, the selection of elements from the population is made according to known probabilities. This method allows no discretion as to which particular elements in the population enter the sample. For example a sample of 100 students was selected from the 1000 students of a college. This sample is a probability sample. Probability sampling has two main advantages :

(1) The sample data can be evaluated by statistical methods.

(2) Personal bias is avoided.

Basic types of random sampling are given below :

(i) Simple Random Sampling

A simple random sample from a population is a sample selected such that each possible sample has an equal probability of being selected and each item of the population has also same probability of being picked.

For example, let us assume that there are five students of M.A. in the population denoted by A, B, C, D and E. If a sample of two students is to be selected from this population, there will be 10 possible sample combinations :

AB, AC, AD, AE, BC, BD, BE, CD, CE, DE.

Probability of each sample combination will be $\frac{1}{10}$ or 0·1 of being chosen as the actual sample.

Above written example as well as definition is of a simple random sampling without replacement. Here the populations element can enter the sample only once. Thus a simple random sampling is one in which each unit has an equal and non-zero chance of being selected. To ensure randomness of selection one may adopt following methods :

(*a*) Lottery Method

(*b*) Random Number Tables

(*a*) Lottery Method : This is a very popular method of sampling. In this method, all items of the population are numbered or named on separate slips of paper of identical size and shape. These slips are then folded and mixed up in a container. A blindfold selection is then made of the number of slips required to constitute the desired size of sample. This method is popularly used in lottery draws to take a decision about prizes.

(*b*) Table of Random Numbers : When the size of population becomes large, the lottery method is not suitable. In this case table of random numbers is used. The random numbers are generally obtained by some mechanism which, when repeated a large number of times, ensures approximately equal frequencies for the numbers from 0 to 9.

Tippett's random number table is most popularly used in practice. This table consists of 41,600 random digits grouped into 10,400 sets of four-digit random numbers.

Illustration of Tippett's table :

2670	7483	3408	2762	3563
1089	6913	7991	0560	5246
1112	6107	6008	8125	4233
8776	2754	9143	1405	9025

The starting point in the table is to be selected randomly so that every unit has an equal chance of being selected. Suppose we have to select five items out of 5000. First, numbered all the items from one to 5000. Then five numbers upto 5000 should be noted down. These numbers will be 2670, 3408, 2762, 3563 and 1089. Items bearing those numbers will be included in the sample.

If the size of universe is less than 1000 then the procedure will be different. Suppose we have to select a sample of 10 items out of 300. So, all items from 1 to 300 should be numbered as 0001 to 0300 as Tippett's numbers are available only in four figures. We may now select 10 numbers from the table which are up to 0300.

Since the selection of items in the sample depends entirely on chance, there is little possibility of personal bias. But it has been clained that the items selected by simple random sampling tend to be too widely dispersed geographically and that the time and cost of collecting data become too large.

(ii) Systematic Sampling

In systematic sampling the initial element is selected from the universe at random and then the other elements are selected at a uniform interval from the population arranged in a systematic order like alphabetical, geographical or numerical order. The space interval of the population is determined by using the following formula :

$$K = \frac{N}{n}$$

where K is the space interval of the population from which the sample is to be drawn.

N is size of the population.

n is size of the sample.

Example **:**

Let us assumed that there are 100 students in a class. A sample of 10 students is to be selected. So, first we will arrange all the 100 students according to their roll numbers from 1 to 100. Space interval will be calculated as $K = \frac{N}{n} = \frac{100}{10} = 10$. After this, one student will be selected at random from the first 10 students. If the selected roll number is 5, then the roll numbers of other nine students will be 15, 25, 35, 45, 55, 65, 75, 85 and 95.

This method is very simple to understand. It saves time as well as labour. But this method has also some limitations as :

(1) If size of universe (N) is not divisible by size of sample (n), that is, K is in fraction, it will become difficult to select elements from population.

(2) It may select unrepresentative items sometimes. Suppose we are interested to get a sample of paper waste produced by households and we decide to have a sample of 50 households every Monday. There will be chance of selecting unrepresentative sample because Monday's trash would very likely include Sunday's newspaper.

(iii) Stratified Sampling

For stratified sampling, first we divide the universe into certain number

(b) Disproportionate Stratified Sample : When the number of elements drawn from the various stratas is independent of the sizes of the stratas, the approach is called disproportionate stratified sampling.

While employing the procedures of stratified random sampling, the researcher must remain assured that no essential groups will be excluded from the sample. Compared with simple random sampling, stratified random sampling is more concentrated geographically, thereby reducing the cost in terms of time and money.

The main limitation of stratified random sampling is that proper stratification of a heterogeneous populations into a number of homogeneous groups is very difficult. Faulty stratification will lead to defective results.

(iv) Cluster Sampling

In cluster sampling, the whole universe is divided into certain subgroups, called clusters. These clusters may be city wards, households, or several geographical or social units. Then, certain clusters are selected at random. All the elements of the selected clusters will constitute the desired sample.

For example, suppose we want to conduct a study on the problems of students of colleges in Punjab. Thus first we will prepare a list of all the colleges in the state (say) 1000. Then we will select a random sample of 100 colleges. Thus all the students of these 100 colleges will constitute the sample.

While using this method of sampling, attention must be paid to the point that the clusters must be as small as possible in size. The limitation of this method is that it is not suitable for an area with widely varying number of the elements.

(v) Multi-Stage Randam Sampling

When the cluster sampling procedure moves through multiple stages say, two, three or more stages, it is known as multi-stage random sampling.

First Stage : The universe is divided into some clusters. A certain number of clusters are selected at random.

2nd Stage : The selected clusters are further divided into subgroups. From these subgroups, some are selected at random.

3rd Stage : The selected sub-groups are again subdivided into some groups from which a certain number of groups are selected at random.

This process of division and sub division will be carried out till the reasonable size of sample is obtained.

4. SIZE OF SAMPLE

Size of sample means the number of units selected from the universe for investigation. This is very important decision that has to be taken in sampling technique that what would be the size of sample. Different experts have given different opinions on this point. It may be noted that only size of the sample does not ensure the representativeness. If the size of sample is very small, it may not represent the population. On the other hand, if the size of sample is very large, it may be very much burdensome financially. Hence the size of sample should neither be too small nor too large. The optimum size of the sample is one that fulfills the requirements of representativeness, reliability and flexibility.

2. PARAMETER AND STATISTICS

Parameter : Population constant which we estimate is called a parameter *e.g.*, per capita income of the population is an unknown constant because unless we resort to complete census we donot have all data about the population and their values (per capita income remain unknown parameter (constant) for us.

The statistical constants of the population like mean (μ) variance (σ^2) skewness (β_1) kurtosis (β_2) moments (μ_r) correlation coefficient (ρ) etc. are known as parameter. Parameters are the function of the population values.

Estimator or Statistic : Statistic are the functions of the sample observations. Mean ($\overline{X}$), Variance (s^2) skewness (b_1) Kurtosis (b_2), Moments (m_r) correlation coefficient (r) etc. are the characteristics of estimator or statistic.

Parameter	Statistic (t)
(1) It comes from population	It comes from sample.
(2) Population Mean (μ).	Sample Mean ($\overline{X}$)
(3) Population variance (σ^2)	Sample variance s^2
(4) Skewness (β_1)	Skewness (b_1)
(5) Kurtosis (β_2)	Kurtosis (b_2)
(6) Moments (μ_r)	Moments (μ_r)
(7) Correlation coefficient (ρ)	Correlation coefficient (r)

Standard Error

Standard deviation of the sampling distribution of a statistic is known a its standard error. Standard Error of the statistic t is given by

$$S.E._t = \sqrt{\text{Variance of } (t)}$$

$$S.E._t = \sqrt{\frac{\sum_{i=1}^{n} (t_i - \bar{t})^2}{K}}$$

Uses of Standard Error

(1) Reciprocal of standard error of statistic gives precision or reliability of the estimate of the parameter.

(2) S.E. is used in testings of hypothesis.

(3) S.E. is used to obtain interval estimate of the population parameter.

5. THEORY OF ESTIMATION

Estimation is concerned with obtaining numerical estimates of the paramer from a sample. Theory of estimation was founded by Prof. R.A. Fisher in 1930. Parameter is unknown since population size is infinite. Even when population is finite the size of population is large and all of them are not known in practice. There is a difference between estimator and estimate. Estimator is a method of making the estimate whereas term estimate is the actual result obtained from a sample *e.g.* sample mean $\overline{X}$ is estimator whereas when we put the sample values and get the result that is called estimate.

For a single parameter many estimators can be there, we have to select one estimator such that their distributions are concentrated as closely as possible round the true parameter values. Theory of estimation can be divided into

two groups.

(*i*) Point Estimation (*ii*) Interval Estimation

(*i*) Point Estimation : It gives single value of the population parameter. Here a single point is obtained from the sample values to estimate values of the population parameter without going into the error of estimation. Since it is simply a point or number so much an estimate is called the point estimate.

Sample mean $\overline{X}$ is a point estimator of the population mean.

(*ii*) Interval Estimation : Single value estimate does not in general coincide with the true value of the parameter. It is preferred to obtain a range of values or interval which may be suspected to cover the true value of the parameter with some probability or the degree of confidence. Such an interval is called interval estimate or confidence interval and the probability or degree of confidence is called the confidence coefficient.

Thus the procedure of determining an interval (a, b) that will include population parameter (θ) with certain probability $(1 - \alpha)$ is known as interval estimate. Here α is the probability that interval does not include the true parameter value.

Prob. $[a < \theta < b] = 1 - \alpha$

Interval (a, b) is also called confidence interval.

Properties of a good Estimator

The goodness of estimator is tested by examining the presence of the following properties.

(*i*) Unbiasedness

(*ii*) Consistency

(*iii*) Efficiency

(*iv*) Sufficiency

(*i*) Unbiasedness : If the average of values of estimator is equal to the parameter then estimator will be called an unbiased one. In other words if we draw many samples of a given size from the same distribution and from each obtain a value $\hat{\theta}$, the arithmetic mean of all values of $\hat{\theta}$ must be close to θ.

Thus if $E(\hat{\theta}) = \theta$ then $\hat{\theta}$ is said to be unbiased.

$E(\hat{\theta}) \neq \theta$ then $\hat{\theta}$ is said to be biased.

$E(\hat{\theta}) > \theta$ then estimator will be called positively biased.

$E(\hat{\theta}) < \theta$ then estimator will be negatively biased.

Unbiased is a desirable property but not particularly important by itself. It becomes important only when combined with a small variance.

***(ii)* Consistency :** The value of the unknown parameter estimated from sample values is often different from the real value of the parameter. If the size of sample increases the probability that estimated value lies close to the value of the parameter tends to unity.

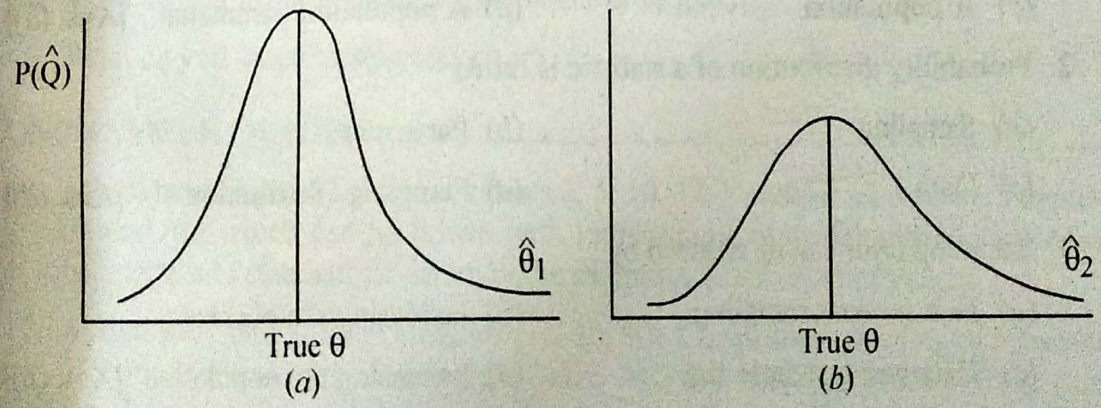
An estimator $\hat{\theta}$ for a parameter θ is said to be consistent if the value of the estimator approaches the value of the parameter as the sample size becomes infinitely large.

$$P(|\hat{\theta} - \theta| < \delta) \to 1 \text{ when } n \to \infty$$

The probability that the estimator $\hat{\theta}$ will differ by less than δ (where δ is a small number) from the true value θ tends to unity as n approaches infinity then we say that $\hat{\theta}$ is a consistent estimator of θ.

Consistency is a limiting property and it refers to the behaviour of the estimator with an increase in the sample size. If the estimator is unbiased it will be consistent but consistent estimator need not be unbiased. This property is concerned with the behaviour of an estimator for large values of the sample size n.

***(iii)* Efficiency :** The efficiency of the estimator is measured by the variance of its sampling distributions. If two estimates based on same sample size and both are unbiased then the one with smaller variance is said to have greater efficiency then other.



In figure (*a*) $\hat{\theta}$ is more efficient because it has smaller variance.

For unbiased estimator, the relative efficiency of two estimators is defined as

$$\text{Efficiency of } \hat{\theta}_1 \text{ compared to } \hat{\theta}_2 = \frac{\text{Variance } \hat{\theta}_2}{\text{Variance } \hat{\theta}_2} < 1$$

***(iv)* Sufficiency :** Estimator is said to be sufficient estimator of parameter if it contains all the information in sample about parameter. It is most important

property defined by R.A. Fisher. Any estimator is said to be sufficient if it utilizes all the information contained in the sample about the parameter. It is expressed in terms of likelihood function.

If $F(x, \theta)$ is density function for a population then likelihood function for random sample is defined as

$$L(x_1, x_2 \ldots x_n, \theta) = \prod_{i=1}^{n} F(x, \theta)$$

Now if it is possible to write

$$L = F(x, \theta)\, F(x_2 \theta) \ldots f(x_n \theta)$$

Now if it is possible to write

$$L = L_1 (t_n \theta)\, L_2 (x_1 x_2 \ldots x_n)$$

i.e. 2nd function does not contain θ then t_n will be said sufficient estimator for θ. If sufficient estimator exist we need not use any other for it. It is most desirable but unfortunately sufficient estimator are the exceptions.