FUNDAMENTALS OF SURVEY SAMPLING

V.K. Gupta, U.C. Sud and Rajender Parsad
I.A.S.R.I., New Delhi-110012

Need for Statistical Data

Since the beginning of the twentieth century the economic and social life of the people and the functional system of industry and business, educational and medical facilities and other activities of the community have undergone substantial changes due to spectacular developments in the field of science and technology. Now the emphasis is on specialization in mass production and utilization of goods and services of a given type with a view to get the maximum possible benefit per unit of cost. Considerable planning is required in a large-scale projects and any rational decision regarding efficient formulation and execution of suitable plans and projects or an objective assessment of their effectiveness, whether in the filed of industry, business or governmental activities, has necessarily to be based on objective data regarding resources and needs. There is, therefore, a need for various types of statistical (quantified) information to be collected and analyzed in an objective manner and presented suitably so as to serve as a sound basis for taking policy decisions in different fields of human activity. In modern times, the primary users of statistical data are the state, industry, business, scientific institutions, public organizations and international agencies.

For instance, to execute its various responsibilities, the state is in need of a variety of information regarding different sectors of the economy, sections of people and geographical regions in the country as well as information on the available resources such as manpower, cultivable land, forests, water, minerals and oil. If the resources were unlimited, planning would be relatively simple as it would consist in just providing each one with what he needs in terms of money, material, employment, education etc. But such a situation is only hypothetical, as in reality the resources are limited and the needs are usually not well defined and are elastic.

Therefore, for the purpose of proper planning fairly detailed data on the available resources and on the needs are to be collected. For example, the country is in need of data on production and consumption of different types of products to enable it to take objective decisions regarding its import and export polices. Statistical information on the cost of living of different categories of people living in various parts of the country is of importance in shaping its policies in respect of wage and price levels.

Complete Enumeration Survey

One way of obtaining the required information at regional and country level is to collect the data for each and every unit (person, household, field, factory, shop, etc as the case may be) belonging to the population or universe, which is the aggregate of all units of a given type under consideration and this procedure of obtaining information is termed complete enumeration survey.

The effort, money and time required for carrying out complete enumeration surveys to obtain the different types of data will, generally, be extremely large. However, if the information is required for each and every unit in the domain of study, a complete enumeration survey is clearly necessary. Examples of such situations are population census, agricultural census, census of retail stores, income tax assessment where the income of each individual is assessed and taxed, preparation of voters' list for election purposes and recruitment of personnel in an establishment etc.

But there are many situations, where only summary figures are required for the domain of study as a whole or for group of units and in such situations collection of data for every unit is only a means to an end and not the end itself. It is worth mentioning that exact planning for the future is not possible, since this would need accurate information on the resources that would be available and on the needs that would have to be satisfied in future. In general, past data are used to forecast the resources and the needs of the future and hence there is some element of uncertainty in planning. Because of this uncertainty, only broad (and not exact) allocations of the resources are usually attempted. Thus some margin of error may be permitted in the data needed for planning, provided this error is not large enough to affect the broad allocations.

Need for Sampling

Considering that some margin of error is permissible in the data needed for practical purposes, an effective alternative to a complete enumeration survey can be a sample survey where only some of the units selected in a suitable manner from the population are surveyed and an inference is drawn about the population on the basis of observations made on the selected units.

Sampling is a day-to-day life experience for all of us. We often go to library and thumb through only a few pages of the book picked up from the shelves of the library to decide if it meets our course requirements. A housewife makes judgment about the quality of cereals she finally buys for domestic use by examination of only a handful of grains taken from the lot offered for sale. A production manager also does the same when s/he closely observes only a few items of raw material to be sure of its quality before placing the bulk purchase order.

It can be easily seen that compared to a sample survey, a complete enumeration survey is time-consuming, expensive, has less scope in the sense of restricted subject coverage and is subject to greater coverage, observational and tabulation errors. In certain investigations, it may be essential to use specialized equipment or highly trained field staff for data collection making it almost impossible to carry out such investigations except on a sampling basis. Besides, in case of destructive surveys, a complete enumeration survey is just not practicable. Thus, if the interest is to obtain the average life of electric bulbs in a batch then one will have to confine the observations, of necessity, to a part (or a sample) of the population or universe and to infer about the population as a whole on the basis of the observations on the sample. However, since an inference is made about the whole from a part in a sample survey, the results are likely to be different from the population values and the differences would depend on the selected part or sample. Thus the information provided by a sample is subject to a kind of error which is known as sampling error. On the other hand, as only a part of the population is to be surveyed, there is greater scope for eliminating the ascertainment or observational errors by proper controls and by employing trained personnel than is possible in a complete enumeration survey. It is of interest to note that if a sample survey is carried out according to certain specified statistical principles, it is possible not only to estimate the value of the characteristic for the population as a whole on the basis of the sample data, but also to get a valid estimate of the sampling error of the estimate.

There are various steps involved in the planning and execution of a sample survey. One of the principal steps in a sample survey relate to methods of data collection.

Methods of Data Collection

The different methods of collecting data are:

1. Physical observation or measurement
2. Personal interview
3. Mail enquiry
4. Telephonic enquiry
5. Web-based enquiry
6. Method of Registration
7. Transcription from records

The first six methods relate to collection of primary data from the units/ respondents directly, while the last one relates to the extraction of secondary data, collected earlier generally by one or more of the first six methods. These methods have their respective merits and therefore sufficient thought should be given in selection of an appropriate method(s) of data collection in any survey. The choice of the method of data collection should be arrived at after careful consideration of accuracy, practicability and cost from among the alternative methods.

a) Physical observations or measurement

Data collection by physical observation or measurement consists in physically examining the units/ respondents and recording data as a result of personal judgment or using a measuring instrument by the investigator. For instance, in a crop cutting experiment for estimating the yield of a crop, the plot is demarcated, the crop in the selected plot is harvested and the produce is weighted to estimate the produce per unit area. Data obtained by this method are likely to be more accurate but may often prove expensive.

b) Personal interview

The method of personal interview consists in contacting the respondents and collecting statistical data by questioning. In this case, the investigator can clearly explain to the respondents the objectives of the survey and the exact nature of the data requirements and persuade them to give the required information thus reducing the possibility of non-response arising from non-cooperation, indifference etc. Further, this method is most suitable for collecting data on conceptually difficult items from respondents. However, this method depends heavily on the availability of well trained interviewer.

c) Mail enquiry

In a mail enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being sent and collected back through an agency such as the postal department. This method is likely to cost much less as compared to above methods. However, the response may not always be satisfactory depending upon the cooperation of the respondents, the type of questionnaire and the design of the questionnaire. In developing countries where a large proportion of the population is illiterate, the method of mailed questionnaire may not even be feasible.

d) Telephonic enquiry

In a telephonic enquiry, data are collected by questioning the respondents. This method provides an opportunity of two-way communication and thus can reduce the possibility of item non-response. However, this method can be used only for those surveys in which all units of target populations have telephone otherwise it will cause bias in the results.

e) Web-based enquiry

The increasing popularity and wide availability of World Wide Web technologies provide a new mode of data collection. In web-based enquiry, data are collected by obtaining questionnaires filled in by the respondents, the questionnaires being posted on the net. One important advantage of using computer technology in data collection is to minimize the loss of data owing to incomplete or incorrectly completed data sets by using Client side validation. In an era of information superhighway, this method is one of the fastest means of data collection. However, in developing countries where a large proportion of the population does not have access to Internet, the method of web-based enquiry may not serve the purpose for most of the surveys. Various Internet sites are using this method for opinion poll on certain issues.

f) Method of registration

In the registration method, the respondents are required to register the required information at specified places. The vital statistics registration system followed in many countries provide an illustration of the registration method. The main difficulty with this method, as in the case of the mail enquiry, is the possibility of non-response due to indifference, reluctance, etc. on the part of informants to visit the place of registration and supply the required data.

g) Transcription from records

The method of transcription from records is used when the data needed for a specific purpose are already available in registers maintained in one or more places, making it no more necessary to collect them directly from the original units at much cost and effort. The method consists in compiling the required information from the registers for the concerned units. This method is extensively used since a good deal of government and business statistics are collected as by-product of routine administrative operations.

Various Concepts and Definitions

a) Population

The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

Population may be finite or infinite. Population is said to be finite when the observations comprising the population are limited and precisely known. Population is infinite when the observations constituting the population are unlimited and unknown. Infinite populations generally involve continuous processes. For example, the output produced by a machine as long as the machine continues to operate under a given set of conditions.

b) Sampling Unit

Elementary units or group of such units which besides being clearly defined, identifiable and observable, are convenient for purpose of sampling are called sampling units. For instance, in a family budget enquiry, usually a family is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit. In a industrial survey of estimating the total production of an item, an industry manufacturing that particular item is a sampling unit.

c) Sampling Frame

A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages in India or counties in the United States. The list of all industries manufacturing a particular item, a list of households in consumption surveys, a list of schools in educational surveys, a list of banks in a particular region for performance appraisal surveys. The frame should be up to date and free from errors of omission and duplication of sampling units.

d) Random Sample

One or more sampling units selected from a population according to some specified procedures are said to constitute a sample. The sample will be considered as random or probability sample, if its selection is governed by ascertainable laws of chance. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a predetermined probability of selection. For example, if a population consists of the N sampling units (clearly defined, identifiable and observable)

$$U_1, U_2,\ldots,U_i,\ldots,U_N$$

then we may select a sample of n units by selecting them unit by unit with equal probability for every unit at each draw with or without replacing the sampling units selected in the previous draws.

e) Non-random sample

A sample selected by a non-random process is termed as non-random sample. A Non-random sample, which is drawn using certain amount of judgment with a view to getting a representative sample is termed as judgment or purposive sample. In purposive sampling units are selected by considering the available auxiliary information more or less subjectively with a view to ensuring a reflection of the population in the sample. This type of sampling is seldom used in large-scale surveys mainly because it is not generally possible to get strictly valid estimates of the population parameters under consideration and of their sampling errors due to the risk of bias in subjective selection and the lack of information on the probabilities of selection of the units.

f) Population parameters

Suppose a finite population consists of the N units $U_1, U_2,\ldots,U_N$ and let Y_i be the value of the variable y, the characteristic under study, for the i^{th} unit U_i, $(i=1,2,\ldots,N)$. For instance, the unit may be a farm and the characteristic under study may be the area under a particular crop. Any function of the values of all the population units (or of all the observations constituting a population) is known as a population parameter or simply a parameter. Some of the important parameters usually required to be estimated in surveys are population total $Y = \sum_{i=1}^{N} Y_i$ and population mean $\overline{Y} = \sum_{i=1}^{N} Y_i / N$.

g) Statistic, Estimator and Estimate

Suppose a sample of *n* units is selected from a population of N units according to some probability scheme and let the sample observations be denoted by $y_1, y_2, \ldots, y_n$. Any function of these values which is free from unknown population parameters is called a statistic.

An estimator is a statistic obtained by a specified procedure for estimating a population parameter. The estimator is a random variable and its value differs from sample to sample and the samples are selected with specified probabilities. The particular value, which the estimator takes for a given sample, is known as an estimate.

h) Sample design

A clear specification of all possible samples of a given type with their corresponding probabilities is said to constitute a sample design. For example, suppose we select a sample of *n* units with equal probability with replacement, the sample design consists of N^n possible samples (taking into account the orders of selection and repetitions of units in the sample) with $1/N^n$ as the probability of selection for each of them, since in each of the *n* draws any one of the N units may get selected.

Similarly, in sampling *n* units with equal probability without replacement, the number of possible samples (ignoring orders of selection of units) is $\binom{N}{n}$ and the probability of selecting each of the samples is $1 \Big/ \binom{N}{n}$.

i) Unbiased Estimator

Let the probability of getting the i-th sample be P_i and let t_i be the estimate, that is, the value of an estimator t of the population parameter θ based on this sample ($i=1,2,\ldots,M_o$), M_o being the total number of possible samples for the specified probability scheme. The expected value or the average of the estimator t is given by $E(t) = \sum_{i=1}^{M_o} t_i P_i$

An estimator t is said to be an unbiased estimator of the population parameter θ if its expected value is equal to θ irrespective of the y-values. In case expected value of the estimator is not equal to population parameter, the estimator t is said to be a biased estimator of θ. The estimator t is said to be positively or negatively biased for population parameter according as the value of the bias is positive or negative.

j) Measures of error

Since a sample design usually gives rise to different samples, the estimates based on the sample observations will, in general, differ from sample to sample and also from the value of the parameter under consideration. The difference between the estimate t_i based on the i^{th} sample and the parameter, namely (t_i - θ), may be called the error of the estimate and this error varies from sample to sample. An average measure of the divergence of the different estimates from the true value is given by the expected value of the squared error, which is

$$M(t) = E(t-\theta)^2 = \sum_{i=1}^{M_0}(t_i-\theta)^2 P_i$$

and this is know as mean square error (MSE) of the estimator. The MSE may be considered to be a measure of the accuracy with which the estimator t estimates the parameter.

The expected value of the squared deviation of the estimator from its expected value is termed sampling variance. It is a measure of the divergence of the estimator from its expected value and is given by

$$V(t) = \sigma^2 t = E\{t - E(t)\}^2 = E(t)^2 - \{E(t)\}^2$$

This measure of variability may be termed as the precision of the estimator t. The MSE of t can be expressed as the sum of the sampling variance and the square of the bias. In case of unbiased estimator, the MSE and the sampling variance are same. The square root of the sampling variance σ(t) is termed as the standard error (SE) of the estimator t. In practice, the actual value of σ(t) is not generally known and hence it is usually estimated from the sample itself.

k) Confidence interval

The frequency distribution of the samples according to the values of the estimator t based on the sample estimates is termed as the sampling distribution of the estimator t. It is important to mention that though the population distribution may not be normal, the sampling distribution of the estimator t is usually close to normal, provided the sample size is sufficiently large. If the estimator t is unbiased and is normally distributed, the interval $\{t \pm KSE(t)\}$ is expected to include the parameter θ in P% of the cases where P is the proportion of the area between –K and +K of the distribution of standard normal variate. The interval considered is said to be a confidence interval for the parameter θ with a confidence coefficient of P% with the confidence limit t – K SE(t) and t + K SE(t). For example, if a random sample of the records of batteries in routine use in a large factory shows an average life t = 394 days, with a standard error SE(t) = 4.6 days, the chances are 99 in 100 that the average life in the population of batteries lies between

t_L = 394 - (2.58)(4.6) = 382 days
t_U = 394 + (2.58)(4.6) = 406 days

The limits, 382 days and 406 days are called lower and upper confidence limits of 99% confidence interval for t. With a single estimate from a single survey, the statement "θ lies between 382 and 406 days" is not certain to be correct. The "99% confidence" figure implies that if the same sampling plan were used may times in a population, a confidence statement being made from each sample, about 99% of these statements would be correct and 1% wrong.

l) Sampling and Non-sampling error

The error arising due to drawing inferences about the population on the basis of observations on a part (sample) of it is termed sampling error. The sampling error is non-existent in a complete enumeration survey since the whole population is surveyed.

The errors other than sampling errors such as those arising through non-response, in-completeness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed.

The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size.

As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

m) Basic Properties of Populations

The basic properties of a population that helps in drawing a representative sample of the population are: Variability in sampling units; Variability within Limits and uniformity in variations. If all units in a population are alike, then a single unit may serve as representative of the population. Too large variability also causes the problem of representative samples.

The Fundamental Principle of Sample Design

With every sample design is associated the cost of the survey and the precision (measured in terms of sampling variance) of the estimates made. The designs should be practical in the sense that it is possible to carry it through according to desired specifications. Out of all these designs, the one to be preferred is that which gives the highest precision for a given cost of the survey or the minimum cost for a specified level of precision. This is the guiding principle of sample design.

The aim of a sample survey is to estimate the unknown population parameters like total, ratio or median based on a random sample drawn by some specified rule from the given population. A sample is a subset of population. The principal advantages of sampling as compared to complete enumeration are reduction in cost, greater speed, wider scope, higher accuracy and more importantly the quantification of uncertainty, i.e., the error.

Sample surveys can be conceptually divided into two broad categories - descriptive and analytical.

In descriptive surveys certain, usually few, population characteristics need to be precisely and efficiently estimated. For example, in a business survey, the average salaries for different occupational groups are to be estimated, based on a sample of business establishments. Statistical efficiency of the sampling design is of great importance.

Analytical surveys, on the other hand, are often multi-purpose so that a variety of subject matters are covered. In the construction of a sampling design for an analytical survey, a feasible overall balance between statistical efficiency and cost efficiency is sought. For example, in a survey where personal interviews are to be carried out, a sampling design can include several stages so that in the final stage all the members in a sample household are interviewed. While this kind of clustering decreases statistical efficiency it often provides the most practical and economical method for data collection. Statistical testing and modeling play more important roles in analytical surveys than in descriptive surveys

In survey sampling a fixed finite population is under consideration, where the population elements are labelled so that each element can be identified. Probability sampling provides a flexible device for the selection of a random sample from such a fixed population. A key property of probability sampling is that for each population element a positive probability of selection is assigned and this probability need not be equal for all the elements. A specific sampling scheme is used in drawing the sample.

The term sampling scheme refers to the collection of techniques or rules for the selection of the sample. The composition of the sample is thus randomized according to the probabilistic definition of the sampling scheme. In principle, a large number of different samples could be drawn from a population using a particular sampling scheme. The actual sample is one of these possible samples. Under a sampling scheme it is possible to state the selection probability for a sample. The sampling distribution generated through a particular sampling design determines the statistical properties (expectation and sampling error) of random quantities such as the sample total, sample ratio etc. calculated from the sample drawn under the actual sampling scheme. Generally, in practice the terms sampling scheme and sampling design are interchangeably used, although some what different definitions have been given for these concepts in the literature. These terms refer roughly to the way in which we draw a sample from the fixed population.

One of the vital issues in sample survey is the choice of proper sampling techniques. In the choice of a sampling method there are some methods of selection while some others are control measures which help in grouping the population before the selection process. The basic sampling techniques which are commonly employed are simple random sampling, systematic sampling and sampling with unequal probabilities of selection of units particularly with probability proportional to size. Among the control measures are procedures such as stratified sampling, cluster sampling and multi-stage sampling, etc.

We shall describe in brief the different procedures of sample selection in the following sections.

Simple Random Sampling

Simple random sampling can be regarded as the basic form of probability sampling applicable to situations where there is no previous information available on the population structure. There are two methods of simple random sampling viz. with replacement and without replacement. Sampling with replacement means that each unit selected in the sample is returned to the population, before the next is drawn. As the population size remains the same after each draw, not only is the probability of each unit being selected in the sample is 1/N at each draw, it remains same even when included in the sample more than once. In case of a random sampling with replacement, at any draw all N members of the population are given an equal chance of being drawn, no matter how often they have already been drawn. The with replacement assumption simplifies the estimation under complex sampling designs and is often adopted, although in practice sampling is usually carried out under a without replacement type scheme. If N denotes the population size and n is sample size. Then the number of all possible samples in simple random sampling with replacement (SRSWR) is N^n. The probability of drawing any of these N^n samples is $1/N^n$. For example, if there is a population of size N = 5, with unit numbers as 1, 2, 3, 4, 5. Then 25 = 5^2 all possible samples selected through SRS WR are (1, 1); (1, 2); ()1, 3); (1, 4); (,1 5); (2,1); (2,2); (2,3); (2,4); (2,5); (3,1); (3,2); (3,3); (3,4); (3,5); (4,1); (4,2); (4,3); (4,4); (4,5); (5,1); (5,2); (5,3); (5,4); (5,5). Each sample can be selected with equal probability of 1/25. Another, way is to select the two units, by selecting one unit at a time. Therefore, probability of selection of one unit at a time is 1/5. Since, each unit is selected 10m times in all possible 25 samples. Therefore, the probability of inclusion of each unit in the sample is 10/25 = n/N.

Simple random sampling without replacement (SRSWOR) is a method of selecting n units out of the N such that every one of the $\begin{pmatrix} N \\ n \end{pmatrix}$ distinct samples has an equal change of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to *N*. A series of random numbers between 1 and *N* is then drawn, either by means of a table of random numbers or by means of a computer program that produces such a table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these numbers constitute the sample. Since a number that has been drawn is removed from the population for all subsequent draws, this method is also called random sampling without replacement. Continuing with the same example as in SRSWR, the number of all possible samples with SRSWOR without giving any importance to the ordering of units are (1,2); (1,3); (1,4); (1,5); (2,3); (2,4); (2, 5); (3,4); (3,5); (4,5). The total number of samples is $\begin{pmatrix} 5 \\ 2 \end{pmatrix} = 10.$ Any of these samples have equal probability of selection. Therefore, the probability of selection of each of the samples is 1/10. Similar to SRSWR, the probability of selection of each unit at any draw is 1/N. For example, if the ith unit is selected at first draw, its probability of selection in 1/N. Probability of drawing the ith unit at 2nd draw is 1/(N-1); Probability of selecting ith unit at 3rd draw is 1/(N-2); similarly the Probability of drawing the ith unit at rth draw is 1/[N-(r-1)]. Therefore, probability of selecting the unit ith unit at rth draw = (1-1/N)(1-1/(N-1))…(1-1/(N-r+2))(1/(N-r+1)=1/N. Its simplification for r=2 is (1-1/N).(1/(N-1)) = $\frac{N-1}{N}.\frac{1}{N-1} = \frac{1}{N}$. Probability of inclusion of any unit in the sample is n/N.

Obviously, the difference between with replacement and without replacement sampling becomes less important when the population size is large and the sample size is noticeably smaller than it.

Simple random sampling serves as a baseline for comparing the relative efficiency of other sampling methods.

Procedure of Selecting a Random Sample

Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

1) Lottery Method
2) Use of Random Number Tables

Lottery Method: Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits/tickets are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits/tickets may be drawn one by one and may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits/tickets and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method. Therefore, this method is generally discouraged.

Use of Random Number Tables: A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1, 2,…,9 appear independent of each other. Some random number tables in common use are:

(a) Tippett's random number Tables
(b) Fisher and Yates Tables
(c) Kendall and Smith Tables
(d) A million random digits Table.

Random number tables are the tables of digits *0, 1, 2, …,* 9 each digit having an equal chance of selection at any draw. In *1927*, Tippett produced *41,600* digits (from *0* to *9*) arranged in sets of *4* in several columns and spread over *26* pages. This was followed by another publication by two great pioneering statisticians, Sir Ronald Alymer A Fisher and Frank Yates, which contained *15,000* digits formed by listing the *15 - 19*th digits in some *20* figure logarithm tables. Rand Corporation (1955) published tables containing 1 million digits. Kendall and Smith (1938) published tables with 100,000 digits.

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digit numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is to select a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The procedure of selection of sample through the use of random numbers is, therefore, modified and some of these modified procedures are:

(i) Remainder Approach
(ii) Quotient Approach

Remainder Approach: Let *N* be an *r*-digit number and let its *r*-digit highest multiple be *N'*. A random number *k* is chosen from *1* to *N'* and the unit with serial number equal to the remainder obtained on dividing *k* by *N* is selected, i.e. the selected number is reduced mod (N). If the remainder is zero, the last unit is selected. As an illustration, let *N = 123*, the highest three-digit multiple of *123* is *984*. For selecting a unit, one random number from *001* to *984* has to be selected. Let the random number selected be *287*. Dividing *287* by *123* gives the remainder as *41*. Hence, the unit with serial number *41* is selected in the sample. Suppose that another random number selected is *245*. Dividing *245* by *123* leaves *122* as remainder. So the unit bearing the serial number *122* is selected. Similarly, if the random number selected is *369*, then dividing *369* by *123* leaves remainder as *0*. So the unit bearing serial number *123* is selected in the sample.

Quotient Approach: Let N be an r-digit number and let its r-digit highest multiple be N^* such that $N^* / N = d$. A random number *k* is chosen from *0* to $\left(N^* - 1\right)$. Dividing *k* by *d*, the quotient *q* is obtained and the unit bearing the serial number (*q - 1*) is selected in the sample. The selected number is reduced mod(N). For example, if q – 1 = -1, then unit bearing serial number N – 1 is selected and if q – 1 = 0, then unit bearing serial number N is selected. As an illustration, let *N = 16* and hence $N^* = 96$ and *d = 96/16 = 6*. Let the two-digit random number chosen be *65* which lies between *0* and *95*. Dividing *65* by *6*, the quotient is *10* and hence the unit bearing serial number (*10 - 1*) = *9* is selected in the sample. Further, if the random number selected is *4*, then the quotient is *4/6 = 0*, and *q – 1 = -1*. The unit selected is *15*. Similarly, if the random number selected is *9*, then the quotient is *9/6 = 1*, and *q – 1 = 0*. The unit selected is *16*.

Estimation of Population Total

Let Y be the character of interest and $Y_1, Y_2, \cdots, Y_k, \cdots, Y_N$ be the values of the character on *N* units of the population. Further, let $y_1, y_2, \cdots, y_k, \cdots, y_n$ be the sample of size *n* selected by simple random sampling without replacement. For the total $T = \sum_{k=1}^{N} Y_k$, we have an estimator $\hat{t} = N \sum_{k=1}^{n} y_k / n = N\bar{y}_n$ i.e., the sample mean $\bar{y}_n$ multiplied by the population size *N*.

The estimator can be expressed as

$\hat{t} = \sum_{k=1}^{n} w_k y_k = \left(N/n\right) \sum_{k=1}^{n} y_k$, where $w_k = N/n$.

The constant N/n is the sampling weight and is the inverse of the sampling fraction n/N. Alternatively, an estimator for the population total can be written by first defining the inclusion probability of a population element. Under SRS, the inclusion probability of a population element *k* is π_k = *n/N* or the same constant for every population element. Based on the inclusion probabilities an estimator of the total can be expressed as a more general Horvitz-Thompson-type estimator

$$\hat{t}_{ht} = \sum_{k=1}^{n} w_k y_k = \sum_{k=1}^{n} \frac{1}{\Pi_k} y_k = \frac{N}{n} \sum_{k=1}^{n} y_k \ .$$

In this case, the estimator $\hat{t}$ *and* $\hat{t}_{ht}$ obviously coincide, because the inclusion probabilities π_k = *n/N* are equal for each k. The Horvitz-Thompson-type estimator is often used for example, with probability-proportional to size sampling where inclusion probabilities vary. The estimator has the statistical property of unbiasedness in relation to the sampling design. A design variance for $\hat{t}$ is

$$V_{SRS}\ (\ \hat{t}\ ) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sum_{k=1}^{N}\left(Y_k - \bar{Y}\right)^2 / (N - 1)$$

where $\bar{Y} = \sum_{k=1}^{N} Y_k / N$ is the population mean and $S^2 = \sum_{k=1}^{N} (Y_k - \bar{Y})^2 / (N-1)$ is the population variance.

An unbiased estimator of the design variance V$_{SRS}$($\hat{t}$) of the estimator $\hat{t}$ of the total is given by

$$\hat{V}_{SRS}\,(\ \hat{t}\ ) = N^2\left(1 - \frac{n}{N}\right)\sum_{k=1}^{n} (y_k - \bar{y})^2 / n(n-1)$$

$$= N^2\left(1 - \frac{n}{N}\right) s^2 / n$$

where $\bar{y} = \sum_{k=1}^{n} y_k / n$ is the sample mean and s^2 is an unbiased estimator of the population variance S^2. The variance of $\bar{y}$ is $Var(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 = \frac{N-n}{N} . \frac{S^2}{n}$. The unbiased estimator of the $Var(SRS)$ is $Var(\bar{y})$ is $V\hat{a}r(SRS) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2$. When $N \to \infty$, the variance reduces to $Var(\bar{y}) = S^2 / n \approx \sigma^2 / n$, which is equal to that of SRSWR. In the expression of variance, $\frac{N-n}{N}$ is called the finite correction factor.

Stratified Random Sampling

Stratified random sampling selects simple random samples from mutually exclusive subpopulations, or strata, of the population.

In stratified random sampling the population is divided into strata such that the data of interest are fairly homogeneous within a given stratum. Therefore, in study of annual incomes of medical doctors in a large city, it might be desirable to stratify the population according to the type of practice or specialty and then sample within these specialties. We would then expect the variability of the incomes within each specialty to be considerably less than within the entire population of incomes. This reduced variability within each stratum will yield a more precise estimate of the population mean so that a stratified random sample of a given size will be more efficient than a simple random sample of the same size. To achieve this homogeneity within the different strata, the stratification must be formed in such a way that there is some relationship between being in a certain stratum and the characteristic under study. In our illustration, the specialty of a doctor is related to annual income.

Stratification of a population results in strata of various sizes. Consideration must therefore be given to the sizes of the random samples selected from these strata. One procedure, using **proportional allocation**, chooses sample sizes proportional to the sizes of the different strata.

Sample sizes for proportional allocation: If we divide a population of size N into L strata of sizes $N_1, N_2, \cdots, N_L,$ and select samples of size $n_1, n_2, \cdots, n_L,$ respectively, from the L strata, the allocation is proportional if

` $$\frac{n_i}{N_i} = \frac{n}{N} \text{ or } n_i = \left(\frac{N_i}{N}\right)n, \qquad \text{for } i = 1,2,\cdots,L,$$

Where n is the total size of the stratified random sample.

Example: At a sample private college the students may be classified according to the following scheme:

Classification	Number of Students
Senior	150
Junior	163
Sophomore	195
Freshman	220

If we use proportional allocation to select a stratified random sample of size $n = 40,$ how large a sample must we take from each stratum?

Solution: For $n = 40,\ N_1 = 150,\ \ N_2 = 163,\ N_3 = 195,\ N_4 = 220,$ and $N = 728,$ it follows that

$$n_1 = \left(\frac{150}{728}\right)40 = 8, \qquad n_2 = \left(\frac{163}{728}\right)40 = 9,$$

$$n_3 = \left(\frac{195}{728}\right)40 = 11, \qquad n_4 = \left(\frac{220}{728}\right)40 = 12.$$

In all cases, we have rounded the calculations to the nearest integer.

Stratification is efficient when the units within the stratum are homogeneous and between strata are heterogeneous.

Cluster Sampling

In many statistical studies we can improve our efficiency over sample random sampling by randomly selecting groups or clusters of elements from a population and then sampling some or all of the elements within the selected cluster. For example, if a shipment of automotive parts consists of 5000 boxes, each containing 10 fuel pumps, and we wish to examine a random sample of 100 of these fuel pumps, it would be difficult to obtain a simple random sample without opening all 5000 boxes. A less costly procedure would be to select perhaps 10 of the boxes at random and examine all 10 fuel pumps in each of these boxes, or we might even select 50 of the 5000 boxes at random and then randomly pick 2 of the 10 fuel pumps from each of the 50 boxes for inspection. Sampling in this manner is referred to as cluster sampling. In the fisrt case, sampling is at one stage and in the second case, it is two stage sampling, first selecting the cluster and then units from the selected cluster.

Cluster sampling selects a sample containing either all, or a random selection, of the elements from clusters that have themselves been selected randomly from the population.

Cluster sampling has the advantage of being more cost efficient when the population is widely scattered. For example, in studying the investment habits of working adults in a given state, it is much cheaper to interview and collect data from individuals living close together in several randomly selected clusters or regions than to select a simple random sample from the entire state. When the clusters are geographic areas, such as regions of a state as we have here, or subdivisions of a large city, this kind of sampling is also called area sampling.

Several or all of the sampling procedures discussed so far may be used in the same study. For instance, if the members of a statistical group for the federal government wish to study voter opinion on the construction of additional nuclear power plants, they might let the voting district within each of the 50 states represent clusters and then use proportional allocation to select a stratified random sample of voting districts. Then they might use simple random sampling or any known sampling design from the voter registration lists to sample voter opinions from within the selected districts.

In cluster sampling, the conditions for composition of groups/ clusters are just opposite to those for stratified sampling. For cluster sampling, it is desired that there is heterogeneity within the clusters and homogeneity between clusters.

Multistage Sampling

We have seen that in the method of cluster sampling all the elements of the selected clusters are enumerated. This scheme as we know is convenient and economical but the method restricts the spread of the sample over the population which generally reduces the efficiency of the estimator. It is, therefore, logical to expect that the efficiency of the estimator can be increased by distributing elements over a larger number of clusters and surveying only a sample of units in each selected cluster instead of completely enumerating all the units in the selected clusters. This logic gives rise to a new sampling procedure.

The procedure of sampling, which consists in first selecting the clusters and then randomly choosing a specified number of units from each selected cluster, is known as two-stage sampling. This is also called sub-sampling.

In such sampling designs, clusters which form the units of sampling at the first stage are called first-stage units (fsu's) or primary stage units (psu's) and the elements within the clusters are called second stage units (ssu's). This procedure can be generalized to three or more stages and is termed as multistage sampling. For example, in crop surveys for estimating yield of a crop in a district, a block may be considered as a first-stage unit, villages within blocks as the second stage units, the crop fields within village as the third-stage units and a plot of specified shape and size within field as the ultimate unit of sampling.

Multistage sampling has been found to be very useful in practice and this procedure is being commonly used in large-scale surveys. This sampling procedure is a compromise between cluster sampling and direct sampling of units. Further, this design is more flexible as it permits the use of different selection procedures at different stages. It may also be mentioned that multi-stage sampling may be the only choice in a number of practical situations where a satisfactory sampling frame of ultimate-stage units is not readily available and the cost of obtaining such a frame is large and time consuming.

Systematic Sampling

Systematic Sampling: Systematic sampling is one of the most frequently used sample selection techniques. Systematic sampling selects every k^{th} element in the population for the sample, with the starting point determined at random from the first *k* elements. In this sampling, list of all population units and assign them the serial numbers from 1 to N. Then a sampling interval *k* is determined as k=N/n. A random number is then selected from 1 to k provide the starting point for selecting a sample of desired size. Let this number be i≤k then a sample of size n will consists of the sampling units bearing serial numbers i, i+k, i+2k, …,i+(n-1)k.

To be clearer, consider that a sample of size n = 20 is to be drawn from a population of size N = 500. Here the sampling interval is k=N/n = 500/20=25. First sampling interval consists of the first 25 population units serially numbered from 1 to 25. Let the randomly selected number is i = 10, the units to be included in the sample are 10, 35, 60, …, 10+(20-1).25=485.

Alternatively, two or more generally, replicated systematic samples can be taken, each of size n/m elements, the length of the sampling interval being $m\times k$. This method is suitable if variance estimation is to be carried out by the so-called replication techniques. In systematic sampling the number of different samples is quite small. If the sampling interval is k=N/n, there will be k separate systematic samples in total. For systematic sampling, the inclusion probability for the population element is equal to that under simple random sampling without replacement. So systematic sampling is also an equal-selection-probability design.

Systematic samples are very easy to obtain and are often used as if they were random samples. In fact, some systematic samples can lead to more precise inferences concerning population parameters simply because the sample values spread evenly over the entire population. However, a real danger in systematic sampling exists if one happens to choose a sampling interval that corresponds to any hidden periodicity. For example, in sampling average monthly gasoline sales, one should not sample every 12th month, since the sample would then include sales always for the same month and this might be a consistently high summer month for gasoline sales.

Systematic sampling may in some cases be more effective than simple random sampling. This will occur, for example, if there is a certain relationship between the ordering of the frame population and the values of the study variable. The most common cases are those where the population is already stratified or a trend exists that follows the population ordering or there is a periodic trend; all these situations can also be reached by appropriate sorting procedures. Periodicity may be harmful in some cases especially if harmonic variation coincides with the sampling interval. For example, a company wishing to estimate the average monthly sales of one of its products on the basis of monthly sales data for the last ten years, may quickly do so by employing the systematic sampling plan. If the sales are subject to periodic fluctuations, says, for being comparatively low during the three-month March-May period, it is possible that systematic sampling method may yield a sample over-represented by the monthly sales data for any of these three months.

A systematic sample also offers a great advantages in organizing control over the field work. In a systematic sample, the relative position in the sample of the different units included in the sample is fixed. There is consequently no risk in the method that any large contiguous part of the population will failed to be represented. Indeed , a method given an evenly spaced sample and is therefore, likely to give a more precise estimate of the population mean than a random sample unless the k^{th} unit constituting the sample happened to be alike or correlated. The method resembles stratified random sampling in the sense that one unit is selected from each stratum of k consecutive units. In reality however, the resemblance is only casual. In stratified sampling the unit to be selected from each stratum is randomly drawn whereas in systematic sampling its position related to the unit in the first stratum is predetermined. Therefore, unless the units in each stratum are randomly listed, a systematic sample will not be equivalent to a stratified random sample.

Systematic sampling strictly resembles with cluster sampling. A systematic sample being equivalent to a sample of one cluster selected out of the k clusters of n units each.

Probability Proportional to Size Sampling

Sampling with probability proportional to size (PPS) provides a practical technique when sampling from populations with large variation in the values of the study variable and often gives considerable gain in efficiency. Under PPS sampling an auxiliary size measure must be available and for efficient estimation the size measure should be strongly related to the study variable. More precisely, a size measure is sought for which ratio to the value of the study variable remains nearly a constant for all the population elements.

These three basic sampling techniques along with control measures can be used to construct a manageable sampling design for a complex sample surveys. In all the techniques, excluding simple random sampling, auxiliary information on the structure of the population is required. As a rule sampling error can be decreased by the proper use of auxiliary information.

Use of Auxiliary Information

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. This has already been described earlier. The knowledge of auxiliary information may also be exploited at the estimation stage. The estimator can be developed in such a way that it makes use of this additional information. Ratio estimator, difference estimator, regression estimator, generalized difference estimators are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sampling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost.

Another way the auxiliary information can be used is at the stage of planning of survey. An example of this is the stratification of the population units by making use of the auxiliary information.

Estimation of population mean and total for various sampling procedures will be illustrated with the help of data by using the SAS package.

References

Cochran, W.G. (1977). Sampling Techniques. Wiley Eastern Limited, New Delhi.

Des Raj and Chandhok, P. (1998). Sampling Survey Theory. Narosa Publishing House, New Delhi.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, New York.

Murthy, M.N. (1977). Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S and Ashok, C. (1984). Sampling Theory of Surveys with Applications. Indian Society of Agricultural Statistics, New Delhi.