Chapter 1

Introduction to Sample Survey Designs

1.1 INTRODUCTION

The objective of a sample survey is to make inferences about a population parameter by observing a portion or sample from it. It is natural to expect that this phenomenon of interpreting about a wider group (a population) on the basis of a sample from it will have some margin of error in the result. How the samples are drawn or selected, while designing a sample survey, is important to ensure that the inference drawn is both valid and reliable. A proper design not only helps obtain a valid idea about the wider population parameters, but also provides a margin of error in an estimate. In addition, the theory can also guide in choosing an alternative design so that the margin of error can be minimized. Different designs that can be employed in a sample survey are discussed in the subsequent chapters. This chapter elaborates few terminologies that will facilitate further discussion on the specific methodologies.

1.2 POPULATION, UNITS AND SAMPLING UNITS

Aggregates of units or elements comprise a population. A study unit is a unit or a member of a study population, as in the case of a study concerning human beings, it can be individual persons, families or households in a given geographical area. A study unit, however, can be any living or non-living subject in a study. For example, if the interest is to estimate the number of fruits in an orchard, the study units will be the fruits in the area. Further, if the interest is to estimate the number of words in a book, words of the book will form the study unit.

A distinction, however, needs to be made between a study unit and a sampling unit. Considering an example of sampling from a human population, it is often preferred to select households before selection of individuals – the study units in a population. In this case, the households, consisting of individuals, which facilitate the process of the sample

selection are called sampling units. However, if our aim is also to estimate household income, then household will be both the study unit and the sampling unit. Therefore, a sampling unit can also be the study unit in a study. Generally, sampling units are a higher level or a group of units. Their purpose is to help selection of the study units. While selecting a sample, as will be discussed later, one may resort to selection at different stages. For example, for selecting households in rural areas of a country like India with state, district, block and village as the administrative subdivisions, one may consider selection at different stages. It is possible to select a few districts at the first stage, few villages from the selected districts at the second stage, and finally few households from the selected villages at the third stage. The sampling units at each stage of this selection are districts at the first stage, villages at the second stage and households at the last stage. Thus, a study unit, which is household in the present case, is unique referring to the ultimate elements or units and is the focus of study, whereas sampling units can vary depending upon the units (generally areal units) that are selected at a particular stage of sampling.

It is also necessary to distinguish between a *study* population, a *target* population and an *external* population. For example, to understand teenage pregnancy in an area, all girls aged 13–19 years residing in the area becomes the target population. However, while implementing the survey design, more often than not, only households that might have girls aged 13–19 years residing in it are considered for selection. Hence, households in which girls aged 13–19 years reside become the study population and it excludes all those staying in institutions like hostels, boarding schools, etc. The difference between a study population and a target population, however, may or may not be there in a particular case.

Sometimes, for sake of convenience or to make a design more efficient in terms of cost, the target population may be very different from the study population. For example, suppose the interest of a study is to estimate the percentage of obese population among the adult population in a city and suppose there exists a big park in the city where all the people come for the purpose of jogging. One way to estimate the obese population could be to draw a sample from among those who visit the park, which would be both economical and convenient. The study population, in this case, is the population that visits the park can be apparently different from the target population, the population of the city. The two populations can be regarded as similar for the purpose of the estimation, if all the factors that influence obesity are similar in the two populations. In other words, if the chances of visiting the park are similar among both obese and non-obese, the two populations could be regarded as similar for the present purpose. Otherwise, if generally the obese are more likely to jog and hence visit the park, the estimate of obesity on the basis of the study population is likely to be positively biased (i.e., overestimated).

The external population, in this case, could be the population of the entire district in which the city is situated or even the country's population. Whether or not the estimate derived from the study population (from which the sample is selected) will hold good or if it can be generalized to the external population will depend on the different factors that influences obesity (the parameter in question). If the study population and

the external population are similar in terms of these factors, a generalization could be made.

Study population, target population and external population can be same or different from each other. A probability sample drawn from a study population will provide a valid estimate of its parameter. But whether it would be so for the target population and external population, if different from the study population, would need further assessment.

To describe a population, there is also a need to have a time reference, particularly if the parameter under study changes with time. In the above example, on measuring obesity in a population, one needs to fix a time reference, such as in year t or during t to $t + k$. It is imperative to have the survey period (the time period when obesity measures are actually collected from a sample) to coincide with the reference period.

1.3 SAMPLING DESIGN

A sampling design is the process to accomplish the following tasks:

- number of units to have in a sample (sample size),
- the procedure of selection of the required units (how to select the required units into the sample),
- the procedure of estimation of the required parameters (to transfer the information collected from the sample into meaningful measure to gauge the population parameter), and
- provide an idea about the likely error that one would commit in making the inference from the sample to the population.

There are different sampling techniques that can be employed to draw a sample. These techniques can be broadly classified into two types: a) probability sampling and b) purposive sampling.

1.4 PROBABILITY AND PURPOSIVE SAMPLING

1.4.1 Probability sampling

In a probability sampling, each and every unit in the study population is given a known and non-zero chance of being included in a sample. Although each and every unit in a population has a chance of getting selected into a sample, the actual selection of a unit, which is carried out mechanically (discussed later), is a chance event and depends on its probability of selection. Hence, the group of units that will ultimately form a sample is not known in advance and there can be a variety of options or alternative samples, once a design, in terms of sample size and the chance of selection of each unit are specified. This phenomenon of exercising the restriction of a known and non-zero probability of selection for each unit into a sample led to the development of sampling theory to provide a basis not only for arriving at a valid estimation procedure but also to have an idea about the error

in estimation. Probability sampling has several applications and, in fact, has been applied extensively – so much so that the application of purposive or non-probability sampling is becoming a rare phenomenon. The present book deals only with the sampling methods that use the probability sampling techniques.

Although the definition given for a probability sampling is simple, the chance of violating it, particularly at the time of sample selection, is high. For example, a violation in providing a chance of selection to each unit can occur in many ways, some of which are very subtle and occur unknowingly. Similarly, one needs to take proper care to ensure that there is no deviation from the proposed scheme of the probability of selection given to each unit. These issues are discussed in Chapters 5 and 10.

1.4.2 Purposive sampling

Purposive sampling is done based on subjective selection of a sample, largely based on a few assumptions. One assumption is that there is homogeneity within a population under study. In other words, the variation in a variable that is being studied is practically negligible. For example, as we all know, few grains of rice (or in other words taking a sample of) from a large bowl are enough to check whether it is cooked properly.

Estimate of population size, in the earlier period, is determined by using this principle. Estimating certain rates or ratios, such as population density (number of population per square kilometres) or ratio of number of school-age children to adult population, or number of civilians per 1000 military population, one could estimate the total population. For example, if total area (X) is known, one could estimate the density of population (d) by counting population in few selected areas as

$$d = \frac{y}{x}$$

where y is the total population size observed in area x. Hence, using simple extrapolation method, the estimated total population size in area X will be

$$\text{Total population} = X \cdot d = X \cdot \frac{y}{x} \tag{1.1}$$

In other words, for the estimate of population of X, extrapolation is done on the basis of that observed in the sample (for area x). For such generalization or extrapolation to be valid, assumption of homogeneity of the distribution should be true for the entire area X. If not true, the extrapolation will be invalid, there will be error in the estimate and its magnitude will depend on how the density of population varies in the area X. It is, however, not possible to provide any idea about the extent of the error, if the sample is drawn purposively.

Another assumption in the purposive sampling is that the distribution of units in a population is random. If population units are distributed randomly, selection of units from any section of such a population would provide a random sample. Simply stated, this means that characteristics of two units placed adjacent to each other are as similar or dissimilar as any two units selected randomly from the population. For example, in a

village, socio-economic conditions of households adjacent to each other are more likely to be similar in terms of certain characteristics. This phenomenon is known as 'clustering' and will be discussed later (Chapters 2 and 3).

In recent years, a type of purposive sampling, known as 'quota sampling' has gained some popularity. In this method, the assumption of homogeneity is relaxed.

Let us discuss this by considering the estimation of total population in area X, discussed above. We may say that the density of population will not be homogeneous throughout but will vary by, say, type of land: a) non-agricultural land, b) agricultural land, and c) fallow land. Let the total area of these three are X_1, X_2, and X_3, respectively, such that $X_1 + X_2 + X_3 = X$. One then needs to subdivide the sample x and select from each of the three categories. Let these be x_1, x_2, and x_3, respectively, which are called quota from each category. Assuming homogeneity within each category and using equation 1.1.

$$\text{Estimated total population} = X_1 \cdot \frac{y_1}{x_1} + X_2 \cdot \frac{y_2}{x_2} + X_3 \cdot \frac{y_3}{x_3} \tag{1.2}$$

where y_i is the estimated population in area x_i.

It can be noticed that, in this case, the extrapolation is done within each category and then added up. For this extrapolation, the required assumption would be that the population is homogeneously distributed within each category of land. This is more tenable than the assumption made earlier for the entire population. As for another example, in an opinion poll survey, suppose it is decided that education, occupation and gender are the three variables that are important in understanding the voting behaviour. One can have different categories with the help of these three variables (for further discussion on creation of categories, see Chapter 2). Interviewers can then be given quotas (x_i), number of persons to be interviewed in a category and obtain each person's preference for an individual political party. If y_{ij} denotes the number of persons in ith category who favoured the jth party, extending equation 1.2, it would be possible to estimate the overall proportion of the population who favours jth political party. However, in many situations, it is not possible to correctly assess the different variables that should be considered to capture the variations in the study variable. Population size in each category (X_i's), which are required for the estimation (see equation 1.2), may also not be available and the assumption of perfect homogeneity within a category may be difficult to ensure. A cost-effective approach is often to adopt a probability sampling design.

1.5 FRAME

A frame is the basis for drawing sampling units in a probability sampling. If sampling units are elements, a frame is supposed to provide the list (detailed address to facilitate contacting a selected unit) of all the elements. A frame, if perfect, provides the list of all the units occurring once and only once and should not include any other elements. That is, if there are N units in a population, the frame should have only N entries in the list giving contact details of each of the units.

If sampling units refer to some macro-level units such as villages and districts, such area units together should cover the entire population under study. Suppose villages are the sampling units. Each village should have a clearly defined boundary; non-overlapping to each other and together they should comprise the total population. It may be noted that the contact details of persons or households within each village is not needed. Such details would be required only for the villages that actually get selected. Auxiliary information on units or sampling units, if available, can (as will be discussed later) strengthen a design. A frame, alternatively called a sampling frame, provides the basis for implementing a probability design.

1.6 BIAS AND ERROR

When a design is specified, as mentioned, there may be several alternative samples that could be selected. Each possible sample will have an error that is defined as the deviation between a sample estimate and the corresponding population parameter. Assuming that the interest is to estimate the population mean, such error would be equal to the difference between a sample mean and the population mean. For a given design, one can have a large number of errors, each corresponding to an alternative sample value. Therefore, it is possible to have one value that will provide a summary measure of error (taking differences from the alternative samples) for a design.

Now the questions are why do we require a summary measure of error for a design and how do we obtain it when, in practice, only one sample value is available.

We require a summary measure of error because as alternative samples are possible, in a probability sampling design, for the actual selection, it is necessary to provide 'protection' against all the possible samples of a design. A summary measure of error can help us provide such an idea for the design.

While it is true that, in practice, only one sample gets selected by applying the design and will be available for the estimation. The sampling theory, however, helps get an estimate of the error for a design from a single sample. The summary measure of error is denoted by mean square error (MSE). The MSE comprises of two components: bias and error.

It needs to be emphasized that the deviation (between a sample estimate and the population parameter) caused by the sample selection is generally referred to as the error due to sampling or simply by 'sampling error'. The bias and error can also occur at several other stages in the entire process of a survey design. For example, a deviation between a sample estimate and the population parameter can happen at the stages of data collection, data entry and its processing. These latter types of causes for the deviations are known as the non-sampling errors (discussed in Chapter 5). Non-sampling errors occur during information generation and they are common to both sample survey and census. Although MSE refers to both the sampling and non-sampling errors, our discussion here only confines to the sampling error.

There is, however, a subtle difference between the terms 'bias' and 'error', both of which tend to increase the MSE. Let us discuss them in the context of the sample mean.

If $\bar{y}_i$ denotes the sample mean in the case of the ith alternative sample with $\bar{y}$ as the mean of all the sample means in a design, then

$$\text{MSE}(\bar{y}) = \frac{\sum_{i=1}^{K}(\bar{y}_i - \bar{Y})^2}{K} \tag{1.3}$$

where $\bar{Y}$ is the population mean and K is the number of alternative samples possible in the design.

It basically takes into account how each sample mean deviates from the overall population mean. Denoting $E(\bar{y}_i)$ as the expected value or mean of all the possible sample means, the above equation can be written as

$$\text{MSE}(\bar{y}) = \frac{\sum_{i=1}^{K}\left((\bar{y}_i - E(\bar{y}_i)) + (E(\bar{y}_i) - \bar{Y})\right)^2}{K}$$

since the product term is zero, it can be written as

$$\begin{aligned} &= \frac{\sum_{1}^{K}\left[\left(\bar{y}_i - E(\bar{y}_i)\right)^2 + \left(E(\bar{y}_i) - \bar{Y}\right)^2\right]}{K} \\ &= \left(\bar{y} - \bar{Y}\right)^2 + \frac{\sum_{i=1}^{K}\left(\bar{y}_i - \bar{y}\right)^2}{K} \end{aligned} \tag{1.4}$$

The first component on the right-hand side of equation 1.4 indicates square of the deviations between mean of all possible sample means in the design from the population parameter and is termed as the bias occurring in the sample estimate due to the sampling design.

Why is it called a bias and not an error? What does the presence of bias in a design signify?

While there are K alternative samples possible in a design, the focus is on their mean value ($\bar{y}$). A desirable property of a design is that $\bar{y}$ should equal the population mean ($\bar{Y}$). In other words, while some of the values of $\bar{y}_i$ will be higher than $\bar{Y}$ and the deviations will be positive, for the others the deviation will be negative and $\bar{y}$ would equal $\bar{Y}$ if the positive and the negative deviations cancel out. This means that the sample mean of the design will be regarded as an unbiased estimate of the corresponding population mean. That is, the first component of equation 1.4 would be zero. It is then that whichever sample actually gets selected, the sample mean of it will be regarded as a valid estimate of the population mean.

If there is a bias, that is, the positive and the negative deviations do not cancel out, it will mean that $\bar{y}$ is either higher or lower than the population mean, $\bar{Y}$. In such a case, the sample mean cannot be regarded as a valid estimate of the corresponding population mean. Hence, one needs to have an estimate of the bias and make appropriate corrections to get a valid estimate.

To clarify further, consider the case of estimation of income in a population consisting of rich as well as poor in equal proportion. That is, half the population is considered rich

and the other half as poor. If a sample design (with a fixed sample size n) provides greater chances of selection of poor people in the sample, it is quite natural to expect that it will end up having poor people in greater proportion into a sample compared to that existed in the population, and hence, the mean of all possible sample means is expected to be lower than the population mean or that it will be negatively biased. Similarly, if the design assigns greater chances of selection to the rich compared with the poor, the mean of all possible sample means would be higher or positively biased.

The sampling theory shows that the mean of all possible sample means would be equal to the population mean or unbiased only when the design is equal probability of selection method (EPSEM); that is, it assigns equal chance of selection to each and every unit to be selected in a sample.

The second component of equation 1.4 gives an idea about the extent to which alternative sample means, in a design, vary from their mean, and is known as the sampling variance of mean and its square root as the standard error of the mean. Hence,

$$\text{MSE}(\bar{y}) = B^2 + \text{sampling variance of the mean} \tag{1.5}$$

where B is the extent of the bias.

The square root of MSE is known as the standard error of the mean and for an unbiased estimate, MSE is equal to the sampling variance of the mean.

Following is an illustration to better understand the meaning of both the bias and the error by considering a small hypothetical population.

Example 1.1

The daily wage of a population consisting of six persons is as follows:

The small population under consideration has six members A–F, whose daily wage is given below.

Person	A	B	C	D	E	F
Daily wage	50	60	1200	100	1400	70

For this small population, the MSE of sample mean is computed for a design where a sample of three individuals are selected one by one without replacement (i.e., once a unit is selected, it will not have another chance of selection) giving equal chance of selection to each units.

$$\text{The population mean } (\bar{Y}) = \frac{(50 + 60 + \cdots + 70)}{6} = \frac{2880}{6} = 480$$

With this sample design, there are 20 $\left(\frac{6!}{3!3!}\right)$ different samples of size 3 that can be chosen. Table 1.1 provides the necessary computations to understand the components of bias and error in the sample mean.

Few characteristics of the population and the distribution of the sample deviations from population mean may be worth noting. Out of the 6 units, 2, C and E, are rich whereas the

Table 1.1 Estimates of average daily wage and deviation between sample mean and the population mean for the 20 alternative samples drawn from the population shown in example 1.1

S. No. of samples	Units in a sample	Sample mean $(\bar{y})$	Deviation $(\bar{y} - \bar{Y})$
1	ABC	436.67	−43.33
2	ABD	70.00	−410.00
3	ABE	503.33	23.33
4	ABF	60.00	−420.00
5	ACD	450.00	−30.00
6	ACE	883.33	403.33
7	ACF	440.00	−40.00
8	ADE	516.67	36.67
9	ADF	73.33	−406.67
10	AEF	506.67	26.67
11	BCD	453.33	−26.67
12	BCE	886.67	406.67
13	BCF	443.33	−36.67
14	BDE	520.00	40.00
15	BDF	76.67	−403.33
16	BEF	510.00	30.00
17	CDE	900.00	420.00
18	CDF	456.67	−23.33
19	CEF	890.00	410.00
20	DEF	523.33	43.33
Sum of deviation			0.0

remaining 4 are relatively poor. The alternative samples, in the design with sample size 3, such as ACE, BCE, CDE and CEF where both the rich individuals get selected, shows a large average wage and consequently their deviations from the population mean are positive and large. Similarly, there are four alternatives, ABD, ABF, ADF and BDF where the samples consist of all the three belonging to the poor. This means that their sample means are way below the population mean. In the remaining 12 samples, the deviations are not too much as they consist of 2 from the poor and 1 from the rich individuals. However, since the design is EPSEM, that is, it gives equal chance to each individual to be selected in the sample, the average of all possible samples is equal to that of the population mean (sum of the deviations of all the possible sample means from that of the population mean is zero or

zero bias). Thus, it gives an unbiased estimate (the proof of this theory can be obtained from any text book on survey sampling). Alternatively, we say that the expected value of sample means $(\bar{y})$ would equal to the population mean $(\bar{Y})$. Hence, the sample mean estimates the population mean without any bias, if the procedure of selection provides equal chance to each and every element in a population to be included in a sample. It is clear from Table 1.1 that none of the alternative sample means is exactly equal to the population mean. So, the property of being unbiased does not in any way guarantee that the sample mean from the sample that actually gets selected by such a design would equal to the population parameter. It would be pertinent to obtain the value of MSE of sample mean.

Since the estimate is unbiased, the first component (equation 1.4) is zero, it will be

$$\text{MSE}\ (\bar{y}) = \frac{(-43.33)^2 + (-410.00)^2 + \cdots + (43.33)^2}{20} = 67953.33$$

Note that this will be equal to its sampling variance as the mean is an unbiased estimate of the population mean. Hence,

$$\text{SE}\ (\bar{y}) = \sqrt{\text{MSE}(\bar{y})} = 260.68$$

where SE $(\bar{y})$ is the standard error of mean in the design.

While the implication of SE $(\bar{y})$ is discussed in Chapter 2, it may be noted here that it is an error and denotes the likely deviation between a sample and the population mean. It can be noted that its magnitude is quite large.

Now, consider another design in which we draw a sample of three units from the six individuals mentioned above. Suppose one decides to select C and E with certainty (probability of one) and one of the remaining four with equal probability. The four possible samples and the corresponding sample means are shown in Table 1.2.

Table 1.2 Estimates of average daily wage and the deviation of sample mean from the population mean for the four alternative samples from the second design (one selected randomly from A, B, D and F and both C and E are included with certainty, see data in example 1.1).

S. No. sample	Units in sample	Sample mean	Deviation from population mean $(\bar{Y})$
1	ACE	883.33	403.33
2	BCE	886.67	406.67
3	CDE	900.00	420.00
4	CEF	890.00	410.00

It can be seen that the deviations are not randomly distributed around the population mean, as in the previous design. The four alternative samples are all on the higher side. In this case,