UNIT 2 SAMPLING

NOTES

Structure

2.0 INTRODUCTION

This unit discusses the purpose, principles and methods of sampling. It also outlines probability and non-probability sampling. The unit further deals with data, its types and sources. It details the tools of data collection, like observation, interview, questionnaire, focus group and case study methods. You will also learn about the measurement and scaling techniques of data and will be able to ascertain reliability and validity of scales.

2.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss advantages of sampling over complete enumeration
- Analyse principal steps in conducting a sample survey
- Elaborate on various methods of drawing a sample from a population

2.2 OVERVIEW OF SAMPLING

NOTES

2011 census had been undertaken in February throughout the country. As is common knowledge, this exercise is conducted once in ten years (decennial census) and takes a head count of every individual in the country (population or aggregate) and also collects certain pre-determined information from every individual. In case we are interested in knowing certain characteristics of the entire population, it might intuitively appear that only a complete enumeration might give the most accurate result but the scientific evidence points to the other direction. In other words, it is possible to estimate population characteristics based on a small sample drawn from it on a scientific basis and using rigorous statistical methodology. The sampling method has several advantages over complete enumeration or census.

In real life, it is said that if we wish to know whether rice has been fully cooked or not, it is enough to take a few rice grains from the cooking vessel and subject them to examination. If the sample so taken is half-cooked, so will be the entire quantity. It will not be inaccurate to say that our attitudes, actions, assessments are based, to a large extent, on samples. This sampling technique is as true in human affairs as in social research or indeed in any scientific research. Before elections, experts conduct pre-poll surveys. When these surveys are conducted using rigorous statistical methodology, one can accurately determine the outcome of the election and also point out the trend. Sample surveys are used for a wide variety of purposes including planning as in statecraft, market research, business, scientific research, etc. Statistical quality control uses sampling techniques extensively. Decisions are made on whether to accept or reject a particular lot depending on whether a sample drawn from it conforms to certain quality specifications. Insofar as social research is concerned, a number of sample surveys are important which include, among others, the National Family and Health Survey (NFHS), educational survey, radio listenership survey, television viewership survey, national readership survey (NRS) of newspapers.

The National Sample Survey Organization in India conducts large-scale sample surveys for the collection of data relating to, among other issues, agricultural production and land use, industrial production, wholesale and retail prices, household income and expenditures (which is used to estimate poverty line) and other socio-economic statistics.

It is often the case that we lack the time, financial and human resources to study more than a fragment of the population. Hence, Cochran states that the sampling method enjoys the following advantages over complete enumeration (1977:1).

NOTES

1. **Reduced cost**

   If data is collected in respect to only a small sample, expenditure is likely to be smaller than if a complete enumeration of all units in the population is attempted. In case of large populations, fairly accurate results could be obtained even by using 1 per cent sample or even smaller samples.

2. **Greater speed**

   Data can be collected and analysed faster with a sample than with a complete count.

3. **Greater scope**

   A complete enumeration is not feasible in cases which require highly trained personnel or specialized equipment, whose availability is limited. Surveys have more scope and flexibility regarding the types of information that can be obtained.

4. **Greater accuracy**

   A sample can lead to deployment of higher quality personnel, more intensive training, more careful supervision of fieldwork and processing of results. For these reasons, a sample produces more accurate results as compared to census.

Statistical inference is concerned about how estimates based on sample can be used to make inferences about population characteristics (population parameters). In case we choose a good sample based on sound statistical methodology, it will lead to good estimates. The process of selecting a sample from the population is called *sampling*. Sampling has many advantages over census or 100 per cent enumeration. In those cases where the research process is destructive in nature, sampling minimizes the destruction. Non-sampling errors are not present in sampling process.

Cochran (1977:4) classifies sample surveys broadly into *descriptive* and *analytical*. The descriptive survey seeks to obtain certain information about large groups: the number of men, women and children who use particular toothpaste. In contrast, 'in an analytical survey, comparisons are made between different subgroups of the population, in order to discover whether differences exist among them and to form or to verify hypotheses about the reasons for these differences.' For instance, in an education survey, it could see whether there exist any significant differences between those getting education from the government institutions and those in private institutions. Many surveys provide data that serve both purposes.

2.2.1 Principal Steps in a Sample Survey

Cochran (1997:5) describes the principal steps in the planning and execution of a survey as follows:

NOTES

1. **Objectives of the survey**

   It is extremely important to define clearly the objectives of the survey. Otherwise there is a danger of straying into uncharted areas while collecting data.

2. **Population to be sampled**

   The aggregate from which the sample is drawn is known as the 'population'. The researcher needs to follow clear rules to decide on the population and also whether the cases belong to this chosen population. The sampled population must match the target population taken.

3. **Data to be collected**

   Questions asked and information collected must be relevant to the purposes of the survey. One must avoid the tendency of collecting data which is not likely to be analysed subsequently. An unduly long questionnaire might result in impairing the quality of answers to important questions.

4. **Degree of precision desired**

   Sampling process is likely to give rise to error as only a part of the population is being measured or on account of errors in measurement. Data user must specify in advance as to how much error in estimates can be tolerated. In some cases, ten per cent error or five per cent error might be acceptable. In extreme and critical cases, 1 per cent or even .001 per cent might be acceptable, viz. as in the case of contamination of intra-venous fluids or lifesaving drugs or defence equipment.

5. **Methods of measurement**

   Selecting methods of measurement is imperative before a research is undertaken. A survey may entail either interviewing people individually or providing a detailed questionnaire to be filled in individually on as a group. It could be by mail, telephone, by personal visit or by a combination of the three. The preliminary work consists of preparing record forms in which the questions and answers are to be entered.

6. **Frame**

   The entire population is divided into smaller groups that could serve as samples. These units should be different and must not overlap. This construction of smaller units is called 'the frame'.

7. **Selection of the sample**

   There is now a variety of plans—simple random sampling, stratified sampling, systematic sampling etc.—by which the sample may be selected. Every plan has its own sampling theory; and according to these theory, sample size is made. Time involved for each plan and relative costs are compared.

NOTES

8. Pretest

The questionnaire as well as filed methods are tried on a pilot basis which points out to the troubles ahead with the full survey and thus leads to improvements in the questionnaire.

9. Organization of fieldwork

Field personnel must be trained in advance about the objectives of the survey, methods of measurement and their work, in particular in initial stages, must be supervised. In case the investigator fails to gather information from the sample units, necessary arrangements must be made for handling the situation.

10. Summary and analysis of data

The first step is to edit the completed questionnaires to correct recording errors, if any, and thereafter estimates are computed. Cochran (1977:7) observes that in the presentation of results, it is good practice to report the amount of error to be expected in the most important estimates.

11. Information gained for future surveys

After completing a sample survey, one gets a good idea about the population which certainly helps in drawing a sample that will give accurate estimates in future. One can also learn from mistakes made.

In the sampling design process, the following steps are crucial:

(a) **Define target population:** This must be based on research objectives.

(b) **Determine sampling frame:** This must be done before the research is undertaken.

(c) **Selection of appropriate sampling technique:** Researcher has to determine whether he will go in for probability or non-probability sampling techniques; with replacement or without replacement. In sampling with replacement, an element is selected from the frame, the required information is obtained, and then the element is placed back in the frame. This way, there is a possibility of the element being selected again in the sample. As compared to this, in sampling without replacement, an element is selected from the frame and not replaced in the frame. In this way, the inclusion of further of the element in the sample is eliminated. (Naval Bajpai, 2010: 260).

(d) **Determine sample size:** Sampling theory seeks to develop methods of sample selection and also precise estimates at low cost. In order to apply the principle of specified precision at minimum cost, Cochran (1977:8) observes that we must be able to predict, for any sampling procedure, the precision and the cost to be expected. He adds that any sampling procedure is judged by

NOTES

examining the frequency distribution generated. This is cross-checked by appliyng the procedure a number of times to the same population. 'With samples of the sizes that are common in practice, there is often good reason to suppose that the sample estimates are approximately normally distributed. With a normally distributed estimate, the whole shape of the frequency distribution is known if we know the mean and the standard deviation (or the variance). A considerable part of sample survey theory is concerned with finding formulas for these means and variances.'

2.3 PROBABILITY AND NON-PROBABILITY

Before we proceed further, it is important to understand the scope of probability. In normal life, we come across either *deterministic* or *probabilistic* phenomena. If an experiment is repeated under similar conditions and if it leads to a unique or certain outcome, it falls under predictable phenomena. Several experiments in physical sciences belong to this category.

On the other hand, if an experiment is repeated under similar conditions but it does not end up in a unique result but one of several possible outcomes, then it is termed as 'unpredictable' or 'probabilistic' phenomena. One finds examples of this type in business, economics and social sciences or in day-to-day life. In common parlance, we come across usages like 'possibly', 'high chance', 'likely' and 'odds' which are indicative of a degree of uncertainty about the happening of an event. 'A numerical measure of uncertainty is provided by a very important branch of Mathematics called "theory of probability". Broadly, there are three possible states of expectation—'certainty', 'impossibility' and 'uncertainty'. The probability theory describes certainty by 1, impossibility by 0 and the various grades of uncertainties by coefficients ranging between 0 and 1.'

The samples used in practice fall into two distinct categories; viz., probability sample and non-probability sample. In a probability sample, one selects the items based on known probabilities. One must use this kind of sampling for it has scientific advantage over non-probability sample. In the case of probability sampling, one can determine unbiased estimators of population parameters and hence can make sound inferences. At times, it might be difficult or not feasible to go in for a probability sample. In such cases, one must strive towards a probability sample and also acknowledge any potential biases that creep in on account of compromises made in scientific selection of samples. The five types of probability samples used are *simple random*, *stratified*, *cluster*, *systematic* and *multi-stage* sampling. Each of these sampling methods vary in their cost, accuracy and complexity.

Ya-Lin Chou observes that 'Probability is the science of decision-making with calculated risks in the face of uncertainty.' It owes its origin to

NOTES

the study of gambling during the 16th century by Europeans but subsequently, several Russian experts played a key role in its development. It finds extensive applications in Statistics, Econometrics, Engineering and many other disciplines.

Probability is defined as the likelihood of the occurrence of an event and it ranges between 0 and 1. A probability of 0 means that the occurrence of that event is impossible while a probability of 1 indicates certainty of the event's occurrence. Normally, its value ranges between 0 and 1.

James Bernoulli gave the following classical definition of probability. He maintains that if a random experiment or a trial results in 'n' exhaustive mutually exclusive and equally likely outcomes [or cases] out of which 'm' are favourable to the occurrence of an event E, then the probability 'p' of occurrence [or happening] of E, usually denoted by P[E] is given by:

$$p = P[E] = \frac{\text{Number of Favourable Cases}}{\text{Total Number of Exhausive Cases}} = \frac{m}{n}$$

Example 1: If a coin is tossed up, the probability of getting head according to the above definition is ½.

Example 2: A pack of 52 cards has 13 cards each of hearts, spades, diamonds and clubs. So when a card is drawn from a pack of 52 cards, the probability of getting a spade is 13/52 = ¼ = 0.25.

It is important to note that the above definition only serves to introduce the subject of probability.

Sampling theory is concerned with the way a sample is selected in order to represent the population and its characteristic. However, in practice, one does not know the population characteristics and for this reason, it is difficult to accurately ascertain whether a sample represents the entire population. The theory of probability described above is very helpful in such a scenario. When a random sample is drawn on a scientific basis using probability, the degree to which a sample represents the population can be calculated in probabilistic terms. The researchers, therefore, will be in a position to assert the degree to which a sample represents a population (Kultar Singh, 2007:900).

In a non-probability sample, one selects the items or individuals without knowing their probabilities of selection. Hence, statistical theory developed in the case of probability sampling cannot be applied. Non-random sampling techniques include quota sampling, convenience sampling, judgment sampling, and snowball sampling. They are described in detail in the following section. Non-probability samples have certain advantages like convenience, speed and low cost. Yet when we go in for them, the above advantages are offset by lack of accuracy due to selection bias. Their results also cannot be

NOTES

generalized. For this reason, non-probability sampling methods are used only for small-scale studies that precede large investigations.

2.3.1 Probability Sampling

Supposing we have a population and our sampling procedure gives rise to a set of distinct samples S_1, S_2,...,S_n. We know which sampling units belong to S_1, to S_2 and so on. Each possible sample S_i has assigned to it a known probability of selection p_i. We select one of the S_i by a random process in which each S_i receives its appropriate probability p_i of being selected. The method for computing the estimate from the sample must be stated and must lead to a unique estimate for any specific sample.

Probability sampling refers to a method of above type of sampling procedure in which we can calculate the frequency distribution of the estimates it generates if repeatedly applied to the same population. The frequency of any particular sample S_i will be selected is known and we can also calculate estimate from data in S_i.

There are two ways of selecting a sample: *random selection* and *purposive selection*. Random selection is an illustration of probability sampling in which each unit in the population had an equal chance of being included in the sample. In this case, the use of sampling theory and normal distribution enable the sampler to predict from sample data, the amount of error to be expected in the estimates made from the sample. In contrast, purposive selection amounts to non-probability sampling.

To sum up, in random sampling, every unit in the population has the same or equal chance (probability) of being selected as part of the sample. In purposive sampling, the researcher's subjectivity comes in and chance factor is given a go-by. On the basis of the selection procedure used, random and non-random sampling techniques are referred to as probability and non-probability sampling. Random sampling methods include:

1. Simple random sampling
2. Stratified sampling
3. Cluster sampling
4. Systematic sampling
5. Multi-stage sampling

1. Simple random sampling

In simple random sampling each member of the population has an equal chance of being included in the sample. It is the most common method of selecting a sample from the population. First, a complete list of all the members of the population is prepared and each element is given a distinct number, say, from 1 to N. Thereafter, 'n' items are selected from a population of size N either using

1. The lottery method or
2. The use of random number tables generated by statisticians or by a computer program

NOTES

Simple random sample only means that the process of selecting a sample should be free from human judgment and therefore bias. (Naval Bajpai, 2010: 261). Simple random sampling is a method of selecting 'n' units out of the N such that every one of the distinct samples has an equal chance of being drawn. In practice, a simple random sample is drawn unit by unit. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. Other methods of sampling are often preferable to simple random sampling on the grounds of convenience or of increased precision. Simple random sampling serves best to introduce sampling theory. (Cochran, 1977: 20).

There are two distinct ways of drawing a sample; viz., sampling with replacement and sampling without replacement. In the case of sampling with replacement, one selects an item after which it is again returned to the frame and thus, it has same probability of being selected again. For instance, if we have a pack of 52 playing cards, we pick up a card. Assuming that it is a Jack of Diamond, we note the selection and then again return that card to the pack. The pack is reshuffled and then one card is selected again. Thus, Jack of Diamond has the same probability of being picked up as any other card, namely, 1/52. This process is carried out till we get a sample of size 'n'. It is possible that one might not like to see the same item being repeated in the sample.

Sampling without replacement means that once an item is selected, it is set aside and is not returned to the population. Thus, it cannot be selected again. In the above example, the probability of drawing a jack of diamond in the first draw is 1/N. The probability of selecting any card not previously selected on second occasion is now 1/N–1. The whole process is repeated until we reach the sample size of 'n'.

Suppose we have N units in the population viz. $y_1, y_2,\ldots,y_N$. Capital letters are normally used to characteristics of the population while lowercase letters to those of the sample. For totals and means we have the following definitions.

Population:

Total $Y = \sum y_i = y_1 + y_2 + \ldots + y_N$

Mean $\bar{y} = \frac{\sum_i^N y_i}{N} = [y_1 + y_2 + \ldots + y_N]/ N$

NOTES

Sample:

Total $\sum_{i}^{n} y_i = y_1 + y_2 + ... + y_n$

Mean $\bar{y} = [y_1 + y_2 + ... + y_n]/ n$

Though sampling is undertaken for many purposes, the following four characteristics of population are of great interest:

1. Mean = $\bar{Y}$ [e.g., the average age of all legislators in the country]
2. Total = Y [e.g., the total number of acres under rice cultivation in a State]
3. Ratio of two totals or means $R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$ [e.g., ratio of boys and girls undergoing schooling in a city]
4. Proportion of units that fall into some defined class [e.g., proportion of people with disabilities in a State]

Sampling theory is concerned with estimation of the above four characteristics of the population. The symbol ^ denotes an estimate of a population characteristic made from a sample. Given below are some simple estimators.

	Estimator
Population mean	Sample mean = $\bar{Y}$
Population Total Y	$N = N \frac{\sum_{i}^{n} y_i}{n}$
Population Ratio	$= \frac{\bar{y}}{\bar{x}} = \frac{\Sigma y_1}{\Sigma x_1}$

When n = N, or when the sample population and the total population is equal, the estimation method is called *consistent*. For simple random sample, it is obvious that $\bar{y}$ and N $\bar{y}$ are consistent estimates of the population mean and total respectively. A method of estimate is regarded as unbiased if the average value of the estimate taken over all possible samples of given size n, is exactly equal to the true population value. It can be proved mathematically that sample mean and total, i.e., $\bar{y}$ and N $\bar{y}$ are unbiased estimates of population mean and population total, i.e., $\bar{Y}$ and Y respectively.

Sampling proportions and percentages

While dealing with qualitative characteristics, sometimes we wish to estimate the total number, the proportion, or the percentage of units in the population that possess some characteristic or attribute or fall into some defined class.

NOTES

Many of the results regularly published from censuses or surveys are of this form, for example, number of unemployed persons, the percentage of the population that is native-born, etc. In case every unit in the population falls into one of the two classes C and C', we use the following notation:

Number of units in C in		Proportion of units of C in	
Population	Sample	Population	Sample
A	a	P = A/N	p = a/n

The sample estimate of P is p and the sample estimate of A is Np or Na/n. The sample proportion p = a/n is an unbiased estimate of the population proportion P = A/N. (Cochran, 1977: 51).

Stratified random sampling

The population of N units is first divided into subpopulations of N_1, N_2, ...N_L units respectively in stratified sampling. These subpopulations are not overlapping and together they comprise the whole of the population, so that $N_1+N_2+...+N_L = N$.

The subpopulations are each called *stratum*. Cochran observes that to obtain the full benefit from stratification, the values of N_h (or in other words the number of stratum) must be known. When the strata have been determined, a sample is drawn from each, the drawings being made independently in different strata. The sample size within the strata are denoted by n_1, n_2, ...,n_L respectively. If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling.

Cochran points out that stratification is a common technique on account of the following reasons:

1. In case of known data, every subdivision should be treated as a unique 'population' in its own right.
2. This stratification of the entire population is often dictated by administrative convenience, as local affices of the surveying organization would be comfortable with regional surveys.
3. Every population may have a differant sampling problem. People living in hospitals, hotels, prisons, etc., are placed on a different stratum than ordinary people. The approach of survey in these two different strata are completely different.
4. A heterogeneous population may stratified into smaller strata which are internally homogeneous. If every stratum is equal, the stratum mean for one would be put to use to calculate mean for the entire population.

NOTES

Cluster sampling

In practice, it is possible that the sampling unit consists of a group or cluster of smaller units that are elements or subunits. Spelling out reasons for the widespread use of cluster sampling, Cochran observes that even if our first intention is to use elements as sampling units,

> ... it is found in many surveys that no reliable list of elements in the population is available and that it would be prohibitively expensive to construct such a list. In many countries, there are no up-to-date lists of people, the houses, or the farms in any large geographic region. From maps of the region, however, it can be divided into areas units such as blocks in the cities and segments of land with readily identifiable boundaries in the rural parts. In the United States these clusters are often chosen because they solve the problem of constructing a list of sampling units.

In cluster sampling, we divide the population into non-overlapping areas or clusters. Unlike stratified sampling where strata are homogeneous, in cluster sampling, clusters are internally heterogeneous. A cluster contains a wide range of elements that are good representatives of the population. Often original clusters are further divided into smaller clusters, which is called *two-stage sampling*.

Bajpai points out that,

> Cluster sampling is very useful in terms of cost and convenience. When compared to a stratum in stratified random sampling, clusters are easy to obtain and focus of the study remains on the cluster instead of the entire population, so cost is also reduced in cluster sampling. In real life, cluster sampling becomes the only available option because of the unavailability of the sample frame. This does not mean cluster sampling is free from drawbacks. Cluster sampling may be statistically inefficient, in cases where elements of the cluster are similar.

In most applications the cluster units (countries, cities, and city blocks) contain different numbers of elements or subunits (areal units, households, persons). Statistical methods have been developed to handle single-stage cluster sampling with clusters of unequal size or sub-sampling with units of unequal size or two-stage sampling.

Systematic sampling

Suppose there are N units in the population that are numbered between 1 to N in some order. To select a sample of n units, we take a unit at random from first k units and every kth unit thereafter. For instance if k = 15 and if the first unit drawn is number 6, the subsequent units are 21, 36, 51,...., and so on. In other words, the selection of first unit, i.e., 6 determines the whole sample. This type is called every kth sample. [Cochran, 1977: 205].

Example 1:

Let us assume that there are 25 units in the population, or N = 25 and that k = 5. Take a systematic sample from the above population.

As K = 5, we take a unit at random from the first five units and thereafter every kth unit is selected. Thus it can give rise to the following samples:

Sample 1:	1	6	11	16	21
Sample 2:	2	7	12	17	22
Sample 3:	3	8	13	18	23
Sample 4:	4	9	14	19	24
Sample 5:	5	10	15	20	25

Cochran outlines advantages of systematic sampling over simple random sampling as follows:

1. It is easier to draw a sample and often easier to execute without mistakes. This is a particular advantage if drawing is done in the field. Even when drawing is done in an office there may be a substantial saving in time.
2. Intuitively, systematic sampling seems likely to be more precise than simple random sampling. In effect, it stratifies the population into 'n' strata which consist of first k units, the second k units, and so on. We might therefore expect the systematic sample to be about as precise as the corresponding stratified random sample with one unit per stratum. The difference is that with systematic sample, the units occur at the same relative position in the stratum, whereas with the stratified random sample, the position in the stratum is determined separately by randomization within each stratum. The systematic sample is spread more evenly over the population, and this fact has sometimes made systematic sample more precise than stratified random sampling.

Multi-stage sampling

It involves selection of units in more than one stage. The population consists of primary stage units and each of these primary stage units consists of secondary stage units. In the process of a multi-stage sampling, first, a sample is taken from the primary stage units and then a sample is taken from the secondary stage units, etc. In a practical situation, stages could correspond to samples taken from States, districts, cities, blocks in different stages.

2.3.2 Non-Probability Sampling

Cochran (1977:10) gives following common examples to illustrate non-probability sampling:

1. The sample is restricted to a part of the population that is readily accessible. A sample of coal from an open wagon may be taken from the top 6 to 9 inches.
2. The sample is selected haphazardly. In picking 10 rabbits from a large cage in a laboratory the investigator may take those that his hands rest on, without conscious planning.

NOTES

3. In case a sampler gets a heterogeneous population, he checks it entirely and then selects a small, typical unit. A typical unit would closely resemble the average population.
4. In case the measuring process create discomforts for the person being measured, organizations actively look for volunteers.

The above methods also give results but 'non-probability sampling' is not amenable to further statistical treatment as random selection is not involved. Non-random sampling techniques include quota sampling, convenience sampling, judgment sampling, and snowball sampling.

Quota sampling

It appears similar to stratified random sampling but is different. In quota sampling, certain sub-classes, such as age, gender, income group, and education level are used as strata. Stratified random sampling involves selection of units from the stratum in a random manner. In contrast, researchers use non-random methods to select sample from a stratum until quota fixed by the researcher is filled. A quota is proportion of sub-classes in the population. For instance, there is a population of 10,000 persons of which 40 per cent are children below the age of 18 years and rest above the age of 18 years. A sample of 100 persons is to be drawn from the above population. The researcher selects 40 children in a non-random manner and remaining 60 people above the age of 18 years. (Naval Bajpai, 2010: 267).

Convenience sampling

Supposing a researcher selects 100 respondents from a block where he is situated and that too those who are living on ground floor, it amounts to convenience sampling. As it is based on the researcher's convenience, there is no randomness involved in the selection of sample. In the examples given above, a sample of coal which is taken from top of the wagon amounts to convenience sampling.

Judgment sampling

In this case, selection of sampling units is based on the judgment of a researcher and thus amounts to non-random sampling. The example of selecting a small 'typical' unit from a heterogeneous group fits into this group.

Snowball sampling

In this case, respondents are selected on the basis of referrals from other survey respondents.

Use of the normal distribution

The following terminology is normally adopted in sampling theory. The word *estimator* is used to denote the rule by which an estimate of some population

NOTES

characteristic '$\bar{Y}$' is calculated from the sample results. The word estimate is used to the value obtained from a specific sample. An estimator $\bar{y}$ of $\bar{Y}$ given by a sampling plan is called *unbiased* if the mean value of $\bar{y}$ taken over all possible samples provided by the sampling plan is equal to $\bar{Y}$.

Samples in surveys are often large enough so that estimates made from them are approximately normally distributed. In the case of a normal distribution, data follow a bell-shaped curve with the mean at the centre. Normal distribution occurs frequently in statistical theory that is used widely. Because of its special mathematical properties, it forms the basis of many statistical tests.

When probability sampling is used, we have formulas that give the mean and variance of the estimates. Cochran (1977:11) says that suppose we have taken a sample by a procedure known to give an unbiased estimator and have computed the sample estimate $\bar{y}$ and its standard deviation $\sigma\bar{y}$. One can then find out how good is the estimate. We cannot know the exact value of the error of the estimate $(\bar{y} - \bar{Y})$ but from the properties of the normal distribution, the chances are:

0.32 (about 1 in 3) that the absolute error of $(\bar{y} - \bar{Y})$ (exceeds $\sigma\bar{y}$

0.05 (1 in 20) that the absolute error of $(\bar{y} - \bar{Y})$ exceeds $\sigma\bar{y}$

0.01 (1 in 100) that the absolute error of $(\bar{y} - \bar{Y})$ exceeds $\sigma\bar{y}$

Cochran further illustrates the above with the help of an example. If a probability sample of the records of batteries in routine use in a large factory shows an average life of $\bar{y} = 394$ days, with a standard deviation $\sigma\bar{y} = 4.6$ days, the chances are 99 in 100 that the average life in the population of batteries lies between:

$394 - (2.58)\ (4.6) = 382$ days

and

$394 + (2.58)\ (4.6) = 406$ days

The limits 382 days and 406 days are called lower and upper confidence limits. With a single estimate from a single survey, the statement that $\bar{Y}$ lies between 382 and 406 days is not certain to be correct. Cochran points out that the 99 per cent confidence figure implies that if the same samling plan were used many times in a population, a confidence statement being made from each sample, about 99 per cent of these statements would be correct and 1 per cent wrong.

NOTES

Sampling and non-sampling errors

At every stage of research, say, collection, tabulation, analysis and interpretation of data, errors can creep into the process. In statistical theory, these have been studied and classified into sampling and non-sampling errors.

Sampling errors

Sampling errors, as the name suggests, are rooted in the sampling process itself. As inferences about population are made on the basis of a sample, one can easily see that there is scope for error. If the sample that we take happens to be a true representative of the population, the scope for error is drastically reduced. If the sample is not a true representative of the population, it gives rise to sampling errors. On the other hand, in a census or complete enumeration there are no sampling errors as we do not take a sample but do a 100 per cent enumeration. It is important to note at this stage that there are other kinds of errors in 100 per cent enumeration.

There are several contributing factors which give rise to sampling errors. Firstly, they could be on account of faulty selection of sample. In pre-election opinion poll, if the researcher wants to show a particular party in good light, he may deliberately choose pockets that are favourable to it in a judgment sampling. Secondly, if there is fresh trouble in Kupwara district in Jammu & Kashmir and on account of this reason, a researcher tries to substitute it with another district, say Jammu, which can be easily surveyed. This can lead to sampling errors as characteristics of substituted district may not possess same characteristics as the original unit, i.e., Kupwara district. Sampling errors also occur when researchers wrongly demarcate sampling units.

A random sample enables us to use statistical methods to compute and analyse sampling errors.

Non-sampling errors

Their origin is not rooted in the sampling process. Non-sampling errors mainly arise at the stage of observation, ascertainment of responses to questionnaire and processing of data. Thus, both sampling and census data are both not free from non-sampling errors. They can occur at any stage and Bajpai (2010: 268) has outlined some common types of such errors.

(a) Faulty designing and planning of survey

As we have seen in the preceding section, it is extremely important to define clearly the objectives of the survey. Otherwise, there is a danger of straying into uncharted areas while collecting data. Questionnaires are to be prepared in conformity with these objectives and when data specification is inconsistent with the questionnaire, it gives rise to non-sampling errors. Hiring of inexperienced and unqualified staff can also lead to errors during the survey process.

NOTES

(b) Response errors

In a consumption expenditure survey, when the interviewer asks the respondent to recall what he consumed during the last one week, they might not like to disclose their poverty. In the same way, in a survey relating to prevalence of disability, a family might not like to disclose the presence of a person with mental illness on account of social stigma. The response error could be either intentional or unintentional. It could arise due to self-interest or prestige bias of the respondent as well as due to the bias of the interviewer. As a result, it ends up in a wrong response.

(c) Non-response bias

At times, the respondent might not be available at home or even if he or she were available might like to answer certain questions. These questions might be crucial for research objectives. This lack of response contributes to non-sampling errors.

(d) Errors in coverage

It can arise on account of two factors. Sometimes, a few sampling units that should not have been included are included in the sample list or alternatively leaving out some important sampling units which ought to have been there. The objectives of research must be borne in mind.

(e) Compiling error and publication error

After collection of data, the following errors can creep into the system due to mistakes committed by data entry operators. Errors could arise in editing and coding of data, tabulation and summarization of data collected during the survey.

It is important to note that statistical techniques are not available to minimize or control non-sampling errors of the above kind. Bajpai (2010: 269) points out that non-sampling errors can be controlled to some extent by employing qualified, well-trained, and experienced personnel and through careful planning and execution of survey.

Check Your Progress

1. What, according to Cochran, are the advantages of the sampling method?
2. What are the principal steps in a sample survey?
3. Define probability.
4. Which methods does the random sampling method include?
5. List a few non-sampling errors.

NOTES

2.4 TOOLS OF DATA COLLECTION

There are several tools of data collection which include, among others, observation, interviews, questionnaires, focus group and case study method. Each of these methods has been described briefly in the following sections.

2.4.1 Observation

It is a matter of public knowledge that over forty Nobel Laureates issued an appeal for the release of Dr Binayak Sen. Subsequently, there was a request from the European Union to observe his trial in Chhattisgarh. A group of Ambassadors from countries of the European Union visited Raipur in order to observe the hearing of his bail plea in the High Court at Raipur.

As can be seen from the above example, observation is a tool of data collection and is valuable for formal and informal action research. Naturalistic observation is generally found in case studies which generate text. Ethnography involves observation in natural settings for long periods to know about particular cultures and the meaning of those cultures to their members. This type of observation is found in social Anthropology.

On the other hand, structured observation is associated with experiments and data collected therein is amenable to statistical analysis, viz., recording of classroom behaviours by educational psychologists. Gerard Guthrie (2010:109) notes that:

> ...observation usually focuses, first, on behaviour and, then, generates ideas about why certain behaviours occur (for example, why interaction occurs between some people and not others, which then leads to an investigation of the cultural explanations for this). It also allows the opportunity for a validity check about whether people do what they say.

Noting that the researcher's role is highly critical to the success of observation, he outlines the following three major roles:

1. **Participant observation:** Here the researcher takes part in the research situation as a member of the group. He observes and collects information while leading life as a full-fledged member of the group.
2. **Non-participant observation:** It requires the researcher to be present, but not to participate in group actions. Here researcher is not distracted by his own role and can give full attention to data collection.
3. **Hidden observation:** It occurs when the observer is out of sight, for example, behind a one-way glass observing a classroom, or where the role has not been revealed to the group being observed. It, however, raises ethical issues like informed consent in research, etc.

There are reliability and validity issues in observation. When we consider meanings attached to observed behaviour, participants can infer

NOTES

meanings quite differently. It is important to note that what we notice is heavily dependent on cultural factors. It is also possible in ethnography, where participants might change their behaviour because of the researcher's presence or even mislead researchers. Guthrie (2010:110) states that the main ways of improving validity are:

1. Mixed methods: Ethnographic case studies usually use interviews as well as observation.
2. Triangulation: It involves asking participants to comment on draft material. It can greatly add to understanding of the reasons for their reported behaviour.

Reliability issues arise from the fact that other researchers could make different observations. To increase reliability, Guthrie (2010: 110) suggest the following two main steps:

1. Adoption of systematic sampling techniques
2. Careful recording of data

Naturalistic observation works best when observing the whole situation over longer periods to get closer to observing the universe of data. On the contrary, structured observation typically samples time, location, people or events.

Once a determination has been made on what is to be observed, the next step is to decide the method of recording the observations. The following are two main types: *relatively unstructured field notes* and *highly structured observation schedules*. They relate to naturalistic observation and structured observation respectively. Structured observation schedules are usually pre-coded sheets with observational categories determined by the research topic, against which behaviours are tallied as they occur (Guthrie, 2011).

2.4.2 Interview

Interviewing is a common method of data collection in social sciences. This preference for interview technique co-exists alongside increasing resistance amongst a section of the people to subject themselves to surveys, interviews and questionnaires as these place demands on precious time. The value of time-old adage that 'time is precious' requires hardly any overemphasis as any intrusion on this scarce commodity is bound to be resisted by those who are busy. Interviews are common in case studies and surveys and are often resorted to in conjunction with other data collection techniques.

There are three types of interviews:

1. Unstructured interviews
2. Semi-structured interviews
3. Structured interviews

NOTES

It is important to note that they generate different types of data. Each of these types is described in detail hereinunder.

1. **Unstructured interviews**

   Guthrie (2010:119) notes that unstructured interviews generate qualitative data by raising issues in conversational form. They can go in-depth into a topic and are appropriate for obtaining sensitive information. They are also suitable for one-off situations with someone holding a particular viewpoint or with those who can provide factual information. Interviewer needs to establish rapport in the first instance especially in cases where sensitive personal topics are involved. Even unstructured interviews require a general plan as the interviewer has to draw out information, keep the interview on track and should speak minimally but nevertheless give cues from time to time to prompt flow of information. Interview notes can be taken or it could even be recorded.

2. **Semi-structured interviews**

   Guthrie (2010:119) points out that semi-structured interviews use guides so that information from different interviews is directly comparable. These guides provide flexibility to vary the order of intervening questions. There are coded closed-response questions like [did you file IT return? Yes/no]. These can be followed up with open-ended questions to get more information. Thus, semi-structured interviews give rise to both quantitative and qualitative data.

   Interviews are conducted one-on-one, but group interviews are possible too. The most common type is *focus groups*, which is a semi-structured technique derived from marketing and advertising. Here a group of people is gathered together in a suitable location and the interviewer asks questions of the group. Guthrie (2010:119) observes that focus groups can be a highly informative representation of a particular group's viewpoints on a particular occasion, but sometimes opinions can be affected by group dynamics, so validity is an issue. It is possible that a dominant person with a strong personality can be a disruptive influence because others do not get enough time to speak. It is also possible that members sometimes try to impress each other rather than provide considered views that a researcher wants. The researcher must possess skills to moderate group interviews and must use open-ended questions from an interview guide to generate discussion. In both unstructured and structured interviews, a researcher must cross-check viewpoints from different respondents who might have different perceptions.

3. Structured interviews

Structured interviews use formal standardized questionnaires. This ensures reliability as all interviews are conducted the same way using set questions and set response codes. Guthrie (2010:122) notes that questionnaire interviews can be used instead of sending them by post as it will increase response rates and decrease 'don't know' answers, especially with children, or when respondents might not be literate. Trained interview teams can be used to manage large numbers of interviews or when time is limited. Questionnaires are often used to seek opinions or perceptions. They usually contain large number of short questions where answers are coded numerically. Qualitative answers come from open-ended questions (why? Can you give examples? etc.) the responses to which are noted down by the interviewer.

Guthrie (2010:123) observes that structured interviews usually have greater coverage than unstructured ones, but lack their depth. They provide a mixture of qualitative and quantitative data.

Conducting interviews

Interviews are used to collect both opinions and information as words and numbers. Conducting interviews is an art. It requires systematic planning and conduct, laying down clear objectives of the interview based on research problem, pre-testing of interview guide, the use of clear and unambiguous language, ethical concerns like informed consent and reasonable duration of interview which should not be unduly prolonged.

Interviewer is expected to explain clearly the purpose of the interview and also assure privacy and confidentiality to those who are interviewed.

Interviewer bias

Guthrie (2010:126) points out that interviewers are prone to bias on account of the influence of the interviewer and suggests certain practical actions to reduce bias:

1. Dress neutrally and do not talk academically
2. Be friendly, but professional.
3. Keep the introductions similar in all interviews to provide a common frame.
4. Start with straightforward questions and save more difficult questions for later.
5. Avoid leading questions that imply answers or body language that might convey an attitude. Be alert to the tendency for interviewees to say what they think the interviewer wants to hear, especially where the interviewer has higher status.

NOTES

6. Be comfortable with silence. Wait a little if candidate hesitates to reply immediately.
7. Use probe questions to gain more understanding of respondents' views, especially to make sure that you do not misinterpret them in light of your own opinions.
8. Take notes all the time. Respondents might be annoyed if you do not, or start telling you what they think you want to hear because you are suddenly busy writing.'

2.4.3 Questionnaire

When a researcher undertakes a project, a questionnaire probably is the first tool that comes to his mind. A questionnaire consists of a series of questions graded from simple to complex in seeking someone's opinion about something. However, designing it is not only time-consuming, but also complicated. One has to be absolutely clear in one's mind about the purpose of the research project that one has undertaken. Before starting on a questionnaire, one should ask oneself, 'what do I need to know?' and 'how can I frame questions so that this comes through easily?'

Advantages

- Since the data gathered is standardized, it is easy to analyse it
- If adequate number of questionnaires is prepared, data can be gathered in a short period of time
- One can compare results with other such questionnaires of other institutes
- Respondents are more honest as they are anonymous
- Online surveys are inexpensive
- One person can execute the entire process once they have the required skill set

Disadvantages

- In self-completing questionnaires, respondents may not understand the questions and hence respond incorrectly
- Before the sample can represent the population, a reasonable sample size needs to be taken
- People might just not feel motivated enough to fill in questionnaires
- Questionnaires may not be liked by many researchers as they are complex and time-consuming to make

Process of designing and using a questionnaire

The clarification of research aim before preparing a questionnaire is vital. Otherwise, it will be a futile exercise as it may not lead to judicious use of

NOTES

the data collected. In other words, we should know how to use the data that we collect. As preparing a new questionnaire is time-consuming, one can try to use a standardized one in case it serves the purpose. The researcher should be ethical enough to use the data collected only for the purpose mentioned on the questionnaire. If confidentiality is mentioned, he should try to maintain it.

The basic process of using a questionnaire:

1. Define aims
2. Identify the population and the size of sample required
3. Decide how to distribute and collect the survey
4. Design the questionnaire
5. Carry out a pilot survey
6. Carry out the main survey
7. Analyse the data
8. Draw conclusions

As already discussed, the basic aims of the questionnaire should be thoroughly and clearly mentioned in it. The population is the entire group chalked out for the study; however, questionnaires are generally distributed to a sample group from within the population. For example, in case one is researching on teachers' use of electronic media to teach in New Delhi, one needs to take into account a reasonable number of teachers from a reasonable number of schools. The questionnaire should then be circulated to these teachers who represent the sample group. However, the researcher should make sure that the sample group typically marks the characteristics of the entire population. The determination of the sample size is often dictated by terms like: funding of the project; importance of the research; willingness of the people to participate, etc. the response rate for any questionnaire is generally taken at 20 per cent; so in case we need 50 responses, we need to send 250 forms to people. Not only is the number important, but 'type' is equally important too; for example, an online survey may want to receive responses from part-time students, the company may do well to send forms to students from all genders, age-groups, social and economic background, etc. A large sample with this mixed group of respondents picked up randomly should be able to take care of this problem.

Method of collection

Questionnaires may be completed by the respondents or by the interviewers in a one-on-one interview. Questionnaires that have to be completed by respondents may be sent either by post or e-mails. A personal interview may be required when detailed information about certain issues have to be collected. In case the questionnaire is a twenty question form with options along with attached tick boxes, a personal interview is not required. Questionnaires sent through posts should include a paid reply card and a

NOTES

covering letter to explain the entire survey process. The researcher needs to mention the deadline for returning the filled-in questionnaire. The questionnaire should be as concise as possible so that the respondents do not feel burdened to complete it. Formatting the questionnaire by using a smaller font, doing away with a cluttered look, deleting unnecessary headings, etc., may actually bring down the page count. The font size, however, should be decreased so that it may be read properly. The researcher should guarantee absolute confidentiality when sending the forms and actually maintain it.

2.4.4 Focus Groups

A focus group is a sample group that is used to gather information on their attitudes towards and opinions on a certain service, product, advertisement, idea, etc. The setting is essentially of a group discussion that helps the researcher broach the topic to be discussed. He then asks the group questions on the topic discussed. Sociologist and Associate Director of the Bureau of Applied Social Research, USA was the first to use a focus group in their research. The term has been coined by Ernest Dichter, a psychologist and marketing expert.

Types of focus groups

Variants of focus groups include:

- **Two-way focus group:** One focus group observes the interactions of another and discusses the positives and negatives
- **Dual moderator focus group:** Two moderators handle the session; one ensures that the session progresses smoothly while the other ensures that all topics are covered
- **Dueling moderator focus group:** Two moderators deliberately argue on the topic and create their own debating teams
- **Respondent moderator focus group:** A focus group where a respondent is asked to act as a moderator
- **Mini focus groups:** Groups are composed of four/five members
- **Teleconference focus groups:** Focus group where a telephone network is used
- **Online focus groups:** Focus group that interacts via computers connected through the internet

Focus groups have traditionally been used more in marketing researches. However, sometimes, when an important product is launched throughout the country, it becomes difficult to access focus groups in different regions. This would also entail boarding and lodging expenses for researchers from the company throughout the country. There are definitely benefits as well as disadvantages of using focus group discussions in research.

NOTES

Benefits/strengths of focus group discussions

- The data and insights produced in a focus group are difficult to come across in other settings. This is also known as the group effect where group members engage in 'a kind of "chaining" or "cascading" effect; talk links to, or tumbles out of, the topics and expressions preceding it' (Lindlof & Taylor, 2002: 182).
- Focus groups help people who feel isolated and lonely by providing them a voice. For example, in a research on workplace bullies, targeted employees often cannot muster courage to walk up to the human resource department of the organization to protest against the bullies. In a focus group discussion, they have been seen to open up and voice their opinions.
- While discussing similar experiences, people in a focus group often use the same vocabulary. The group forms a 'native language' or 'vernacular speech' to talk about situations they were in.

Disadvantages

- The researcher has less control over the group.
- The issues might get lost in irrelevant conversation.
- Moderators need to be highly trained to be able to strain the required data and information.
- Focus groups cannot be large; this is a vital disadvantage as the researchers need to use this tool continuously in order to achieve the required sample population.
- Since the validity of the moderator's observations is subjective, data gathered after the focus groups have discussed a topic may not be completely valid.
- Heisenberg's 'uncertainty principle' maintains that the participants of the focus groups answer according to the questions asked; the setting of the group; the intelligence level of the participants; the level of the questions framed, etc. Thus, in case a researcher decides that the data gathered from a focus group needs to be used, he has to make sure that the group, settings and the participants chosen are impeccable and apt for the research.
- The other disadvantage of a focus group discussion is that participants are no longer anonymous as in the usage of a questionnaire, etc. Participants may, thus, become unwilling to share their honest opinions with an unknown group that easily.
- The next important factor for the focus group is the appropriate setting. The setting should be as comfortable as possible, so that people do not close up on being seated in a laboratory with a serious-looking scientist