

Design of Sample Surveys

4.1 INTRODUCTION

All items in any field of inquiry constitute a "Universe" or "Population." A complete enumeration of all items in the "population" is known as a census inquiry or census survey. It is obvious that for any study or investigation census survey is rather infeasible. For example, to have an idea of average per capita monthly income of the people in India, we will have to enumerate all the earning individuals in the country, which is a very difficult task.

Census survey is impossible in the situations when population is infinite. In some cases when population is finite but the units are destroyed while inspected, census survey is not at all desirable e.g., inspection of crackers. Further, many a time it is not possible to examine every item in the population, and sometimes it is possible to obtain sufficiently accurate results by studying only a part of total population. In such cases there is no utility of census surveys.

However, it needs to be emphasised that when the universe is a small one, it is no use resorting to a sample survey. When field studies are undertaken in practical life, considerations of time and cost almost invariably lead to a selection of respondents i.e., selection of only a few items. The respondents selected should be as representative of the total population as possible in order to produce a miniature cross-section. The selected respondents constitute what is technically called a 'sample' and the selection process is called 'sampling technique.' The survey so conducted is known as 'sample survey'. Algebraically, let the population size be N and if a part of size n (which is $< N$) of this population is selected according to some rule for studying some characteristic of the population, the group consisting of these n units is known as 'sample'. Researcher must prepare a sample design for his study i.e., he must plan how a sample should be selected and of what size such a sample would be.

4.2 SAMPLE DESIGN

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of items to be included in the sample i.e., the size of the

sample. Sample design is determined before data are collected. There are many sample designs from which a researcher can choose. Some designs are relatively more precise and easier to apply than others. Researcher must select/prepare a sample design which should be reliable and appropriate for his research study. The main steps of sampling design are as follows:

- (i) **Objective:** The first step of sampling design is to define the objectives of survey in clear and concrete terms. The sponsors or the researchers of the survey should confirm that the objectives are commensurate with the money, manpower and time limit available for the survey.
- (ii) **Population:** In order to meet the objectives of the survey, what should be the population? This question should be answered in the second step. The population should be clearly defined.
- (iii) **Sampling units and frame:** A decision has to be taken concerning a sampling unit before selecting sample. Sampling unit may be a geographical one such as state, district, village, etc., or a construction unit such as house, flat, etc., or it may be a social unit such as family, club, school, etc., or it may be an individual. The researcher will have to decide one or more of such units that he has to select for his study. The list of sampling units is called as 'frame' or sampling frame. Sampling frame contains the names of all items of a universe (in case of finite universe only). Such a list should be comprehensive, correct, reliable and appropriate. It is extremely important for the source list to be as representative of the population as possible.
- (iv) **Size of sample:** This refers to the number of items to be selected from the universe to constitute a sample. This is a major problem before a researcher. The size of sample should neither be excessively large, nor too small. It should be optimum. An optimum sample is one which fulfills the requirements of efficiency, representativeness, reliability and flexibility. While deciding the size of sample, researcher must determine the desired precision as also an acceptable confidence level for the estimate. The size of population variance needs to be considered as in case of larger variance usually a bigger sample is needed. The size of population must be kept in view for this also limits the sample size. The parameters of interest in a research study must be kept in view, while deciding the size of the sample. Costs too dictate the size of sample that we can draw. As such, budgetary constraint must invariably be taken into consideration when we decide the sample size. The formulation of sample size determination is provided in Chapter 10.
- (v) **Parameters of interest:** Statistical constants of the population are called as parameters, e.g., population mean, population proportion etc. When we do census survey we get the actual value of parameters. On the other hand, when we do sample survey we get the estimates of unknown population parameters in place of their actual values.

In determining the sample design, one must consider the question of the specific population parameters which are of interest. For instance, we may be interested in estimating the proportion of persons with some characteristic in the population, or we may be interested in knowing some average or the other measure concerning the population. There may also be important sub-groups in the population about whom we would like to make estimates. All this has a strong impact upon the sample design we would accept.

- (vi) **Data collection:** No irrelevant information should be collected and no essential information should be discarded. The objectives of the survey should be very much clear in the mind of surveyor. The methods of data collection are discussed in Chapter 6.
- (vii) **Non-respondents:** Because of practical difficulties, data may not be collected for all the sampled units. This non-response tends to change the results. The reasons for non-response should be recorded by the investigator. Such cases should be handled with caution.
- (viii) **Selection of proper sampling design:** The researcher must decide the type of sample he will use i.e., he must decide about the technique to be used in selecting the items for the sample. In fact, this technique or procedure stands for the sample design itself. There are several sample designs (explained in the pages that follow) out of which the researcher must choose one for his study. Obviously, he must select that design which, for a given sample size and for a given cost, has a smaller sampling error.
- (ix) **Organizing field work:** The success of a survey depends on the reliable field work. There should be efficient supervisory staff and trained personnels for the field work.
- (x) **Pilot survey:** It is always helpful to try out the research design on a small scale before going to the field. This is called as 'pilot survey' or 'pretest'. It might give the better idea of practical problems and troubles.
- (xi) **Budgetary constraint:** Cost considerations, from practical point of view, have a major impact upon decisions relating to not only the size of the sample but also to the type of sample. This fact can even lead to the use of a non-probability sample.

4.3 SAMPLING AND NON-SAMPLING ERRORS

The errors involved in the collection of data are classified into sampling and non-sampling errors.

4.3.1 Sampling Errors

Sampling errors arise due to the fact that only a part of the population has been used to estimate population parameters and to draw inferences about the population. Sampling errors are absent in a census survey. *Sampling error* can be measured for a given sample design and size. The measurement of sampling error is usually called the 'precision of the sampling plan'. If we increase the sample size, the precision can be improved. But increasing the size of the sample has its own limitations viz., a large sized sample increases the cost of collecting data and also enhances the systematic bias. Thus the effective way to increase precision is usually to select a better sampling design which has a smaller sampling error for a given sample size at a given cost. In practice, however, people prefer a less precise design because it is easier to adopt the same and also because of the fact that systematic bias can be controlled in a better way in such a design.

In brief, while selecting a sampling procedure, researcher must ensure that the procedure causes a relatively small sampling error and helps to control the systematic bias in a better way.

4.3.2 Non-sampling Errors

Non-sampling errors arise at the stage of collection and preparation of data and thus are present in both the sample survey as well as the census survey. Thus the data obtained in census survey is free from sampling errors, however subjected to non-sampling errors. Non-sampling errors can be reduced by defining the sampling units, frame and the population correctly and by employing efficient people in the investigations.

4.4 SAMPLE SURVEY VS. CENSUS SURVEY

In a sample survey, since we study only a subpart of the whole population, requires less money and less time. Most of the times, non-sampling errors are so much large that the results of sample survey are much more accurate than those of census survey. Non-sampling errors arise due to a number of factors such as inefficiency of field workers, non-response, bias due to interviewers, etc. These errors are likely to grow when the number of units inspected increases.

However, if the objective of study is very serious in nature and information is required about each and every sampling unit, there is no way out but to resort to census survey. Moreover, if time and money are not important factors or if population is not so large, a census survey may provide better results than any sample survey, provided efficient and trained staff is employed.

4.5 TYPES OF SAMPLING DESIGNS

The method of selecting a sample is of fundamental importance and depends upon the nature of data and investigation. The techniques of selecting a sample are classified as – ‘non - probability sampling’ and ‘probability sampling.’ We take up these two designs separately.

4.5.1 Non-probability Sampling

Non-probability sampling is that sampling procedure which does not afford any basis for estimating the probability that each item in the population has of being included in the sample. Non-probability sampling is also known by different names such as deliberate sampling, purposive sampling and judgement sampling. In this type of sampling, items for the sample are selected deliberately by the researcher; his choice concerning the items remain supreme. In other words, under non-probability sampling the organisers of the inquiry purposively choose the particular units of the universe for constituting a sample on the basis that the small mass that they so select out of a huge one will be typical or representative of the whole. For instance, if economic conditions of people living in a state are to be studied, a few towns and villages may be purposively selected for intensive study on the principle that they can be representative of the entire state. Thus, the judgement of the organisers of the study plays an important part in this sampling design.

In such a design, personal element has a great chance of entering into the selection of the sample. The investigator may select a sample which shall yield results favourable to his point of view and if that happens, the entire inquiry may get vitiated. Thus, there is always the danger of bias entering into this type of sampling technique. But in the investigators are impartial, work without bias

and have the necessary experience so as to take sound judgement, the results obtained from an analysis of deliberately selected sample may be tolerably reliable. However, in such a sampling, there is no assurance that every element has some specifiable chance of being included. Sampling error in this type of sampling cannot be estimated and the element of bias, great or small, is always there. As such this sampling design is rarely adopted in large inquiries of importance. However, in small inquiries and researches by individuals, this design may be adopted because of the relative advantage of time and money inherent in this method of sampling. *Quota sampling* is also an example of non-probability sampling. Under quota sampling the interviewers are simply given quotas to be filled from the different strata, with some restrictions on how they are to be filled. In other words, the actual selection of the items for the sample is left to the interviewer's discretion. This type of sampling is very convenient and is relatively inexpensive. But the samples so selected certainly do not possess the characteristic of random samples. Quota samples are essentially judgement samples and inferences drawn on their basis are not amenable to statistical treatment in a formal way.

4.5.2 Probability Sampling

Probability sampling is also known as 'random sampling' or 'chance sampling'. Under this sampling design, every item of the universe has an equal chance of inclusion in the sample. It is, so to say, a lottery method in which individual units are picked up from the whole group not deliberately but by some mechanical process. Here it is blind chance alone that determines whether one item or the other is selected. The results obtained from probability or random sampling can be assured in terms of probability i.e., we can measure the errors of estimation or the significance of results obtained from a random sample, and this fact brings out the superiority of random sampling design over the deliberate sampling design. Random sampling ensures the law of Statistical Regularity which states that if on an average the sample chosen is a random one, the sample will have the same composition and characteristics as the universe. This is the reason why random sampling is considered as the best technique of selecting a representative sample.

4.5.2.1 Simple Random Sampling

Simple random sampling from a finite population refers to that method of sample selection which gives each possible sample combination an equal probability of being picked up and each item in the entire population to have an equal chance of being included in the sample. This applies to sampling without replacement i.e., once an item is selected for the sample, it cannot appear in the sample again (sampling with replacement is used frequently less in which procedure the element selected for the sample is returned to the population before the next element is selected. In such a situation the same element could appear twice in the same sample before the second element is chosen). In brief, the implications of random sampling (or simple random sampling) are:

- (a) It gives each element in the population an equal probability of getting into the sample; and all choices are independent of one another.
- (b) It gives each possible sample combination an equal probability of being chosen.

Keeping this in view we can define a simple random sample (or simply a random sample) from a finite population as a sample which is chosen in such a way that each of the ${}^N C_n$ possible samples has the same probability, $1/{}^N C_n$, of being selected. To make it more clear we take a certain finite

population consisting of six elements (say a, b, c, d, e, f) i.e., $N = 6$. Suppose that we want to take a sample of size $n = 3$ from it. Then there are ${}^6C_3 = 20$ possible distinct samples of the required size, and they consist of the elements $abc, abd, abe, abf, acd, ace,acf, ade,adf, aef, bcd, bce, bcf, bde, bdf, bef, cde, cdf,cef$, and def . If we choose one of these samples in such a way that each has the probability $1/20$ of being chosen, we will then call this a random sample or a simple random sample.

With regard to the question of how to take a random sample in actual practice, we could, in simple cases like the one above, write each of the possible samples on a slip of paper, mix these slips thoroughly in a container and then draw as a lottery either blindfolded or by rotating a drum or by any other similar device. Such a procedure is obviously impractical, if not altogether impossible in complex problems of sampling. In fact, the practical utility of such a method is very much limited.

Fortunately, we can take a random sample in a relatively easier way without taking the trouble of enlisting all possible samples on paper-slips as explained above. Instead of this, we can write the name of each element of a finite population on a slip of paper, put the slips of paper so prepared into a box or a bag and mix them thoroughly and then draw (without looking) the required number of slips for the sample one after the other without replacement. In doing so we must make sure that in successive drawings each of the remaining elements of the population has the same chance of being selected. This procedure will also result in the same probability for each possible sample. We can verify this by taking the above example. Since we have a finite population of 6 elements and we want to select a sample of size 3, the probability of drawing any one element for our sample in the first draw is $3/6$, the probability of drawing one more element in the second draw is $2/5$ (the first element drawn is not replaced) and similarly the probability of drawing one more element in the third draw is $1/4$. Since these draws are independent, the joint probability of the three elements which constitute our sample is the product of their individual probabilities and this works out to $3/6 \times 2/5 \times 1/4 = 1/20$. This verifies our earlier calculation.

Even this relatively easy method of obtaining a random sample can be simplified in actual practice by the use of random number tables. Various statisticians like Tippett, Yates, Fisher have prepared tables of random numbers which can be used for selecting a random sample. Generally, Tippett's random number tables are used for the purpose. Tippett gave 10,400 four figure numbers. He selected 41,600 digits from the census reports and combined them into fours to give his random numbers which may be used to obtain a random sample.

We can illustrate the procedure by an example. First of all we reproduce the first thirty sets of Tippett's numbers

2952	6641	3992	9792	7979	5911
3170	5624	4167	9525	1545	1396
7203	5356	1300	2693	2370	7483
3408	2769	3563	6107	6913	7691
0560	5246	1112	9025	6008	8126

Suppose we are interested in taking a sample of 10 units from a population of 5000 units, bearing numbers from 3001 to 8000. We shall select 10 such figures from the above random numbers which are not less than 3001 and not greater than 8000. If we randomly decide to read the table numbers from left to right, starting from the first row itself, we obtain the following numbers: 6641, 3992, 7979, 5911, 3170, 5624, 4167, 7203, 5356, and 7483.

The units bearing the above serial numbers would then constitute our required random sample.

One may note that it is easy to draw random samples from finite populations with the aid of random number tables only when lists are available and items are readily numbered. But in some situations it is often impossible to proceed in the way we have narrated above. For example, if we want to estimate the mean height of trees in a forest, it would not be possible to number the trees, and choose random numbers to select a random sample. In such situations what we should do is to select some trees for the sample haphazardly without aim or purpose, and should treat the sample as a random sample for study purposes.

Other alternative may be to divide the whole forest into a finite number of equal-sized blocks and consider these blocks as sampling units. Height of each and every tree falling into selected blocks should be measured.

Simple random sample, we discussed above in 'simple random sample without replacement'. Here, a unit which is selected once is not replaced back to the population, i.e., one sampling unit can be selected only once. Other alternative to this design is 'simple random sample with replacement'. Because of the possibility of over-representation of some sampling units, this design is not of much practical usage.

So far we have talked about random sampling, keeping in view only the finite populations. But what about random sampling in context of infinite populations? It is relatively difficult to explain the concept of random sampling from an infinite population. However, a few examples will show the basic characteristic of such a sample. Suppose we consider the 20 throws of a fair dice as a sample from the hypothetically infinite population which consists of the results of all possible throws of the dice. If the probability of getting a particular number, say 1, is the same for each throw and the 20 throws are all independent, then we say that the sample is random. Similarly, it would be said to be sampling from an infinite population if we sample with replacement from a finite population and our sample would be considered as a random sample if in each draw all elements of the population have the same probability of being selected and successive draws happen to be independent. In brief, one can say that the selection of each item in a random sample from an infinite population is controlled by the same probabilities and that successive selections are independent of one another.

4.5.3 Complex Random Sampling Designs

Some complex random sampling designs, which are the mixture of probability and non-probability sampling methods, are given below.

4.5.3.1 Systematic Sampling

In some instances, the most practical way of sampling is to select every i th item on a list. Sampling of this type is known as systematic sampling. An element of randomness is introduced into this kind of sampling by using random numbers to pick up the unit with which to start. For instance, if a 4 per cent sample is desired, the first item would be selected randomly from the first twenty-five and thereafter every 25th item would automatically be included in the sample. Thus, in systematic sampling only the first unit is selected randomly and the remaining units of the sample are selected at fixed intervals. Although a systematic sample is not a random sample in the strict sense of the term, but it is often considered reasonable to treat systematic sample as if it were a random sample.

Systematic sampling has certain plus points. It can be taken as an improvement over a simple random sample in as much as the systematic sample is spread more evenly over the entire population. It is an easier and less costlier method of sampling and can be conveniently used even in case of large populations. But there are certain dangers too in using this type of sampling. If there is a hidden periodicity in the population, systematic sampling will prove to be an inefficient method of sampling. For instance, every 25th item produced by a certain production process is defective. If we are to select a 4% sample of the items of this process in a systematic manner, we would either get all defective items or all good items in our sample depending upon the random starting position. If all elements of the universe are ordered in a manner representative of the total population, i.e., the population list is in random order, systematic sampling is considered equivalent to random sampling. But if this is not so, then the results of such sampling may, at times, not be very reliable. In practice, systematic sampling is used when lists of population are available and they are of considerable length.

4.5.3.2 Stratified Sampling

If a population from which a sample is to be drawn does not constitute a homogeneous group, stratified sampling technique is generally applied in order to obtain a representative sample. Under stratified sampling the population is divided into several sub-populations that are individually more homogeneous than the total population (the different sub-populations are called 'strata') and then we select items from each stratum to constitute a sample. Since each stratum is more homogeneous than the total population, we are able to get more precise estimates for each stratum and by estimating more accurately each of the component parts, we get a better estimate of the whole. In brief, stratified sampling results in more reliable and detailed information.

The following three questions are highly relevant in the context of stratified sampling:

- (a) How to form strata?
- (b) How should items be selected from each stratum?
- (c) How many items be selected from each stratum or how to allocate the sample size of each stratum?

Regarding the first question, we can say that the strata be formed on the basis of common characteristic(s) of the items to be put in each stratum. This means that various strata be formed in such a way as to ensure elements being most homogeneous within each stratum and most heterogeneous between the different strata. Thus, strata are purposively formed and are usually based on past experience and personal judgement of the researcher. One should always remember that careful consideration of the relationship between the characteristics of the population and the characteristics to be estimated are normally used to define the strata. At times, pilot study may be conducted for determining a more appropriate and efficient stratification plan. We can do so by taking small samples of equal size from each of the proposed strata and then examining the variances within and among the possible stratifications, we can decide an appropriate stratification plan for our inquiry.

In respect of the second question, we can say that the usual method, for selection of items for the sample from each stratum, resorted to is that of simple random sampling. Systematic sampling can be used if it is considered more appropriate in certain situations.

Regarding the third question, we usually follow the method of proportional allocation under which the sizes of the samples from the different strata are kept proportional to the sizes of the strata. That is, if P_i represents the proportion of population included in stratum i , and n represents the total sample size, the number of elements selected from stratum i is $n \cdot P_i$. To illustrate it, let us suppose that we want a sample of size $n = 30$ to be drawn from a population of size $N = 8000$ which is divided into three strata of size $N_1 = 4000$, $N_2 = 2400$ and $N_3 = 1600$. Adopting proportional allocation, we shall get the sample sizes as under for the different strata:

For strata with $N_1 = 4000$, we have $P_1 = 4000/8000$
and hence $n_1 = n \cdot P_1 = 30 (4000/8000) = 15$

Similarly, for strata with $N_2 = 2400$, we have

$$n_2 = n \cdot P_2 = 30 (2400/8000) = 9, \text{ and}$$

for strata with $N_3 = 1600$, we have

$$n_3 = n \cdot P_3 = 30 (1600/8000) = 6.$$

Thus, using proportional allocation, the sample sizes for different strata are 15, 9 and 6 respectively which is in proportion to the sizes of the strata viz., 4000 : 2400 : 1600. Proportional allocation is considered most efficient and an optimal design when the cost of selecting an item is equal for each stratum, there is no difference in within-stratum variances, and the purpose of sampling happens to estimate the population value of some characteristic. But in case the purpose happens to compare the differences among the strata, then equal sample selection from each stratum would be more efficient even if the strata differ in sizes. In cases where strata differ not only in size but also in variability and it is considered reasonable to take larger samples from the more variable strata and smaller samples from the less variable strata, we can then account for both (differences in stratum size and differences in stratum variability) by using disproportionate sampling design by requiring:

$$n_1/N_1\sigma_1 = n_2/N_2\sigma_2 = \dots = n_k/N_k\sigma_k$$

where $\sigma_1, \sigma_2, \dots$ and σ_k denote the standard deviations of the k strata, N_1, N_2, \dots, N_k denote the sizes of the k strata and n_1, n_2, \dots, n_k denote the sample sizes of k strata. This is called '*optimum allocation*' in the context of disproportionate sampling. The allocation in such a situation results in the following formula for determining the sample sizes different strata:

$$n_i = \frac{n \cdot N_i \sigma_i}{N_1 \sigma_1 + N_2 \sigma_2 + \dots + N_k \sigma_k} \quad \text{for } i = 1, 2, \dots \text{ and } k.$$

We may illustrate the use of this by an example.

Example 4.1: A population is divided into three strata so that $N_1 = 5000$, $N_2 = 2000$ and $N_3 = 3000$. Respective standard deviations are:

$$\sigma_1 = 15, \sigma_2 = 18 \text{ and } \sigma_3 = 5.$$

How should a sample of size $n = 84$ be allocated to the three strata, if we want optimum allocation using disproportionate sampling design?

Solution: Using the disproportionate sampling design for optimum allocation, the sample sizes for different strata will be determined as under:

Sample size for strata with $N_1 = 5000$

$$n_1 = \frac{84(5000)(15)}{(5000)(15) + (2000)(18) + (3000)(5)}$$

$$= 6300000/126000 = 50$$

Sample size for strata with $N_2 = 2000$

$$n_2 = \frac{84(2000)(18)}{(5000)(15) + (2000)(18) + (3000)(5)}$$

$$= 3024000/126000 = 24$$

Sample size for strata with $N_3 = 3000$

$$n_3 = \frac{84(3000)(5)}{(5000)(15) + (2000)(18) + (3000)(5)}$$

$$= 1260000/126000 = 10$$

In addition to differences in stratum size and differences in stratum variability, we may have differences in stratum sampling cost, then we can have cost optimal disproportionate sampling design by requiring

$$\frac{n_1}{N_1 \sigma_1 \sqrt{C_1}} = \frac{n_2}{N_2 \sigma_2 \sqrt{C_2}} = \dots = \frac{n_k}{N_k \sigma_k \sqrt{C_k}}$$

where

C_1 = Cost of sampling in stratum 1

C_2 = Cost of sampling in stratum 2

⋮

C_k = Cost of sampling in stratum k

and all other terms remain the same as explained earlier. The allocation in such a situation results in the following formula for determining the sample sizes for different strata:

$$n_i = \frac{n \cdot N_i \sigma_i / \sqrt{C_i}}{N_1 \sigma_1 / \sqrt{C_1} + N_2 \sigma_2 / \sqrt{C_2} + \dots + N_k \sigma_k / \sqrt{C_k}} \quad \text{for } i = 1, 2, \dots, k$$

It is not necessary that stratification be done keeping in view a single characteristic. Populations are often stratified according to several characteristics. For example, a system-wide survey designed to determine the attitude of students towards a new teaching plan, a state college system with 20 colleges might stratify the students with respect to class, section and college. Stratification of this type is known as *cross-stratification*, and up to a point such stratification increases the reliability of estimates and is much used in opinion surveys.

From what has been stated above in respect of stratified sampling, we can say that the sample so constituted is the result of successive application of purposive (involved in stratification of items) and random sampling methods. As such it is an example of mixed sampling. The procedure wherein we first have stratification and then simple random sampling is known as stratified random sampling.

4.5.3.3 Cluster Sampling

If the total area of interest happens to be a big one, a convenient way in which a sample can be taken is to divide the area into a number of smaller non-overlapping areas and then to randomly select a number of these smaller areas (usually called clusters), with the ultimate sample consisting of all (or samples of) units in these small areas or clusters.

Thus, in cluster sampling the total population is divided into a number of relatively small subdivisions which are themselves clusters of still smaller units and then some of these clusters are randomly selected for inclusion in the overall sample. Suppose, we want to estimate the proportion of machine-parts in an inventory which are defective. Also assume that there are 20,000 machine parts in the inventory at a given point of time, stored in 400 cases of 50 each. Now, using a cluster sampling, we would consider the 400 cases as clusters and randomly select ' n ' cases and examine all the machine-parts in each randomly selected case.

Cluster sampling, no doubt, reduces cost by concentrating surveys in selected clusters. But certainly it is less precise than random sampling. There is also not as much information in ' n ' observations within a cluster as there happens to be in ' n ' randomly drawn observations. Cluster sampling is used only because of the economic advantage it possesses; estimates based on cluster samples are usually more reliable per unit cost.

If clusters happen to be some geographic sub-divisions, in that case cluster sampling is better known as area sampling. In other words, cluster designs, where the primary sampling unit represents a cluster of units based on geographic area, are distinguished as area sampling. The plus and minus points of cluster sampling are also applicable to area sampling.

4.5.3.4 Multi-stage Sampling

Multi-stage sampling is a further development of the principle of cluster sampling. Suppose, we want to investigate the working efficiency of nationalised banks in India and we want to take a sample of few banks for this purpose. The first stage is to select large primary sampling unit such as states in a country. Then we may select certain districts and interview all banks in the chosen districts. This would represent a two-stage sampling design with the ultimate sampling units being clusters of districts.

If instead of taking a census of all banks within the selected districts, we select certain towns and interview all banks in the chosen towns. This would represent a three-stage sampling design. If instead of taking a census of all banks within the selected towns, we randomly sample banks from each selected town, then it is a case of using a four-stage sampling plan. If we select randomly at all stages, we will have what is known as 'multistage random sampling design'.

Ordinarily multi-stage sampling is applied in big inquiries extending to a considerable large geographical area, say, the entire country. There are two advantages of this sampling design viz., (a) It is easier to administer than most single stage designs mainly because of the fact that sampling frame under multi-stage sampling is developed in partial units. (b) A large number of units can be sampled for a given cost under multi-stage sampling because of sequential clustering, whereas this is not possible in most of the simple designs.

4.5.3.5 Sampling with Probability Proportional to Size

In case the cluster sampling units do not have the same number or approximately the same number of elements, it is considered appropriate to use a random selection process where the probability of each cluster being included in the sample is proportional to the size of the cluster. For this purpose, we have to list the number of elements in each cluster irrespective of the method of ordering the cluster. Then we must sample systematically the appropriate number of elements from the cumulative totals. The actual numbers selected in this way do not refer to individual elements, but indicate which clusters and how many from the cluster are to be selected by simple random sampling or by systematic sampling. The results of this type of sampling are equivalent to those of a simple random sample and the method is less cumbersome and is also relatively less expensive. We can illustrate this with the help of an example.

Example 4.2: The following are the number of departmental stores in 15 cities: 35, 17, 10, 32, 70, 28, 26, 19, 26, 66, 37, 44, 33, 29 and 28. If we want to select a sample of 10 stores, using cities as clusters and selecting within clusters proportional to size, how many stores from each city should be chosen? (Use a starting point of 10).

Solution: Let us put the information as under (Table 4.1):

Since in the given problem, we have 500 departmental stores from which we have to select a sample of 10 stores, the appropriate sampling interval is 50. As we have to use the starting point of 10, so we add successively increments of 50 till 10 numbers have been selected. The numbers, thus, obtained are: 10, 60, 110, 160, 210, 260, 310, 360, 410 and 460 which have been shown in the last column of the table (Table 4.1) against the concerning cumulative totals. From this we can say that two stores should be selected randomly from city number five and one each from city number 1, 3, 7, 9, 10, 11, 12, and 14. This sample of 10 stores is the sample with probability proportional to size.

Table 4.1

City number	No. of departmental stores	Cumulative total	Sample
1	35	35	10
2	17	52	
3	10	62	60
4	32	94	
5	70	164	110 160
6	28	192	
7	26	218	210
8	19	237	
9	26	263	260
10	66	329	310
11	37	366	360
12	44	410	410
13	33	443	
14	29	472	460
15	28	500	

*If the starting point is not mentioned, then the same can randomly be selected.

4.5.3.6 Sequential Sampling

This sampling design is somewhat complex sample design. The ultimate size of the sample under this technique is not fixed in advance, but is determined according to mathematical decision rules on the basis of information yielded as survey progresses. This is usually adopted in case of acceptance sampling plan in context of statistical quality control. When a particular lot is to be accepted or rejected on the basis of a single sample, it is known as single sampling; when the decision is to be taken on the basis of two samples, it is known as double sampling and in case the decision rests on the basis of more than two samples but the number of samples is certain and decided in advance, the sampling is known as multiple sampling. But when the number of samples is more than two but it is neither certain nor decided in advance, this type of system is often referred to as sequential sampling. Thus, in brief, we can say that in sequential sampling, one can go on taking samples one after another as long as one desires to do so.

REMARK

From a brief description of the various sample designs presented above, we can say that normally one should resort to simple random sampling because under it bias is generally eliminated and the sampling error can be estimated. But purposive sampling is considered more appropriate when the universe happens to be small and a known characteristic of it is to be studied intensively. There are situations in real life under which sample designs other than simple random samples may be considered better (say easier to obtain, cheaper or more informative) and as such the same may be used. In a situation when random sampling is not possible, then we have to use necessarily a sampling design other than random sampling. At times, several methods of sampling may well be used in the same study.

PROBLEMS

1. What do you mean by 'Sample Design'? What points should be taken into consideration by a researcher in developing a sample design for this research project.
2. How would you differentiate between simple random sampling and complex random sampling designs? Explain clearly giving examples.
3. Why probability sampling is generally preferred in comparison to non-probability sampling? Explain the procedure of selecting a simple random sample.
4. Under what circumstances stratified random sampling design is considered appropriate? How would you select such sample? Explain by means of an example.
5. Distinguish between:
 - (a) Restricted and unrestricted sampling;
 - (b) Convenience and purposive sampling;
 - (c) Systematic and stratified sampling;
 - (d) Cluster and area sampling.

6. Under what circumstances would you recommend:
 - (a) A probability sample?
 - (b) A non-probability sample?
 - (c) A stratified sample?
 - (d) A cluster sample?
7. Explain and illustrate the procedure of selecting a random sample.
8. How will you optimize sampling and non-sampling errors?
9. "A systematic bias results from errors in the sampling procedures". What do you mean by such a systematic bias? Describe the important causes responsible for such a bias.
10. (a) The following are the number of departmental stores in 10 cities: 35, 27, 24, 32, 42, 30, 34, 40, 29 and 38. If we want to select a sample of 15 stores using cities as clusters and selecting within clusters proportional to size, how many stores from each city should be chosen? (Use a starting point of 4).
(b) What sampling design might be used to estimate the weight of a group of men and women?
11. A certain population is divided into five strata so that $N_1 = 2000$, $N_2 = 2000$, $N_3 = 1800$, $N_4 = 1700$, and $N_5 = 2500$. Respective standard deviations are: $\sigma_1 = 1.6$, $\sigma_2 = 2.0$, $\sigma_3 = 4.4$, $\sigma_4 = 4.8$, $\sigma_5 = 6.0$ and further the expected sampling cost in the first two strata is Rs 4 per interview and in the remaining three strata the sampling cost is Rs 6 per interview. How should a sample of size $n = 226$ be allocated to five strata if we adopt proportionate sampling design; if we adopt disproportionate sampling design considering (i) only the differences in stratum variability (ii) differences in stratum variability as well as the differences in stratum sampling costs.