UNIT 4 SAMPLING

STRUCTURE

4.0 OBJECTIVES

After studying this Unit, you should be able to:

- distinguish between census and sampling study,
- explain various reasons for opting for the sample method,
- explain the different methods of sampling and their advantages and disadvantages,
- describe the sampling and non-sampling errors and minimize them, and
- design a representative sample from a population keeping both cost and precision in mind.

4.1 INTRODUCTION

In the previous Unit 3, we have studied the types of data (primary and secondary data) and various methods and techniques of collecting the primary data. The desired data may be collected by selecting either census method or sampling method.

Researchers usually cannot make direct observations of every unit of the population they are studying for a variety of reasons. Instead, they collect data from a subset of population – a sample – and use these observations drawn to make inferences about the entire population. Ideally, the characteristics of a sample should correspond to the characteristics of a population from which the sample was drawn. In that case, the conclusions drawn from a sample are probably applicable to the entire population.

In this Unit, we shall discuss the basics of sampling, particularly how to get a sample that is representative of a population. It covers different methods of drawing samples which can save a lot of time, money and manpower in a

variety of situations. These include random sampling methods, such as, simple random sampling, stratified sampling, systematic sampling, multistage sampling, cluster sampling methods (and non-random sampling methods viz., convenience sampling, judgement sampling and quota sampling. The advantages and disadvantages of sampling and census are covered. How to determine the sample size of a given population is also discussed.

4.2 CENSUS AND SAMPLE

Let us try to understand the terms 'census' and 'sample' with the help of an illustration. Suppose you wish to study the 'impact of T.V. advertisements on children in Delhi, then you have to collect relevant information from the children residing in Delhi who view T.V. Alternatively, we can say this is the population (statistical terminology) for your study. If you collect the data from all of them not leaving a single child, it known as Census method of data collection. This means studying the whole population. Otherwise, if you select only some children from among them for gathering the desired information for the study, because it is not feasible to gather the information from all the children, then it is known as **Sample** for data collection. Therefore, a sample is a subset of a statistical population whose characteristics are studied to know the information about the whole population. When dealing with people, it can be defined as a set of respondents (people) selected from a population for the purpose of a survey. A population is a group of individual persons, objects, items or any other units from which samples are taken for measurement.

The numerical characteristics of a population are called **parameters**. They are fixed and usually of unknown quantity. For example, the average (μ) height of all Indian male adults is a population parameter. The numerical characteristics of the sample data such as the mean, variance or proportion are called **sample statistics**. It can be used to provide estimates of the corresponding population parameters. For example, the average ($\bar{x}$) height of a sample of 1000 Indian male adults residing in Delhi is a sample statistic. The process of selecting a representative sample for the purpose of inferring the characteristics of population is called sampling.

Webster defines a survey as 'the action of ascertaining facts regarding conditions or the condition of something to provide *exact information* especially to persons responsible or interested' and as 'a systematic collection and analysis of data on some aspect of an area or group.' Unless the researcher makes a systematic collection of data followed by careful analysis and interpretation of data, the data cannot become exact information. Surveys can be divided into two categories on the basis of their extensiveness, namely, census and sample survey. A complete survey of population is called a census. It involves covering all respondents, items, or units of the population. For example, if we want to know the wage structure of the textile industry in the country, then one approach is to collect the data on the wages of each and every worker in the textile industry. On the other hand, a sample is a representative subset of population. Thus in a sample survey we cover only a sample of respondents, items or units of population we are interested in and then draw inferences about the whole population.

The following are the advantages of census:

1) In a census each and every respondent of the population is considered and various population parameters are compiled for information.

2) The information obtained on the basis of census data is more reliable and accurate. It is an adopted method of collecting data on exceptional matters like child labour, distribution by sex, educational level of the people etc.

3) If we are conducting a survey for the first time we can have a census instead of sample survey. The information based on this census method becomes a base for future studies. Similarly, some of the studies of special importance like population data are obtained only through census.

4.3 WHY SAMPLING?

One of the decisions to be made by a researcher in conducting a survey is whether to go for a census or a sample survey. We obtain a sample rather than a complete enumeration (a census) of the population for many reasons. The most important considerations for this are: cost, size of the population, accuracy of data, accessibility of population, timeliness, and destructive observations.

1) **Cost:** The cost of conducting surveys through census method would be prohibitive and sampling helps in substantial cost reduction of surveys. Since most often the financial resources available to conduct a survey are scarce, it is imperative to go for a sample survey than census.

2) **Size of the Population:** If the size of the population is very large it is difficult to conduct a census if not impossible. In such situations sample survey is the only way to analyse the characteristics of a population.

3) **Accuracy of Data:** Although reliable information can be obtained through census, sometime the accuracy of information may be lost because of a large population. Sampling involves a small part of the population and a few trained people can be involved to collect accurate data. On the other hand, a lot of people are required to enumerate all the observations. Often it becomes difficult to involve trained manpower in large numbers to collect the data thereby compromising accuracy of data collected. In such a situation a sample may be more accurate than a census. A sloppily conducted census can provide less reliable information than a carefully obtained sample.

4) **Accessibility of Population:** There are some populations that are so difficult to get access to that only a sample can be used, e.g., people in prison, birds migrating from one place to another place etc. The inaccessibility may be economic or time related. In a particular study, population may be so costly to reach, like the population of planets, that only a sample can be used.

5) **Timeliness:** Since we are covering a small portion of a large population through sampling, it is possible to collect the data in far less time than covering the entire population. Not only does it take less time to collect the data through sampling but the data processing and analysis also takes less time because fewer observations need to be covered. Suppose a company wants to get a quick feedback from its consumers on assessing their perceptions about a new improved detergent in comparison to an existing version of the detergent. Here the time factor is very significant. In such situations it is better to go for a sample survey rather than census because it reduces a lot of time and product launch decision can be taken quickly.

6) **Destructive Observations:** Sometimes the very act of observing the desired characteristics of a unit of the population destroys it for the intended

use. Good examples of this occur in quality control. For example, to test the quality of a bulb, to determine whether it is defective, it must be destroyed. To obtain a census of the quality of a lorry load of bulbs, you have to destroy all of them. This is contrary to the purpose served by quality-control testing. In this case, only a sample should be used to assess the quality of the bulbs. Another example is blood test of a patient.

The disadvantages of sampling are few but the researcher must be cautious. These are risk, lack of representativeness and insufficient sample size each of which can cause errors. If researcher don't pay attention to these flaws it may invalidate the results.

1) **Risk:** Using a sample from a population and drawing inferences about the entire population involves risk. In other words the risk results from dealing with a part of a population. If the risk is not acceptable in seeking a solution to a problem then a census must be conducted.
2) **Lack of representativeness:** Determining the representativeness of the sample is the researcher's greatest problem. By definition, 'sample' means a representative part of an entire population. It is necessary to obtain a sample that meets the requirement of representativeness otherwise the sample will be biased. The inferences drawn from nonreprentative samples will be misleading and potentially dangerous.
3) **Insufficient sample size:** The other significant problem in sampling is to determine the size of the sample. The size of the sample for a valid sample depends on several factors such as extent of risk that the researcher is willing to accept and the characteristics of the population itself.

4.4 ESSENTIALS OF A GOOD SAMPLE

It is important that the sampling results must reflect the characteristics of the population. Therefore, while selecting the sample from the population under investigation it should be ensured that the sample has the following characteristics:

1) A sample must represent a true picture of the population from which it is drawn.
2) A sample must be unbiased by the sampling procedure.
3) A sample must be taken at random so that every member of the population of data has an equal chance of selection.
4) A sample must be sufficiently large but as economical as possible.
5) A sample must be accurate and complete. It should not leave any information incomplete and should include all the respondents, units or items included in the sample.
6) Adequate sample size must be taken considering the degree of precision required in the results of inquiry.

Self Assessment Exercise A

1) What do you mean by census and sample methods for data collection?

.....

.....

.....

.....

2) Explain whether census or sample is more appropriate in the following situations?

   a) To test the quality of a soft drink.

   .....

   .....

   b) To enumerate eligible voters of an assembly constituency.

   .....

   .....

   c) To know the opinion of consumers on launching a new product.

   .....

   .....

3) Fill in the blanks

   a) If the sample does not represent the population characteristics, we call it a ———— sample.

   b) One of the major advantages of sampling is that it helps in ——— reduction.

   c) A sample must be ——— large but as ——— as possible.

4.5 METHODS OF SAMPLING

If money, time, trained manpower and other resources were not a concern, the researcher could get most accurate data from surveying the entire population of interest. Since most often the resources are scarce, the researcher is forced to go for sampling. But the real purpose of the survey is to know the characteristics of the population. Then the question is with what level of confidence will the researcher be able to say that the characteristics of a sample represent the entire population. Using a combination of tasks of hypotheses and unbiased sampling methods, the researcher can collect data that actually represents the characteristics of the entire population from which the sample was taken. To ensure a high level of confidence that the sample represents the population it is necessary that the sample is unbiased and sufficiently large.

It was scientifically proved that if we increase the sample size we shall be that much closer to the characteristics of the population. Ultimately, if we cover each and every unit of the population, the characteristics of the sample will be equal to the characteristics of the population. That is why in a census there is no sampling error. Thus, “generally speaking, the larger the sample size, the less sampling error we have.”

The statistical meaning of bias is error. The sample must be error free to make it an unbiased sample. In practice, it is impossible to achieve an error free sample even using unbiased sampling methods. However, we can minimize the error by employing appropriate sampling methods.

The various sampling methods can be classified into two categories. These are random sampling methods and non-random sampling methods. Let us discuss them in detail.

4.5.1 Random Sampling Methods

The random sampling method is also often called probability sampling. In random sampling all units or items in the population have a chance of being chosen in the sample. In other words a random sample is a sample in which each element of the population has a known and non-zero chance of being selected. Random sampling always produces the smallest possible sampling error. In the real sense, the size of the sampling error in a random sample is affected only by a random chance. Because a random sample contains the least amount of sampling error, we may say that it is an unbiased sample. Remember that we are not saying that a random sample contains no error, but rather the minimum possible amount of error. The major advantage of random sampling is that it is possible to quantify the magnitude of the likely error in the inference made and this will help in building confidence in drawing inferences.

The following are the important methods of random sampling:

1) Simple Random Sampling
2) Systematic Sampling
3) Stratified Random Sampling
4) Cluster Sampling
5) Multistage Sampling

1. Simple Random Sampling: The most commonly used random sampling method is simple random sampling method. A simple random sample is one in which each item in the total population has an equal chance of being included in the sample. In addition, the selection of one item for inclusion in the sample should in no way influence the selection of another item. Simple random sampling should be used with a homogeneous population, that is, a population consisting of items that possess the same attributes that the researcher is interested in. The characteristics of homogeneity may include such as age, sex, income, social/religious/political affiliation, geographical region etc.

The best way to choose a simple random sample is to use random number table. A random sampling method should meet the following criteria.

a) Every member of the population must have an equal chance of inclusion in the sample.
b) The selection of one member is not affected by the selection of previous members.

The random numbers are a collection of digits generated through a probabilistic mechanism. The random numbers have the following properties:

i) The probability that each digit (0,1,2,3,4,5,6,7,8,or 9) will appear at any place is the same. That is 1/10.
ii) The occurrence of any two digits in any two places is independent of each other.

Each member of a population is assigned a unique number. The members of the population chosen for the sample will be those whose numbers are identical to the ones extracted from the random number table in succession until the desired sample size is reached. An example of a random number table is given below.

Table 1: Table of Random Numbers

	1	2	3	4	5	6	7	8	9	10
1	96268	11860	83699	38631	90045	69696	48572	05917	51905	10052
2	03550	59144	59468	37984	77892	89766	86489	46619	50236	91136
3	22188	81205	99699	84260	19693	36701	43233	62719	53117	71153
4	63759	61429	14043	44095	84746	22018	19014	76781	61086	90216
5	55006	17765	15013	77707	54317	48862	53823	52905	70754	68212
6	81972	45644	12600	01951	72166	52682	37598	11955	73018	23528
7	06344	50136	33122	31794	86723	58037	36065	32190	31367	96007
8	92363	99784	94169	03652	80824	33407	40837	97749	18361	72666
9	96083	16943	89916	55159	62184	86206	09764	20244	88388	98675
10	92993	10747	08985	44999	35785	65036	05933	77378	92339	96151
11	95083	70292	50394	61947	65591	09774	16216	63561	59751	78771
12	77308	60721	96057	86031	83148	34970	30892	53489	44999	18021
13	11913	49624	28519	27311	61586	28576	43092	69971	44220	80410
14	70648	47484	05095	92335	55299	27161	64486	71307	85883	69610
15	92771	99203	37786	81142	44271	36433	31726	74879	89384	76886
16	78816	20975	13043	55921	82774	62745	48338	88348	61211	88074
17	79934	35392	56097	87613	94627	63622	08110	16611	88599	02890
18	64698	83376	87527	36897	17215	74339	69856	43622	22567	11518
19	44212	12995	03581	37618	94851	63020	65348	55857	91742	79508
20	89292	00204	00579	70630	37136	50922	83387	15014	51838	81760
21	08692	87237	87879	01629	72184	33853	95144	67943	19345	03469
22	67927	76855	50702	78555	97442	78809	40575	79714	06201	34576
23	62167	94213	52971	85794	68067	78814	40103	70759	92129	46716
24	45828	45441	74220	84157	23241	49332	23646	09390	13031	51569
25	01164	35307	26526	80335	58090	85871	07205	31749	40571	51755
26	29283	31581	04359	45538	41435	61103	32428	94042	39971	63678
27	19868	49978	81699	84904	50163	22652	07845	71308	00859	87984
28	14292	93587	55960	23159	07370	65065	06580	46285	07884	83928
29	77410	52135	29495	23032	83242	89938	40516	27252	55565	64714
30	36580	06921	35675	81645	60479	71035	99380	59759	42161	93440
31	07780	18093	31258	78156	07871	20369	53977	08534	39433	57216
32	07548	08454	36674	46255	80541	42903	37366	21164	97516	66181
33	22023	60448	69344	44260	90570	01632	21002	24413	04671	05665
34	20827	37210	57797	34660	32510	71558	78228	42304	77197	79168
35	47802	79270	48805	59480	88092	11441	96016	76091	51823	94442
36	76730	86591	18978	25479	77684	88439	34112	26052	57112	91653
37	26439	02903	20935	76297	15290	84688	74002	09467	41111	19194
38	32927	83426	07848	59372	44422	53372	27823	25417	27150	21750
39	51484	05286	77103	47284	00578	88774	15293	50740	07932	87633
40	45142	96804	92834	26886	70002	96643	36008	02239	93563	66429

To select a random sample using simple random sampling method we should follow the steps given below:

i) Determine the population size (N).

ii) Determine the sample size (n).

iii) Number each member of the population under investigation in serial order. Suppose there are 100 members number them from 00 to 99.

iv) Determine the starting point of selecting sample by randomly picking up a page from random number tables and dropping your finger on the page blindly.

v) Choose the direction in which you want to read the numbers (from left to right, or right to left, or down or up).

vi) Select the first 'n' numbers whose X digits are between 0 and N. If N = 100 then X would be 2, if N is a four digit number then X would be 3 and so on.

vii) Once a number is chosen, do not use it again.

viii) If you reach the end point of the table before obtaining 'n' numbers, pick another starting point and read in a different direction and then use the first X digit instead of the last X digits and continue until the desired sample is selected.

Example: Suppose you have a list of 80 students and want to select a sample of 20 students using simple random sampling method. First assign each student a number from 00 to 79. To draw a sample of 20 students using random number table, you need to find 20 two-digit numbers in the range 00 to 79. You can begin any where and go in any direction. For example, start from the 6th row and 1st column of the random number table given in this Unit. Read the last two digits of the numbers. If the number is within the range (00 to 79) include the number in the sample. Otherwise skip the number and read the next number in some identified direction. If a number is already selected omit it. In the example starting from 6th row and 1st column and moving from left to right direction the following numbers are considered to selected 20 numbers for sample.

819**72**	456**44**	126**00**	019**51**	721**66**	52682	37598	119**55**	730**18**	235**28**
06344	501**36**	33122	31794	867**23**	580**37**	360**65**	32190	313**67**	960**07**
923**63**	99784	941**69**	036**52**	808**24**	33407	40837	977**49**	18361	72666

The bold faced digits in the one's and ten's place value indicate the selected numbers for the sample. Therefore, the following are the 20 numbers chosen as sample.

72	44	00	51	66	55	18	28
36	22	23	37	65	67	07	63
69	52	24	49				

Advantages

i) The simple random sample requires less knowledge about the characteristics of the population.

ii) Since sample is selected at random giving each member of the population equal chance of being selected the sample can be called as unbiased sample. Bias due to human preferences and influences is eliminated.

iii) Assessment of the accuracy of the results is possible by sample error estimation.

iv) It is a simple and practical sampling method provided population size is not large.

Limitations

i) If the population size is large, a great deal of time must be spent listing and numbering the members of the population.

ii) A simple random sample will not adequately represent many population characteristics unless the sample is very large. That is, if the researcher is interested in choosing a sample on the basis of the distribution in the population of gender, age, social status, a simple random sample needs to be very large to ensure all these distributions are representative of the population. To obtain a representative sample across multiple population attributes we should use stratified random sampling.

2. Systematic Sampling: In systematic sampling the sample units are selected from the population at equal intervals in terms of time, space or order. The selection of a sample using systematic sampling method is very simple. From a population of 'N' units, a sample of 'n' units may be selected by following the steps given below:

   i) Arrange all the units in the population in an order by giving serial numbers from 1 to N.

   ii) Determine the sampling interval by dividing the population by the sample size. That is, K=N/n.

   iii) Select the first sample unit at random from the first sampling interval (1 to K).

   iv) Select the subsequent sample units at equal regular intervals.

For example, we want to have a sample of 100 units from a population of 1000 units. First arrange the population units in some serial order by giving numbers from 1 to 1000. The sample interval size is K=1000/100=10. Select the first sample unit at random from the first 10 units (i.e. from 1 to 10). Suppose the first sample unit selected is 5, then the subsequent sample units are 15, 25, 35,.....995. Thus, in the systematic sampling the first sample unit is selected at random and this sample unit in turn determines the subsequent sample units that are to be selected.

Advantages

i) The main advantage of using systematic sample is that it is more expeditious to collect a sample systematically since the time taken and work involved is less than in simple random sampling. For example, it is frequently used in exit polls and store consumers.

ii) This method can be used even when no formal list of the population units is available. For example, suppose if we are interested in knowing the opinion of consumers on improving the services offered by a store we may simply choose

every k^{th} (say 6^{th}) consumer visiting a store provided that we know how many consumers are visiting the store daily (say 1000 consumers visit and we want to have 100 consumers as sample size).

Limitations

i) If there is periodicity in the occurrence of elements of a population, the selection of sample using systematic sample could give a highly un-representative sample. For example, suppose the sales of a consumer store are arranged chronologically and using systematic sampling we select sample for 1^{st} of every month. The 1^{st} day of a month can not be a representative sample for the whole month. Thus in systematic sampling there is a danger of order bias.

ii) Every unit of the population does not have an equal chance of being selected and the selection of units for the sample depends on the initial unit selection. Regardless how we select the first unit of sample, subsequent units are automatically determined lacking complete randomness.

3. Stratified Random Sampling: The stratified sampling method is used when the population is heterogeneous rather than homogeneous. A heterogeneous population is composed of unlike elements such as male/female, rural/urban, literate/illiterate, high income/low income groups, etc. In such cases, use of simple random sampling may not always provide a representative sample of the population. In stratified sampling, we divide the population into relatively homogenous groups called strata. Then we select a sample using simple random sampling from each stratum. There are two approaches to decide the sample size from each stratum, namely, proportional stratified sample and disproportional stratified sample. With either approach, the stratified sampling guarantees that every unit in the population has a chance of being selected. We will now discuss these two approaches of selecting samples.

i) ***Proportional Stratified Sample:*** If the number of sampling units drawn from each stratum is in proportion to the corresponding stratum population size, we say the sample is proportional stratified sample. For example, let us say we want to draw a stratified random sample from a heterogeneous population (on some characteristics) consisting of rural/urban and male/female respondents. So we have to create 4 homogeneous sub groups called stratums as follows:

Urban		Rural	
Male	Female	Male	Female

To ensure each stratum in the sample will represent the corresponding stratum in the population we must ensure each stratum in the sample is represented in the same proportion to the stratums as they are in the population. Let us assume that we know (or can estimate) the population distribution as follows: 65% male, 35% female and 30% urban and 70% rural. Now we can determine the approximate proportions of our 4 stratums in the population as shown below.

Urban		Rural	
Male	Female	Male	Female
0.30 × 0.65 = 0.195	0.30 × 0.35 = 0.105	0.70 × 0.65 = 0.455	0.70 × 0.35 = 0.245

Thus a representative sample would be composed of 19.5% urban-males, 10.5% urban-females, 45.5% rural-males and 24.5% rural females. Each percentage should be multiplied by the total sample size needed to arrive at the actual

sample size required from each stratum. Suppose we require 1000 samples then the required sample in each stratum is as follows:

Urban-male	0.195 × 1000 = 195
Urban-female	0.105 × 1000 = 105
Rural-male	0.455 × 1000 = 455
Rural-female	0.245 × 1000 = 245
Total:	1,000

ii) ***Disproportional Stratified Sample:*** In a disproportional stratified sample, sample size for each stratum is not allocated on a proportional basis with the population size, but by analytical considerations of the researcher such as stratum variance, stratum population, time and financial constraints etc. For example, if the researcher is interested in finding differences among different stratums, disproportional sampling should be used. Consider the example of income distribution of households. There is a small percentage of households within the high income brackets and a large percentage of households within the low income brackets. The income among higher income group households has higher variance than the variance among the lower income group households. To avoid under-representation of higher income groups in the sample, a disproportional sample is taken. This indicates that as the variability within the stratum increases sample size must increase to provide accurate estimates and vice-versa.

Suppose in our example of urban/rural and male/female stratum populations, the stratum estimated variances (s^2) are as follows. However, the variance is discussed in Unit 9 of this course.

Urban-male 3.0; Urban-female 5.5; Rural-males 2.5; Rural-females 1.75.

The above figures are, normally, estimated on the basis of previous knowledge of a researcher.

Then the allocation of sample size of 1000 for each strata using disproportional stratified sampling method will be as shown in the following table:

Stratum	Stratum population proportion (P_i)	Stratum variance (σ_i^2)	Stratum standard deviation (σ_i)	$P_i \times \sigma_i$	Sample size $(P_i \times \sigma_i \times 1000)/ \Sigma P_i \sigma_i$)
Urban-male	0.195	3.0	1.73	0.338	207
Urban-female	0.105	5.5	2.35	0.246	151
Rural-male	0.455	2.5	1.58	0.719	442
Rural-female	0.245	1.75	1.32	0.324	199
			Total	1.628	1000

Advantages

a) Since the sample are drawn from each of the stratums of the population, stratified sampling is more representative and thus more accurately reflects characteristics of the population from which they are chosen.

b) It is more precise and to a great extent avoids bias.

c) Since sample size can be less in this method, it saves a lot of time, money and other resources for data collection.

Limitations

a) Stratified sampling requires a detailed knowledge of the distribution of attributes or characteristics of interest in the population to determine the homogeneous groups that lie within it. If we cannot accurately identify the homogeneous groups, it is better to use simple random sample since improper stratification can lead to serious errors.

b) Preparing a stratified list is a difficult task as the lists may not be readily available.

4. **Cluster Sampling:** In cluster sampling we divide the population into groups having heterogenous characteristics called clusters and then select a sample of clusters using simple random sampling. We assume that each of the clusters is representative of the population as a whole. This sampling is widely used for geographical studies of many issues. For example if we are interested in finding the consumers' (residing in Delhi) attitudes towards a new product of a company, the whole city of Delhi can be divided into 20 blocks. We assume that each of these blocks will represent the attitudes of consumers of Delhi as a whole, we might use cluster sampling treating each block as a cluster. We will then select a sample of 2 or 3 clusters and obtain the information from consumers covering all of them. The principles that are basic to the cluster sampling are as follows:

   i) The differences or variability within a cluster should be as large as possible. As far as possible the variability within each cluster should be the same as that of the population.

   ii) The variability between clusters should be as small as possible. Once the clusters are selected, all the units in the selected clusters are covered for obtaining data.

Advantages

a) The cluster sampling provides significant gains in data collection costs, since traveling costs are smaller.

b) Since the researcher need not cover all the clusters and only a sample of clusters are covered, it becomes a more practical method which facilitates fieldwork.

Limitations

a) The cluster sampling method is less precise than sampling of units from the whole population since the latter is expected to provide a better cross-section of the population than the former, due to the usual tendency of units in a cluster to be homogeneous.

b) The sampling efficiency of cluster sampling is likely to decrease with the decrease in cluster size or increase in number of clusters.

The above advantages or limitations of cluster sampling suggest that, in practical situations where sampling efficiency is less important but the cost is of greater significance, the cluster sampling method is extensively used. If the division of clusters is based on the geographic sub-divisions, it is known as area sampling. In cluster sampling instead of covering all the units in each cluster we can resort to sub-sampling as two-stage sampling. Here, the clusters are termed as primary units and the units within the selected clusters are taken as secondary units.

5. **Multistage Sampling:** We have already covered two stage sampling. Multi stage sampling is a generalisation of two stage sampling. As the name suggests, multi stage sampling is carried out in different stages. In each stage progressively smaller (population) geographic areas will be randomly selected.

   A political pollster interested in assembly elections in Uttar Pradesh may first divide the state into different assembly units and a sample of assembly constituencies may be selected in the first stage. In the second stage, each of the sampled assembly constituents are divided into a number of segments and a second stage sampled assembly segments may be selected. In the third stage within each sampled assembly segment either all the house-holds or a sample random of households would be interviewed. In this sampling method, it is possible to take as many stages as are necessary to achieve a representative sample. Each stage results in a reduction of sample size.

   In a multi stage sampling at each stage of sampling a suitable method of sampling is used. More number of stages are used to arrive at a sample of desired sampling units.

Advantages

a) Multistage sampling provides cost gains by reducing the data collection on costs.

b) Multistage sampling is more flexible and allows us to use different sampling procedures in different stages of sampling.

c) If the population is spread over a very wide geographical area, multistage sampling is the only sampling method available in a number of practical situations.

Limitations

a) If the sampling units selected at different stages are not representative multistage sampling becomes less precise and efficient.

4.5.2 Non-Random Sampling Methods

The non-random sampling methods are also often called non-probability sampling methods. In a non-random sampling method the probability of any particular unit of the population being chosen is unknown. Here the method of selection of sampling units is quite arbitrary as the researchers rely heavily on personal judgment. Non-random sampling methods usually do not produce samples that are representative of the general population from which they are drawn. The greatest error occurs when the researcher attempts to generalise the results on the basis of a sample to the entire population. Such an error is insidious because it is not at all obvious from merely looking at the data, or even from looking at the sample. The easiest way to recognise whether a sample is representative or not is to determine whether the sample is selected randomly or not. Nevertheless, there are occasions where non-random samples are best suited for the researcher's purpose.The various non-random sampling methods commonly used are:

1) Convenience Sampling;

2) Judgement Sampling; and

3) Quota Sampling.

Let us discuss these methods in detail.

1) **Convenience Sampling:** Convenience sampling refers to the method of obtaining a sample that is most conveniently available to the researcher. For example, if we are interested in finding the overtime wage paid to employees working in call centres, it may be convenient and economical to sample

employees of call centres in a nearby area. Also, on various issues of public interest like budget, election, price rise etc., the television channels often present on-the-street interviews with people to reflect public opinion. It may be cautioned that the generalisation of results based on convenience sampling beyond that particular sample may not be appropriate. Convenience samples are best used for exploratory research when additional research will be subsequently conducted with a random sample. Convenience sampling is also useful in testing the questionnaires designed on a pilot basis. Convenience sampling is extensively used in marketing studies.

2) **Judgement Sampling:** Judgement sampling method is also known as purposive sampling. In this method of sampling the selection of sample is based on the researcher's judgment about some appropriate characteristic required of the sample units. For example, the calculation of consumer price index is based on a judgment sample of a basket of consumer items, and other related commodities and services which are expected to reflect a representative sample of items consumed by the people. The prices of these items are collected from selected cities which are viewed as typical cities with demographic profiles matching the national profile. In business judgment sampling is often used to measure the performance of salesmen/saleswomen. The salesmen/saleswomen are grouped into high, medium or low performers based on certain specified qualities. Then the sales manager may actually classify the salesmen/saleswomen working under him/her who in his/her opinion will fall in which group. Judgment sampling is also often used in forecasting election results. We may often wonder how a pollster can predict an election based on only 2% to 3% of votes covered. It is needless to say the method is biased and does not have any scientific basis. However, in the absence of any representative data, one may resort to this kind of non-random sampling.

3) **Quota Sampling:** The quota sampling method is commonly used in marketing research studies. The samples are selected on the basis of some parameters such as age, sex, geographical region, education, income, occupation etc, in order to make them as representative samples. The investigators, then, are assigned fixed quotas of the sample meeting these population characteristics. The purpose of quota sampling is to ensure that various sub-groups of the population are represented on pertinent sample characteristics to the extent that the investigator desires. The stratified random sampling also has this objective but should not be confused with quota sampling. In the stratified sampling method the researcher selects a random sample from each group of the population, where as, in quota sampling, the interviewer has a quota fixed for him/her to achieve. For example, if a city has 10 market centres, a soft drink company may decide to interview 50 consumers from each of these 10 market centres to elicit information on their products. It is entirely left to the investigator whom he/she will interview at each of the market centres and the time of interview. The interview may take place in the morning, mid day, or evening or it may be in the winter or summer.

   Quota sampling has the advantage that the sample confirms the selected characteristics of the population that the researcher desires. Also, the cost and time involved in collecting the data are also greatly reduced. However, quota sampling has many limitations, as given below:

a) In quota sampling the respondents are selected according to the convenience of the field investigator rather than on a random basis. This kind of selection of sample may be biased. Suppose in our example of soft drinks, after the sample is taken it was found that most of the respondents belong to the lower income group then the purpose of conducting the survey becomes useless and the results may not reflect the actual situation.

b) If the number of parameters, on which basis the quotas are fixed, are larger then it becomes difficult for the researcher to fix the quota for each sub-group.

c) The field workers have the tendency to cover the quota by going to those places where the respondents may be willing to provide information and avoid those with unwilling respondents. For example, the investigators may avoid places where high income group respondents stay and cover only low income group areas.

Self Assessment Exercise B

1) Suppose there are 900 families residing in a colony. You are asked to select a sample of families using simple random sampling for knowing the average income. The families are identified with serial numbers 001 to 900.

   i) Select a random sample using the following random table.

29283	31581	04359	45538	41435	61103	32428	94042	39971	63678
19868	49978	81699	84904	50163	22652	07845	71308	00859	87984
14292	93587	55960	23159	07370	65065	06580	46285	07884	83928
77410	52135	29495	23032	83242	89938	40516	27252	55565	64714
36580	06921	35675	81645	60479	71035	99380	59759	42161	93440

   ii) While selecting the random sample in the above example, what are the random numbers you have rejected and why?

   .....

   .....

   .....

   .....

   .....

   .....

2) There are 4 (A,B,C, and D) sections in class X of a secondary school. You are asked to find the average income of the parents of the students of section A and C. Which sampling method will be used from the following?

   a) Simple random sampling; b) Systematic sampling; c) Stratified sampling; d) Cluster sampling.

   .....

3) The employees of a company are classified into 4 groups (A,B,C and D) on the basis of their salary structure. You are asked to find the average salary income of the employees working in the company. What is the sampling method to be used?

   a) Simple random sampling; b) Systematic sampling; c) Stratified sampling; d) Quota sampling.

   .....

4) State true or false.

   a) A systematic sampling can be used even if all the units of the population are not available.

   b) A budget has been announced by the government. A TV journalist recorded the views of the people residing near his house. The sampling method that the TV journalist used is quota sampling.

.....

4.6 SAMPLE SIZE

The question of how large a sample should be is a difficult one. Sample size can be determined by various factors (like time, funds, manpower, population size, purpose of study etc. For example, if the available funds for study are limited then the researcher may not be able to spend more than a fixed proportion of the total fund available with him/her. In general, sample size depends on the nature of the analysis to be performed, the desired precision of the estimates one wishes to achieve, number of variables that have to be examined simultaneously and how heterogeneous is the population spread. Moreover, technical considerations suggest that the required sample size is a function of the precision of the estimates one wishes to achieve, the variance of the population and statistical level of confidence one wishes to use. The higher the precision and confidence level required, the larger the sample size should be. Typical confidence levels are 95% and 99%, while a typical precision (significance) value is 1% or 5%. You will learn more about the confidence and precision levels in Unit 16 and Unit 17 of this course.

Once the researcher determines the desired degree of precision and confidence level, there are several formulas he/she can use to determine the sample size and interpretation of results depending on the plan of the study. Here we will discuss three of them.

1) If the researcher wishes to report the results as proportions of the sample responses, use the following formula.

$$n = \frac{P(1-P)}{\frac{A^2}{Z^2} + \frac{P(1-P)}{N}}$$

   Where, n = Sample size.

   P = Estimated percentage of the population possessing attribute of interest.

   A = Accuracy desired, usually expressed as a decimal (i.e. 0.01, 0.05, etc.)

   Z = Standardization value indicating a confidence level (Z=1.96 at 95% confidence level and Z = 2.56 at 99% confidence level. See Unit 16 for more details.

   N = Population size (known or estimated)

2) If the researcher wishes to report the results as means of the sample responses, use the following formula.

$$n = \frac{\sigma^2}{\frac{A^2}{Z^2} + \frac{\sigma^2}{N}}$$

Where, n = Sample size.

P = Estimated percentage of the population possessing attribute of interest.

A = Accuracy desired, usually expressed as a decimal (i.e. 0.01, 0.05, etc.)

Z = Standardization value indicating a confidence level (Z=1.96 at 95% confidence level and Z = 2.56 at 99% confidence level. See Unit 16 for more details.

N = Population size (known or estimated)

3) If the researcher plans the results in a variety of ways or if he/she has difficulty in estimating the proportion or standard deviation of the attribute of interest, the following formula may be more useful.

$$n = \frac{NZ^2 \times .25}{[d^2 \times (N-1)] + [Z^2 \times .25]}$$

Where, n = Sample size required

d = Accuracy precision level (i.e. 0.01, 0.05, 0.10 etc.)

Z = Standardization value indicating a confidence level (Z = 1.96 at 95% confidence level and Z = 2.56 at 99% confidence level. See Unit 16 for more details.

N = Population size (known or estimated).

For example, if the population size (N) is 1000 and you wish a 95% confidence level and ±5% precision level (d=0.05 and Z=1.96) then the sample size (n):

$$n = \frac{1000 \times 1.96^2 \times 0.25}{(0.05^2 \times 999) + (1.96^2 \times 0.25)} = 277.7 \text{ or say } 280$$

4.7 SAMPLING AND NON-SAMPLING ERRORS

The quality of a research project depends on the accuracy of the data collected and its representation to the population. There are two broad sources of errors. These are sampling errors and non-sampling errors.

4.7.1 Sampling Errors

The principal sources of sampling errors are the sampling method applied, and the sample size. This is due to the fact that only a part of the population is covered in the sample. The magnitude of the sampling error varies from one sampling method to the other, even for the same sample size. For example, the sampling error associated with simple random sampling will be greater than stratified random sampling if the population is heterogeneous in nature.

Intuitively, we know that the larger the sample the more accurate the research. In fact, the sampling error varies with samples of different sizes. Increasing the sample size decreases the sampling error.

The following Figure gives an approximate relationship between sample size and sampling error. Study the following figure carefully.

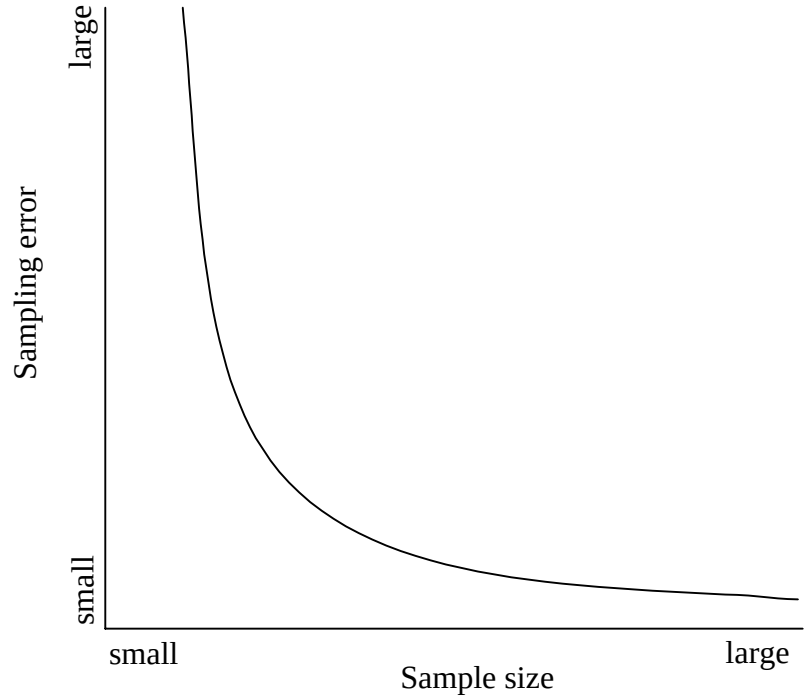

Fig.: 4.1

4.7.2 Non-Sampling Errors

The non-sampling errors arise from faulty research design and mistakes in executing research. There are many sources of non-sampling errors which may be broadly classified as: (a) respondent errors, and (b) administrative errors.

a) **Respondent Errors**: If the respondents co-operate and give the correct information the objectives of the researcher can be easily accomplished. However, in practice, this may not happen. The respondents may either refuse to provide information or even if he/she provides information it may be biased.

   If the respondent fails to provide information, we call it as ***non-response error***. Although this problem is present in all types of surveys, the problem is more acute in mailed surveys. Non-response also leads to some extreme situations like those respondents who are willing to provide information are over represented while those who are indifferent are under-represented in the sample. In order to minimise the non-response error the researcher often seeks to re-contact with the non-respondents if they were not available earlier.

   If the researcher finds that the non-response rate is more in a particular group of respondents (for example, higher income groups) additional efforts should be made to obtain data from these under-represented groups of the population. For example, for these people who are not responding to the mailed questionnaires, personal interviews may be conducted to obtain data. In a mailed questionnaire the researcher never knows whether the respondent really refused to provide data or was simply indifferent. There are several techniques which help to encourage respondents to reply. You must have already learned these techniques in Unit 3 of this course.

   Response bias occurs when the respondent may not give the correct information and try to mislead the investigator in a certain direction. The

respondents may consciously or unconsciously misrepresent the truth. For example, if the investigator asks a question on the income of the respondent he/she may not give the correct information for obvious reasons. Or the investigator may not be able to put a question that is sensitive (thus avoiding embarrassment). This may arise from the problems in designing the questionaire and the content of questions. Respondents who must understand the questions may unconsciously provide biased information.

The response bias may also occur because the interviewer's presence influences respondents to give untrue or modified answers. The respondents/interviewers tendency is to please the other person rather than provide/elicit the correct information.

b) Administrative Errors: The errors that have arisen due to improper administration of the research process are called administrative errors. There are four types of administrative errors. These are as follows:

i) sample selection error,

ii) investigator error,

iii) investigator cheating, and

iv. data processing error.

i) **Sample Selection Error:** It is difficult to execute a sampling plan. For example, we may plan to use systematic sampling plan in a market research study of a new product and decide to interview every 5^{th} customer coming out of a consumer store. If the day of interview happened to be a working day then we are excluding all those consumers who are working. This may lead to an error because of the unrepresentative sample selection.

ii) **Investigator Error:** When the investigator interviews the respondent, he/she may fail to record the information correctly or may fail to cross check the information provided by the respondent. Therefore, the error may arise due to the way the investigator records the information.

iii) **Investigator Cheating:** Some times the investigator may try to fake the data even without meeting the concerned respondents. There should be some mechanism to crosscheck this type of faking by the investigator.

iv) **Data Processing Error:** Once the data is collected the next job the researcher does is edit, code and enter the data into a computer for further processing and analysis. The errors can be minimised by careful editing, coding and entering the data into a computer.

4.7.3 Control of Errors

In the above two sections we have identified the most significant sources of errors. It is not possible to eliminate completely the sources of errors. However, the researcher's objective and effort should be to minimise these sources of errors as much as possible. There are ways of reducing the errors. Some of these are:
(a) designing and executing a good questionnaire; (b) selection of appropriate sampling method; (c) adequate sample size; (d) employing trained investigators to collect the data; and (e) care in editing, coding and entering the data into the computer. You have already learned the above ways of controlling the errors in Unit 3 and in this Unit.

Self Assessment Exercise C

1) The size of a population is 10000. You wish to have a 99% confidence level and ±5% precision level. What is the sample size required?

   .....

   .....

   .....

2) As the sample size increases, the sampling error:

   a) Increases b) Decreases c) Remains constant

   .....

3) The sampling errors arise due to:

   a) The investigator's bias b) The data processing problem

   c) The respondent's bias d) The sampling method applied

   .....

4.8 LET US SUM UP

A sample is a subset of population whose characteristics are studied to know the information about the population. A complete survey of population is called census. When compared with census, sampling is less expensive, requires less time and other resources and is more accurate when samples are taken properly. Also, sampling is the only alternative when the measurement of population units is destructive in nature.

There are two broad categories of sampling methods. These are: (a) random sampling methods, and (b) non-random sampling methods. The random sampling methods are based on the chance of including the units of population in a sample.

Some of the sampling methods covered in this Unit are: (a) simple random sampling, (b) systematic random sampling, (c) stratified random sampling, (d) cluster sampling, and (e) multistage sampling. With an appropriate sampling plan and selection of random sampling method the sampling error can be minimised. The non-random sampling methods include: (a) convenience sampling, (b) judgment sampling, and (c) Quota sampling. These methods may be convenient to the researcher to apply. These methods may not provide a representative sample to the population and there are no scientific ways to check the sampling errors.

There are two major sources of errors in survey research. These are: (a) sampling errors, and (b) non-sampling errors. The sampling errors arise because of the fact that the sample may not be a representative sample to the population. Two major sources of non-sampling errors are due to: (a) non-response on the part of respondent and/or respondent's bias in providing correct information, and (b) administrative errors like design and implementation of questionnaire, investigators' bias, and data processing errors.

It may not be possible to completely eliminate the sampling and non-sampling errors. However, there are some ways to minimise these errors. These are:

(a) designing a good questionnaire, (b) selection of appropriate sampling method, (c) adequate sample size, (d) employing trained investigators and, (e) care in data processing.

4.9 KEY WORDS

Administrative Errors : The administrative errors arise due to improper administration of the research.

Census : A complete survey of population is called census.

Convenient Sampling : Here the units of the population are included in the sample as per the convenience of the researcher.

Cluster Sampling: In cluster sampling method we divide the population into groups called clusters, selective sample of clusters using simple random sampling and then cover all the units in each of the clusters included in the sample.

Judgment Sampling: In this sampling method the selection of sample is based on the researcher's judgment about some appropriate characteristics required of the sample units.

Multi-stage Sampling: Here we select the sample units in a number of stages using one or more random sampling methods.

Non-sampling Errors : The non-sampling errors arise from faulty research design and mistakes in executing the research.

Non-random Sampling/Non-Probability Sampling : In this sampling method the probability of any particular unit of the population being included in the sample is unknown.

Parameters : The numerical characteristics of a population are called parameters.

Quota Sampling : In this sampling method the samples are selected on the basis of some parameters such as age, gender, geographical region, education, income, occupation etc.

Random Sampling/Probability Sampling : If all the units of the population have a chance of being chosen in the sample, the sampling method is called random sampling/probability sampling.

Respondent Errors : The respondent errors arise due to failure of the respondent to provide correct information.

Sample : A sample is a representative set of population.

Sampling Errors : The sampling errors arise because we cover only a part of the population.

Simple Random Sampling : This is one of the basic methods of random sampling where each unit in the population has equal chance of being included in the sample.

Stratified Sampling : The stratified sampling method is used when the population is heterogeneous. Here the population is divided into some homogeneous groups called stratums.

Systematic Sampling : In systematic sampling the sample units are selected from the population at equal intervals in terms of time, space or order.

4.10 ANSWERS TO SELF ASSESSMENT EXERCISES

A. 2) a) sample survey; b) census; c) sample survey.

3) a) biased; b) cost; c) sufficiently, economical.

B. 1) i) Selected sample using simple random sampling

283, 581, 359, 538, 435, 103, 428, 042, 678, 868,
699, 163, 652, 845, 308, 859, 292, 587, 960, 159,
370, 065, 580, 285, 884, 410, 135, 495, 032, 242

ii) 39**971**, 49**978**, 84**904**, 87**984**, 55**960**, 83**928**

The population size is 900 and these random numbers fall outside the population range of 000 to 899.

2) Cluster sampling

3) Stratified sampling

4) a) true

b) false, it is convenience sampling

C. 1) The required sample size is 370

2) Decreases

3) Sampling method applied

4.11 TERMINAL QUESTIONS

1) What is the difference between random sampling and non-random sampling?
2) List some of the situations where (a) sampling is more appropriate than census and (b) census is more appropriate than sampling.
3) What are the advantages and disadvantages of stratified random sampling?
4) What are the ways to control survey errors?
5) What are the advantages of sampling over census?
6) Discuss the method of cluster sampling. What is the difference between cluster sampling and stratified random sampling?
7) The total population is 5000 and you wish a 99% confidence level and a ±5% precision level. What is the sample size required?
8) A certain population is divided into 4 stratums so that $N_1 = 4000$, $N_2 = 6000$, $N_3 = 7000$, $N_4 = 3000$. The respective stratum standard deviations are $\sigma_1 = 2.0$, $\sigma_2 = 4.0$, $\sigma_3 = 3.0$, $\sigma_4 = 6.0$. How should a sample size of 300 be allocated to four stratums using: (a) proportional and (b) disproportional methods.
9) Discuss the sources of sampling and non-sampling errors.
10) What are the essentials of a good sample?

Note: These questions/exercises will help you to understand the unit better. Try to write answers for them. But do not submit your answers to the university for assessment. These are for your practice only.

4.12 FURTHER READING

The following text books may be used for more indepth study on the topics dealt with in this unit.

Gupta, C.B., & Vijay Gupta, *An Introduction to Statistical Methods,* Vikas Publishing House Pvt. Ltd., New Delhi.

Kothari, C.R.(2004) *Research Methodology Methods and Techniques,* New Age International (P) Ltd., New Delhi.

Levin, R.I. and D.S. Rubin. (1999) *Statistics for Management,* Prentice-Hall of India, New Delhi

Mustafi, C.K.(1981) *Statistical Methods in Managerial Decisions,* Macmillan, New Delhi