Chapter 15

Fundamental survey sampling techniques and its applications in nutritioal labelling

C.G. Joshy

Scientist

Central Institute of Fisheries Technology, Cochin - 682 029

The generation of nutrient data for nutritional labeling compliance is a complex task comprised of several steps. The foremost important step is the development of a well-designed and economically feasible sampling plan by taking in to account the nutrient categories established by different authorities. In addition, the sampling plan must take into consideration those factors responsible for variation in the nutrient composition of foods as well as analytical variation.

The first step in determining the values to be declared on the nutrition label is to characterize any product in terms of nutrient content and nutrient variability. This information is critical for determining the nutrients for which laboratory analysis should be performed and the number and the source of the samples to be analyzed. Characterization of the product will accomplish following objectives:

1. Determining the nutrients present in the product and approximate levels for each.
2. Estimating the variability of nutrient content of the product.
3. Identifying potential sources of nutrient variability (eg. ingredients, processing methods and post-processing storage and handling).
4. Identifying nutrients of special significance for the product (eg. Class I nutrients and/or nutrients for which nutrient content or health claims will be made).

To accomplish the above objectives, a scientific oriented data collection through a

significant sampling scheme is required. This will help the examiner timely end up of the examination and reduce the cost by increasing the accuracy. Once the characterization of the nutrient profile of a food product is done, it is also important to quantify the variability in nutrient content of the food product.

Major sources of variability in nutrient composition

Foods are inherently variable in composition, and the approach to sampling and the design of the sampling and analytical protocols need to take account of these factors.

1. Geographical samples

In a single country there may be a wide diversity of soil and climatic conditions, resulting in significant variations in food composition. Variations in food marketing and food preparation within different parts of a country–or among countries in the case of a multicountry database may also produce notable variance. For these reasons, geographically-specific data may be presented in the database as a supplement to nationwide and/or region wide averages. In other countries, the variations may be of similar magnitude to those due to other causes, in which case the national sample could be weighted according to the proportions of the population living in the regions or the proportions of the total consumption of the foods.

2. Seasonal samples

Seasonal variations in nutrient composition need to be accommodated in the combined protocols. Plant foods are especially prone to variation, particularly in their water, carbohydrate and vitamin content. Fish also show seasonal variations, especially in fat content, and milk and milk products exhibit variations in vitamin content primarily due to seasonal differences in feeding patterns. The collection of samples needs to be organized, in terms of timing and frequency, to reflect these variations. In some cases, seasonal data need to be given separately in the database. The analytical measurements of the seasonal samples can often be restricted to those nutrients showing variation.

3. Physiological state and maturity

The states of maturity of plants and animal foods cause variation in composition in the concentrations of sugars, organic acids and vitamins in many plants, and of fats and some minerals in animal foods. Some of these variations are a consequence of seasonal effects.

4. Cultivar and breed

These may be a significant source of variation for some nutrients and the combined protocols will need to provide for this variation. It is desirable to document this cultivar or breed variation within the database. Some research organizations sample specifically to capture cultivar and breed differences. The significance of the differences attributable to cultivar or breed can only be ascertained by controlling for other factors that can influence variation, and by sampling and analyzing individually, not in composite, a large number of samples.

Measures of nutrient variability

There are several measures of variability generally associated with analytical data.

1. Variance
2. Standard deviation
3. Coefficient of variation
4. Standard error of the mean

1. Variance

The simple variance is calculated from the formula $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

2. Standard deviation

Standard deviation is the square root of the variance

3. Coefficient of variation (CV)

The CV is the ratio of Standard deviation to the mean multiplied by 100

4. Standard error of the mean

The standard error of the mean is actually a measure of the precision of the value of the mean given in the database rather than the variability of the nutrient content of the food. This is obtained by dividing standard deviation by the square root of the number of samples.

Before selecting a sampling plan to select samples of the product for laboratory analysis, one should examine the potential sources of nutrient variability in that food. There are numbers of factors which contribute to the nutrient variability of a food. The nutrient content of the ingredients and the relative amount of each ingredient in the product formulation will contribute to the overall nutrient composition and variability of the product. External factors such as conditions under which foods are processed or stored may also affect the nutrient content of the final product. Major sources of nutrient variability in foods include the following.

1. Inherent variability
2. Environmental factors
3. Processing
4. Post-processing handling and storage

Those factors which are responsible for nutrient variability of a food product may be reasonably controlled for in the sampling plan and it should be distinguished from those over which the producer has no control. This kind of information will help in designing an efficient sampling plan.

Designing a sampling plan

This stage includes all aspects of specifying how many and which units of the product will be subjected to laboratory analysis of nutrient content.

A. Considerations in sampling plan

In planning a sample, four criteria must be balanced.

1. Design the sample plan to meet the research objectives
2. The plan must allow computation of valid estimates of sampling variability
3. The plan must be practical, simple and economical
4. The plan must achieve the level of precision required

B. Determining the sampling frame/sampling point

A key decision in the sampling plan in the food production and distribution is the identification of list sample units to be sampled

C. Determining the sample size

1. The precision target

Based on 95% confidence level as the standard, the sample size is calculated using the formula

$$n = \frac{(1.96)^2 s^2}{m^2 p^2}$$

Where 'n' is the number of sample units must be analyzed to achieve the target precision

's' is the standard deviation of the nutrient content of the product

'm' is the mean nutrient content of the product and

'p' is the precision goal expressed in decimal form

D. Estimating the different sources of nutrient variability

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

In general, 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 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.

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, food items 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 food retail shop survey on compliance of food nutrition labeling. Usually a retail shop is considered as the sampling unit since it is found to be convenient for sampling and for ascertaining the required information. In an 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, a list of

retail shops in a nutrient compliance survey etc. 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 laws of probability. In other words, a random or probability sample is a sample drawn in such a manner that each unit in the population has a pre-determined probability of selection. For example, if a population consists of *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 of selection 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 judgement or personal expertize 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 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,...,N). For instance, the unit may be nutrient contents of a food or may be the farm and the characteristic under study or 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 sample 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 $1/\binom{N}{n}$.

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 to the fact that 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 $(t_i - \theta)$, 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 known 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 sample (a fraction of population) is termed sampling error. The sampling error is zero in a complete enumeration survey since the whole population is surveyed.

The errors other than sampling errors such as those arising through non-response, incompleteness 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.

We shall now describe different sampling methods

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. There are two types of methods to draw sample units from the population using simple random sampling

1. Simple random sampling with replacement (SRSWR)
2. Simple random sampling without replacement (SRSWOR)

In Simple random sampling with replacement each unit selected in the sample is returned to the population before next unit is drawn from the population. The probability of selecting each unit being selected in the sample is

$\frac{1}{N}$ at each drawand the population size remains same after each draw; it remains same even when included in the sample more than once. If N denotes the population size and n is the sample size, then the number of all possible samples in SRSWR is N^n. The probability of drawing any of these N^n is $1/N^n$. For example, if there is a population of size N=4, with unit numbers as 1,2,3,4. Then $4^2=16$. All possible samples selected through SRSWR are (1,1),(1,2),(1,3),(1,4);(2,1),(2,2),(2,3),(2,4);(3,1),(3,2),(3,3),(3,4); (4,1),(4,2),(4,3),(4,4). Each sample can be selected with equal probability of 1/16. Another way is to select by selecting one unit at a time. Therefore, probability of selection of one unit at a time is ¼. Since, each unit

is selected 4 times in all possible 16 samples. Therefore, probability of inclusion of each unit in the sample is 4/16 = n/N.

In simple random without replacement a unit that has been drawn from the population will be removed for the subsequent draws. Based on this method, n units out of the N units can be selected from every one of the

$\binom{N}{n}$ distinct samples and each have an equal chance of being drawn. 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 random number table or by lottery method especially when the population size is small. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. Continuing with the same example of 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);(2,3),(2,4);(3,4). The total number of samples is

$\binom{5}{2}$ =6. Any of these samples have probability of selection. Therefore, the probability of selection of each of the samples is 1/6. Similar to SRSWR, the probability of selectionof each unit at any draw is 1/N. For example, if the i^{th} unit is selected at the first draw, its probability of selection is 1/N. Probability of selecting the i^{th} unit at the second draw is 1/N-1; Probability of selecting the i^{th} unit at the third draw is 1/N-2; similarly Probability of selecting the ith unit at the r^{th} draw is 1/[N-(r-1)]; Therefore, probability of selecting the i^{th} unit at the r^{th} draw is = (1-1/N)(1-1/(N-1))(1-1/(N-r+2))(1/(N-r+1))=1/N. Its simplification for r=3 is (1-1/N).(1-1/(N-1)).(1/(N-2)) =

$\frac{N-1}{N}.\frac{N-2}{N-1}.\frac{1}{N-2}=\frac{1}{N}$. Probability of inclusion of any unit in the sample is n/N.

Simple random sampling serves as a baseline for comparing the relative efficiency of other sampling methods and the difference between with replacement and without replacement becomes less important when the population size is large and the sample size is noticeably smaller than it.

Procedure of selecting a random sample

There are mainly two procedures used for selecting simple random sample

1. Lottery method
2. Using random number table

Lottery Method: Each unit in the population with unique identification mark from 1 to N will be associated with a chit/ticket. All the chits/tickets are in a material where a thorough mixing is possible before each draw. Chits/tickets will be drawn one by one and continued until required sample size is obtained. When the population size is large, this method becomes big task and human bias may also creep. Therefore, this method is generally discouraged.

Use of Random Number Tables: A random number table is a linear or rectangular arrangement of digits 0 to 9, 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. Random number tables are the tables of digits 0, 1, 2, ...,9 each digit having an equal chance of selection at any draw. This was much more effective than manually selecting the random samples (with dice, cards, etc.). Nowadays, tables of random numbers have been replaced by computational random number generators. 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.

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 three-digit numbers, we can obtain numbers from 000 to 999, all having the same frequency. Similarly, four or more digit numbers may be obtained by combining four 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 is 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 be highest multiple 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 $N = 225$, the highest three-digit multiple of 225 is 900. For selecting a unit, one random number from 001 to 900 has to be selected. Let the random number selected be 325. Dividing 325 by 225, gives the remainder as 100. Hence, the unit with serial number 100 is selected in the sample. Suppose that another random number selected is 845. Divide 845 by 225, which leave 170 as remainder. So the unit bearing the serial number 170 is selected. Similarly, if the random number selected is 450, then divide 450 by 225, it leaves remainder as 0. So the unit bearing serial number 225 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 0to $(N^* - 1)$. 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 $N = 12$ and hence $N^* = 96$ and $d = 96/12 = 8$. Let the two-digit random number chosen be 65 which lies between 0 and 95. Dividing 65 by 8, the quotient is 8 and hence the unit bearing serial number (8 - 1) = 7 is selected in the sample. Further, if the random number selected is 2, then the quotient is $2/8 = 0$, and $q - 1 = -1$. The unit selected is 11. Similarly, if the random number selected is 9, then the quotient is $9/8 = 1$, and $q - 1 = 0$. The unit selected is 12.

Estimation of Population Total

Let Y be the character of interest and $Y_1, Y_2, \cdots, Y_i, \cdots, Y_N$ be the values of the character on N units of the population. Further, let $y_1, y_2, \cdots, y_i, \cdots, y_n$ be the sample of size n selected by simple random sampling without replacement. For the total

$$T = \sum_{i=1}^{N} Y_i \text{ and Sample mean } \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

we have an estimator $\hat{t} = N\sum_{i=1}^{n} y_i / n = N\bar{y}_n$ i.e., the sample mean $\bar{y}_n$ is multiplied by the population size N.

The estimator can be expressed as

$$\hat{t} = \sum_{i=1}^{n} w_i y_i = (N/n)\sum_{i=1}^{n} y_i \text{ , where } w_i = N/n.$$

The constant N/n is the sampling weight and is the inverse of the sampling fraction n/N.

A design variance for $\hat{t}$ is

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

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ is the population mean and $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ 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

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

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

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

is the sample mean and s^2 is an unbiased estimator of the population variance S^2. The variance of $\bar{y}$ is

$$Var(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 = \frac{N-n}{N}\cdot\frac{S^2}{n}.$$

The unbiased estimator of the $Var(SRS)$ is $Var(\bar{y})$ is

$\hat{Var}(SRS) = \left(\frac{1}{n} - \frac{1}{N}\right)s^2$. 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 sampling is a method of collecting sample from a mutually exclusive sub-populations, or strata, of the population. In this method units in the population are divided into a different group or strata such that units within the strata are homogeneous and between strata are heterogeneous. Then random or systematic sampling is applied within each stratum to collect required sample size n. This method often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population. Stratification is efficient when the units within the stratum are homogeneous and between strata are heterogeneous.

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 'k' strata of sizes $N_1, N_2, \cdots, N_k$, and select samples of size $n_1, n_2, \cdots, n_k$, respectively, from the 'k' strata, the allocation is proportional if

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

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

Sample mean $\bar{y}_{st} = \frac{1}{n}\sum_{i=1}^{n} w_i \bar{y}_i$, *where* $w_i = \frac{n_i}{n}$

Sample variance $V(\bar{y}_{st}) = \sum_{i=1}^{k} W_i^2 \frac{N_i - n_i}{N_i} \frac{s_i^2}{n_i}$ *where* $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$

Example: Stratified sampling can be applied to study the presence of heavy metal content in a particular species of fish in the Kerala state by dividing different districts into different strata to accommodate the variability in the water bodies in different places.

Cluster Sampling

Cluster Sampling is a sampling technique used when "natural" groupings are evident in a population. In this technique, the total population is divided into different groups (or clusters) and a sample of the groups is selected. Then the required information is collected from the elements within each selected group. This may be done for every element in these

groups or a sub-sample of elements may be selected within each of these groups. Elements within a cluster should ideally be as heterogeneous as possible, but there should be homogeneity between cluster means. Each cluster should be a small scale representation of the total population. The clusters should be mutually exclusive and collectively exhaustive. A random sampling technique is then used on any relevant clusters to choose which clusters to include in the study. In single-stage cluster sampling, all the elements from each of the selected clusters are used. In two-stage cluster sampling, a random sampling technique is applied to the elements from each of the selected clusters.

The main difference between cluster sampling and stratified sampling is that in cluster sampling the cluster is treated as the sampling unit. So analysis is done on a population of clusters (at least in the first stage). In stratified sampling, the analysis is done on elements within strata. In stratified sampling, a random sample is drawn from each of the strata, whereas in cluster sampling only the selected clusters are studied. The main objective of cluster sampling is to reduce costs by increasing sampling efficiency. This contrasts with stratified sampling where the main objective is to increase precision.

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 first 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.

Estimation of population mean

Population mean $\bar{Y}=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=}^{M}Y_{ij}$ *where* $i=1,2,\cdots N,\ j=1,2,\cdots M$

Mean square between elements in the ith cluster $S_i^2=\frac{1}{M-1}\sum_{j=1}^{M}(Y_{ij}-\bar{Y}_i)^2$

Mean square within clusters $S_w^2=\frac{1}{N}\sum_{i=1}^{N}S_i^2$

Mean square between cluster means $S_b^2=\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_{i.}-\bar{y})^2$

Sample mean of cluster means $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$

Variance of ample mean $V(\bar{y})=\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2$, *Wheres*$_b^2=\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.}-\bar{y})$

Multistage Sampling

Multistage sampling is a complex form of cluster sampling. In 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.

Example: In a food retail shop survey on compliance of nutritional labeling information for raw fruit and vegetables and raw fish of a particular state, each district is considered as first stage sampling units, from each districts some corporations are selected as secondary stage sampling units, finally from each corporation some retail shops are selected as tertiary units (final stage units).This will help to segregate the whole retail shops into different groups to conduct the survey in a smooth way.

Systematic Sampling

In systematic sampling procedure only the first unit is selected at random, the rest of the n^{-1} units will be selected automatically according to a predetermined manner. In this sampling technique total units (N) in the population will be serially numbered from 1 to N. Then a sampling interval *k* is determined as k=N/n. Draw a random number less than or equal to k, say 'i' and select the unit with the corresponding serial number and every k^{th} unit in the population thereafter. The sample will contain the n units with serial numbers i, i+k, i+2k, ——, i+(n-1)k and such a sample is known as systematic sample.

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.

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 12^{th} 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.

A systematic sample also offers 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 fail 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

In probability proportional to size (PPS) sampling procedure units are selected with probabilities proportional to the size of the unit if the units vary in size and the variable under study is *prima facie* correlated with size. It will take into account the possible importance of larger units in the population.

Sample selection

The following procedure will explain how to draw a sample of size n from a population of size N with probability proportional to size with replacement. If X_i is an integer proportional to the size of the i^{th} unit i=1,2,——,N, we form successive cumulative totals $X_1, X_1+X_2, X_1+X_2+X_3, \cdots, \sum_{i=1}^{n} X_i$ and draw random number R not exceeding $\sum_{i=1}^{n} X_i$ from a random numbers. If $X_1+X_2+X_3\cdots+X_{i-1} < R \leq X_1+X_2+X_3\cdots+X_i$, the i^{th} unit is selected. Repeat the procedure n times to get a sample of size n. The main drawback of this procedure

is that it involves writing down the successive cumulative totals which is time consuming and tedious especially if the number of units in the population is large.

To overcome the above said drawbacks, Lahiri suggested another alternative method which avoids the necessity of writing down the cumulative totals. In this approach a pair of random numbers, say (i, j) is selected such that $1 \leq i \leq N$ *and* $1 \leq l \leq M$, where M is the maximum of the sizes of the N units in the population. If $j \leq X_i$, the i-th unit is selected; otherwise it is rejected and another pair of random numbers is chosen. For selecting a sample of n units with probability proportional to size and with replacement, the procedure is to be repeated till 'n' units are selected.

References

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

Shapiro, R. (1995). *Nutrition Labeling Handbook*. Marcel Dekker, Inc. New York.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S and Ashok, C. (1984). *Sampling Theory of Surveys with Applications*.3rd edition.Indian Society of Agricultural Statistics, New Delhi.