UNIT 13 SAMPLING METHODS

Objectives

On successful completion of this unit, you should be able to:

- appreciate why sampling is so common in managerial situations
- identify the potential sampling errors
- list the various sampling methods with their strengths and weaknesses
- distinguish between probability and non-probability sampling
- know when to use the proportional or the disproportional stratified sampling
- understand the role of multi-stage and multi-phase sampling in large sampling studies
- appreciate why and how non-probability sampling is used in spite of its theoretical weaknesses
- recognise the factors which affect the sample size decision.

Structure

13.1 INTRODUCTION

Let us take a look at the following five situations to find out the common features among them, if any:

i) An inspector from the Weights &Measures department of the government goes to a unit manufacturing vanaspati. He picks up a small number of packed containers from the day's production, pours out the contents from each of these selected containers and weighs them individually to determine if the .manufacturing unit is packing enough vanaspati in its containers to conform to what is claimed as the net weight in the label.

ii) The personnel department of a large bank wants to measure the level of employee motivation and morale so that it can initiate appropriate measures to help improve the same. It administers a questionnaire to about 250 employees from different branches and offices all over India selected from a total of about

30,000 employees and analysis the information contained in these 250 filled-in questionnaires to assess the morale and motivation levels of all employees.

iii) The product development department of a consumer products company has developed a "new improved" version of its talcum powder. Before launching the new product, the marketing department gives a container of the old version first and after a week, a container of the new version to a group of 400 consumers and gets the feedback of these consumers on various attributes of the products. These consumer responses will form the basis for assessing the consumer perception of the new talcum powder as compared to the old talcum powder.

iv) The quality control department of a company manufacturing fluorescent tubes checks the life of its products by picking up 15 of its tubes at random and letting them burn till each one of them fuses. The life of all its products is assessed based on the performance of these 15 tubes.

v) An industrial engineer takes 100 rounds of the shop floor over a period of six clays and based on these 100 observations, assesses the machine utilisation on the shop floor.

What is Sampling

On the face of it, there is little that is common among the five situations described above. Each one refers to a different functional area and the nature of the problem also is quite different from one situation to another. However, on closer observation, it appears that in all these situations one is interested in measuring some attribute of a large or infinite group of elements by studying only a part of that group. This process of inferring something about a large group of elements by studying only a part of it, is referred to as sampling.

Most of us use sampling in our daily life, e.g. when we go to buy provisions from a grocery. We might sample a few grains of rice or wheat to infer the quality of a whole bag of it. In this unit we shall study why sampling works and the various methods of sampling available so that we can make the process of sampling more efficient.

Some Basic Concepts

We shall refer to the collection of all elements about which some inference is to be made as the population. For example, in situation (ii) above,, the population is the set of 30,000 employees working in the bank and in situation (iii), the population comprises of all the consumers of talcum powder in the country.

We are basically interested in measuring some characteristics of the population. This could be the average life of a fluorescent tube, the percentage of consumers of talcum powder who prefer the "new improved" talcum powder to the old one or the percentage of time a machine is being used as in situation (v) above. Any characteristic of a population will be referred to as a **paramete**r of the population.

In sampling, some population parameter is inferred by studying only a part of the population. We shall refer to the part of the population that has been chosen as a sample. Sampling, therefore, refers to the process of choosing a sample from the population so that some inference about the population can be made by studying the sample. For example, the sample in situation (ii) consists of the 250 employees from different branches and offices of the bank.

Any characteristic of a sample is called a **statistic.** For example, the mean life of the sample of 15 tubes in situation (iv) above is a sample statistic.

Conventionally, population parameters are denoted by Greek or capital letters and sample statistics by lower case Roman letters. There can be exceptions to this form of notation, e.g. population proportions is usually denoted by p and the sample proportion by p.

Figure I shows the concept of a population and a sample in the form of the Venn diagram, where the population is shown as the universal set and a sample is shown as a true subset of the population. The characteristics of a population and a sample and some symbols for these are presented in Table 1.

Figure I: Population and Sample

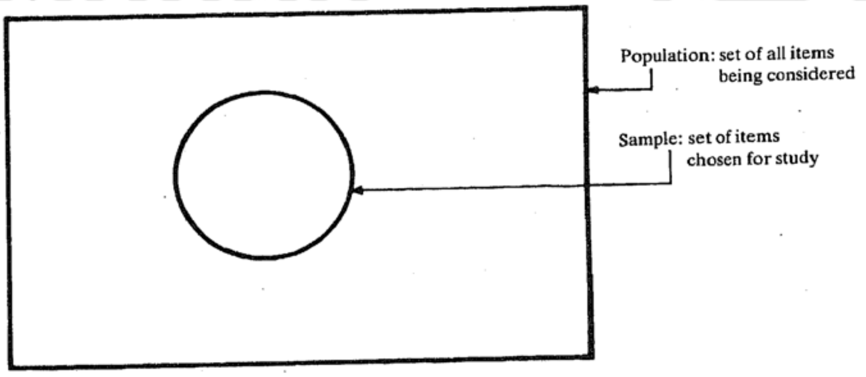

Population: set of all items being considered
Sample: set of items chosen for study

Table 1: Symbols for Population and Samples.

	POPULATION	SAMPLE
Characteristic	Parameter	Statistic
Symbols	Population size = N	Sample size = n
	Population mean = μ	Sample mean = $\bar{x}$
	Population s.d. = σ	Sample s.d. = s
	Population proportion = p	Sample proportion = $\bar{p}$

Sampling is not the only process available for making inferences about a population. For small populations, it may be feasible and practical, and sometimes desirable to examine every member of the population e.g. for inspection of some aircraft ,components. This process is referred to as **census** or complete enumeration of the. population.

13.2 WHY SAMPLING?

In the example situations given in section 13.1 above, the reasons for resorting to sampling should be very clear. We give below the various reasons which make sampling a desirable, and in many cases, the only course open for making an inference about a population.

Time taken for the Study i

Inferring from a sample can be much faster than from a complete enumeration of the population because fewer elements are being studied. In situation (iii) above in section 13. 1, a complete enumeration of all consumers, even if feasible, would perhaps take so much time that it is unacceptable for product launch decisions.

Cost involved for the Study

Sampling also helps in substantial cost reductions as compared to censuses and as we shall see later in this unit, a better sample design could reduce the cost of the study further. In many cases, like in situation (ii) above in section .13.1, it may be too costly, although feasible, to contact all the employees in the bank and get information from them.

Physical Impossibility of Complete Enumeration

In many situations the element being studied gets destroyed while being tested. The fluorescent tubes in situation (iv) of section 13. 1, which are chosen for testing their lives, get destroyed while being tested. In such cases, a complete enumeration is impossible as there would be no population left after such an enumeration.

Practical Infeasibility of Complete Enumeration

Quite often it is practically infeasible to do a complete enumeration due to many practical difficulties. For example, in situation (iii) of section 13.1, it would be infeasible to collect information from all the consumers of talcum powder in India. Some consumers would have moved from one place to another during the period of study, some others would have stopped consuming talcum powder just before the period of study whereas some others would have been users of talcum powder during the period of study but would have stopped using it some time later. In such situations, although it is theoretically possible to do a complete enumeration, it is practically infeasible to do so.

Enough Reliability of Inference based on Sampling

In many eases, sampling provides adequate information so that not much additional reliability can be gained with complete enumeration in spite of spending large amounts of additional money and time. It is also possible to quantify the magnitude of possible error on using; some types of sampling as will be explained later.

Quality of Data Collected

For large populations, complete enumeration also suffers from the possibility of spurious or unreliable data collected by the enumerators. On the other hand, there is greater confidence on the purity of the data collected in sampling as there can be better interviewing, better training and supervision of enumerators, better analysis of missing data and so on.

Activity A

When would you prefer complete enumeration to sampling?

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………

Activity B

Name two decisions in each of the following functional areas, where sampling can be of use:

Functional Area	**Decision**
Manufacturing	1) Inspection of components
	2)
Personnel	1)
	2)
Marketing	1)
	2)
Finance	1)
	2)

13.3 TYPES OF SAMPLING

There are two basic types of sampling depending on who or what is allowed to govern the selection of the sample. We shall call them by the names of probability sampling and non-probability sampling.

Probability Sampling

In probability sampling the decision whether a particular element is included in the sample or not, is governed by chance alone. All probability sampling designs ensure that each element in the population has some nonzero probability of getting included in the sample. This would mean defining a procedure for picking up the sample, based on chance, and avoiding changes in the sample except by way of a pre-defined process again. The picking up of the sample is therefore totally insulated against the judgment, convenience or whims of any person involved with the study. That is why probability sampling procedures tend to become rigorous and at times quite time-consuming to ensure that each element has a nonzero probability of getting included in the sample. On-the other hand, when probability sampling designs are used, it is possible to quantify the magnitude of the likely error in inference made and this is of great help in many situations in building up confidence in the inference.

Non-probability Sampling

Any sampling process which does not ensure some nonzero probability for each element in the population to be included in the sample would belong to the category of non-probability sampling. In this case, samples may be picked up based on the judgment or convenience of the enumerator. Usually, the complete sample is not decided at the beginning of the study but it evolves as the study progresses.

However, the very same factors which govern the selection of a sample e.g. judgment or convenience, can also introduce biases in the study. Moreover, there is no way that the magnitude of errors can be quantified when non-probability sampling designs are used.

Many times samples are selected by interviewers or enumerators "at random" meaning that the actual sample selection is left to the discretion of the enumerators. Such a sampling design would also belong to the non-probability sampling category and not the category of probability or random sampling.

13.4 PROBABILITY SAMPLING METHODS

In the category of probability sampling, we shall discuss the following four designs:

i) Simple Random Sampling

ii) Systematic Sampling

iii) Stratified Sampling

iv) Cluster Sampling

One can also use sampling designs which are combinations of the above listed ones.

Simple Random Sampling

Conceptually, simple random sampling is one of the simplest sampling designs and can work well for relatively small populations. However, there are many practical problems when one tries to use simple random sampling for large populations.

What is simple random sampling?: Suppose we have a population having N elements and that we want to pick up a sample of size n (< N). Obviously, there are many possible samples of size n.

Simple random sampling is a process which ensures that each of the samples of size n has an equal probability of being picked up as the chosen sample.

As we shall see later in this section this also implies that under simple random sampling, each element of the population has an equal probability of getting included in the sample.

All other forms of probability sampling use this basic concept of simple random sampling but applied to a part of the population at a time and not to the whole population.

Let us consider a small example to illustrate what simple random sampling is. Our population is a family of five members, two adults and three children, viz. A, B, C, D and E respectively. There are 10 different samples possible of size three as listed in Table 2 below. As we have shown in the same Table, if each of the 10 samples has an equal probability of 1/10 of being picked up, this implies that the probability that any particular element, say A or B, is included in the sample is the same.

In general, there are $\binom{N}{n}$ different samples of size n that can be picked up from a population of size N. Simple random sampling ensures that any of these samples has the same probability of bet g picked up viz. $\frac{1}{\binom{N}{n}}$

Table 2: Simple Random Sampling

Population of size 5: (A, B, C, D and E)

Let P [ABC] be the probability that the sample of size 3 containing elements A, B and C, is chosen.

Simple Random Sampling ensures that

P[ABC] =1/10		P[ADE] =	1/10
P[ABD] =1/10		P[BCD] =	1/10
P[ABE] =1/10		P[BCE] =	1/10
P[ACD] =1/10		P[BDE] =	1/10
P[ACE] =1/10		P[CDE] =	1/10

∴ Probability that element A

is in the sample, P(A) = P[ACC] + P[ABD] + P[ABE] + P[ACD] + P[ACE] + [ADE]
= 6/10

and P(B) = P[ABC] + P[ABD] + P[ABE]
+ P[BCD] + P[BCE] + P[BDE]
= 6/10

Similarly P(C)= 6/10
P(D)= 6/10
and P(E)= 6/10

If we want to find the probability that element A (or any other element for that matter) is included in the sample picked up, we have to find the number of different samples in which this element A occurs. There are (n -1) positions available in the sample (since one is occupied by A) which can be picked up from any of the (N-1) elements of the population (since A is not available to be picked up) and so there are $\binom{N}{n}$ different samples in which element A occurs.

Therefore, the probability that element A is included in

$$\text{the sample} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!}{(n-1)!\,(N-n)!} \times \frac{n!\,(N-n)!}{N!}$$

$$= \frac{n}{N}.$$

The fact that every element of the population has an equal probability of getting included in the sample is made use of in actually picking up simple random samples.

Sampling with and without replacement: We have implicitly assumed above that we are sampling without replacement, i.e. if an element is picked up-once, it is not available to be picked up again. This is how most practical samples are, but as a concept, it is possible to think in terms of sampling with replacement in which case an element, after being picked up and included in the sample, is replaced in the population so that it can be picked up again.

What is important for us to note at this stage is that even in the case of simple random sampling with replacement, each element has an equal probability of getting included in the sample.

How is simple random sampling done?: It is imperative to have a list of all the members of the population before a simple random sample can be picked up. Such an exhaustive list of all population members is called a sampling frame.

Suppose we write the name of one such member on a chit of paper and thus have N chits in a bowl, one chit for each member of the population. We can then mix the chits well and pick up one chit at random to represent one member of the sample. If we want a sample of size n, we have to repeat this process n times and we shall have a simple random sample of size n consisting of the names of members appearing on the chits picked.

It is easy to see that if we replace the chits in the bowl after noting down the name of the element, we will have a simple random sample with replacement and one without replacement if we do not.

As the population size increases, it becomes more and more difficult to work with chits and one can simulate this process on a computer or by using a table of random numbers. We can associate a serial number with each member of our population and then instruct a computer to pick up a member from 1 through N using its pseudo-random number generator. This ensures that every number from 1 through N has an equal probability of getting picked up and so the sample selected is a simple random sample.

We can also use a table of random numbers to pick up a simple random sample. In a table-of random numbers there is an equal probability for any digit from 0 to 9 to appear in any particular position. In table 3 we have a page of five digit random numbers containing 100 such numbers. The most important thing in using a random number table is to specify to the minutest detail the sequence of steps that has been decided before the table is actually referred to. We shall demonstrate this with an example.

Suppose we have a population of size 900 with each number being given a serial number ranging from 000 through 899 and we want to pick up a simple random sample of size 20. We proceed by defining a procedure.

1 Starting point and direction of movement. We may decide to start with the top left hand number and consider the first three digits (from left) as the three-digited random number picked up e.g. the first number would then be 121. We also specify that we shall move down a column to pick up further numbers-e.g. the second number would be 073, If there is no further number down the column, we shall go to the top of the next column of five-digited numbers and pick up the first three digits (from left)-e.g. after 851 our next number shall be 651.

2 Checking the number picked up. If the number picked up is in the range 000 to 899, we accept the number but if it is outside this range, we shall discard it and pick up the next number-e.g. after the third number 703, we discard 934 and the fourth member of the sample would be 740. Similarly, if we are doing sampling without replacement and a number is picked up again, it is discarded and we move to the next three-digited number.

   Using this process, if we want a sample of size 10, our sample would contain members with the following numbers: 121, 073, 703, 740, 736, 513, 464, 571, 379 and 412.

Simple random sampling in practice: Simple random sampling, as described here, is not the most efficient sampling design either statistically or economically in all practical situations. However, it forms the basis for Al other forms of probability sampling which are used on parts of the population or sub-population and not on the population as a whole.

Table 3: Table of five-digited random numbers

12135	65186	86886	72976	79885
07369	49031	45451	10724	95051
70387	53186	97116	32093	95612
93451	53493	56442	67121	70257
74077	66687	45394	33414	15685
73627	54287	42596	05544	76826
51353	56404	74106	66185	23145
46426	12855	48497	05532	36299
57126	99010	29015	65778	93911
37997	89034	79788	94676	32307
41283	42498	73173	21938	22024
76374	68251	71593	93397	26245
51668	47244	13732	48369	60907
17698	32685	24490	56983	81152
12448	00902	07263	16764	71261
52515	93269	61210	55526	71912
43501	10248	34219	83416	91239
45279	19382	82151	57365	84915
11437	98102	58168	61534	69495
85183	38161	22848	06673	35293

As mentioned earlier, in listing all members of the population viz. a frame is required before a simple random sample can be chosen. In many situations the frame is not available nor is it practical to prepare the frame in a time and cost-effective manner. Obviously, under such conditions simple random sampling is not a viable sampling design.

Most large populations are not homogeneous and can be broken down into more homogeneous units. In such conditions one can design sampling schemes which are statistically more efficient, meaning that they allow the same precision from smaller sample sizes.

Similarly by picking up members from geographically closer areas the cost efficiency of the sampling design can be improved. Cluster sampling is based on this concept.

The process of picking up a simple random sample through using a table of random numbers or any other such aids as discussed earlier, is rather cumbersome and not very purposeful to the uninitiated interviewer. Simpler forms of sampling overcomes this handicap of simple random sampling.

Activity C

There are 20 elements in a population, each identified by a letter of the English alphabet from A through T. Using the random number table given in Table 3, describe how you would pick up a sample of size 5 when sampling is done without replacement.

………………………………………………………………………………………
………………………………………………………………………………………
………………………………………………………………………………………
………………………………………………………………………………………
………………………………………………………………………………………
………………………………………………………………………………………

Systematic Sampling

Systematic sampling proceeds by picking up one element after a fixed interval depending on the sampling ratio. For example, if we want to have a sample of size 10 from a population of size 100, our sampling ratio would be n/N = 10/100 = 1/10. We would, therefore, have to decide where to start from among the first 10 names in our frame. If this number happens to be 7 for example, then the sample would contain members having serial numbers 7,17,27, ...97 in the frame. It is to be noted-that the random process establishes only the first member of the sample-the rest are pre-ordained automatic because of the known sampling ratio.

Systematic sampling in the previous example would choose one out of ten possible samples each starting with either number 1, or number 2, or ....number 10. This is usually decided by allowing chance to play its role-e.g. by using a table of random numbers.

Systematic sampling is relatively much easier to implement compared to simple random sampling. However, there is one possibility that should be guarded against while using systematic sampling-the possibility of a strong bias in the results if there is any periodicity in the frame that parallels the sampling ratio. One can give some ridiculously simple example to highlight the point. If you were making studies on the demand for various banking transactions in a bank branch by studying the demand on some days randomly selected by systematic sampling-be sure that your sampling ratio is not 1/7 or 1/14 etc. Otherwise you would always be studying the demand on the same day of the week and your inferences could be biased depending on whether the day selected is a Monday or a Friday and so on. Similarly, when the frame contains addresses of flats in buildings all alike and having say 12 flats in one building, systematic sampling with a sampling ratio of 1/6, 1/60 or any other such fraction would bias your sample with flats of only one type-e.g. a ground floor corner flat i.e., all types of flats would not be members of your sample; and this might lead to biases in the inference made.

I F the frame is arranged in an order, ascending or descending, of some attribute then the location of the first sample element may affect the result of the study. For example, if our frame contains a list of students arranged in a descending order of their percentage in the previous examination and we are picking a systematic sample with a sampling ratio of 1/50. If the first number picked is 1 or 2, then the sample chosen will be academically much better off compared to another systematic sample with the first number chosen as 49 or 50. In such situations, one should devise ways of nullifying the effect of bias due to starting number by insisting on multiple starts after a small cycle or other such means.

On the other hand, if the frame is so arranged that similar elements are grouped together, then systematic sampling produces almost a proportional stratified sample and would be, therefore, more statistically efficient than simple random sampling.

Systematic sampling is perhaps the most commonly used method among the probability sampling designs and for many purposes e.g. for estimating the precision of the results, systematic samples are treated as simple random samples.

Stratified Sampling

Stratified sampling is more complex than simple random sampling, but where applied properly, stratification can significantly increase the statistical efficiency of sampling.

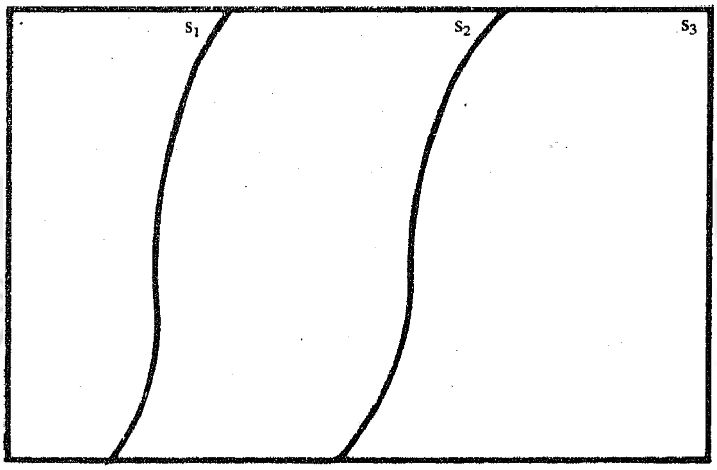
The concept: Suppose we are interested in estimating the demand of non-aerated beverages in a residential colony. We know that the consumption of these beverages has some relationship with the family income and that the families residing in this colony can be classified into three categories-viz., high income, middle income and low income families. If we are doing a sampling study we would like to make sure that our sample does have some members from each of the three categories-perhaps in the same proportion as the total number of families belonging to that category-in which case we would have used proportional stratified sampling. On the other hand, if we know that the variation in the consumption of these beverages from one family to another is relatively large for the low income category whereas there is not much variation in the high income category, we would perhaps pick up a smaller than proportional sample from the high income category and a larger than proportional sample from-the low income category. This is what is done in disproportional stratified sampling.

The basis for using stratified sampling is the existence of strata such that each stratum is more homogeneous within and markedly different from another stratum. The higher the homogeneity within each stratum, the higher the gain in statistical efficiency due to stratification.

What are strata?: The strata are so defined that they constitute a partition of the population-i.e., they are mutually exclusive and collectively exhaustive. Every element of the population belongs to one stratum and not more than one stratum, by definition. This is shown in Figure II in the form of a Venn diagram, where three strata have been shown.

A stratum can therefore he conceived of as a sub-population which is more homogeneous than the complete population-the members of a stratum, are similar to each other and are different from the members of another stratum in the characteristics that we are measuring.

Figure II: A Population with three strata



Proportional stratified sampling: After defining the strata, a simple random sample is picked up from each of the strata. If we want to have a total sample of size 100, this number is allocated to the different strata-either in proportion to the size of the stratum in the population or otherwise.

If the different strata have similar variances of the characteristic being measured, then the statistical efficiency will be the highest if the sample sizes for different strata are in the same proportion as the size of the respective stratum in the population. Such a design is called proportional stratified sampling and is shown in Table 4 below.

If we want to pick up a proportional stratified sample of size n from a population of size N, which has been stratified to p different strata with sizes $N_1, N_2, \ldots\ldots\ldots\ldots\ldots N_p$ respectively, then the sample sizes for different strata, viz $n_1, n_2, \ldots\ldots n_p$ will be given by

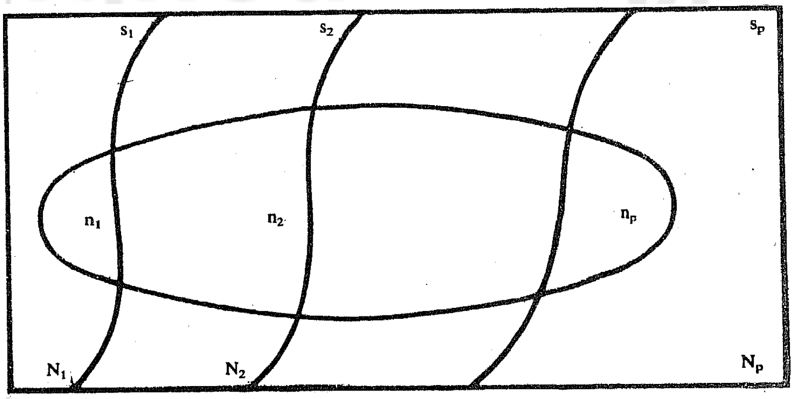
$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_p}{N_p} = \frac{n}{N}$$

Table 4: Proportional Stratified Sampling

Stratum No. (i)	No. of Elements in stratum (N_i)	Sample size (n_i)	Sampling Ratio (n_i/N_i)
1	200	10	1/20
2	300	15	1/20
3	500	25	1/20
Total	1000	50	1/20

The strata and the samples from each stratum are shown in the form of a Venn diagram in Figure III below, where S_I, S etc. refer to the stratum number 1. stratum number 2 etc. respectively.

Figure III: Stratified Sampling



Disproportional stratified sampling: If the different strata in the population have unequal variances of the characteristic being measured, then the sample size allocation decision should consider the variance as well. It would be logical to have a smaller sample from a stratum where the variance is smaller than from another stratum where the variance is higher. In fact, if $\sigma_1{}^2, \sigma_2{}^2, ..., \sigma_p{}^2$ are the variance of the p strata respectively, then the statistical efficiency is the highest when

$$\frac{n_1}{N_1\sigma_1} = \frac{n_2}{N_2\sigma_2} = \dots = \frac{n_p}{N_p\sigma_p}$$

where the other symbols have the same meaning as in the previous example.

Suppose the variances of the characteristic we are measuring were different for each of the three strata of the earlier example and were actually as shown in Table 5. If the total sample size was still restricted to 50, the statistically optimal

allocation would be as given in Table 5 and one can compare this Table with Table 4 above to find that the sampling ratio would fall for Stratum-3 as the variance is smaller here and would go up for Stratum-2 where the variance is larger.

Table 5: Disproportional Stratified Sampling

Stratum No. (i)	No. of Elements in Stratum (N_i)	Stratum Variance (σ_i^2)	Stratum s.d. (σ_i)	Sample size (n_i)	Sampling Ratio (n_i/N_i)
1	200	2.25	1.5	13	0.065
2	300	4.00	2.0	26	0.087
3	500	0.25	0.5	11	0.022
Total	1000			50	

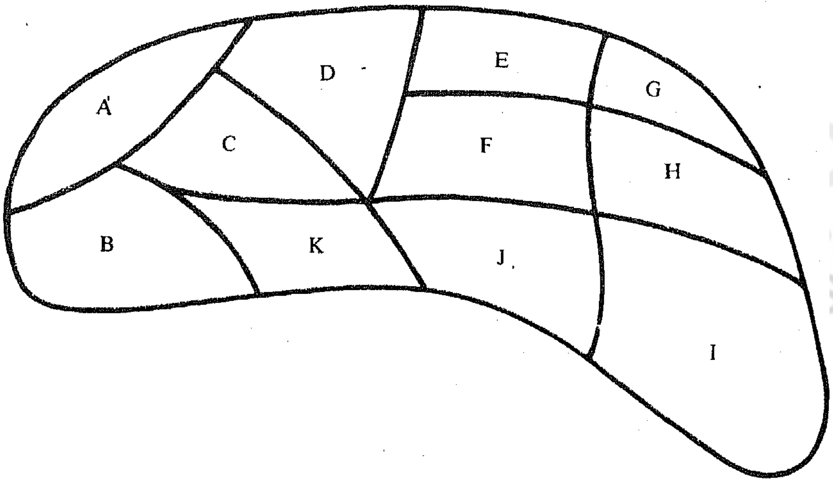
Stratified sampling in practice: Stratification of the population is quite common in managerial applications because it also allows to draw separate conclusions for each stratum. For example, if we are estimating the demand for a non-aerated beverage in a residential colony and have stratified the population based on the family income, then we would have data pertaining to each stratum which might be useful in making many marketing decisions.

Stratification requires us to identify the strata such that the intra-stratum differences are as small as possible and inter-strata differences as large as possible. However, whether a stratum is homogeneous or not-in the characteristic that we are measuring e.g. consumption of non-aerated beverage in the family in the previous example-can be known only at the end of the study whereas stratification is to be done at the beginning of the study and that is why some other variable like family income is to be used for stratification. This is based on the implicit assumption that family income and consumption of non-aerated beverages are very closely associated with each other. If this assumption is true, stratification would increase the statistical efficiency of sampling. In many studies, it is not easy to find such associated variables which can be used as the basis for stratification and then stratification may not help in increasing the statistical efficiency, although the cost of the study goes up due to the additional costs of stratification.

Cluster Sampling

Let us take up the situation where we are interested in estimating the demand for a non-aerated beverage in a residential colony again. The colony is divided into 11 blocks, called Block A through Block K as shown in Figure IV below.

Figure IV: Blocks in a residential colony



We might use cluster sampling in this situation by treating each block as a cluster. We will then select 2 blocks out of the 11 blocks at random and then collect information from all families residing in those 2 blocks.

Cluster vs stratum: We can now compare cluster sampling with stratified sampling. Stratification is done to make the strata homogeneous within and different from other strata. Clusters, on the other hand, should be heterogeneous within and the different clusters should be similar to each other. A clusture, ideally, is a mini-population and has all the features of the population.

The criterion used for stratification is a variable which is closely associated with the characteristic we are measuring e.g. income level when we are measuring the family consumption of non-aerated beverages in the example quoted earlier. On the other hand, convenience of data collection is usually the basis for cluster definitions.

Geographic contiguity is quite often used for clusture definitions, like in Figure IV above and in such cases, cluster sampling is also known as Area Sampling.

There are very fewer strata and one requires to pick up a random sample from each of the strata for drawing inferences. In cluster sampling, there are many clusters out of which only a few are picked up by random sampling and then the clusters are completely enumerated.

Cluster sampling in practice: Cluster sampling is used primarily because it allows for great economies in data collection costs since the travel related costs etc. are smaller. Although it is statistically less efficient than simple random sampling in most cases, this deficiency may be more than offset by the high economic efficiency that it offers. For example, to get a certain precision level one might need a sample size of 100 under simple random sampling and a sample size of 175 under cluster sampling. However if the cost of data collection is Rs. 20 under simple random sampling and only Rs. 5 under cluster sampling, it would be cost-effective to use cluster sampling.

Cluster sampling is rarely used in single-stage sampling plans. In a national survey, a district might be treated as a cluster and cluster sampling used in the first stage to pick up 15 districts in the country. Some other form of probability sampling like stratified sampling cluster sampling etc. is then used to go to a smaller sampling unit.

If a frame has to be developed, then cluster sampling allows us to save on the cost of developing a frame because frames need to be developed only for the selected clusters and not for the whole population.

Multi-stage and Multi-phase Sampling

In most large surveys one uses multi-stage sampling where the sampling unit is something larger than an individual element of the population in all stages but the final. For example, in a national survey on the demand of fertilizers one might use stratified sampling in the first stage with a district as a sampling unit and the average rainfall in the district as the criterion for stratification. Having obtained 20 districts from this stage, cluster sampling may be used in the second stage to pick up 10 villages in each of the selected districts. Finally, in the third stage, stratified sampling may be used in each village to pick up frames in each of the strata defined with land holding as the criterion.

Multi-phase sampling, on the other hand, is designed to make use of the information collected in one phase to develop a sampling design in a subsequent phase. A study with two phases is often called **double sampling.** The first phase of the study might reveal a relationship between the family consumption of non-aerated beverages and the family income and this information would then be used in the second phase to stratify the population with family income as the criterion.

Activity D

Using a calendar for the current year, identify a systematic sample of size 10 when the sampling ratio is 1/20. (Tomorrow is the first possible member of the sample.)

……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………

Activity E

A lot of debate is going on regarding the grant of statehood to Delhi. If you plan to do a sample survey of 3000 residents in Delhi on this question, what kind of sampling design would you use? In Delhi, many colonies are posh and many others are poor and you believe that the response on statehood is highly dependent on the income level of the respondent.

……………………………………………………………………………………
……………………………………………………………………………………
……………………………………………………………………………………

13.5 NON-PROBABILITY SAMPLING METHODS

Probability sampling has some theoretical advantages over non-probability sampling. The bias introduced due to sampling could be completely eliminated and it is possible to set a confidence interval for the population parameter that is being studied. In spite of these advantages of probability sampling, non-probability sampling is used quite frequently in many sampling surveys. This is so because all are based on practical considerations.

Probability sampling requires a list of all the sampling units and this frame is not available in many situations nor is it practically feasible to develop a frame of say all the households in a city or zone or ward of a city. Sometimes the objective of the study may not be to draw a statistical inference about the population but to get familiar with extreme cases or other such objectives. In a dealer survey, our objective may be to get familiar with the problems faced by our dealers so that we can take some corrective actions, wherever possible. Probability sampling is rigorous and this rigour e.g. in selecting samples, adds to the cost of the study. And finally, even when we are doing probability sampling, there are chances of deviations from the laid out process especially where some samples are selected by the interviewers at site-say after reaching a village. Also, some of the sample members may not agree to be interviewed or not available to be interviewed and our sample may turn out to be a non-probability sample in the strictest sense of the term.

Convenience Sampling

In this type of non-probability sampling, the choice of the sample is left completely to the convenience of the interviewer. The cost involved in picking up the sample is minimum and the cost of data collection is also generally low, e.g. the interviewer can go to some retail shops and interview some shoppers while studying the demand for non-aerated beverages.

However, such samples can suffer from excessive bias from known or unknown sources and also there is no way that the possible errors can be quantified.

Purposive Sampling

Inconvenience sampling, any member of the population can be included in the sample without any restriction. When some restrictions are put on the possible inclusion of a member in the sample, the sampling is called purposive.

Judgment Sampling: In judgment sampling, the judgment or opinion of some experts forms the basis for sample selection. The experts are persons who are believed to have information on the population which can help in giving us better samples. Such sampling is very useful when we want to study rare events, or when members have extreme positions, or even when the objective of the study is to collect a wide cross-section of views from one extreme to the other.

Quota Sampling: Even when we are using non-probability sampling, we might want our sample to be representative of the population in some defined ways. This is sought to be achieved in quota sampling so that the bias introduced by sampling could be reduced.

If in our population, 20% of the members belong to the high income group, 30% to the middle income group and 50% to the low income group and we are using quota sampling, we would specify that the sample should also contain members in the same proportion as in the population e.g. 20% of the sample members would belong to the high income group and so on.

The criteria used to set quotas could be many. For example, family size could be another criterion and we can set quotas for families with family size upto 3, between 4 and 5, and above 5. However, if the number of such criteria is large, it becomes difficult to locate sample members satisfying the combination of the criteria. In such cases, the overall relative frequency of each criterion in the sample is matched with the overall relative frequency of the criterion in the population.

13.6 THE SAMPLE SIZE

How large a sample should be taken in a study? So far in this unit we have not addressed ourselves to this question. At this stage, we will only mention some factors affect the sample size decision and in later units some of these ideas will be gone into in more depth.

One of the most important factors that affect the sample size is the extent of variability in the population. Taking an extreme case, if there is no variability, i.e. if all the members of the population are exactly identical, a sample of size 1 is as good as a sample of 100 or any other number. Therefore, the larger the variability, the larger is the sample size required.

A second consideration is the confidence in the inference made-the larger the sample size the higher is the confidence. In many situations, the confidence level is used as the basis to decide sample size as we shall see in the next unit.

In many real life situations, the factor of overriding importance is the cost of the study and the problem then becomes one of designing a sampling scheme to achieve the highest statistical efficiency subject to the budget for the study. It is here that cluster sampling and convenience sampling score over other more statistically efficient methods of sampling, since the unit cost of data collection is lower.

13.7 SUMMARY

In this unit we have looked at various sampling methods available when one wants to make some inferences about a population without enumerating it completely. We started by looking at some situations where sampling was being done and then found that in many situations sampling may be the only feasible way of knowing something about the population-either because of the time or cost involved, or because of the physical impossibility or practical infeasibility of observing the complete population. Also, sampling can give us adequate results in many applications and can be preferred over complete enumeration as it ensures a higher purity of the data collected, especially when the population is large.

We noted that there are two basic methods of sampling-probability sampling which ensures that every member of the population has a calculable nonzero probability getting included in the sample and non-probability sampling where there is no such assurance. Probability sampling is theoretically superior to non-probability sampling as it helps us in reducing the bias and also allows us to quantify the possible error involved, but non-probability sampling is less rigorous, easy to use, practically feasible and gives adequate results in some applications.

Among the probability sampling methods, simple random sampling works the best when the population is homogeneous but may have many practical limitations when the population is large. Simple random sampling ensures that each of the possible samples of a particular size has an equal probability of getting picked up as the sample selected and it also implies that each element of the population has an equal probability of being included in the sample. Systematic sampling starts with a random start and picking up members after a fixed interval down a list of all members called the sampling frame. If the population can be broken down into smaller, more homogeneous sub-populations or strata, then stratified sampling should be used which allows higher economic efficiency as the cost of data collection per element is reduced if members are physically or otherwise closer to each other as they are in a cluster. Most large studies are based on multi-stage sampling where different sampling methods are used at each stage. In some studies multi-phase sample is also used, especially where the information collected in one phase is used in the sampling design of a later phase.

We have also discussed some of the non-probability sampling methods used in practice. If any member of the population could be included in the sample, we would get a convenience sample. On the other hand, if the entry is subject to the judgment of some expert or experts who have a better knowledge of the population, we would have used judgment sampling and if the sample is made representative of the population by setting quotas for elements satisfying different criteria, this is called quota sampling. Purposive sampling is a genuine name for all non-probability sampling methods where restrictions are used on entry. We have looked at all of these sampling methods to gauge their strengths and weaknesses and also to find their applicability under different conditions.

13.8 SELF-ASSESSMENT EXERCISES

1 List the various reasons that make sampling so attractive in drawing conclusions about the population.

2 What is the major difference between probability and non-probability sampling?

3 A study aimes to quantify the organisational climate in any organisation by administering a questionnaire to a sample of its employees. There are 1000 employees in a company with 100 executives, 200 supervisors and 700 workers. If the employees are **stratified** based on this classification and a sample of 100 employees is required, what should the sample size be from each **stratum,** if proportional **stratified** sampling is used?

4 In question 3 above, if it is known that the standard deviation of the response for executives is 1.9, for supervisors is 3.2 and for workers is 2.1, what should the respective sample sizes be?

Please state for each of the following statements, which of the given response is the most correct:

5 To determine the salary, the sex and the working hours structure in a large multi-storeyed office building, a survey was conducted in which all the employees working on the third, the eighth and the thirteenth floors were contacted. The sampling scheme used was:

a) simple random sampling

b) stratified sampling

c) cluster sampling

d) convenience sampling

6 We do not use extremely large sample sizes because

a) the unit cost of data collection and data analysis increases as the sample size increases-e.g. it costs more to collect the thousandth sample member as compared to the first.

b) the sample becomes unrepresentative as the sample size is increased.

c) it becomes more difficult to store information about large sample size.

d) As the sample size increases, the gain in having an additional sample element falls and so after a point, is less than the cost involved in having an additional sample element:

7 If it is known that a population has groups which have a wide amount of variation within them, but only a small variation among the groups themselves, which of the following sampling schemes would you consider appropriate:

a) cluster sampling

b) stratified sampling

c) simple random sampling

d) systematic sampling

8 One of the major drawbacks of judgement sampling is that

a) the method is cumbersome and difficult to use

b) there is no way of quantifying the magnitude of the error involved

c) it depends **on** only one individual for sample selection

d) it gives us small sample sizes

13.9 FURTHER READINGS

Levin, R.I;, 1987. *Statistics for Management,* Prentice Hall of India: New Delhi..

Mason, R.D., 1986. *Statistical Techniques in Business and Economics,* Richard D. Irwin, Inc: Homewood.

Mendenhall, W.,R.L. Scheaffer and D.D. Wackerly, 1981. *Mathematical Statistics with Applications,* Danbury Press: Boston.

Plane, D.R. and E.B. Oppermann, 1986. *Business and Economic Statistics;* Business Publications, Inc: Plano.