SAMPLING TECHNIQUES

Basic concepts of sampling

Essentially, sampling consists of obtaining information from only a part of a large group or population so as to infer about the whole population. The object of sampling is thus to secure a sample which will represent the population and reproduce the important characteristics of the population under study as closely as possible.

The principal advantages of sampling as compared to complete enumeration of the population are reduced cost, greater speed, greater scope and improved accuracy. Many who insist that the only accurate way to survey a population is to make a complete enumeration, overlook the fact that there are many sources of errors in a complete enumeration and that a hundred per cent enumeration can be highly erroneous as well as nearly impossible to achieve. In fact, a sample can yield more accurate results because the sources of errors connected with reliability and training of field workers, clarity of instruction, mistakes in measurement and recording, badly kept measuring instruments, misidentification of sampling units, biases of the enumerators and mistakes in the processing and analysis of the data can be controlled more effectively. The smaller size of the sample makes the supervision more effective. Moreover, it is important to note that the precision of the estimates obtained from certain types of samples can be estimated from the sample itself. The net effect of a sample survey as compared to a complete enumeration is often a more accurate answer achieved with fewer personnel and less work at a low cost in a short time.

The most ‘convenient’ method of sampling is that in which the investigator selects a number of sampling units which he considers ‘representative’ of the whole population. For example, in estimating the whole volume of a forest stand, he may select a few trees which may appear to be of average dimensions and typical of the area and measure their volume. A walk over the forest area with an occasional stop and flinging a stone with the eyes closed or some other simple way that apparently avoids any deliberate choice of the sampling units is very tempting in its simplicity. However, it is clear that such methods of selection are likely to be biased by the investigator’s judgement and the results will thus be biased and unreliable. Even if the investigator can be trusted to be completely objective, considerable conscious or unconscious errors of judgement, not frequently recognized, may occur and such errors due to bias may far outweigh any supposed increase in accuracy resulting from deliberate or purposive selection of the units. Apart from the above points, subjective sampling does not permit the evaluation of the precision of the estimates calculated from samples. Subjective sampling is statistically unsound and should be discouraged.

When sampling is performed so that every unit in the population has some chance of being selected in the sample and the probability of selection of every unit is known, the method of sampling is called probability sampling. An example of probability sampling is random selection, which should be clearly distinguished from haphazard selection, which implies a strict process of selection equivalent to that of drawing lots. In this manual, any reference to sampling, unless otherwise stated, will relate to some form of probability sampling. The probability that

any sampling unit will be selected in the sample depends on the sampling procedure used. The important point to note is that the precision and reliability of the estimates obtained from a sample can be evaluated only for a probability sample. Thus the errors of sampling can be controlled satisfactorily in this case.

The object of designing a sample survey is to minimise the error in the final estimates. Any forest survey involving data collection and analysis of the data is subject to a variety of errors. The errors may be classified into two groups *viz*., (i) non-sampling errors (ii) sampling errors. The non-sampling errors like the errors in location of the units, measurement of the characteristics, recording mistakes, biases of enumerators and faulty methods of analysis may contribute substantially to the total error of the final results to both complete enumeration and sample surveys. The magnitude is likely to be larger in complete enumeration since the smaller size of the sample project makes it possible to be more selective in assignment of personnel for the survey operations, to be more thorough in their training and to be able to concentrate to a much greater degree on reduction of non-sampling errors. Sampling errors arise from the fact that only a fraction of the forest area is enumerated. Even if the sample is a probability sample, the sample being based on observations on a part of the population cannot, in general, exactly represent the population. The average magnitude of the sampling errors of most of the probability samples can be estimated from the data collected. The magnitude of the sampling errors, depends on the size of the sample, the variability within the population and the sampling method adopted. Thus if a probability sample is used, it is possible to predetermine the size of the sample needed to obtain desired and specified degree of precision.

A sampling scheme is determined by the size of sampling units, number of sampling units to be used, the distribution of the sampling units over the entire area to be sampled, the type and method of measurement in the selected units and the statistical procedures for analysing the survey data. A variety of sampling methods and estimating techniques developed to meet the varying demands of the survey statistician accord the user a wide selection for specific situations. One can choose the method or combination of methods that will yield a desired degree of precision at minimum cost. Additional references are Chacko (1965) and Sukhatme *et al, (1984)*

The principal steps in a sample survey

In any sample survey, we must first decide on the type of data to be collected and determine how adequate the results should be. Secondly, we must formulate the sampling plan for each of the characters for which data are to be collected. We must also know how to combine the sampling procedures for the various characters so that no duplication of field work occurs. Thirdly, the field work must be efficiently organised with adequate provision for supervising the work of the field staff. Lastly, the analysis of the data collected should be carried out using appropriate statistical techniques and the report should be drafted giving full details of the basic assumptions made, the sampling plan and the results of the statistical analysis. The report should contain estimate of the margin of the sampling errors of the results and may also include the possible effects of the non-sampling errors. Some of these steps are elaborated further in the following.

(i) Specification of the objectives of the survey: Careful consideration must be given at the outset to the purposes for which the survey is to be undertaken. For example, in a forest survey, the area

to be covered should be decided. The characteristics on which information is to be collected and the degree of detail to be attempted should be fixed. If it is a survey of trees, it must be decided as to what species of trees are to be enumerated, whether only estimation of the number of trees under specified diameter classes or, in addition, whether the volume of trees is also proposed to be estimated. It must also be decided at the outset what accuracy is desired for the estimates.

(*ii*) *Construction of a frame of units :* The first requirement of probability sample of any nature is the establishment of a frame. The structure of a sample survey is determined to a large extent by the frame. A frame is a list of sampling units which may be unambiguously defined and identified in the population. The sampling units may be compartments, topographical sections, strips of a fixed width or plots of a definite shape and size.

The construction of a frame suitable for the purposes of a survey requires experience and may very well constitute a major part of the work of planning the survey. This is particularly true in forest surveys since an artificial frame composed of sampling units of topographical sections, strips or plots may have to be constructed. For instance, the basic component of a sampling frame in a forest survey may be a proper map of the forest area. The choice of sampling units must be one that permits the identification in the field of a particular sampling unit which has to be selected in the sample. In forest surveys, there is considerable choice in the type and size of sampling units. The proper choice of the sampling units depends on a number of factors; the purpose of the survey, the characteristics to be observed in the selected units, the variability among sampling units of a given size, the sampling design, the field work plan and the total cost of the survey. The choice is also determined by practical convenience. For example, in hilly areas it may not be practicable to take strips as sampling units. Compartments or topographical sections may be more convenient. In general, at a given intensity of sampling (proportion of area enumerated) the smaller the sampling units employed the more representative will be the sample and the results are likely to be more accurate.

(iii) Choice of a sampling design: If it is agreed that the sampling design should be such that it should provide a statistically meaningful measure of the precision of the final estimates, then the sample should be a probability sample, in that every unit in the population should have a known probability of being selected in the sample. The choice of units to be enumerated from the frame of units should be based on some objective rule which leaves nothing to the opinion of the field worker. The determination of the number of units to be included in the sample and the method of selection is also governed by the allowable cost of the survey and the accuracy in the final estimates.

(iv) Organisation of the field work : The entire success of a sampling survey depends on the reliability of the field work. In forest surveys, the organization of the field work should receive the utmost attention, because even with the best sampling design, without proper organization the sample results may be incomplete and misleading. Proper selection of the personnel, intensive training, clear instructions and proper supervision of the fieldwork are essential to obtain satisfactory results. The field parties should correctly locate the selected units and record the necessary measurements according to the specific instruction given. The supervising staff should check a part of their work in the field and satisfy that the survey carried out in its entirety as planned.

(v) Analysis of the data : Depending on the sampling design used and the information collected, proper formulae should be used in obtaining the estimates and the precision of the estimates should be computed. Double check of the computations is desired to safeguard accuracy in the analysis.

(vi) Preliminary survey (pilot trials) : The design of a sampling scheme for a forest survey requires both knowledge of the statistical theory and experience with data regarding the nature of the forest area, the pattern of variability and operational cost. If prior knowledge in these matters is not available, a statistically planned small scale 'pilot survey' may have to be conducted before undertaking any large scale survey in the forest area. Such exploratory or pilot surveys will provide adequate knowledge regarding the variability of the material and will afford opportunities to test and improve field procedures, train field workers and study the operational efficiency of a design. A pilot survey will also provide data for estimating the various components of cost of operations in a survey like time of travel, time of location and enumeration of sampling units, etc. The above information will be of great help in deciding the proper type of design and intensity of sampling that will be appropriate for achieving the objects of the survey.

Sampling terminology

Although the basic concepts and steps involved in sampling are explained above, some of the general terms involved are further clarified in this section so as to facilitate the discussion on individual sampling schemes dealt with in later sections.

Population : The word population is defined as the aggregate of units from which a sample is chosen. If a forest area is divided into a number of compartments and the compartments are the units of sampling, these compartments will form the population of sampling units. On the other hand, if the forest area is divided into, say, a thousand strips each 20 m wide, then the thousand strips will form the population. Likewise if the forest area is divided into plots of, say, one-half hectare each, the totality of such plots is called the population of plots.

Sampling units : Sampling units may be administrative units or natural units like topographical sections and subcompartments or it may be artificial units like strips of a certain width, or plots of a definite shape and size. The unit must be a well defined element or group of elements identifiable in the forest area on which observations on the characteristics under study could be made. The population is thus sub-divided into suitable units for the purpose of sampling and these are called sampling units.

Sampling frame : A list of sampling units will be called a 'frame'. A population of units is said to be finite if the number units in it is finite.

Sample : One or more sampling units selected from a population according to some specified procedure will constitute a sample.

Sampling intensity : Intensity of sampling is defined as the ratio of the number of units in the sample to the number of units in the population.

Population total : Suppose a finite population consists of units $U_1, U_2, \ldots, U_N$. Let the value of the characteristic for the *i*th unit be denoted by y_i. For example the units may be strips and the characteristic may be the number of trees of a certain species in a strip. The total of the values y_i ($i = 1, 2, \ldots, N$), namely,

$$Y = \sum_{i=1}^{N} y_i \quad (5.1)$$

is called the population total which in the above example is the total number of trees of the particular species in the population.

Population mean : The arithmetic mean

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (5.2)$$

is called the population mean which, in the example considered, is the average number of trees of the species per strip.

Population variance : A measure of the variation between units of the population is provided by the population variance

$$S_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{Y})^2 = \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \overline{Y}^2 \quad (5.3)$$

which in the example considered measures the variation in number of trees of the particular species among the strips. Large values of the population variance indicate large variation between units in the population and small values indicate that the values of the characteristic for the units are close to the population mean. The square root of the variance is known as *standard deviation.*

Coefficient of variation : The ratio of the standard deviation to the value of the mean is called the coefficient of variation, which is usually expressed in percentage.

$$C.V. = \frac{S_y}{\overline{Y}} \quad (5.4)$$

The coefficient of variation, being dimensionless, is a valuable tool to compare the variation between two or more populations or sets of observations.

Parameter : A function of the values of the units in the population will be called a parameter. The population mean, variance, coefficient of variation, etc., are examples of population

parameters. The problem in sampling theory is to estimate the parameters from a sample by a procedure that makes it possible to measure the precision of the estimates.

Estimator, estimate : Let us denote the sample observations of size n by $y_1, y_2, \ldots, y_n$. Any function of the sample observations will be called a *statistic*. When a statistic is used to estimate a population parameter, the statistic will be called an estimator. For example, the sample mean is an estimator of the population mean. Any particular value of an estimator computed from an observed sample will be called an estimate.

Bias in estimation : A statistic t is said to be an unbiased estimator of a population parameter q if its expected value, denoted by E(t), is equal to q . A sampling procedure based on a probability scheme gives rise to a number of possible samples by repetition of the sampling procedure. If the values of the statistic t are computed for each of the possible samples and if the average of the values is equal to the population value q , then t is said to be an unbiased estimator of q based on sampling procedure. Notice that the repetition of the procedure and computing the values of t for each sample is only conceptual, not actual, but the idea of generating all possible estimates by repetition of the sampling process is fundamental to the study of bias and of the assessment of sampling error. In case E(t) is not equal to q , the statistic t is said to be a biased estimator of q and the bia s is given by, bias = E(t) - q . The introduction of a genuinely random process in selecting a sample is an important step in avoiding bias. Samples selected subjectively will usually be very seriously biased. In forest surveys, the tendency of forest officers to select typical forest areas for enumerations, however honest the intention may be, is bound to result in biased estimates.

Sampling variance : The difference between a sample estimate and the population value is called the sampling error of the estimate, but this is naturally unknown since the population value is unknown. Since the sampling scheme gives rise to different possible samples, the estimates will differ from sample to sample. Based on these possible estimates, a measure of the average magnitude over all possible samples of the squares of the sampling error can be obtained and is known as the *mean square error* (*MSE*) of the estimate which is essentially a measure of the divergence of an estimator from the true population value. Symbolically, $MSE = E[t - q]^2$. The sampling variance ($V(t)$)is a measure of the divergence of the estimate from its expected value. It is defined as the average magnitude over all possible samples of the squares of deviations of the estimator from its expected value and is given by $V(t) = E[t - E(t)]^2$.

Notice that the sampling variance coincides with the mean square error when t is an unbiased estimator. Generally, the magnitude of the estimate of the sampling variance computed from a sample is taken as indicating whether a sample estimate is useful for the purpose. The larger the sample and the smaller the variability between units in the population, the smaller will be the sampling error and the greater will be the confidence in the results.

Standard error of an estimator : The square root of the sampling variance of an estimator is known as the standard error of the estimator. The standard error of an estimate divided by the value of the estimate is called relative standard error which is usually expressed in percentage.

Accuracy and precision : The standard error of an estimate, as obtained from a sample, does not include the contribution of the bias. Thus we may speak of the standard error or the sampling variance of the estimate as measuring on the inverse scale, the precision of the estimate, rather than its *accuracy.* Accuracy usually refers to the size of the deviations of the sample estimate from the mean $m = E(t)$ obtained by repeated application of the sampling procedure, the bias being thus measured by $m - q$.

It is the accuracy of the sample estimate in which we are chiefly interested; it is the precision with which we are able to measure in most instances. We strive to design the survey and attempt to analyse the data using appropriate statistical methods in such a way that the precision is increased to the maximum and bias is reduced to the minimum.

Confidence limits : If the estimator *t* is normally distributed (which assumption is generally valid for large samples), a confidence interval defined by a lower and upper limit can expected to include the population parameter q with a specified probability level. The limits are given by

$$\text{Lower limit} = t - z\sqrt{\hat{V}(t)} \quad (5.5)$$

$$\text{Upper limit} = t + z\sqrt{\hat{V}(t)} \quad (5.6)$$

where $\hat{V}(t)$ is the estimate of the variance of *t* and *z* is the value of the normal deviate corresponding to a desired P % confidence probability. For example, when *z* is taken as 1.96, we say that the chance of the true value of q being contained in the random interval defined by the lower and upper confidence limits is 95 per cent. The confidence limits specify the range of variation expected in the population mean and also stipulate the degree of confidence we should place in our sample results. If the sample size is less than 30, the value of *k* in the formula for the lower and upper confidence limits should be taken from the percentage points of Student's *t* distribution (See Appendix 2) with degrees of freedom of the sum of squares in the estimate of the variance of *t*. Moderate departures of the distribution from normality does not affect appreciably the formula for the confidence limits. On the other hand, when the distribution is very much different from normal, special methods are needed. For example, if we use small area sampling units to estimate the average number of trees in higher diameter classes, the distribution may have a large skewness. In such cases, the above formula for calculating the lower and upper confidence limits may not be directly applicable.

Some general remarks : In the sections to follow, capital letters will usually be used to denote population values and small letters to denote sample values. The symbol 'cap' (^) above a symbol for a population value denotes its estimate based on sample observations. Other special notations used will be explained as and when they are introduced.

While describing the different sampling methods below, the formulae for estimating only population mean and its sampling variance are given. Two related parameters are population total and ratio of the character under study (*y*) to some auxiliary variable (*x*). These related statistics can always be obtained from the mean by using the following general relations.

$$\hat{Y} = N\hat{\bar{Y}} \quad (5.7)$$

$$V(\hat{Y}) = N^2 V(\hat{\bar{Y}}) \quad (5.8)$$

$$\hat{R} = \frac{\hat{Y}}{X} \quad (5.9)$$

$$V(\hat{R}) = \frac{V(\hat{Y})}{X^2} \quad (5.10)$$

where $\hat{Y}$ = Estimate of the population total

N = Total number of units in the population

$\hat{R}$ = Estimate of the population ratio

X = Population total of the auxiliary variable

Simple random sampling

A sampling procedure such that each possible combination of sampling units out of the population has the same chance of being selected is referred to as simple random sampling. From theoretical considerations, simple random sampling is the simplest form of sampling and is the basis for many other sampling methods. Simple random sampling is most applicable for the initial survey in an investigation and for studies which involve sampling from a small area where the sample size is relatively small. When the investigator has some knowledge regarding the population sampled, other methods which are likely to be more efficient and convenient for organising the survey in the field, may be adopted. The irregular distribution of the sampling units in the forest area in simple random sampling may be of great disadvantage in forest areas where accessibility is poor and the costs of travel and locating the plots are considerably higher than the cost of enumerating the plot.

Selection of sampling units

In practice, a random sample is selected unit by unit. Two methods of random selection for simple random sampling without replacement are explained in this section.

(i) Lottery method : The units in the population are numbered 1 to *N*. If *N* identical counters with numberings 1 to *N* are obtained and one counter is chosen at random after shuffling the counters, then the probability of selecting any counter is the same for all the counters. The process is repeated *n* times without replacing the counters selected. The units which correspond to the numbers on the chosen counters form a simple random sample of size *n* from the population of *N* units.

(ii) Selection based on random number tables : The procedure of selection using the lottery method, obviously becomes rather inconvenient when *N* is large. To overcome this difficulty, we may use a table of random numbers such as those published by Fisher and Yates (1963) a sample of which is given in Appendix 6. The tables of random numbers have been developed in such a way that the digits 0 to 9 appear independent of each other and approximately equal number of times in the table. The simplest way of selecting a random sample of required size consists in selecting a set of *n* random numbers one by one, from 1 to *N* in the random number table and, then, taking the units bearing those numbers. This procedure may involve a number of rejections since all the numbers more than *N* appearing in the table are not considered for selection. In such cases, the procedure is modified as follows. If *N* is a *d* digited number, we first determine the highest *d* digited multiple of *N*, say *N*'. Then a random number *r* is chosen from 1 to *N'* and the unit having the serial number equal to the remainder obtained on dividing *r* by *N*, is considered as selected. If remainder is zero, the last unit is selected. A numerical example is given below.

Suppose that we are to select a simple random sample of 5 units from a serially numbered list of 40 units. Consulting Appendix 6 : Table of random numbers, and taking column (5) containing two-digited numbers, the following numbers are obtained:

39, 27, 00, 74, 07

In order to give equal chances of selection to all the 100 units, we are to reject all numbers above 79 and consider (00) equivalent to 80. We now divide the above numbers in turn by 40 and take the remainders as the selected strip numbers for our sample, rejecting the remainders that are repeated. We thus get the following 16 strip numbers as our sample :

39, 27, 40, 34, 7.

Parameter estimation

Let $y_1, y_2, \ldots, y_n$ be the measurements on a particular characteristic on *n* selected units in a sample from a population of *N* sampling units. It can be shown in the case of simple random sampling without replacement that the sample mean,

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (5.11)$$

is an unbiased estimator of the population mean, $\bar{Y}$. An unbiased estimate of the sampling variance of $\bar{y}$ is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{N-n}{Nn} s_y^2 \quad (5.12)$$

where $s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ (5.13)

Assuming that the estimate $\bar{y}$ is normally distributed, a confidence interval on the population mean $\bar{Y}$ can be set with the lower and upper confidence limits defined by,

Lower limit $\hat{\bar{Y}}_L = \bar{y} - z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$ (5.14)

Upper limit $\hat{\bar{Y}}_U = \bar{y} + z\frac{s_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$ (5.15)

where *z* is the table value which depends on how many observations there are in the sample. If there are 30 or more observations we can read the values from the table of the normal distribution (Appendix 1). If there are less than 30 observations, the table value should be read from the table of *t* distribution (Appendix 2), using *n* - 1 degree of freedom.

The computations are illustrated with the following example. Suppose that a forest has been divided up into 1000 plots of 0.1 hectare each and a simple random sample of 25 plots has been selected. For each of these sample plots the wood volumes in m^3 were recorded. The wood volumes were,

7 10 7 4 7

8 8 8 7 5

2 6 9 7 8

6 7 11 8 8

7 3 8 7 7

If the wood volume on the *i*th sampling unit is designated as y_i , an unbiased estimator of the population mean, is obtained using Equation (5.11) as,

$$\hat{\bar{Y}} = \bar{y} = \frac{7+8+2+\ldots+7}{25} = \frac{175}{25}$$

$= 7\ m^3$

which is the mean wood volume per plot of 0.1 ha in the forest area.

An estimate (s_y^2) of the variance of individual values of *y* is obtained using Equation (5.13).

$$s_y^2 = \frac{(7-7)^2 + (8-7)^2 + \ldots + (7-7)^2}{25-1}$$

$$= \frac{82}{24} = 3.833$$

Then unbiased estimate of sampling variance of is

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1000-25}{(1000)(25)}\right)3.833$$

$$= 0.1495 \ (m^3)^2$$

$$SE(\hat{\bar{Y}}) = \sqrt{0.1495} = 0.3867 \ m^3$$

The relative standard error which is $\frac{SE(\hat{\bar{Y}})}{\hat{\bar{Y}}}(100)$ is a more common expression. Thus,

$$RSE(\hat{\bar{Y}}) = \frac{\sqrt{0.1495}}{7}(100) = 5.52\ \%$$

The confidence limits on the population mean are obtained using Equations (5.14) and (5.15).

Lower limit $\hat{\bar{Y}}_L = 7 - (2.064)\sqrt{0.1495}$

$$= 6.20 \text{ cords}$$

Upper limit $\hat{\bar{Y}}_U = 7 + (2.064)\sqrt{0.1495}$

$$= 7.80 \text{ cords}$$

The 95% confidence interval for the population mean is (6.20, 7.80) m^3. Thus, we are 95% confident that the confidence interval (6.20, 7.80) m^3 would include the population mean.

An estimate of the total wood volume in the forest area sampled can easily be obtained by multiplying the estimate of the mean by the total number of plots in the population. Thus,

$\hat{Y} = 7(1000) = 7000\ \mathrm{m}^3$

with a confidence interval of (6200, 7800) obtained by multiplying the confidence limits on the mean by $N = 1000$. The RSE of $\hat{Y}$, however, will not be changed by this operation.

Systematic sampling

Systematic sampling employs a simple rule of selecting every *k*th unit starting with a number chosen at random from 1 to *k* as the random start. Let us assume that *N* sampling units in the population are numbered 1 to *N*. To select a systematic sample of *n* units, we take a unit at random from the first *k* units and then every *k*th sampling unit is selected to form the sample. The constant *k* is known as the *sampling interval* and is taken as the integer nearest to *N / n*, the inverse of the sampling fraction. Measurement of every *k*th tree along a certain compass bearing is an example of systematic sampling. A common sampling unit in forest surveys is a narrow strip at right angles to a base line and running completely across the forest. If the sampling units are strips, then the scheme is known as systematic sampling by strips. Another possibility is known as systematic line plot sampling where plots of a fixed size and shape are taken at equal intervals along equally spaced parallel lines. In the latter case, the sample could as well be systematic in two directions.

Systematic sampling certainly has an intuitive appeal, apart from being easier to select and carry out in the field, through spreading the sample evenly over the forest area and ensuring a certain amount of representation of different parts of the area. This type of sampling is often convenient in exercising control over field work. Apart from these operational considerations, the procedure of systematic sampling is observed to provide estimators more efficient than simple random sampling under normal forest conditions. The property of the systematic sample in spreading the sampling units evenly over the population can be taken advantage of by listing the units so that homogeneous units are put together or such that the values of the characteristic for the units are in ascending or descending order of magnitude. For example, knowing the fertility trend of the forest area the units (for example strips) may be listed along the fertility trend.

If the population exhibits a regular pattern of variation and if the sampling interval of the systematic sample coincides with this regularity, a systematic sample will not give precise estimates. It must, however, be mentioned that no clear case of periodicity has been reported in a forest area. But the fact that systematic sampling may give poor precision when unsuspected periodicity is present should not be lost sight of when planning a survey.

Selection of a systematic sample

To illustrate the selection of a systematic sample, consider a population of $N = 48$ units. A sample of $n = 4$ units is needed. Here, $k = 12$. If the random number selected from the set of numbers from 1 to 12 is 11, then the units associated with serial numbers 11, 23, 35 and 47 will

be selected. In situations where *N* is not fully divisible by *n*, *k* is calculated as the integer nearest to *N / n*. In this situation, the sample size is not necessarily *n* and in some cases it may be *n* -1.

5.3.2. Parameter estimatio n

The estimate for the population mean per unit is given by the sample mean

$$\hat{\bar{Y}} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (5.16)$$

where *n* is the number of units in the sample.

In the case of systematic strip surveys or, in general, any one dimensional systematic sampling, an approximation to the standard error may be obtained from the differences between pairs of successive units. If there are *n* units enumerated in the systematic sample, there will be (*n*-1) differences. The variance per unit is therefore, given by the sum of squares of the differences divided by twice the number of differences. Thus if $y_1, y_2, \ldots, y_n$ are the observed values (say volume) for the *n* units in the systematic sample and defining the first difference $d(y_i)$ as given below,

$$d(y_i) = y_{(i+1)} - y_{(i)}; \ (i = 1, 2, \ldots, n-1), \quad (5.17)$$

the approximate variance per unit is estimated as

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} \left[d(y_i) \right]^2 \quad (5.18)$$

As an example, Table 5.1 gives the observed diameters of 10 trees selected by systematic selection of 1 in 20 trees from a stand containing 195 trees in rows of 15 trees. The first tree was selected as the 8th tree from one of the outside edges of the stand starting from one corner and the remaining trees were selected systematically by taking every 20th tree switching to the nearest tree of the next row after the last tree in any row is encountered.

Table 5.1. Tree diameter recorded on a systematic sample of

10 trees from a plot.

Selected tree number	Diameter at breast-height(cm) y_i	First difference $d(y_i)$
8	14.8	
28	12.0	-2.8
48	13.6	+1.6
68	14.2	+0.6
88	11.8	-2.4
108	14.1	+2.3
128	11.6	-2.5
148	9.0	-2.6
168	10.1	+1.1
188	9.5	-0.6

Average diameter is equal to

$$\hat{\bar{Y}} = \frac{1}{10}(14.8 + 12.0 + \ldots\ldots + 9.5) = 12.07$$

The nine first differences can be obtained as shown in column (3) of the Table 5.1. The error variance of the mean per unit is thus

$$\hat{V}(\hat{\bar{Y}}) = \frac{(-2.8)^2 + (1.6)^2 + \ldots + (-0.6)^2}{2x9x10} = \frac{36.9}{180}$$

= 0.202167

A difficulty with systematic sampling is that one systematic sample by itself will not furnish valid assessment of the precision of the estimates. With a view to have valid estimates of the precision, one may resort to partially systematic samples. A theoretically valid method of using the idea of systematic samples and at the same time leading to unbiased estimates of the sampling error is to draw a minimum of two systematic samples with independent random starts.

If $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_m$ are *m* estimates of the population mean based on *m* independent systematic samples, the combined estimate is

$$\bar{y} = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i \quad (5.19)$$

The estimate of the variance of $\bar{y}$ is given by

$$\hat{V}(\bar{y}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \quad (5.20)$$

Notice that the precision increases with the number of independent systematic samples.

As an example, consider the data given in Table 5.1 along with another systematic sample selected with independent random starts. In the second sample, the first tree was selected as the 10th tree. Data for the two independent samples are given in Table 5.2.

Table 5.2. Tree diameter recorded on two independent systematic samples of 10 trees from a plot.

Sample 1		Sample 2	
Selected tree number	Diameter at breast-height(cm)	Selected tree number	Diameter at breast-height(cm)
	y_i		y_i
8	14.8	10	13.6
28	12.0	30	10.0
48	13.6	50	14.8
68	14.2	70	14.2
88	11.8	90	13.8
108	14.1	110	14.5
128	11.6	130	12.0
148	9.0	150	10.0
168	10.1	170	10.5

188	9.5	190	8.5

The average diameter for the first sample is $\bar{y}_1 = 12.07$. The average diameter for the first sample is $\bar{y}_2 = 12.19$. Combined estimate of population mean ($\bar{y}$) is obtained by using Equation (5.19) as,

$$\bar{y} = \frac{1}{2}(12.07 + 12.19)$$

= 12.13

The estimate of the variance of is obtained by using Equation (5.20).

$$\hat{V}(\bar{y}) = \frac{1}{2(2-1)}(12.07 - 12.13)^2(12.19 - 12.13)^2$$ = 0.0036

$$SE(\bar{y}) = \sqrt{0.0036}$$ = 0.06

One additional variant of systematic sampling is that sampling may as well be systematic in two directions. For example, in plantations, a systematic sample of rows and measurements on every tenth tree in each selected row may be adopted with a view to estimate the volume of the stand. In a forest survey, one may take a series of equidistant parallel strips extending over the whole width of the forest and the enumeration in each strip may be done by taking a systematic sample of plots or trees in each strip. Forming rectangular grids of (p x q) metres and selecting a systematic sample of rows and columns with a fixed size plot of prescribed shape at each intersection is another example.

In the case of two dimensional systematic sample, a method of obtaining the estimates and approximation to the sampling error is based on stratification and the method is similar to the stratified sampling given in section 5.4. For example, the sample may be arbitrarily divided into sets of four in 2 x 2 units and each set may be taken to form a stratum with the further assumption that the observations within each stratum are independently and randomly chosen. With a view to make border adjustments, overlapping strata may be taken at the boundaries of the forest area.

Stratified sampling

The basic idea in stratified random sampling is to divide a heterogeneous population into sub-populations, usually known as strata, each of which is internally homogeneous in which case a precise estimate of any stratum mean can be obtained based on a small sample from that stratum and by combining such estimates, a precise estimate for the whole population can be obtained.

Stratified sampling provides a better cross section of the population than the procedure of simple random sampling. It may also simplify the organisation of the field work. Geographical proximity is sometimes taken as the basis of stratification. The assumption here is that geographically contiguous areas are often more alike than areas that are far apart. Administrative convenience may also dictate the basis on which the stratification is made. For example, the staff already available in each range of a forest division may have to supervise the survey in the area under their jurisdiction. Thus, compact geographical regions may form the strata. A fairly effective method of stratification is to conduct a quick reconnaissance survey of the area or pool the information already at hand and stratify the forest area according to forest types, stand density, site quality etc. If the characteristic under study is known to be correlated with a supplementary variable for which actual data or at least good estimates are available for the units in the population, the stratification may be done using the information on the supplementary variable. For instance, the volume estimates obtained at a previous inventory of the forest area may be used for stratification of the population.

In stratified sampling, the variance of the estimator consists of only the 'within strata' variation. Thus the larger the number of strata into which a population is divided, the higher, in general, the precision, since it is likely that, in this case, the units within a stratum will be more homogeneous. For estimating the variance within strata, there should be a minimum of 2 units in each stratum. The larger the number of strata the higher will, in general, be the cost of enumeration. So, depending on administrative convenience, cost of the survey and variability of the characteristic under study in the area, a decision on the number of strata will have to be arrived at.

Allocation and selection of the sample within strata

Assume that the population is divided into k strata of $N_1, N_2, \ldots, N_k$ units respectively, and that a sample of n units is to be drawn from the population. The problem of allocation concerns the choice of the sample sizes in the respective strata, *i.e.*, how many units should be taken from each stratum such that the total sample is n.

Other things being equal, a larger sample may be taken from a stratum with a larger variance so that the variance of the estimates of strata means gets reduced. The application of the above principle requires advance estimates of the variation within each stratum. These may be available from a previous survey or may be based on pilot surveys of a restricted nature. Thus if this information is available, the sampling fraction in each stratum may be taken proportional to the standard deviation of each stratum.

In case the cost per unit of conducting the survey in each stratum is known and is varying from stratum to stratum an efficient method of allocation for minimum cost will be to take large samples from the stratum where sampling is cheaper and variability is higher. To apply this procedure one needs information on variability and cost of observation per unit in the different strata.

Where information regarding the relative variances within strata and cost of operations are not available, the allocation in the different strata may be made in proportion to the number of units

in them or the total area of each stratum. This method is usually known as ‘proportional allocation’.

For the selection of units within strata, In general, any method which is based on a probability selection of units can be adopted. But the selection should be independent in each stratum. If independent random samples are taken from each stratum, the sampling procedure will be known as ‘stratified random sampling’. Other modes of selection of sampling such as systematic sampling can also be adopted within the different strata.

Estimation of mean and variance

We shall assume that the population of N units is first divided into k strata of $N_1, N_2, \ldots, N_k$ units respectively. These strata are non-overlapping and together they comprise the whole population, so that

$N_1 + N_2 + \ldots\ldots + N_k = N.$ (5.21)

When the strata have been determined, a sample is drawn from each stratum, the selection being made independently in each stratum. The sample sizes within the strata are denoted by $n_1, n_2, \ldots, n_k$ respectively, so that

$n_1 + n_2 + \ldots\ldots + n_3 = n$ (5.22)

Let y_{tj} ($j = 1, 2, \ldots, N_t$; $t = 1, 2, \ldots k$) be the value of the characteristic under study for the j the unit in the tth stratum. In this case, the population mean in the tth stratum is given by

$$\overline{Y}_t = \frac{1}{N_t}\sum_{j=1}^{N_t} y_{tj}, (t = 1, 2, \ldots, k) \quad (5.23)$$

The overall population mean is given by

$$\overline{Y} = \frac{1}{N}\sum_{t=1}^{k} N_t \overline{Y}_t \quad (5.24)$$

The estimate of the population mean $\overline{Y}$, in this case will be obtained by

$$\hat{\overline{Y}} = \frac{\sum_{t=1}^{k} N_t \bar{y}_t}{N} \quad (5.25)$$

where $$\bar{y}_t = \sum_{j=1}^{n_t} \frac{y_{tj}}{n_t} \quad (5.26)$$

Estimate of the variance of $\hat{\bar{Y}}$ is given by

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2}\sum_{t=1}^{k} N_t(N_t - n_t)\frac{s_{t(y)}^2}{n_t} \quad (5.27)$$

where $$s_{t(y)}^2 = \sum_{j=1}^{n_t}\frac{(y_{tj} - \bar{y}_t)^2}{n_t - 1} \quad (5.28)$$

Stratification, if properly done as explained in the previous sections, will usually give lower variance for the estimated population total or mean than a simple random sample of the same size. However, a stratified sample taken without due care and planning may not be better than a simple random sample.

Numerical illustration of calculating the estimate of mean volume per hectare of a particular species and its standard error from a stratified random sample of compartments selected independently with equal probability in each stratum is given below.

A forest area consisting of 69 compartments was divided into three strata containing compartments 1-29, compartments 30-45, and compartments 46 to 69 and 10, 5 and 8 compartments respectively were chosen at random from the three strata. The serial numbers of the selected compartments in each stratum are given in column (4) of Table 5.3. The corresponding observed volume of the particular species in each selected compartment in m^3/ha is shown in column (5).

Table 5.3. Illustration of estimation of parameters under stratified sampling

Stratum number	Total number of units in the stratum (N_t)	Number of units sampled (n_t)	Selected sampling unit number	Volume (m^3/ha) (y_{t_s})	($y_{t_s}^2$)
(1)	(2)	(3)	(4)	(5)	(6)
			1	5.40	29.16
			18	4.87	23.72
			28	4.61	21.25
I			12	3.26	10.63

			20	4.96	24.60
			19	4.73	22.37
			9	4.39	19.27
			6	2.34	5.48
			17	4.74	22.47
			7	2.85	8.12
Total	29	10	..	42.15	187.07
			43	4.79	22.94
II			42	4.57	20.88
			36	4.89	23.91
			45	4.42	19.54
			39	3.44	11.83
Total	16	5	..	22.11	99.10
			59	7.41	54.91
			50	3.70	13.69
			49	5.45	29.70
III			58	7.01	49.14
			54	3.83	14.67
			69	5.25	27.56
			52	4.50	20.25
			47	6.51	42.38
Total	24	8	..	43.66	252.30

Step 1. Compute the following quantities.

$N = (29 + 16 + 24) = 69$

$n = (10 + 5 + 8) = 23$

$\bar{y}_t = 4.215, \quad = 4.422, \quad = 5.458$

Step 2. Estimate of the population mean $\bar{Y}$ using Equation (3) is

$$\hat{\bar{Y}} = \frac{\sum_{t=1}^{3} N_t \bar{y}_t}{N} = \frac{(29 \times 4.215) + (16 \times 4.422) + (24 \times 5.458)}{69} = \frac{323.979}{69} = 4.70$$

Step 3. Estimate of the variance of $\hat{\bar{Y}}$ using Equation (5) as

$$\hat{V}(\hat{\bar{Y}}) = \frac{1}{N^2} \sum_{t=1}^{3} N_t (N_t - n_t) \frac{s_{t(y)}^2}{n_t}$$

In this example,

$$s_{1(y)}^2 = \frac{187.07 - \frac{(42.15)^2}{10}}{9} = \frac{9.41}{9} = 1.046$$

$$s_{2(y)}^2 = \frac{99.10 - \frac{(22.11)^2}{5}}{4} = \frac{1.33}{4} = 0.333$$

$$s_{3(y)}^2 = \frac{252.30 - \frac{(43.66)^2}{8}}{7} = \frac{14.03}{7} = 2.004$$

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1}{69}\right)^2 \left[\left(\frac{29 \times 19}{10} \times 1.046\right) + \left(\frac{16 \times 11}{5} \times 0.333\right) + \left(\frac{24 \times 16}{8} \times 2.004\right)\right]$$

$$= \frac{165.5482}{4761} = 0.03477$$

$$SE(\hat{\bar{Y}}) = \sqrt{0.03477} = 0.1865$$

$$RSE(\hat{\bar{Y}}) = \frac{SE(\hat{\bar{Y}}) \times 100}{\hat{\bar{Y}}} \quad (5.29)$$

$$= \frac{0.1865 \times 100}{4.70} = 3.97\%$$

Now, if we ignore the strata and assume that the same sample of size $n = 23$, formed a simple random sample from the population of $N = 69$, the estimate of the population mean would reduce to

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{42.15 + 22.11 + 43.66}{23} = \frac{107.92}{23} = 4.69$$

Estimate of the variance of the mean $\bar{y}$ is

$$\hat{V}(\bar{y}) = \frac{N - n}{Nn} s^2$$

where

$$s^2 = \frac{538.47 - \frac{(107.92)^2}{23}}{22}$$

$$= \frac{32.09}{22} = 1.4586$$

so that

$$\hat{V}(\bar{y}) = \frac{(69 - 23)}{69 \times 23} \times 1.4586$$

$$= \frac{2.9172}{69} = 0.04230$$

$$SE(\bar{y}) = \sqrt{0.04230} = 0.2057$$

$$RSE(\bar{y}) = \frac{0.2057 \times 100}{4.69} = 4.39\%$$

The gain in precision due to stratification is computed by

$$\frac{\hat{V}(\hat{\bar{Y}})_{srs}}{\hat{V}(\hat{\bar{Y}})_{st}} \times 100 = \frac{0.04230}{0.03477} \times 100$$

= 121.8

Thus the gain in precision is 21.8%.

Multistage sampling

With a view to reduce cost and/or to concentrate the field operations around selected points and at the same time obtain precise estimates, sampling is sometimes carried out in stages. The procedure of first selecting large sized units and then choosing a specified number of sub-units from the selected large units is known as sub-sampling. The large units are called 'first stage units' and the sub-units the 'second stage units'. The procedure can be easily generalised to three stage or multistage samples. For example, the sampling of a forest area may be done in three stages, firstly by selecting a sample of compartments as first stage units, secondly, by choosing a sample of topographical sections in each selected compartment and lastly, by taking a number of sample plots of a specified size and shape in each selected topographical section.

The multistage sampling scheme has the advantage of concentrating the sample around several 'sample points' rather than spreading it over the entire area to be surveyed. This reduces considerably the cost of operations of the survey and helps to reduce the non-sampling errors by efficient supervision. Moreover, in forest surveys it often happens that detailed information may be easily available only for groups of sampling units but not for individual units. Thus, for example, a list of compartments with details of area may be available but the details of the topographical sections in each compartment may not be available. Hence if compartments are selected as first stage units, it may be practicable to collect details regarding the topographical sections for selected compartments only and thus use a two-stage sampling scheme without attempting to make a frame of the topographical sections in all compartments. The multistage sampling scheme, thus, enables one to use an incomplete sampling frame of all the sampling units and to properly utilise the information already available at every stage in an efficient manner.

The selection at each stage, in general may be either simple random or any other probability sampling method and the method may be different at the different stages. For example one may select a simple random sample of compartments and take a systematic line plot survey or strip survey with a random start in the selected compartments.

Two-stage simple random sampling

When at both stages the selection is by simple random sampling, method is known as two stage simple random sampling. For example, in estimating the weight of grass in a forest area, consisting of 40 compartments, the compartments may be considered as primary sampling units.

Out of these 40 compartments, $n = 8$ compartments may be selected randomly using simple random sampling procedure as illustrated in Section 5.2.1. A random sample of plots either equal or unequal in number may be selected from each selected compartment for the measurement of the quantity of grass through the procedure of selecting a simple random sample. It is then possible to develop estimates of either mean or total quantity of grass available in the forest area through appropriate formulae.

Parameter estimation under two-stage simple random sampling

Let the population consists of N first stage units and let M_i be the number of second stage units in the ith first stage unit. Let n first stage units be selected and from the ith selected first stage unit let m_i second stage units be chosen to form a sample of $m = \sum_{i=1}^{n} m_i$ units. Let y_{ij} be the value of the character for the jth second stage unit in the ith first stage unit.

$$\overline{Y} = \frac{\sum_{i}^{N}\sum_{j}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i}$$

An unbiased estimator of the population mean is obtained by Equation (5.30).

$$\hat{\overline{Y}} = \frac{1}{n\overline{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} \quad (5.30)$$

where $\overline{M} = \frac{\sum_{i=1}^{N} M_i}{N}$. (5.31)

The estimate of the variance of $\hat{\overline{Y}}$ is given by

$$\hat{V}(\hat{\overline{Y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nN} \sum_{i=1}^{n} \left(\frac{M_i}{\overline{M}}\right)^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{w_i}^2 \quad (5.32)$$

where $s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i}{\overline{M}} \overline{y}_i - \overline{y}\right)^2$ (5.33)

$$s_{w_i}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \overline{y}_i)^2 \quad (5.34)$$

The variance of $\hat{\bar{Y}}$ here can be noticed to be composed of two components. The first is a measure of variation between first stage units and the second, a measure of variation within first stage units. If $m_i = M_i$, the variance is given by the first component only. The second term, thus represents the contribution due to sub-sampling.

An example of the analysis of a two stage sample is given below. Table 5.4 gives data on weight of grass (mixed species) in kg from plots of size 0.025 ha selected from 8 compartments which were selected randomly out of 40 compartments from a forest area. The total forest area was 1800ha.

Table 5.4. Weight of grass in kg in plots selected through a two stage sampling procedure.

Plot	Compartment number								Total
	I	II	III	IV	V	VI	VII	VIII	
1	96	98	135	142	118	80	76	110	
2	100	142	88	130	95	73	62	125	
3	113	143	87	106	109	96	105	77	
4	112	84	108	96	147	113	125	62	
5	88	89	145	91	91	125	99	70	
6	139	90	129	88	125	68	64	98	
7	140	89	84	99	115	130	135	65	
8	143	94	96	140	132	76	78	97	
9	131	125	..	98	148	84	..	106	
10	..	116	..	..	..	105	..	..	
Total	1062	1070	872	990	1080	950	744	810	7578
m_i	9	10	8	9	9	10	8	9	72
Mean $(\bar{y}_i)$	118	107	109	110	120	95	93	90	842
M_i	1760	1975	1615	1785	1775	2050	1680	1865	14505
$s_{w_i}^2$	436.00	515.78	584.57	455.75	412.25	496.67	754.86	496.50	4152

$\frac{s_{w_i}^2}{m_i}$	48.44	51.578	73.07	50.63	45.80	49.667	94.35	55.167

Step1. Estimate the mean weight of grass in kg per plot using the formula in Equation (5.30).

$$\hat{\bar{Y}} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i = \frac{1}{40} \left(\frac{1800}{0.025}\right)$$

= 1800

Since $\sum M_i$ indicates the total number of second stage units, it can be obtained by dividing the total area (1800 ha) by the size of a second stage unit (0.025 ha).

Estimate of the population mean calculated using Equation (5.30) is

$$\hat{\bar{Y}} = \frac{1}{n\bar{M}} \sum_{i=1}^{n} \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$= \frac{1523230}{(8)(40)} = 105.78$$

$$s_b^2 = \frac{1}{(8-1)} \left[\left(\frac{1760}{1800} \times 118 - 105.25 \right)^2 + \left(\frac{1975}{1800} \times 107 - 105.25 \right)^2 + \ldots + \left(\frac{1865}{1800} \times 90 - 105.25 \right)^2 \right]$$

=140.36

Estimate of variance of $\hat{\bar{Y}}$ obtained by Equation (5.32) is

$$\hat{V}(\hat{\bar{Y}}) = \left(\frac{1}{8} - \frac{1}{40} \right) 140.3572 + \frac{1}{(8)(40)} (465.1024)$$

=15.4892

$$SE(\hat{\bar{Y}}) = \sqrt{15.4892} = 3.9356$$

$$RSE(\hat{\bar{Y}}) = \frac{3.9356 \times 100}{105.78} = 3.72\%$$

Multiphase sampling

Multiphase sampling plays a vital role in forest surveys with its application extending over continuous forest inventory to estimation of growing stock through remote sensing. The essential idea in multiphase sampling is that of conducting separate sampling investigations in a sequence of phases starting with a large number of sampling units in the first phase and taking only a subset of the sampling units in each successive phase for measurement so as to estimate the parameter of interest with added precision at relatively lower cost utilizing the relation between characters measured at different phases. In order to keep things simple, further discussion in this section is restricted to only two phase sampling.

A sampling technique which involves sampling in just two phases (occasions) is known as two phase sampling. This technique is also referred to as double sampling. Double sampling is particularly useful in situations in which the enumeration of the character under study (main character) involves much cost or labour whereas an auxiliary character correlated with the main character can be easily observed. Thus it can be convenient and economical to take a large sample for the auxiliary variable in the first phase leading to precise estimates of the population total or mean of the auxiliary variable. In the second phase, a small sample, usually a sub-sample, is taken wherein both the main character and the auxiliary character may be observed and using the first phase sampling as supplementary information and utilising the ratio or regression estimates, precise estimates for the main character can be obtained. It may be also possible to increase the precision of the final estimates by including instead of one, a number of correlated auxiliary variables. For example, in estimating the volume of a stand, we may use diameter or girth of trees and height as auxiliary variables. In estimating the yield of tannin materials from bark of trees certain physical measurements like the girth, height, number of shoots, etc., can be taken as auxiliary variables.

Like many other kinds of sampling, double sampling is a technique useful in reducing the cost of enumerations and increasing the accuracy of the estimates. This technique can be used very advantageously in resurveys of forest areas. After an initial survey of an area, the estimate of growing stock at a subsequent, period, say 10 or 15 years later, and estimate of the change in growing stock can be obtained based on a relatively small sample using double sampling technique.

Another use of double sampling is in stratification of a population. A first stage sample for an auxiliary character may be used to sub-divide the population into strata in which the second (main) character varies little so that if the two characters are correlated, precise estimates of the main character can be obtained from a rather small second sample for the main character.

It may be mentioned that it is possible to couple with double sampling other methods of sampling like multistage sampling (sub-sampling)known for economy and enhancing the

accuracy of the estimates. For example, in estimating the availability of grasses, canes, reeds, etc., a two-stage sample of compartments (or ranges) and topographical sections (or blocks) may be taken for the estimation of the effective area under the species and a sub-sample of topographical sections, blocks or plots may be taken for estimating the yield.

Selection of sampling units

In the simplest case of two phase sampling, simple random sampling can be employed in both the phases. In the first step, the population is divided into well identified sampling units and a sample is drawn as in the case of simple random sampling. The character *x* is measured on all the sampling units thus selected. Next, a sub-sample is taken from the already selected units using the method of simple random sampling and the main character of interest (*y*) is measured on the units selected. The whole procedure can also be executed in combination with other modes of sampling such as stratification or multistage sampling schemes.

Parameter estimation

Regression estimate in double sampling :

Let us assume that a random sample of *n* units has been taken from the population of *N* units at the initial phase to observe the auxiliary variable *x* and that a random sub-sample of size *m* is taken where both *x* and the main character *y* are observed.

Let $\bar{x}_{(n)}$ = mean of *x* in the first large sample = $\bar{x}_{(n)} = \sum_{i=1}^{n} \frac{x_i}{n}$ (5.35)

$\bar{x}_{(m)}$ = mean of *x* in the second sample = $\bar{x}_{(m)} = \sum_{i=1}^{m} \frac{x_i}{m}$ (5.36)

$\bar{y}$ = mean of y in the second sample = $\bar{y} = \sum_{i=1}^{m} \frac{y_i}{m}$ (5.37)

We may take $\bar{y}$ as an estimate of the population mean $\bar{Y}$. However utilising the previous information on the units sampled, a more precise estimate of $\bar{Y}$ can be obtained by calculating the regression of *y* on *x* and using the first sample as providing supplementary information. The regression estimate of $\bar{Y}$ is given by

$$\bar{y}_{(drg)} = \bar{y} + b(\bar{x}_{(n)} - \bar{x}_{(m)}) \quad (5.38)$$

where the suffix *(drg)* denotes the regression estimate using double sampling and *b* is the regression coefficient of *y* on *x* computed from the units included in the second sample of size *m*. Thus

$$b = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}_{(m)}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{m}\left(x_i - \bar{x}_{(m)}\right)^2} \quad (5.39)$$

The variance of the estimate is approximately given by,

$$V(\bar{y})_{(drg)} = \frac{s_{y.x}^2}{m} + \frac{s_{y.x}^2 - s_y^2}{n} \quad (5.40)$$

where $$s_{y.x}^2 = \frac{1}{m-2}\left[\sum_{i=1}^{m}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{m}(x_i - \bar{x}_{(m)})^2\right] \quad (5.41)$$

$$s_y^2 = \sum_{i=1}^{m}\frac{(y_i - \bar{y})^2}{m-1} \quad (5.42)$$

(ii) Ratio estimate in double sampling

Ratio estimate is used mainly when the intercept in the regression line between *y* and *x* is understood to be zero. The ratio estimate of the population mean $\bar{Y}$ is given by

$$\bar{y}_{(dra)} = \frac{\bar{y}}{\bar{x}_{(m)}}\bar{x}_{(n)} \quad (5.43)$$

where $\bar{y}_{dra}$ denotes the ratio estimate using double sampling. The variance of the estimate is approximately given by

$$V(\bar{y}_{dra}) = \frac{s_y^2 - 2\hat{R}s_{yx} + \hat{R}^2 s_x^2}{m} + \frac{2\hat{R}s_{yx} - \hat{R}^2 s_x^2}{n} \quad (5.44)$$

where

$$s_y^2 = \frac{\sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2}{m-1} \quad (5.45)$$

$$s_{yx} = \frac{\sum_{i=1}^{m}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}_{(m)}\right)}{m-1} \quad (5.46)$$

$$s_x^2 = \frac{\sum_{i-1}^{m}\left(x_i - \bar{x}_m\right)^2}{m-1} \quad (5.47)$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}_{(m)}} \quad (5.48)$$

An example of analysis of data from double sampling using regression and ratio estimate is given below. Table 5.5 gives data on the number of clumps and the corresponding weight of grass in plots of size 0.025 ha, obtained from a random sub-sample of 40 plots taken from a preliminary sample of 200 plots where only the number of clumps was counted.

Table 5.5. Data on the number of clumps and weight of grass in plots selected through a two phase sampling procedure.

< TD WIDTH="16%" VALIGN="TOP">

60

Serial number	Number of clumps	Weight in kg	Serial number	Number of clumps	Weight in kg
	(*x*)	(*y*)		(*x*)	(*y*)
1	459	68	21	245	25
2	388	65	22	185	50
3	314	44	23	59	16
4	35	15	24	114	22
5	120	34	25	354	59
6	136	30	26	476	63
7	367	54	27	818	92
8	568	69	28	709	64

9	764	72	29	526	72
10	607	65	30	329	46
11	886	95	31	169	33
12	507	32	648	74	
13	417	72	33	446	61
14	389	60	34	86	32
15	258	50	35	191	35
16	214	30	36	342	40
17	674	70	37	227	40
18	395	57	38	462	66
19	260	45	39	592	68
20	281	36	40	402	55

Here, $n = 200$, $m = 40$. The mean number of clumps per plot as observed from the preliminary sample of 200 plots was $\bar{x}_{(n)} = 374.4$.

$$\sum_{i=1}^{40} x_i = 15419, \quad \sum_{i=1}^{40} y_i = 2104,$$

$$\sum_{i=1}^{40} x_i^2 = 7744481, \quad \sum_{i=1}^{40} y_i^2 = 125346, \quad \sum_{i=1}^{40} x_i y_i = 960320$$

$$\sum_{1}^{40} (x_i - \bar{x}_{(m)})^2 = \sum_{i=1}^{40} x_i^2 - \frac{\left(\sum_{1}^{40} x_i\right)^2}{40} = 7744481 - \frac{\left(15419^2\right)}{40} = 1800842$$

$$\sum_{1}^{40} (y_i - \bar{y})^2 = \sum_{i=1}^{40} y_i^2 - \frac{\left(\sum_{i=1}^{40} y_i\right)^2}{40} = 125346 - \frac{(2104)^2}{40} = 146756$$

$$\sum_{1}^{40}\left(x_i - \bar{x}_{(m)}\right)\left(y_i - \bar{y}\right) = \sum_{1}^{40} x_i y_i - \frac{\sum_{1}^{40} x_i \sum_{1}^{40} y_i}{40} = 960320 - \frac{15419 \times 2104}{40} = 149280.6$$

Mean number of clumps per plot from the sub-sample of 40 plots is

$$\bar{x}_{(m)} = \frac{15419}{40} = 385.5$$

Mean weight of clumps per plot from the sub-sample of 40 plots

$$\bar{y} = \frac{2104}{40} = 52.6$$

The regression estimate of the mean weight of grass in kg per plot is obtained by using Equation (5.38) where the regression coefficient *b* obtained using Equation (5.39) is

$$b = \frac{149280.6}{1800842} = 0.08$$

Hence, $\bar{y}_{(drg)} = 52.6 + 0.08(374.4 - 385.5)$

= 52.6 - 0.89

= 51.7 kg /plot

$$s^2_{y.x} = \frac{1}{40-2}\left[14675.6 - (0.08)^2(1800842)\right]$$

= 82.9

$$s^2_y = \frac{14675.6}{39}$$

=376.297

The variance of the estimate is approximately given by Equation (5.40)

$$V(\bar{y})_{(drg)} = \frac{82.9}{40} + \frac{82.9 - 376.297}{200} \quad (5.40)$$

= 3.5395

The ratio estimate of the mean weight of grass in kg per plot is given by Equation (5.43)

$$\bar{y}_{(ra)} = \frac{52.6}{385.5}(374.4)$$

= 51.085

$$s_{yx} = \frac{149280.6}{40-1}$$

= 3827.708

$$s_x^2 = \frac{1800842}{40-1}$$

= 46175.436

$$\hat{R} = \frac{52.6}{385.5}$$

= 0.1364

The variance of the estimate is approximately given by Equation (5.44) is

$$V(\bar{y}_{ra}) = \frac{376.297 - 2(0.1364)(3827.708) + (0.1364)^2(46175.436)}{40}$$

$$+\frac{(2)(0.1364)(3827.708) - (0.1364)^2(46175.436)}{200}$$

= 5.67

Probability Proportional to Size (PPS) sampling

In many instances, the sampling units vary considerably in size and simple random sampling may not be effective in such cases as it does not take into account the possible importance of the larger units in the population. In such cases, it has been found that ancillary information about the size of the units can be gainfully utilised in selecting the sample so as to get a more efficient estimator of the population parameters. One such method is to assign unequal probabilities for selection to different units of the population. For example, villages with larger geographical area are likely to have larger area under food crops and in estimating the production, it would be desirable to adopt a sampling scheme in which villages are selected with probability proportional

to geographical area. When units vary in their size and the variable under study is directly related with the size of the unit, the probabilities may be assigned proportional to the size of the unit. This type of sampling where the probability of selection is proportion to the size of the unit is known as 'PPS Sampling'. While sampling successive units from the population, the units already selected can be replaced back in the population or not. In the following, PPS sampling with replacement of sampling units is discussed as this scheme is simpler compared to the latter.

Methods of selecting a pps sample with replacement

The procedure of selecting the sample consists in associating with each unit a number or numbers equal to its size and selecting the unit corresponding to a number chosen at random from the totality of numbers associated. There are two methods of selection which are discussed below:

(i) Cumulative total method: Let the size of the *i*th unit be x_i, ($i = 1, 2, \ldots, N$). We associate the numbers 1 to x_i with the first unit, the numbers (x_1+1) to (x_1+x_2) with the second unit and so on such that the total of the numbers so associated is $X = x_1 + x_2 + \ldots + x_N$. Then a random number r is chosen at random from 1 to X and the unit with which this number is associated is selected.

For example, a village has 8 orchards containing 50, 30, 25, 40, 26, 44, 20 and 35 trees respectively. A sample of 3 orchards has to be selected with replacement and with probability proportional to number of trees in the orchards. We prepare the following cumulative total table:

Serial number of the orchard	Size (x_i)	Cumulative size	Numbers associated
1	50	50	1 - 50
2	30	80	51 - 80
3	25	105	81 -105
4	40	145	106 -145
5	26	171	146 - 171
6	44	215	172 - 215
7	20	235	216 - 235
8	35	270	236 - 270

Now, we select three random numbers between 1 and 270. The random numbers selected are 200, 116 and 47. The units associated with these three numbers are 6^{th}, 4^{th}, and 1^{st} respectively. And hence, the sample so selected contains units with serial numbers, 1, 4 and 6.

(ii) Lahiri's Method: We have noticed that the cumulative total method involves writing down the successive cumulative totals which is time consuming and tedious, especially with large populations. Lahiri in 1951 suggested an alternative procedure which avoids the necessity of writing down the cumulative totals. Lahiri's method consists in selecting a pair of random numbers, say (i,j) such that $1 \leq i \leq N$ and $1 \leq j \leq M$; where M is the maximum of the sizes of the N units of the population. If $j \leq X_i$, the ith unit is selected: otherwise, the pair of random number is rejected and another pair is chosen. For selecting a sample of n units, the procedure is to be repeated till n units are selected. This procedure leads to the required probabilities of selection.

For instance, to select a sample of 3 orchards from the population in the previous example in this section, by Lahiri's method by PPS with replacement, as $N = 8$, $M = 50$ and $n = 3$, we have to select three pairs of random numbers such that the first random number is less than or equal to 8 and the second random number is less than or equal to 50. Referring to the random number table, three pairs selected are (2, 23) (7,8) and (3, 30). As in the third pair $j > X_i$, a fresh pair has to be selected. The next pair of random numbers from the same table is (2, 18) and hence, the sample so selected consists of the units with serial numbers 2, 7 and 2. Since the sampling unit 2 gets repeated in the sample, the effective sample size is two in this case. In order to get an effective sample size of three, one may repeat the sampling procedure to get another distinct unit.

Estimation procedure

Let a sample of n units be drawn from a population consisting of N units by PPS with replacement. Further, let (y_i, p_i) be the value and the probability of selection of the ith unit of the sample, $i = 1, 2, 3, \ldots, n$.

An unbiased estimator of population mean is given by

$$\hat{\bar{Y}} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{p_i} \quad (5.49)$$

An estimator of the variance of above estimator is given by

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{n(n-1\_N^2}\left(\sum_{i=1}^{N}\left(\frac{y_i}{p_i}\right)^2 - n\hat{Y}^2\right) \quad (5.50)$$

where $p_i = \frac{x_i}{X}$, $\hat{Y} = N\hat{\bar{Y}}$

For illustration, consider the following example. A random sample 23 units out of 69 units were selected with probability proportional to size of the unit (compartment) from a forest area in U.P. The total area of 69 units was 14079 ha. The volume of timber determined for each selected compartment are given in Table 5.6 along with the area of the compartment.

Table 5. 6. Volume of timber and size of the sampling unit for a PPS sample of forest compartments.

Serialno.	Size in ha (x_i)	Relative size (x_i/X)	Volume in m^3 (y_i)	$\frac{y_i}{p_i} = v_i$	$(v_i)^2$
1	135	0.0096	608	63407.644	4020529373.993
2	368	0.0261	3263	124836.351	15584114417.014
3	374	0.0266	877	33014.126	1089932493.652
4	303	0.0215	1824	84752.792	7183035765.221
5	198	0.0141	819	58235.864	3391415813.473
6	152	0.0108	495	45849.375	2102165187.891
7	264	0.0188	1249	66608.602	4436705896.726
8	235	0.0167	1093	65482.328	4287935235.716
9	467	0.0332	1432	43171 .580	1863785345.581
10	458	0.0325	3045	93603.832	8761677342.194
11	144	0.0102	410	40086.042	1606890736.502
12	210	0.0149	1460	97882.571	9580997789.469
13	467	0.0332	1432	43171.580	1863785345.581
14	458	0.0325	3045	93603.832	8761677342.194
15	184	0.0131	1003	76745.853	5889925992.739
16	174	0.0124	834	67482.103	4553834285.804
17	184	0.0131	1003	76745.853	5889925992.739
18	285	0.0202	2852	140888.800	19849653965.440
19	621	0.0441	4528	102656.541	10538365422.979
20	111	0.0079	632	80161.514	6425868248.777
21	374	0.0266	877	33014.126	1089932493.652

22	64	0.0045	589	129570.797	16788591402.823
23	516	0.0367	1553	42373.424	1795507096.959
				1703345.530	147356252987.120

Total area X = 14079 ha.

An unbiased estimator of population mean is obtained by using Equation (5.49).

$$\hat{\bar{Y}} = \frac{1}{(23)(69)}(1703345.530)$$

= 1073.312

An estimate of the variance of $\hat{\bar{Y}}$ is obtained through Equation (5.50).

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{1}{23(23-1)(69)^2}(147356252987.120-(23)(67618.632))$$

= 17514.6

And the standard error of $\hat{\bar{Y}}$ is $\sqrt{17514.6}$ = 132.343.

--------*D Guha*-------