13. SAMPLING THEORY

13.1 Introduction

In a statistical investigation the interest lies in some characteristics relating to a group of individuals and such a group under study is referred to as the *population* or *universe*. The members of a population may be the employees of an industry, the apples in a basket, the various cultivable plots in a village, and so on. The number of members in a population represents its size.

Quite often, it may not be practicable to study the whole population due to limitations of time, money and man-power, or due to the population being infinite, and, as such, we are to depend on the study of a part of the population for determination of the population characteristics. A part of the population which is meant to represent the whole population is called a *sample* and its selection is termed as *sampling*.

13.2 Basic principles of sample surveys

The two basic principles of sample surveys are *validity* and *optimisation*.

By validity of a sample design we mean that the sample should be so selected that the results obtained from it can be interpreted objectively in terms of probability. Validity is ensured by selecting a random (probability) sample so that each of the population members has a definite preassigned probability of being included in the sample.

The principle of optimisation takes into consideration the factors of efficiency and cost. Efficiency is measured by the inverse of the sampling variance of the estimator (which is a function of the sample values used for guessing about an unknown population characteristic) and cost is measured by money or man-hours spent in the whole process. The principle of optimisation is satisfied if a given level of efficiency is reached with minimum cost or maximum efficiency is attained with a given level of cost.

13.3 Different steps in a sample survey

Conducting a sample survey involves three main stages, namely, ***planning stage***, *execution stage* and ***analysis and reporting stage***.

The planning stage contains the following steps :

(a) *Defining the purpose* : The purpose of the survey must be unambiguously stated and the planner should consider the available resources in terms of money and man-power together with the purpose.

(b) *Defining the population* : The aggregate of individuals, called the population, must be clearly specified so that in terms of the coverage of the survey there remains no scope of confusion about the geographical and other boundaries of the population.

(c) *Deciding on the nature of data* : Depending on the purpose of the survey the type of data to be gathered is determined and subsequently a questionnaire or a schedule of enquiry is formed. The set of questions forming a questionnaire should contain brief, simple and relevant questions.

(d) *Selection of method of collection of data* : Decision about the suitable method, mainly between interview method and mail-questionnaire method, is to be taken for collection of necessary data.

(e) *Selection of sampling unit* : The final unit to be sampled for the purpose of the survey is called the sampling unit. For instance, in a socio-economic survey, it is to be decided whether an individual of a family or the family as a whole will be the sampling unit.

(f) *Determination of sampling frame* : On defining the sampling unit, one must verify the availability of the sampling frame, a frame of all members in the population. For example, a list of individuals or households surveyed in a population census form a frame.

(g) *Designing the survey* : Here one must decide the nature of sampling (e.g. random sampling) to be adopted. Also, the details of a pilot survey, if necessary for the main design, is decided. The cost and time factors should be considered before making a final selection of the design.

(h) *Training of personnel* : Before the survey is actually conducted the enumerators and supervisors should be imparted suitable training.

In the execution stage the sampled members in the field are identified and the questionnaires are suitably filled.

The analysis and reporting stage involves the following steps :

(a) *Scrutiny of data* : To verify the authenticity of data, the filled-in questionnaires should be scruitinised with due care. Whenever doubt arises in any questionnaire, it should be sent back to the field personnel for re-survey.

(b) *Tabulation of data* : Tabulation is to be carried out manually or by machine depending on the quantity of data and time alloted for the purpose.

(c) *Analysis of data* : Estimation of population characteristics (*i.e.* parameters) and testing of hypotheses are made after tabulation.

(d) *Reporting* : The details about all the stages of the survey should be presented in the report and proper interpretation of the data together with relevant conclusions should be given.

(e) *Storing of information* : Necessary arrangements should be made for storing of the information usuable for any future survey.

13.4 Advantages of sample survey over complete enumeration

A survey conducted on a suitable sample is called a *sample survey*, while in *complete enumeration* or *complete census* the entire populaton is surveyed. Generally, sample survey is preferred to complete enumeration for the following reasons :

(a) Reduction of cost :

The cost, either in terms of money or in terms of man-hours, is less in a sample survey than in a complete census. Although the cost per unit is usually greater in a sample survey, the total cost is likely to be smaller because a sample comprises only a part of the population.

(b) Greater speed :

Since a sample comprises only some of the population members, the data can be collected and processed more rapidly in sampling than in complete enumeration. This is a vital consideration when the information are urgently needed.

(c) Greater scope :

Sample survey has usually greater scope as compared to complete enumeration because in some cases, highly trained personnel or specialised equipment may be required for collection of data and thus, complete census becomes almost impracticable. Again, a sample survey may have greater coverage as the smaller scale of a sampling process and the better training of the enumerators enable the enquirer to collect more information and to include more areas in the survey.

(d) Greater accuracy :

A sample survey generally gives more accurate and reliable data than a complete census due to the possibility of employing better-trained personnel, effecting better supervision or using better equipments.

Again, the informants are expected to furnish correct information readily when they know that they have been selected from the whole population for the study. The combined effect of sampling and non-sampling errors in sample survey is likely to be much less than the effect of non-sampling errors in complete enumeration.

The error arising due to drawing inferences about the population on the basis of sample is termed sampling error, and the errors mainly arising at the stages of collection and processing of data are termed non-sampling errors. Non-sampling errors are common to both complete enumeration and sample surveys. Generally, the non-sampling error increases, but the sampling error decreases with increase in sample size.

(e) Measure of accuracy :

It should also be noted that complete census generally involves a large amount of non-*sampling errors* and there is no way of measuring the magnitude of such errors. On the other hand, a properly selected sample will give an idea of the magnitude of the *sampling errors* involved in the estimates.

(f) Greater applicability :

In some situations, complete enumeration is ruled out by the nature of the population. If the population is infinite or hypothetical, sampling is the only course available. Again, when the enumeration is a destructive process (i.e., when the character of an item can be determined only by destroying the item in the process, as in testing the life of an electric bulb), sampling is the only method to be used.

However, when the population size is not large or when time and cost are not important factors or when detailed information is required for different segments of the population, a complete census is better than any sampling procedure. Again if information is wanted for every member of the population, complete enumeration is a must. But even in such cases, a small-scale sample survey may be conducted simultaneously to assess the quality of the census data.

13.5 Biases in surveys

Bias is an effect that deprives a statistical result of representativeness by systematically distorting it. It is different from a random error which may distort the result on any occasion but balances out on the average. The different types of bias that are likely to arise in a survey may be broadly classified under two heads :

(1) *Procedural biases*, and

(2) *Sampling biases.*

(1) Procedural Biases : These arise in a complete census as well as in a sample survey. The following are the various kinds of such biases :

(i) *Response biases* : These biases arise from wrong responses of the informants. For instance, one may deliberately over-state one's education or under-state one's age. In many cases, one may furnish wrong answers to guard one's self-interest; such as, a person may give an under-statement of his income, personal savings, etc. This type of bias may also arise due to preference of the people for certain numbers, like multiples of 5 or even numbers.

(ii) *Interviewer biases* : Answers given by the informants after receiving suggestions from the interviewer are likely to be affected by the interviewer's own beliefs and prejudices.

(iii) *Non-response biases* : These may arise in a survey if the respondent is not found at home even after repeated calls or if he either fails or refuses to answer certain questions.

(iv) *Observational biases* : In case of finding the variable value by observation, biases may occur due to the influence of psychological factors. For instance, people with different political affiliations may say widely different figures for size of the gathering in an open meeting of a political party.

(2) Sampling biases : These occur only in sampling. Here the following types may be recognised.

(i) *Biases due to defective sampling method* : If a proper random procedure is not strictly followed, the investigator may use his own judgement or discretion in selecting the units, and, thereby, introduce bias.

(ii) *Biases due to substitution* : When difficulties arise in enumerating a member originally included in the sample, investigators generally substitute it by another convenient member. Clearly, this leads to some bias due to the difference in the characteristics of the substituted and the original members.

(iii) *Biases due to faulty demarcation of sampling units* : In a crop-cutting survey, for example, there is an inclination on the part of the investigators to include some good plants in the sample when they are situated near but outside the selected area; this results in over-estimation.

(iv) *Biases due to wrong selection of the statistic* : For estimating a parameter, one should use the appropriate statistic; otherwise, a constant bias may occur. For example, sample variance is a biased estimator of population variance in simple random sampling with replacement.

13.6 Types of population and types of sampling

A population may be *finite* or *infinite* depending on whether it contains a finite or an infinite number of members. For example, the population of houses in a certain locality is finite, while the population of atmospheric pressures at different points in a room is infinite. Again, the population may be *existent* or *hypothetical.* If the population consists of concrete existent individuals, it is called an existent population. On the other hand, if the individuals of the population do not exist in reality, but exist only in imagination, then the population is hypothetical. For example, the outcome of an infinite number of tosses of a coin represents a hypothetical population of heads and tails.

Sampling is broadly classified as *subjective* and *objective*. In subjective

(ii) *Biases due to substitution* : When difficulties arise in enumerating a member originally included in the sample, investigators generally substitute it by another convenient member. Clearly, this leads to some bias due to the difference in the characteristics of the substituted and the original members.

(iii) *Biases due to faulty demarcation of sampling units* : In a crop-cutting survey, for example, there is an inclination on the part of the investigators to include some good plants in the sample when they are situated near but outside the selected area; this results in over-estimation.

(iv) *Biases due to wrong selection of the statistic* : For estimating a parameter, one should use the appropriate statistic; otherwise, a constant bias may occur. For example, sample variance is a biased estimator of population variance in simple random sampling with replacement.

13.6 Types of population and types of sampling

A population may be *finite* or *infinite* depending on whether it contains a finite or an infinite number of members. For example, the population of houses in a certain locality is finite, while the population of atmospheric pressures at different points in a room is infinite. Again, the population may be *existent* or *hypothetical.* If the population consists of concrete existent individuals, it is called an existent population. On the other hand, if the individuals of the population do not exist in reality, but exist only in imagination, then the population is hypothetical. For example, the outcome of an infinite number of tosses of a coin represents a hypothetical population of heads and tails.

Sampling is broadly classified as *subjective* and *objective*. In subjective sampling, the selection of members depends on the personal judgement or discretion of the sampler, while in objective sampling, there is a specific rule of selection and the selection is independent of the judgement or discretion of the sampler. It should be noted that any haphazard or motivated selection will lead to subjective sampling and such a sampling generally involves a bias of unknown magnitude.

Objective sampling, in its turn, is classified as *non-probabilistic, probabilistic* (or *random*) and *mixed.* When there is no probability assigned to the mode of selection but the sampling is done according to a specific rule, then the sampling is called non-probabilistic. In random sampling, each member has a definite preassigned probability of being selected. Again, if the sampling is partly probabilistic and partly non-probabilistic, it is called mixed sampling.

Suppose, for instance, that we are to select 10 trees from a row of 80 trees in a forest. Starting with the first tree, if we select every 8th tree, then sampling becomes non-probabilistic. If 10 trees are so selected that each of 80 trees has a definite preassigned probability of being included in the sample, then the sampling is probabilistic. If, however, one tree is selected at random from the first 8 trees, and then every 8th tree is taken starting from the selected one, the sampling becomes mixed.

In random sampling, if each member of the population has equal chance of being selected, then the sampling is called simple random sampling. Simple random sampling may be with or without replacement according as the individual selected at any drawing is returned to the population or not before the next drawing.

13.7 Random sampling

If a sample is drawn from a given population in such a manner that each member of the population has a definite pre-assigned probability of being included in the sample, the sampling is called *random sampling* or *probability sampling*. In particular, when each member of the population has an equal chance of being selected, then the sampling is called *simple random sampling*, the sample, thus obtained, is called a *simple random sample*. Simple random sampling may be with or without replacement. To explain these two types, let us consider the selection of a sample of size n from a population of size N.

The sampling is called *simple random sampling with replacement* (**SRSWR**) when the n members of the sample are drawn from the population one by one, and after each drawing the selected member is returned to the population, so that at each drawing any member of the population has the same probability 1/N of being selected. Here the population remains the same both in size and composition throughout the sampling procedure and any of the population members may appear more than once in the sample. It is evident that, in SRSWR total number of possible samples is N^n and each sample has the probability $1/N^n$ of being selected.

Again, the sampling is said to be *simple random sampling without replacement* (**SRSWOR**) if the n members of the sample are drawn from the population one by one, and the member selected at any drawing is not returned to the population, so that at each drawing, each of the remaining members of the population gets the same chance of being included in the sample. Thus, for instance, the probability of selecting any of the remaining N – $(i - 1)$ members at the ith drawing is $1/(N - i + 1)$. In this case, the population size goes on diminishing and the population composition varies as the sampling operation continues. Moreover, no member of the population can occur more than once in the sample, which means that the sample members are necessarily distinct. In SRSWOR, total number of possible samples is ${}^NC_n$, provided the order in which the sample members are obtained is ignored, and each sample has probability

$$\frac{n}{N}\cdot\frac{n-1}{N-1}\cdot\frac{n-2}{N-2}\cdots\frac{1}{N-n+1}=1/{}^NC_n$$

of being selected here, at the kth drawing one has to choose from $N - k + 1$ members one of the $n - k + 1$ members to be included in the sample. When order is considered, the number of possible samples is ${}^NP_n$ and each has the probability

$$\frac{1}{N}\cdot\frac{1}{N-1}\cdot\frac{1}{N-2}\cdots\frac{1}{N-n+1}=1/{}^NP_n$$

to materialise. It may be noted that in SRSWOR the probability that a particular member, say the αth member, of the population will be selected at any drawing, say at the ith drawing, is

$$\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\frac{N-3}{N-2}\cdots\frac{N-i+1}{N-i+2}\cdot\frac{1}{N-i+1}=\frac{1}{N},$$

which is same as in SRSWR.

Illustration 13.6

Show that, whatever be the population proportion, the standard error of sample proportion in SRSWR cannot exceed the value $1/(2\sqrt{n})$, n being the sample size.

It p be sample proportion, then in SRSWR

$$\begin{aligned}\text{Var}(p) &= \frac{\text{PQ}}{n}, \quad \text{where P is population proportion and Q} = 1 - \text{P}.\\ &= \frac{1}{4n}[(\text{P} + \text{Q})^2 - (\text{P} - \text{Q})^2] = \frac{1}{4n}[1 - (\text{P} - \text{Q})^2]\\ &\le \frac{1}{4n}, \ [\because (\text{P} - \text{Q})^2 \ge 0]\end{aligned}$$

$$\therefore \ \text{S.E.}\ (p) \le \frac{1}{2\sqrt{n}}.$$

13.13 Stratified random sampling

In stratified sampling, at first the population is divided into several disjoint subpopulations (or groups), called strata, which are relatively homogeneous within themselves. Then samples are drawn independently from the different strata. If a simple random sample is selected from each stratum, the method is called stratified random sampling.

The main advantages of stratified random sampling are :

(i) This method usually gives better estimates than random sampling; this is because, the strata being homogeneous, the estimates obtained from them are likely to be very satisfactory even when the sizes of the samples from different strata are small. These good estimates, when suitably combined, give a good estimate for the whole population.

(ii) The sampling procedure here ensures that some members from each stratum are included in the study; so the combined sample observations serve as a very good representative of the population under enquiry.

AP

(iii) Stratified random sampling supplies not only an estimate for the whole population, but also separate estimates for the individual strata.

(iv) Sometimes stratified sampling is administratively very convenient. The existing administrative set-up at different zones may be used to facilitate the organisation of field work.

(v) In some situations, different segments of the population demand different sampling procedures. For example, in human populations, people living in hostels, hospitals, prisons, etc. should not be treated in the same way as those living in their homes. Stratified sampling is very useful in such cases.

Suppose the population of size N with mean μ and standard deviation σ for the variable Y is stratified into k strata; the size, mean and standard deviation of hth stratum being respectively N_h, μ_h and σ_h. Thus

$$\sum_{h=1}^{k} \text{N}_h = \text{N}, \ \mu = \frac{\sum_{h=1}^{k} \text{N}_h \mu_h}{\text{N}}$$

$$\mathrm{Var}(\overline{y}_{st}) = \frac{1}{N^2}\sum_h \frac{N_h^2 S_h^2}{n_h}(1 - f_h) \simeq \frac{1}{N^2}\sum_h \frac{N_h^2 S_h^2}{n_h}$$

$$\text{So, } v_{\text{opt}} = \frac{1}{N^2.n}\left(\sum_h N_h S_h\right)^2 \quad \left(\text{since } n_h = n.\frac{N_h S_h}{\sum\limits_h N_h S_h}\right).$$

$$\text{Now } v_{\text{ran}} - v_{\text{prop}} = \frac{S^2}{n} - \frac{\sum\limits_h N_h S_h^2}{Nn}$$

$$= \frac{1}{n}\left(S^2 - \frac{\sum\limits_h N_h S_h^2}{N}\right) = \frac{1}{n}\left(\frac{NS^2 - \sum\limits_h N_h S_h^2}{N}\right)$$

$$= \frac{1}{nN}\sum_h N_h(\mu_h - \mu)^2 \geq 0$$

or, $v_{\text{ran}} \geq v_{\text{prop}}$.

The equality sign holds if and only if $\mu_h = \mu$, for each h, i.e. iff the stratum means are equal.

Again, we have,

$$v_{\text{prop}} - v_{\text{opt}} = \frac{\sum\limits_h N_h S_h^2}{Nn} - \frac{\left(\sum\limits_h N_h S_h\right)^2}{N^2 n} = \frac{1}{Nn}\left[\sum_h N_h S_h^2 - \frac{\left(\sum\limits_h N_h S_h\right)^2}{N}\right]$$

AP

$$= \frac{1}{Nn}\sum_h N_h\left(S_h - \overline{S}\right)^2 \geq 0 \quad \left(\overline{S} = \frac{\sum\limits_h N_h S_h}{N}\right)$$

or, $v_{\text{prop}} \geq v_{\text{opt}}$

Here the equality sign holds iff $S_h = \overline{S}$ for all h, i.e. iff the stratum variances are all equal.

Thus finally, we have,

$$v_{\text{opt}} \leq v_{\text{prop}} \leq v_{\text{ran}}.$$

13.14 Systematic sampling

When a list of sampling units is available, this sampling technique is frequently used because it is operationally very convenient and at the same time ensures for every unit equal probability of being included in the sample. Suppose there are N units in the population which are numbered from 1 to N, and a sample of n units is to be chosen such that $\frac{N}{n} = k$ is an integer. Systematic sampling (more specifically, *lineàr systematic sampling*) then consists in selecting one unit at random from the first k units and then selecting every subsequent kth unit, k is called the sampling interval. This is a case of mixed sampling, which is partly probabilistic and partly non-probabilistic.

Suppose a population consists of $N = 6$ observations $Y_1, Y_2, ..., Y_6$ and we want a linear systematic sample of size $n = 2$. So, sampling interval is $k = \frac{N}{n} = 3$. Then the possible samples are as follows :

Sample No.	Sample observations	Sample mean ($\overline{y}$)
1	Y_1, Y_4	$(Y_1 + Y_4)/2$
2	Y_2, Y_5	$(Y_2 + Y_5)/2$
3	Y_3, Y_6	$(Y_3 + Y_6)/2$

Each sample has probability $\frac{1}{3}$ to occur. So, $E(\overline{y}) = \frac{1}{6}(Y_1 + Y_2 + ... + Y_6) = \overline{Y}$, the population mean.

This shows that in linear systematic sampling, the sample mean is unbiased for the population mean.

It is much easier and quicker to draw a linear systematic sample and the work may be done even by laymen. Again, this sampling procedure provides very precise estimates when the units similar to one another with respect to the variable under consideration or an associated variable are put side by side in the list of sampling units. But if there be some periodic feature in the list and the sampling interval is equal to (or a multiple of) the period, then this method may give highly biased estimates.

This method of sampling cannot be used when the sampling interval $\frac{N}{n}$ is not an integer. In such a situation the technique applied is termed as *circular systematic sampling*. Here at first one unit is selected at random from N units of the population and then alongwith this unit every k th unit thereafter (k being the integer nearest to $\frac{N}{n}$) is included in a cyclical procedure. The selection process is continued until n sampling units are obtained.

It is noteworthy that circular systematic sampling is more general than the linear systematic sampling due to the fact that the former reduces to the linear form when $\frac{N}{n}$ is an integer.

13.15 Cluster sampling

This method of sampling consists in forming suitable clusters of sampling units and surveying all the units in a sample of clusters, chosen according to some apprqpriate sampling scheme. For example, in a human population survey where each person is a sampling unit, one may select a sample of households and then collect data from all persons in the selected households. Usually clusters are formed by grouping nearby units or units which can be conveniently surveyed together.

Cluster sampling is resorted to in many surveys because it is operationally more convenient and less costly than sampling of individual units due to the possible saving in time for journey, identification, contact, etc. In general, for a given total number of sampling units, cluster sampling is less efficient than sampling of individual units since the latter is expected to cover a better cross-section of the population than the former due to the tendency of the units in a cluster to be more or less similar.

13.16 Sampling distributions associated with binomial and Poisson population

If $x_1, x_2, ..., x_n$ be independent and random sample observations from a population, then we know that each of the observations has the same marginal distribution and this common distribution is identical with the distribution of the variable in the population.

We shall find the sampling distribution of sample total $x_1 + x_2 + \cdots + x_n = \text{X(say)}$ when the sample observations are from binomial and Poisson populations. In doing so, we shall use the moment-generating function which, when exists, is unique and completely determines the distribution of the random variable.

Case I : **Samples from binomial population :**

Suppose the sample observations $x_1, x_2, ..., x_n$ are randomly and independently taken from a binomial distribution with parameters m and p. Then for each i, x_i follows binomial distribution with parameters m and p.

AP

The m.g.f. of the distribution of x_i, for each i, is

$$\text{M}_{x_i}(t) = \text{E}(e^{tx_i}) = \sum_{x_i=0}^{m} e^{tx_i} \cdot \binom{m}{x_i} p^{x_i} q^{m-x_i}, \quad q = 1 - p$$

$$= \sum_{x_i=0}^{m} \binom{m}{x_i} (pe^t)^{x_i} q^{m-x_i} = (q + pe^t)^m$$

$$\therefore \text{M}_x(t) = \text{E}(e^{tx}) = \text{E}\left(e^{t\sum_{i=1}^{n} x_i}\right) = \text{E}\left(\prod_{i=1}^{n} e^{tx_i}\right) = \prod_{i=1}^{n} \text{E}(e^{tx_i}) = \prod_{i=1}^{n} (q + pe^t)^m = (q + pe^t)^{nm},$$

which is m.g.f. of a binomial distribution with parameters nm and p. So, the sampling distribution of the sample total $\text{X} = x_1 + x_2 + \cdots + x_n$ is binomial with parameters nm and p.

Note : If $x_1, x_2, ..., x_n$ follow independently binomial distribution with parameters (m, p),