14·1. INTRODUCTION

Before giving the notion of sampling we will first define *population*. In a statistical investigation the interest usually lies in the assessment of the general magnitude and the study of variation with respect to one or more characteristics relating to individuals belonging to a group. This group of individuals under study is called *population or universe*. Thus in statistics, population is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

It is obvious that for any statistical investigation complete enumeration of the population is rather impracticable. For example, if we want to have an idea of the average per capita (monthly) income of the people in India, we will have to enumerate all the earning individuals in the country, which is rather a very difficult task.

If the population is infinite, complete enumeration is not possible. Also if the units are destroyed in the course of inspection (*e.g.*, inspection of crackers, explosive materials, etc.), 100% inspection, though possible, is not at all desirable. But even if the population is finite or the inspection is not destructive, 100% inspection is not taken recourse to because of multiplicity of causes, *viz.*, administrative and financial implications, time factor, etc., and we take the help of *sampling*.

A finite subset of statistical individuals in a population is called a sample and the number of individuals in a sample is called the sample size.

For the purpose of determining population characteristics, instead of enumerating the entire population, the individuals in the sample only are observed. Then the sample characteristics are utilised to approximately determine or estimate the population. For example, on examining the sample of a particular stuff we arrive at a decision of purchasing or rejecting that stuff. The error involved in such approximation is known as *sampling error* and is inherent and unavoidable in any and every sampling scheme. But sampling results in considerable gains, especially in time and cost, not only in respect of making observations of characteristics but also in the subsequent handling of the data.

Sampling is quite often used in our day-to day practical life. For example, in a shop we assess the quality of sugar, wheat or any other commodity by taking a handful of it from the bag and then decide to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt.

14·2. TYPES OF SAMPLING

Some of the commonly known and frequently used types of sampling are :

(*i*) *Purposive sampling*, (*ii*) *Random sampling*,
(*iii*) *Stratified sampling*, (*iv*) *Systematic sampling*.

Below we will precisely explain these terms, without entering into detailed discussion.

14·2·1. Purposive Sampling Purposive sampling is one in which the sample units are selected with definite purpose in view. For example, if we want to give the picture that the standard of living has increased in the city of New Delhi, we may take individuals in the sample from rich and posh localities like Defence Colony, South Extension, Golf Links, Jor Bagh, Chanakyapuri, Greater Kailash etc. and ignore the

localities where low income group and the middle class families live. This sampling suffers from the drawback of favouritism and nepotism and does not give a representative sample of the population.

14·2·2. Random Sampling. In this case the sample units are selected at random and the drawback of purposive sampling, *viz.*, favouritism or subjective element, is completely overcome. *A random sample is one in which each unit of population has an equal chance of being included in it.*

Suppose we take a sample of size n from a finite population of size N. Then there are ${}^{N}C_{n}$ possible samples. A sampling technique in which each of the ${}^{N}C_{n}$ samples has an equal chance of being selected is known as *random sampling* and the sample obtained by this technique is termed as a *random sample*.

Proper care has to be taken to ensure that the selected sample is random. Human bias, which varies from individual to individual, is inherent in any sampling scheme administered by human being. Fairly good random samples can be obtained by the use of *Tippet's random number tables* or by throwing of a dice, draw of a lottery, etc.

The simplest method, which is normally used is the *lottery system* which is illustrated below by means of an example.

Suppose we want to select 'r' candidates out of n. We assign numbers one to n, one number to each candidate and write these numbers (1 to n) on n slips which are made as homogeneous as possible in shape, size, etc. These slips are then put in a bag and thoroughly shuffled and then 'r' slips are drawn one by one. The 'r' candidates corresponding to the numbers on the slips drawn, will constitute the random sample.

Remark. *Tippet's Random Numbers.* L.H.C. Tippet's random number tables consist of 10400 four-digited numbers, giving in all 10400×4, *i.e.*, 41600 digits, taken from the British census reports. These tables have proved to be fairly random in character. Any page of the table is selected at random and the numbers in any row or column or diagonal selected at random may be taken to constitute the sample.

14·2·3. Simple Sampling. Simple sampling is random sampling in which each unit of the population has an equal chance, say p, of being included in the sample and that this probability is independent of the previous drawings. Thus a simple sample of size n from a population may be identified with series of n independent trials with constant probability 'p' of success for each trial.

Remark. It may be pointed out that random sampling does not necessarily imply simple sampling though, obviously, the converse is true. For example, if an urn contains 'a' white balls and 'b' black balls, the probability of drawing a white ball at the first draw is $[a/(a+b)] = p_1$, (say) and if this ball is not replaced the probability of getting a white ball in the second draw is $[(a-1)/(a+b-1)] = p_2 \neq p_1$, the sampling is not simple. But since in the first draw each white ball has the same chance, *viz.*, $a/(a+b)$, of being drawn and in the second draw again each white ball has the same chance, *viz.*, $(a-1)/(a+b-1)$, of being drawn, the sampling is random. Hence in this case, the sampling, though random, is not simple. To ensure that sampling is simple. it must be done with replacement, if population is finite. However, in case of infinite population no replacement is necessary.

14·2·4. Stratified Sampling. Here the entire heterogeneous population is divided into a number of homogeneous groups, usually termed as *strata*, which differ from one another but each of these groups is homogeneous within itself. Then units are sampled at random from each of these stratum, the sample size in each stratum varies according to the relative importance of the stratum in the population. The sample,

which is the aggregate of the sampled units of each of the stratum, is termed as *stratified sample* and the technique of drawing this sample is known as *stratified sampling*. Such a sample is by far the best and can safely be considered as representative of the population from which it has been drawn.

14·3. PARAMETER AND STATISTIC

In order to avoid verbal confusion with the statistical constants of the population, *viz.*, mean (μ), variance σ^2, etc., which are usually referred to as *parameters*, statistical measures computed from the sample observations alone, *e.g.*, mean ($\bar{x}$), variance (s^2), etc., have been termed by Professor R.A. Fisher as *statistics*.

In practice parameter values are not known and the estimates based on the sample values are generally used. Thus, statistic which may be regarded as an estimate of parameter, obtained from the sample, is a function of the sample values only. It may be pointed out that a statistic, as it is based on sample values and as there are multiple choices of the samples that can be drawn from a population, varies from sample to sample. The determination or the characterisation of the variation (in the values of the statistic obtained from different samples) that may be attributed to chance or fluctuations of sampling, is one of the fundamental problems of the sampling theory.

Remarks **1.** Now onwards, μ and σ^2 will refer to the population mean and variance respectively while the sample mean and variance will be denoted by $\bar{x}$ and s^2 respectively.

2. *Unbiased Estimate.* A statistic $t = t(x_1, x_2, ..., x_n)$, a function of the sample values $x_1, x_2, ..., x_n$ is an unbiased estimate of the population parameter θ, if $E(t) = \theta$. In other words, if :

$$E(\text{Statistic}) = \text{Parameter}, \qquad \text{... (14·1)}$$

then statistic is said to be an unbiased estimate of the parameter.

14·3·1. Sampling Distribution of a Statistic. If we draw a sample of size n from a given finite population of size N, then the total number of possible samples is :

$$^{N}C_{n} = \frac{N!}{n!\,(N-n)!} = k, \text{ (say)}.$$

For each of these k samples we can compute some statistic $t = t(x_1, x_2, ..., x_n)$, in particular the mean $\bar{x}$, the variance s^2, etc., as given below.

Sample Number	Statistic		
	t	$\bar{x}$	s^2
1	t_1	$\bar{x}_1$	s_1^2
2	t_2	$\bar{x}_2$	s_2^2
3	t_3	$\bar{x}_3$	s_3^2
⋮	⋮	⋮	⋮
k	t_k	$\bar{x}_k$	s_k^2

The set of the values of the statistic so obtained, one for each sample, constitutes what is called the *sampling distribution* of the statistic. For example, the values $t_1, t_2, t_3, ..., t_k$ determine the sampling distribution of the statistic t. In other words, statistic t may be regarded as a random variable which can take the values $t_1, t_2, t_3, ..., t_k$ and we can compute the various statistical constants like mean variance, skewness, kurtosis,

etc., for its distribution. For example, the mean and variance of the sampling distribution of the statistic t are given by :

$$\bar{t} = \frac{1}{k}(t_1 + t_2 + ... + t_k) = \frac{1}{k}\sum_{i=1}^{k} t_i$$

and $$\text{Var}(t) = \frac{1}{k}\left[(t_1 - \bar{t})^2 + (t_2 - \bar{t})^2 + ... + (t_k - \bar{t})^2\right] = \frac{1}{k}\sum_{i=1}^{k}(t_i - \bar{t})^2.$$

14·3·2. Standard Error. The standard deviation of the sampling distribution of a statistic is known as its *Standard Error*, abbreviated as S.E. The standard errors of some of the well-known statistics, *for large samples*, are given below, where n is the sample size, σ^2 the population variance, and P the population proportion, and $Q = 1 - P$; n_1 and n_2 represent the sizes of two independent random samples respectively drawn from the given population (s).

S. No.	Statistic	Standard Error
1.	Sample mean : $\bar{x}$	$\sigma/\sqrt{n}$
2.	Observed sample proportion 'p'	$\sqrt{PQ/n}$
3.	Sample *s.d.* : s	$\sqrt{\sigma^2/2n}$
4.	Sample variance : s^2	$\sigma^2\sqrt{2/n}$
5.	Sample quartiles	$1{\cdot}36263\,\sigma/\sqrt{n}$
6.	Sample median	$1{\cdot}25331\,\sigma/\sqrt{n}$
7.	Sample correlation coefficient (r)	$(1-\rho^2)/\sqrt{n}$, ρ being the population correlation coefficient
8.	Sample moment : μ_3	$\sigma^3\sqrt{96/n}$
9.	Sample moment : μ_4	$\sigma^4\sqrt{96/n}$
10.	Sample coefficient of variation (v)	$\frac{v}{\sqrt{2n}}\sqrt{1+\frac{2v^2}{10^4}} \simeq \frac{v}{\sqrt{2n}}$
11.	Difference of two sample means : $(\bar{x}_1 - \bar{x}_2)$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.	Difference of two sample *s.d.*'s : $(s_1 - s_2)$	$\sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$
13.	Difference of two sample proportions : $(p_1 - p_2)$	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$

Utility of Standard Error. S.E. plays a very important role in the large sample theory and forms the basis of the testing of hypothesis. If t is any statistic, then for large samples :

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} \sim N(0, 1) \qquad \text{(c.f. § 14·5)}$$

$$\Rightarrow \quad Z = \frac{t - E(t)}{\text{S.E.}(t)} \sim N(0, 1), \text{ for large samples.}$$