U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 103/ MASTAT – 103 Survey Sampling

Course Design Committee

Course Preparation Committee

Block: 1 ***Random Sampling Procedures - I***

PGSTAT – 103/ MASTAT – 103 **SURVEY SAMPLING**
©UPRTOU
First Edition: ***July 2021***
ISBN : : 978-93-94487-03-1

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Blocks & Units Introduction

The present SLM on ***Survey Sampling*** consists of ten Units with three Blocks.

The ***Block - 1 – Random Sampling Procedures - I,*** is the first block, which is divided into three units. It describes the concept of probability sampling with its application in sample surveys. The basic techniques and methodologies of some important sampling designs such as simple random sampling, stratified random sampling, systematic random sampling and others are explained with their properties.

In ***Unit – 1 – Basics of Sampling Theory***, is discussed about the related definitions and theories and also advantages of sampling over complete enumeration.

In ***Unit – 2- Simple Random Sampling***, the both methods of simple random sampling with and without replacement, their needs and advantages are explained.

Unit – 3 – Systematic Sampling is being introduced about the systematic sampling with its mean and variance.

The ***Block - 2 – Random Sampling Procedures – II*** is the second block with four units. It explains the stratified random sampling with its advance concepts and use of auxiliary information in the estimation of population parameters. Ratio and regression methods of estimation are explained to understand the application of auxiliary information with suitable examples. Further cluster and multi- stage sampling along with their properties have been explained. This unit also deals with the consequences and effect of non- sampling errors in sample surveys.

Unit – 4 – Stratified Sampling and Use of Auxiliary Information is discussed about the stratified sampling, its mean and variance and many more other theorems related with this.

Unit – 5 – Ratio and Regression Methods of Estimation have been introduced to the ratio and regression sampling with related theorems.

Unit – 6 – Cluster and Multi-Stage Sampling dealt with cluster and multi-stage sampling with related theories.

Unit – 7 – Response and Non Response Sampling dealt with the theory of response and non-response sampling.

The ***Block - 3 – Varying Probability Sampling*** has three units. It is consists of varying probability sampling in sixth and seventh units. The first unit of present block explains the procedure of selecting a sample and estimation of population mean, under probability proportional to size, with and without replacement. Des- Raj's estimator in ordered estimator is discussed in the second part of this unit. This block is also deals with the estimation of unordered estimators. Horwitz-Thompson Estimator in this class of unordered estimators is explained to

estimate the population mean. Midzuno System and Narain Method of sampling are also given with examples.

Unit – 8 – Sampling on Probability Proportional to Size, comprises the methods of selection of units in the sample and related theorems.

In ***Unit – 9 – Ordered Estimators***, ordered estimators with their requirements and properties are explained.

In ***Unit – 10 – Unordered Estimators***, Horvitz- Thompson estimators along with Midzuno system and Narain method of sampling have been given with suitable example to understand the unordered estimators.

At the end of every block/unit the summary, self assessment questions and further readings are given.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 103/ MASTAT – 103 Survey Sampling

Block: 1 Random Sampling Procedures - I

Course Design Committee

Course Preparation Committee

Block: 1 ***Random Sampling Procedures - I***

PGSTAT – 103/ MASTAT – 103 **SURVEY SAMPLING**
©UPRTOU
First Edition: ***July 2021***
ISBN : : 978-93-94487-03-1

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The ***Block - 1 – Random Sampling Procedures - I,*** is the first block, which is divided into three units. It describes the concept of probability sampling with its application in sample surveys. The basic techniques and methodologies of some important sampling designs such as simple random sampling, stratified random sampling, systematic random sampling and others are explained with their properties.

In ***Unit – 1 – Basics of Sampling Theory***, is discussed about the related definitions and theories and also advantages of sampling over complete enumeration.

In ***Unit – 2- Simple Random Sampling***, the both methods of simple random sampling with and without replacement, their needs and advantages are explained.

Unit – 3 – Systematic Sampling is being introduced about the systematic sampling with its mean and variance.

At the end of unit the summary, self assessment questions and further readings are given.

Unit-1: Basics of Sampling Theory

Structure

1.0 Introduction

A population is the assortment of entity about whom information is required and a sample is a miniature of that collection. In day to day existence well beyond likewise in scientific research or any sort of enquiry about a population, or attitudes and actions are often based on samples. For example, when information about the proportion of good items in a lot of manufactured product, total no. of graduates in state, items in a lot of manufactured product, total no. of graduates in a state, average consumer expenditure in a city, total turnover of sales of an item, total area under a crop, total number of fish in a lake is required, often the estimates are observed on the basis of suitably selected sample.

Prior to studying the theory of sampling, it is essential to know about the *parameter* and *statistic.* The summery measures (like as mean, variance, correlation coefficient, etc.) of population are known *parameters* of the population. Same as summery measures which drawn from sample is known as *statistic*. Hence parameter and statistic are the important component of any further statistical analysis.

It is supposed that the population is well clear and consists of a finite number N (usually fairly huge) of individuals, called units, a_1, a_2 ……a_N. The meaning of a population might be simple, as if there should be an occurrence of electric bulbs fabricated in a manufacturing plant, however not so as in the event of structures or fields in towns where clear determination is important to deal with fringe or farfetched cases. Likewise the units comprising populace should be characterized appropriately relying on the kind of data looked for. For instance, in the event of a populace of individuals the unit might be individual people or a group of

people or a gathering of families living in a region and so forth It is exceptionally valuable to have a total rundown of units in a populace which is known as a casing or an inspecting outline. In some cases it is difficult to control an edge as if there should be an occurrence of the fish populace in a lake.

It is likewise assumed that each unit in the population can be estimated, quantitatively or subjectively, concerning some trademark under study,, say, y may who for a_i is y_i (i=1, 2, --, N). For example, the unit may be person and y may be his age or income; the unit may be a family and y may be the total consumer expenditure in a particular month; or the unit may be a field and y may be its area of yield of a crop sown in it. Any function of the values of all the population units is known as parameter which is to be calculated. Two common parameters are 'total' and 'mean' and our basic problem is estimate them. These are:

Population Mean Total $Y= (y_1 + y_2 \text{ ------- } + y_n) / N$

There are two different ways of tracking down the worth of Y or $\bar{Y}$, either by estimating each unit in the population, called complete enumeration (or census) or estimating them on the basis of sampling or sample survey, which consist of selecting a sample suitably and using the sample values.

1.1 Objectives

After studying this unit you should be able to understand.

- The need of Sampling Theory
- Advantages of Sampling Theory
- Sampling and Non-sampling Errors

1.2 Advantages of Sampling over Complete Enumeration

(a) Greater Speed: Seeing as a sample consists of only few number of units, sampling requires less time than complete enumeration.

(b) Reduced Cost: Again, as a result of little size, sampling requires less expenditure than complete enumeration. The entire expenditure of complete enumeration is substantial for it involves huge administrative mechanism for the purpose.

(c) Greater Accuracy: The sampling technique includes few workforces who can be given proficient preparation and facilities to assemble information more correctly than in case of a huge enumeration for the total population.

(d) Greater Scope: The more modest extent of sampling process and the better preparation to which the investigators might be bare likewise permit the enquirer to called information on a larger number of items than it is feasible for a complete enumeration.

(e) Greater Applicability: Sometimes it is not effectively perhaps to afford suitable facilities to workforce for complete enumeration and then the only method accessible is sampling. Additionally when enumeration is a critical process sampling is a must.

(f) Provision of Measure of Accuracy: The complete enumeration includes a group of errors of measurement and no estimate of accuracy of result is feasible, even as sampling, done statically offer us with a measure of the accuracy of the estimate obtained.

1.3 Types of Errors

The errors emerging in the phases of measurement and handling of information are termed as sampling and non-sampling errors. These may common to both complete enumeration and sampling. The major kinds of such errors are the accompanying:

Errors of Measurement: At the point when the units are estimation by perception for example eye-evaluation of a reap district or yield, the assessment will depend upon the singular judgment of the enumerator and will be subject to mistakes. Commonly, it has been seen that eye-evaluation of yield is under appraisal while that of typical is over appraisal.

Such errors may in like manner arise due to respond tendencies. For example, individuals met may offer wrong reaction concerning one's tutoring or age offer under-articulation of pay or over clarification of expenses.

Sometimes examiner tendencies could manage in due to answer given by thoughts from the questioner or due to effect of examiner's conviction and predispositions in translating a couple of requests.

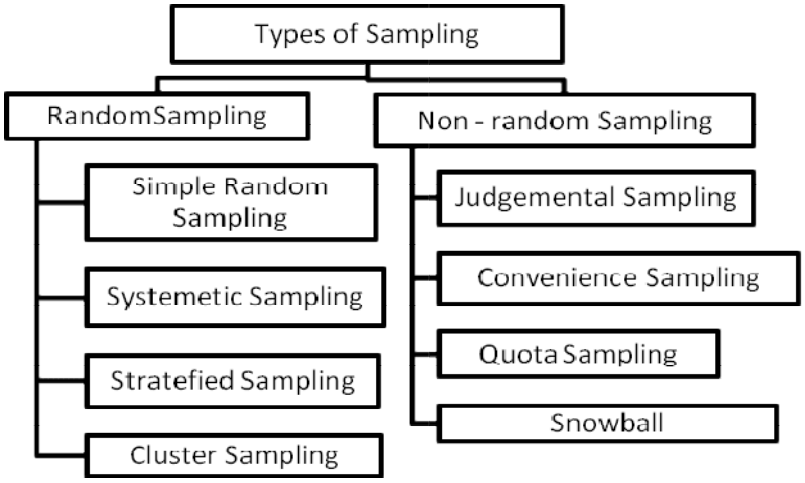
Errors of Non-Response: Sometimes examiner tendencies could manage in due to answer given by thoughts from the questioner or in view of effect of questioner's conviction and predispositions in unraveling a couple of requests.

Errors of Tabulation: arise due to deficient examination of the data, goofs in coding, computation and plan of data and slip-ups submitted during show and printing of results.

On the other hand, “sampling error” arise solely in light of sampling fluctuations. Since an sample is just a small part of the complete population,

there will normally be a qualification between the populace boundary and its sampling estimate. Complete enumeration is freed from such errors while the non-sampling errors for it are oftentimes significantly more unmistakable than in sample survey. The sampling error frequently diminishes with extension in sample size.

1.4 Types of Sampling



There are two types of sampling: Random Sampling and Non Random Sampling.

Random Sampling:

A sampling is supposed to be *random,* when it follows the scientific rules for choosing the samples as per laws of possibility. In which every unit of the population has an equivalent chance of being elected in the sample.

Simple Random Sampling (SRS):

In which sampling procedure, each unit of the population has equivalent independent opportunity to be elected in the sample. There are two kinds of simple random sampling, when

- The sample units are elected without any substitution or replacement (no element can be chosen more than once in the same sample), is known as *Simple random sampling without replacement (SRSWOR).*
- The sample units are elected with any substitution or replacement (an element may come out multiple times in the same sample), is known as *Simple random sampling with replacement (SRSWR).* (essentially it isn't utilized for additional investigation).

in a crate of apples, looking through the spoiled or harmed apple; first inquiry them and subsequent to looking through them quickly return them to the bushel and proceeding until test size not finished; this is a SRSWR plan, on the grounds that may wind up the review, estimating a similar spoiled or

harmed apple at least a couple of times. Be that as it may, on the off chance that it doesn't put get once again to the bin, this turns into a SRSWOR plan.

Systematic Sampling:

A sampling method, select each k^{th} component from a listing of population elements, after the first element has been randomly selected. In other words, only first unit has been selected at random and then all rest units being selected automatically according to predetermined pattern involving regular spacing *k* e.g. if k = 5 and sample size is 10, then after selecting the first unit randomly, every 5^{th} unit will going to be selected in the sample, until all10 units are not selected.

Hence systematic procedure is followed to choose a sample by taking every k^{th} individual where *k* refers to the sample interval, which is calculated by the formula;

$$k = \frac{\text{Total population}}{\text{Sample size desired}}$$

In field study, when the population is large and homogeneous, then systematic sampling is very simple and easy to adopt.

Stratified Sampling:

This is used when the population is heterogeneous (not homogeneous). A sample which is drawn after stratification follows two steps mainly; the primary step is to split the whole heterogeneous population into numerous non-overlapping sub-population which is internally homogeneous (as compared to whole population), these homogeneous groups called *strata* and secondly the units are drawn through SRS, from each *stratum* at random in proportion to its size and then finally combining all units together. This method gives more representative sample than SRS in a large heterogeneous population. In other words, it gives a proportionate representative sample from each group is secured and it gives greater accuracy.

Cluster Sampling:

In this sampling population is divided into sub population (separate groups of units) say *clusters,* where each unit of the population belongs to one and only one cluster. A SRS is taken from each cluster. Whenever the elements within cluster are heterogeneous, it tends to provide best results. Each cluster is the representative small scale version of the entire population. It is also known as ***Area Sampling*.**

Non-random Sampling:

A sampling is said to be *non-random* when it do not follow the scientific methods for selecting the samples and in which each unit of the

population has not an equal opportunity of being selected in the sample. In other words the sample units are selected without use of randomization.

Judgmental Sampling:

In this sampling the selection process is partially biased and the choice of sample units depends on the decision of the investigator e.g. for the cultural activity if sample of 100 students is to be selected form a school of 1000 students, the investigator would select 100 students, who, in his own opinion, are represent the 1000 students.

Convenience Sampling:

In this sampling the fraction of the population is being investigated which is selected without any rule, but by convenience of the investigator. Sometimes it is also known as *grab* or *opportunity sampling* or *accidental* or *haphazard sampling.* Sometimes it is used in exploratory study where the investigator is interested for receiving low-cost approximation.

Quota Sampling:

In this sampling, the selection of the samples is partially biased and also based on the preference of the investigator. Here data is collected from the homogeneous group but the sample units may not be selected randomly. It is a non-random version of stratified sampling e.g. for a study of health of 100 children, investigator choose children only from 1-5 and 10-15 age groups other may not be taken by investigator.

Snowball Sampling:

This is also known as *chain sampling* or *referral sampling* or *chain referral sampling*. In this sampling method the initial respondents are chosen by randomly or on the preference of investigator, after that in sequence collect the information from the additional respondents who were proposed by the initial respondents (respondents themselves recruit each other). It is used when required respondents are tough to find and the population is hard to reach.

1.5 Sampling Distribution

An sample ought to be a legitimate delegate of related population. The gauge in light of it ought to be close the reasonable worth of the parameter which is being anticipated. By and by we are not particularly focused on expecting the differentiation between the substantial and assessed esteem is enormous for a specific sample as long as such contrast are minute (or immaterial) in successive examples. Hence the inspecting strategy is basic. As a substitute of purposive or haphazard determination, method of probability or random sampling is adopted. This has the accompanying steps:

(i) We can describe the arrangement of foreordained samples S_1, S_2, S_3, which the system is operational for choosing from the population,

(ii) Every feasible sample Si has consigned to it a known probability of selection,

(iii) We choose one of the samples by scheme such that the i^{th}-sample Si receives its appropriate probability of being elected, For example, we may allocate equivalent likelihood to all conceivable Sample.

(iv) The technique for computing the estimate from the sample should be expressed and should prompt a remarkable unique estimate for any specific sample, For instance, the average of sample value be an estimate of the population mean.

For any sampling method, there is a frequency distribution of the estimator which it generates when the procedure is over and again applied to the similar population. This is known as the sampling distribution of the estimate.

1.6 Bias of an Estimator

Assume the parameter of the population is θ (Y or Y). Let Z be an estimate for which Z=Zi when the sample selected is Si (i=1,2,. , v) with probability π_i . The bias of the estimator is given by

$$B(Z) = E(Z) - \theta$$

$$= \sum_{i=1}^{v} \pi_i Z_i - \theta$$

The estimator is unbiased if B(Z) = 0 or E (Z) = θ

1.7 Measures of Sampling Error

Since probability sampling prompts different various samples, the etimate in view of the sample perceptions will vary from one sample to another and, likewise, go amiss from the value of the parameter. The distinction between the estimate Zi in light of the ith sample Si, and the parameter i.e.,(Zi=θ) , may be known as the the error of the estimate and this error shifts from one sample to another. A standard measure of the divergence of the estimator form the true value is given by

$$M(Z) = \sum_{i=1}^{v} (Zi - \theta)^2 \, \pi_i$$

This is known as the mean square error (M.S.E.) of the estimator. The M.S.E. can be seen as a measure of precision with which the estimator Z estimates the parameter. The square root of the M.S.E. is termed as "root mean square error." The precision is inversely proportional to M.E.S. The sampling variance of the estimator is characterized by

$$V(Z) = \sum_{i=1}^{v} [Zi - E(z)^2]\, \pi_i$$

So,

$$M(z) = V(z) + [B(z)]2$$

If the estimator is unbiased, M(z) = V(z). The positive square root of the variance of an estimator Z is called Standard error of Z. The sampling variance (or Standard error) may have been considered as a measure of precision of the estimator. The precision is inversely proportional to sampling variance.

1.8 Exercises

1. Define random sampling. Why is it preferred to other methods of selection of sample? How will you estimate the mean and its sampling error from a random sample from a finite population?

2. What are the advantages of sampling methods?

3. Define precision?
4. Discuss about the errors?
5. Explain different types of sampling methods?

1.9 Summary

The above unit explained about the definition of sample, population, complete enumeration and sampling errors, types of sampling procedures etc.

1.10 Further Readings

1. Cocharan, W.G. (1993). *Sampling Techniques*, Third Edition. Wiley Eastern Limited, New Delhi.
2. Sigh, D and Chaudhary, F.S. (1999). *Theory and Analysis of sample Survey Designs.* New Age International (P) Limited, Publishers, New Delhi.
3. Sukhatme, P.V., B.V., S and Asok, C (1984). *Sampling Theory of Surveys with Applications,* Third Revised Edition. Iowa State

University Press, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi-110112

4. Swain, A.K.P.C. (2003). *Finite Population Sampling Theory and Methods*. South Asian Publishers Pvt. Ltd., New Delhi-110014.

Unit-2: Simple Random Sampling

Structure

2 Introduction

The sampling procedure, in which every unit of the population, has equal autonomous chance to be selected in the sample. There are two types of simple random sampling, when the sample units are chosen without any replacement (no element can be chosen more than once in the same sample), is known as *Simple random sampling without replacement (SRSWOR).* The sample units are elected with any substitution or replacement (an element may come out multiple times in the same sample), is known as *Simple random sampling with replacement (SRSWR).* (essentially it isn't utilized for additional investigation). For example in a crate of apples, looking through the spoiled or harmed apple; first inquiry them and subsequent to looking through them quickly return them to the bushel and proceeding until test size not finished; this is a SRSWR plan, on the grounds that may wind up the review, estimating a similar spoiled or harmed apple at least a couple of times. Be that as it may, on the off chance that it doesn't put get once again to the bin, this turns into a SRSWOR plan.

2.1 Objectives

After studying this unit you should be able to understand

- SRSWOR (Simple Random Sampling without Replacement)
- SRSWR (Simple Random Sampling with Replacement)
- Mean and variance of SRS.

2.2 SIMPLE RANDOM SAMPLING

The simplest type of probability sampling where the probabilities associated with different possible samples are equal is called simple random

sampling procedure. In this procedure, the sample is drawn unit by unit with equal probability of selection for every unit in each draw. Suppose a simple random sample of n units is to be drawn from a population of N unites U_1, U_2, ----U_N, for which the values of the characteristic y under study are y_1, y_2, ----, y_N. The sample may be drawn in following two different ways.

2.3 Simple Random Sampling without Replacement (SRSWOR)

In such case, the n units of the sample are draw from the population one by one, the unit obtained at any draw not being replaced in the population, in such a way the probability of any unit in the first draw is 1/N, that of any unit in the second drawn is 1(N/-I),……., that of any unit in the rth draw is 1/(N-r+1) and so on. Therefore, the probability of drawing a sample of n units in SRSWOR is

$$\frac{n!}{N(N-1)\ldots\ldots(N-n+1)} = \frac{1}{\binom{N}{n}}$$

This means that there are $\binom{N}{n}$ possible samples; the probability of drawing each of these is the same.

The probability that a specified unit is selected at the rth draw in SRSWOR is

$$\frac{N-1}{N}, \frac{N-2}{N-1} - - - - - - \frac{N-r+1}{N-R+1}, \frac{1}{N-r+1} = \frac{1}{N}$$

Thus the probability that a specified unit is included in the sample at any draw is $\frac{1}{N}$ and probability of selecting the unit in the sample is n/N.

2.4 Simple Random Sampling with Replacement (SRSWR)

In such case, the n units of the sample are drawn from the population one, the units obtained at any draw being replaced in the population, in such a way that the probability of drawing any unit in any draws is 1/N

The probability of drawing a sample of n units in SRSWR is

$$\frac{1}{N^n}$$

This means that there are N^n possible samples; the probability of drawing each of these is the same. The probability that a specified unit is selected at any draw is 1/N and the probability that a specified unit is included in the sample is n/N.

2.5 Procedure of Selecting a Simple Random Sample

The first step is to prepare a list of all (N) units in the population and number them serially from 1 to N. Then a sample of n units is taken by any of the following methods.

(a) **Lottery Method:** Taken N cards or tickets or counters bearing numbers from 1 to N, these are entirely blended and n tickets are drawn, individually, from this set (either without replacement or with replacement) and these numbers noted, blending the tickets completely after each draw, afterward, the sample of n units is chosen from the population which bears the numbers on these tickets. This way is unwieldy and does not ensure that units will be chosen with equivalent probability. Human predisposition and bias may likewise creep in the technique.

(b) **Use of Random Numbers Tables:** A random number table is a collection of digits 0 to 9 in either a linear\ rectangle pattern, such that every ten numbers show, separately of one another, with the same frequency. By combining the numbers 00 to 99, 000 to 999 and 0000 to 9999 which arise with approximately the same frequency. Some random number tables in common use are.

(i) Tippet's Yates table
(ii) Fisher & Yates table
(iii) Rand corporation series
(iv) Cordell & smith Series
(v) I.S.I Series (in Table of Ras-Mibe & Matter)

The easiest approach to choosing an sample of n units from a population of N units is to take a gander at any row (or column) of the tables and select n random numbers 1 and N, then, taking the population units bearing those numbers. The procedure includes number of rejections since all numbers greater than N showing up in the table is dismissed for consideration. The utilization of random numbers is, in this manner adjusted by utilizing either remainder approach or quotient approach as follows:

Remainder Approach: Suppose; N be a r-digit number and its r-digit highest multiple be K. A random number R is chosen from 1 to K and the unit equal to the remainder obtained on dividing R by N is selected. If the remainder is Zero, the last unit is selected. e.g., let N = 123, the highest these digit multiple is 984, For selecting a unit a random number from 001 to 984 is selected, say 287 which on division by 123 given 41 as the remainder. Hence the unit with serial number 41 is selected in the sample.

Quotient approach: Suppose; N be r-digit and let its r-digit highest multiple by N' such N'IN = q, A random number k is chosen from 0 to (N'I) and quotient r is obtained on dividing, k by q. Then the unit bearing the serial number (r-1) is selected in the sample. For example, let N = 123 and N' = 984 or that q = 8. If the random number selected is 287 than r=35, Hence the unit with serial number 34 is included in the sample.

2.6 Estimation of Population Mean or Total

Let us suppose that the value of y, the character under study, is y_i for the i^{th} population unit U_1 (i = 1, 2,….., N). We define

$$Population\ mean\ \bar{y} = \sum_{i=1}^{N} y_i / N$$

$$Population\ total\ y = \sum_{i=1}^{N} y_i = N\bar{Y}$$

$$Population\ variance\ \sigma^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / N$$

$$and\ S^2 = \sum_{i=1}^{N} (y_i - \bar{Y})^2 / (N-1) = \frac{N\sigma^2}{N-1}$$

Also, for a sample of size n assume the sample values be y_1, y_2, ……., y_n. It doesn't mean that only first n population units are selected in the sample but, without any lose of generally, these may be considered to be values of the units selected in that order. Evidently r^{th} sample unit value y_r, may be any of the population values y_1, y_2, ……., y_n. We denote by

$$\bar{y} = \sum_{i=1}^{N} y_i / N = the\ \ sample\ mean$$

$$y = \sum_{i=1}^{N} y_i = n\bar{y}\ \ = sample\ total$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2 = sample\ variance$$

Case I: Simple Random Sampling: Without Replacement (SRSWOR):

Theorem 1: $\bar{y}$ is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n} = (1-f))\frac{S^2}{n}$$

Where, f = n/N is the sampling fraction and (1 - f) is finite population correction (f.p.c.) factor.

Proof: We have

$$E(\bar{y}) = E\left[\sum_{i=1}^{N} y_i / n\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(y_i)$$

$$Here\ \ E(y_i) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}$$

$$Hence\ E\ (\bar{y}) = \frac{1}{n}\sum_{i=1}^{n} \bar{Y} = \frac{1}{n} \times n\bar{Y} = \bar{Y}$$

This shows that $\bar{y}$ is an unbiased estimate of $\bar{Y}$.

We have

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2$$

$$= E\ \left[\{1/n\sum_{r=1}^{n} y_r - \bar{Y})\}^2\right]$$

$$= \frac{1}{n^2} E\left\{\sum_{i=1}^{n} (y_n - \bar{Y})\right\}^2$$

$$= \frac{1}{n^2}\left[E\left\{\sum_{i=1}^{n} (y_n - \bar{Y})^2 + \sum_{r\neq}^{n}\sum_{s-1}^{n} (y_r - \bar{Y})(y_s - \bar{Y})\right\}\right]$$

Now,

$$E\left\{\sum_{r=1}^{n}(y_r-\bar{Y})^2\right\}=\sum_{r=1}^{n}E\,(y_r-\bar{Y})^2$$

$$=\sum_{r=1}^{n}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_r-\bar{Y})^2\right\}$$

$$=n\sigma^2$$

$$Also,\qquad E\left[\sum_{r\neq}^{n}\sum_{\substack{s-1\\ r\neq s}}^{n}(y_r-\bar{Y})(y_s-\bar{Y})\right]=\sum_{r\neq}^{n}\sum_{\substack{s-1\\ r\neq s}}^{n}(y_r-\bar{Y})(y_s-\bar{Y})$$

$$\sum_{r=1}^{n}\sum_{\substack{s=1\\ r\neq s}}^{n}\left(\frac{1}{N(N-1)}\sum_{i\neq}^{N}\sum_{j}^{N}(y_i-\bar{Y})(y_j-\bar{Y})\right)$$

$$=\frac{1}{N(N-1)}\sum_{r=1}^{n}\sum_{\substack{s=1\\ r\neq s}}^{n}\left(\left\{\sum_{i=1}^{N}(y_r-\bar{Y})\right\}^2-\sum_{i=1}^{N}(y_i-\bar{y})^2\right)$$

$$=\frac{1}{N(N-1)}\sum_{r=1}^{n}\sum_{\substack{s=1\\ r\neq s}}^{n}\left(\sum_{i=1}^{N}(y_i-\bar{y})^2\right)=\frac{1}{N(N-1)}\sum_{r=1}^{N}\sum_{\substack{s=1\\ r\neq S}}^{N}(N\sigma^2)$$

$$=\frac{-n(n-1)\sigma^2}{N-1}$$

Therefore

$$V(\bar{y})=\frac{1}{n^2}\left(n\sigma^2-\frac{n(n-1)\sigma^2}{N-1}\right)$$

$$=\frac{N-n}{N-1}\frac{\sigma^2}{n}$$

$$=\left(\frac{N-n}{N}\right)\frac{S^2}{n}$$

$$The\ standard\ error\ of\ \bar{y}\ is\ \sigma_y=\frac{N-n}{N}\frac{S}{\sqrt{n}}$$

Corollary: The unbiased estimate of population total $\bar{Y}$ is $N\bar{y}$ having variance.

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

Theorem 2: S^2 is an unbiased estimator of S^2.

Proof: We have

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{r=1}^{n}(y_r - \bar{y})^2\right)$$

$$\frac{1}{n-1}E\left(\sum_{r=1}^{n}\{(y_r - \bar{Y}) - (\bar{y} - \bar{Y})\}^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{N}(y_r - \bar{Y})^2 - 2(\bar{y} - \bar{Y})\sum_{r=1}^{n}(y_r - \bar{Y}) + n(\bar{y} - \bar{Y})^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{N}(y_r - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{N}(y_r - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{N}\frac{1}{N}\sum_{r=1}^{N}(y_r - \bar{Y})^2 - nV(\bar{y})\right)$$

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

$$= \frac{1}{n-1}\left(\frac{S^2}{nN}\{n^2(N-1) - n(N+n)\}\right)$$

$$= \frac{1}{n-1}\ \frac{nN(n-1)}{nN}S^2$$

$$= S^2$$

Case II: Simple Random Sampling with Replacement (SRSWR)

Theorem 3:

$(i) E(\bar{y}) = \bar{Y}$

$(ii) V(\bar{y}) = \frac{\sigma^2}{n} = \frac{N-1}{n}\frac{S^2}{n}$

$(iii) E(s^2) = \sigma^2$

Solution: We have

$$E(\bar{y}) = E\left(\frac{1}{n}\sum_{r=1}^{n}\bar{y} = \bar{Y}\right)$$

Also we have,

$$V(\bar{y}) = E(\bar{y} - \bar{Y})^2$$

$$= \frac{1}{n^2}E\left(\sum_{r=1}^{n}(y_r - \bar{Y})\right)^2$$

$$= \frac{1}{n^2}\sum_{r=1}^{n}E\,(y_r - \bar{Y})^2$$

$$= \frac{1}{n^2}n\sigma^2$$

$$= \frac{\sigma^2}{n}$$

$$= \frac{N-1}{n}\frac{S^2}{n}$$

Again

$$E(s^2) = \frac{1}{n-1}E\left(\sum_{r=1}^{n}E\,(y_r - \bar{Y})^2\right)$$

$$= \frac{1}{n-1}\sum_{r=1}^{n}E\,(y_r - \bar{Y})^2$$

$$= \frac{1}{n-1}\sum_{r=1}^{n}E\{(y_r - \bar{Y}) - (\bar{y} - \bar{Y})\}^2$$

$$= \frac{1}{n-1}\sum_{r=1}^{n}\{E(y_r - \bar{Y})^2 - (\bar{y} - \bar{Y})^2\}$$

$$= \frac{1}{n-1}\left(\sum_{r=1}^{N}\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2 - nV(\bar{y})\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right)$$

$$= \sigma^2$$

The standard error of $\bar{y}$ *is* $\sigma_y = \sigma/\sqrt{n}$ whose unbiased estimator is.

Corollary: The unbiased estimator of population total Y is N $\bar{y}$ and its variance is (N-1) NS^2/n.

2.7 Estimation of Population Proportion

In some cases, the characteristic under study is subjective/qualitative or an attribute and each unit in the population are characterized into one of two classes viz, (a) those having the attribute and (b) those not having the attribute. Assume M out of N units in the population posses the attribute So that the population proportion is P = M/N. We are keen on estimating P on the basis of a simple random sample of size n.

Suppose the number of units in the sample which posses the attribute be m and, consequently the sample proportion is p=m/n.

Theorem 4: For SRSWOR, p is the unbiased estimator of P and its variance is given by

$$V(p) = \frac{N-n}{N-1}\ \frac{PQ}{n} \qquad where\ Q = 1 - p$$

Proof: We know that m has the hyper geometric distribution given by p.m.f. = $= \frac{n\,(N-n)}{N-1}\ PQ$

$$f(x) = P(m = x) = \frac{\binom{M}{N}\binom{N-M}{n-x}; \underbrace{x = 0,1,2,\ldots n}}{\binom{N}{x}}$$

Whose mean and variance are known to be

$$E\,(m) = n\binom{M}{N} = np$$

$$V\,(m) = \frac{n(N-1)}{N-1}\frac{M}{N}\left(1-\frac{m}{N}\right)$$

$$= \frac{n(N-n)}{N-1}PQ$$

Therefore,

$$E\,(p) = E\left(\frac{m}{n}\right) = P$$

$$V(p) = V\left(\frac{m}{n}\right) = \frac{N-n}{N-1}\frac{PQ}{n}$$

Alternative Proof: Suppose we associate a measurement y to each population unit such that

$$y_i = \begin{cases} 1 & if\ the\ unit\ U, possesses\ the\ attribute \\ 0 & otherwise \end{cases}$$

Then we have the population value $y_i \ldots\ldots, y_N$ and sample values $y_i \ldots\ldots, y_N$ such that

$$P = \sum_{l}^{N} y_i/N(M/N) = \bar{Y}$$

$$p = \sum_{l}^{n} y_i/n(m/n) = \bar{y}$$

$$S^2 = \sum_{l}^{N}\frac{(y_i-\bar{Y})^2}{N-1} = \sum_{l}^{N}\frac{{y_i}^2 - N(\bar{Y})^2}{N-1}$$

$$= \frac{NP-NP^2}{N-1} = \frac{NPQ}{N-1}$$

$$S^2 = \frac{npq}{n-1}$$

Therefore, by Theorem 1 we get

$$E\,(p) = E(\bar{y}) = \bar{Y} = P.$$

$$V(p) = V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n} = \frac{N-n}{N-1}\frac{PQ}{n}$$

Theorem 5: An unbiased estimator of V(p) is

$$v(p) = \frac{N-n}{(n-1)N}pq$$

Proof: Since $E(S^2) = S^2$, therefore

$$E\left(\frac{npq}{n-1}\right) = \frac{NPQ}{N-1} \; or \; E(pq) = \frac{n-1}{N-1}.\frac{N}{n}PQ$$

So that

$$E\big(v(p)\big) = \frac{N-n}{(n-1)N}E(pq)$$

$$= \frac{n-1}{N-1}\frac{PQ}{n} = V(p).$$

Theorem 6: For SRSWOR, p is an unbiased estimator of P and its variance is given by

$$V(p) = \frac{PQ}{n}$$

Its unbiased estimator is $V(p) = \frac{PQ}{n-1}$

Proof: In this case, p has binomial distribution given by p.m.f.

$$f(x) = P(m = x) = \binom{N}{x}\binom{M}{N}^x\left(\frac{1-M}{n}\right)^{n-x}$$

$$= \binom{n}{x}p^x(1-P)^{n-x}$$

Whose means and variance are given by

E(m) = np and V(m) = nPQ

Therefore,

$E\,(p) = E\left(\frac{m}{n}\right) = P$ $\quad and \quad$ $V(p) = V\left(\frac{m}{n}\right) = \frac{PQ}{n}$

Alternative Proof: It follows from theorem 3 that

$$V(p) = V(\bar{y}) = \frac{\sigma^2}{n}$$

$$= \frac{N-1}{N}\quad\frac{S^2}{n}$$

$$= \frac{N-1}{N} \quad \frac{NPQ}{(N-1)n} = \frac{PQ}{n}$$

Since $E(S^2) = \sigma^2$, therefore

$$E\left(\frac{npq}{n-1}\right) = PQ$$

$$or\ E(pq) = PQ\left(\frac{n-1}{n}\right)$$

So that

$$E\big(v(p)\big) = E\left(\frac{pq}{n-1}\right) = \frac{PQ}{n} = V(p)$$

2.8 Exercises

1. Explain simple random sampling?

2. Calculate mean and variance of SRSWOR?

3. Calculate mean and variance of SRSWR?

2.9 Summary

After studying the above unit we can able to understand about the simple random sampling, its mean and variances.

2.5 Further Readings

1. Cocharan, W.G. (1993). *Sampling Techniques*, Third Edition. Wiley Eastern Limited, New Delhi.

2. Sigh, D and Chaudhary, F.S. (1999). *Theory and Analysis of sample Survey Designs.* New Age International (P) Limited, Publishers, New Delhi.

3. Sukhatme, P.V., B.V., S and Asok, C (1984). *Sampling Theory of Surveys with Applications,* Third Revised Edition. Iowa State University Press, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi-110112

4. Swain, A.K.P.C. (2003). *Finite Population Sampling Theory and Methods*. South Asian Publishers Pvt. Ltd., New Delhi-110014.

Unit-3: Systematic Sampling

Structure

3.0 Introduction

A sampling method, select each k^{th} component from a listing of population elements, after the first element has been randomly selected. In other words, only first unit has been selected at random and then all rest units being selected automatically according to predetermined pattern involving regular spacing k e.g. if k = 5 and sample size is 10, then after selecting the first unit randomly, every 5^{th} unit will going to be selected in the sample, until all10 units are not selected.

Hence systematic procedure is followed to choose a sample by taking every k^{th} individual where k refers to the sample interval, which is calculated by the formula;

$$k = \frac{\text{Total population}}{\text{Sample size desired}}$$

In field study, when the population is large and homogeneous, then systematic sampling is very simple and easy to adopt.

3.1 Objectives

After studying this unit, you should be able to understand:

- Systematic random sampling
- Mean and Variance .

3.2 Systematic Random Sampling

Suppose there be n units in the population sequentially numbered from 1 to N, and we wishes to draw a random sample of size n where N/n=k i.e. N is completely divisible by n, k is the quotient. First of all, one unit is chosen at random from first k units of the population and after that each k^{th} unit is selected in the sample mechanically. Suppose that i^{th} unit is chosen at random from first units of the population i.e.,$1 \leq i \leq k$. Then, sample of size n consists of the units sequentially numbered i, i+k, i+2k,…., i+(n-1)k

in the population. Sample, elected form the population is known as a systematic random sample and this technique of electing the sample from the population is known as Systematic Random Sampling.

Suppose N = 100 and n = 20. Then k = 5, suppose 3rd unit is chosen at random from first five units, then our systematic sample consists of the units serially numbered 3,8,13,18,23,….98 in the population.

In fact, systematic sampling resembles cluster sampling. We can express that there are k clusters each of size n and we have to choose one cluster at random from these k clusters as is clear from the accompanying schematic diagram-

1	2	i	k
1+k	2+k	i+k	2k
1+(n-1)k	2+(n-1)k	i+(n-1)k	nk

; the i^{th} cluster here consists of the units serially numbered i, i+k, …..,i+(n-1)k in the population.

Notations:

Let Y_{ij} be the observation on the characteristic under study for the unit of the population serially numbered as $i + (j - 1)k;\ 1 \leq i \leq k, 1 \leq j \leq n$

Thus, as per schematic diagram these values may be represented as

y_{11}	y_{21}	y_{i1}	y_{k1}
y_{12}	y_{22}	y_{i2}	y_{k2}
y_{1n}	y_{2n}	y_{in}	y_{kn}

We denote by

$$\bar{y}_i = \frac{1}{n_i} \sum_{i=1}^{k} y_{ij}$$

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

$$= \frac{1}{nk} \sum_{i=1}^{k} \sum_{i=1}^{k} y_{ij}; i.e. populatin\ mean$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{i=1}^{k}(\bar{Y}-\bar{Y})^2$$

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{i=1}^{k}(\bar{Y}-\bar{Y})^2$$

If $\bar{Y}$ is not known, it is proposed to be estimated by

$$\bar{y}_{sy} = \bar{Y}_\iota = \frac{1}{n}\sum_{i=1}^{k} y_{ij}$$

Unbiasedness of y_{sj}

$$E(\bar{y}_{sy}) = E(\bar{y}_i)$$

Here $\bar{y}_i$ is a r-v which may take any of the values $\bar{y}_1, \bar{y}_2 \ldots\ldots\ldots\ldots \bar{y}_k$ each with probability 1/k. Hence,

$$E(\bar{y}_i) = \frac{1}{k}\bar{y}_1 + L_k\overline{y_2} + \cdots .. + L_k\overline{y_k}$$

$$= \frac{1}{k}\sum_{i=1}^{k}\bar{y} = \bar{Y}$$

This shows that $\bar{y}_{sy}$ is an unbiased estimator of $\bar{Y}$

Variance of $\bar{y}_{sy}$.

$$V_r(\bar{y}_{sy}) = E\left[\bar{y}_{sy} - E(\bar{y}_{sy})\right]^2$$

$$= E\left[\bar{Y}_\iota - \bar{Y}\right]^2$$

$$= \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_\iota - \bar{Y})^2$$

$$= \frac{k-1}{k}\frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_\iota - \bar{Y})^2$$

$$= \frac{k-1}{k}S_c^2$$

Remarks

1. Unbiasedness and variance of $\bar{y}_{sy}$ may also be directly derived from cluster sampling.

2. The case N= nk + r (r<n) has not been discussed here.

3.4 Exercises

1. A population is stratified into two strata with sizes as 400 and 600 with population mean as 81 and 16 respectively. We wish to draw a sample of size 200 form this population using (i) proportional allocation and (ii) Neyman allocation. Allocation the sample sizes that are to be drawn from the two strata under the two methods. Also obtain the variances of the stratified mean under proportional and Neyman allocation.

3.12 Summary

After studying this unit you should able to understand about systematic random sampling with its mean and variance.

3.13 Further Readings

1. Cocharan, W.G. (1993). *Sampling Techniques*, Third Edition. Wiley Eastern Limited, New Delhi.
2. Sigh, D and Chaudhary, F.S. (1999). *Theory and Analysis of sample Survey Designs.* New Age International (P) Limited, Publishers, New Delhi.
3. Sukhatme, P.V., B.V., S and Asok, C (1984). *Sampling Theory of Surveys with Applications,* Third Revised Edition. Iowa State University Press, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi-110112
4. Swain, A.K.P.C. (2003). *Finite Population Sampling Theory and Methods*. South Asian Publishers Pvt. Ltd., New Delhi-110014.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 103/ MASTAT – 103 Survey Sampling

Block: 2 Random Sampling Procedures - II

Course Design Committee

Course Preparation Committee (Block)

Block: 2 Random Sampling Procedure - II

Dr. R. R. Sinha **Writer**
Sr. Assistant Professor, Department of Mathematics
Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab

Prof. Vineeta Singh **Editor**
Department of Statistics, Institute of Social Sciences
Dr. B. R. Ambedkar University, Agra

Dr. Shruti **Course / SLM Coordinator**
School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 103/ MASTAT – 103 **SURVEY SAMPLING**
©UPRTOU
First Edition: ***July 2021***
ISBN : 978-93-94487-03-1

Printed By: K. C. Printing & Allied Works, Panchwati, Mathura - 281003
.

Block & Units Introduction

The ***Block - 2 – Random Sampling Procedures – II*** is the second block with four units. It explains the stratified random sampling with its advance concepts and use of auxiliary information in the estimation of population parameters. Unbiased estimate of population mean in stratified random sampling and its properties are explained. It also includes the allocation of sample size in different strata as well as post and deep stratification. Further, ratio and regression methods of estimation are explained to understand the application of auxiliary information with suitable examples. This is also deals with the cluster sampling and its properties are revealed along with the examples.

Unit – 4 – Stratified Sampling and Use of Auxiliary Information is discussed about the stratified sampling, its mean and variance and many more other theorems related with this.

Unit – 5 – Ratio and Regression Methods of Estimation have been introduced to the ratio and regression sampling with related theorems.

Unit – 6 – Cluster and Multi-Stage Sampling dealt with cluster and multi-stage sampling with related theories.

Unit – 7 – Response and Non - Response Sampling dealt with the theory of response and non-response sampling.

Unit-4 Stratified Sampling and Use of Auxiliary Information

Structure

4.1 Introduction

It has been observed that the precision of a sample estimate of the population mean depends not only upon the sample size and the sampling fraction but also on the variability or heterogeneity of the population. In stratified sampling the population of N units is first divided into sub-population of N_1, N_2,.....,N_k units so that $N_1 + N_2 + \cdots + N_k = N$. The sub-populations are called strata. When the strata have been determined, a sample is drawn from each, the drawing being made independently in different strata. The sample sizes within the strata are denoted by n_1,n_2,....,n_k respectively. If a random sample is taken in each stratum, the whole procedure is described as stratified random sampling.

4.2 Objectives

After this unit, the learner would be able to understand about:

- The selection of a sample population that would be the best representative of the entire population under study,
- The division of the entire population into homogeneous groups or strata in such a manner so as to gain a higher degree of relative precision.
- The sample allocation under different situations of fixed and precision.
- The necessitate of post and deep stratifications and their properties.

4.3 Principles of Stratification

The principles which are followed in stratifying a population are given as follows:

a) The stratification of population should be done in such a way that strata are homogeneous within themselves, with respect to the characteristic under study.
b) The strata should be non-overlapping and should comprise the whole population.
c) Administrative convenience may be considered as the basis for stratification when it is difficult to stratify with respect to the characteristic under study.
d) If the limit of precision for certain sub-population is given, it will be better to treat each sub-population as a stratum.

4.4 Advantages of Stratification

There are many advantages of stratification. The principle ones are as follows:

a) Stratification by natural characteristics helps in improving the sampling design.
b) Stratification may be desired for administrative convenience.
c) Stratification makes it possible to use different sampling design in different strata.
d) Stratification ensures adequate representation to various groups of the population, which may be of some interest or importance.
e) Stratification is particularly more effective when there are extreme values in the population which can be divided into separate strata, thereby reducing the variability within strata.

f) Stratification ensures selection of a better cross-section of the population than that under unstratified population.
g) Stratification brings a gain in the precision in estimation of a characteristic of a population.

4.5 Notations and Properties of the Estimates

For the population mean per unit, the estimate used in stratified sampling is $\bar{y}_w$ (w- for stratified), where

$$\bar{y}_w = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i} = \sum_{i=1}^{k} p_i \bar{y}_{n_i} \qquad (1.5.1)$$

where $N = N_1 + N_2 + \cdots + N_k$, $p_i = \frac{N_i}{N}$ (stratum weight)

and $\bar{y}_{n_i} = \frac{1}{n_i}\sum_{j}^{n_i} y_{ij}$ →Mean of sample units from i^{th} stratum.

$\sum_{j}^{n_i} y_{ij}$ →Sum taken over all selected unit in j^{th} unit from i^{th} stratum $\begin{pmatrix} i = 1,2,\ldots,k \\ \mathrm{j} = 1,2,\ldots,n_i \end{pmatrix}$

The difference is that in $\bar{y}_w$ the estimate from the individual strata receive their correct weights N_i/N.

$\bar{y}_{n_i}$ coincides with $\bar{y}_w$ provided that in every stratum $\frac{N_i}{N} = \frac{n_i}{n}$ or $\frac{n_i}{N_i} = \frac{n}{N}$.

The principle properties of the estimate $\bar{y}_w$ are as follows:

a) If in every stratum the sample estimate $\bar{y}_{n_i}$ is unbiased, then $\bar{y}_w$ is an unbiased estimate of the population mean $\bar{y}_N$.
b) If the sample are drawn independently in different strata, then $V(\bar{y}_w) = \sum_{i=1}^{k} p_i^2 V(\bar{y}_{n_i})$; where $V(\bar{y}_{n_i})$ is the variance of $\bar{y}_{n_i}$ over repeated samples from stratum n_i.

Population mean of i^{th} stratum $\bar{y}_{N_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$

and population mean $\bar{y}_N = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} y_{ij} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{N_i} = \sum_{i=1}^{k} p_i \bar{y}_{N_i}$

Let us assume that n be the sample size and according to simple random sampling without replacement (SRSWOR) n_1,n_2,....,n_k be the number of units to be selected from respective strata i.e. $n = \sum_{i=1}^{k} n_i$

Mean of sample units from i^{th} stratum $\bar{y}_{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$

and sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{n} n_i \bar{y}_{n_i}$

4.6 Expected Value of Sample Mean ($\bar{y}_n$)

Consider the sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} n_i \bar{y}_{n_i}$

$$E(\bar{y}_n) = E\left\{\frac{1}{n}\sum_{i=1}^{n} n_i \bar{y}_{n_i}\right\}$$

$$= \frac{1}{n}\sum_{i=1}^{k} n_i E(\bar{y}_{n_i})$$

Since n_i units were drawn from i^{th} stratum according to SRSWOR and in SRSWOR sample mean is an unbiased estimate of population mean. i.e. $E(\bar{y}_{n_i}) = \bar{y}_{N_i}$.

Therefore $$E(\bar{y}_n) = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_{N_i} \neq \bar{y}_N \left(= \frac{1}{N}\sum_{i=1}^{k} N_i\, \bar{y}_{N_i}\right)$$

Hence, sample mean is not an unbiased estimate of population mean in general.

4.7 Expected Value of $\bar{y}_w$

In stratified random sampling, the estimator $\bar{y}_w$ for the population mean is given by

$$\bar{y}_w = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_{n_i} = \sum_{i=1}^{k} p_i\, \bar{y}_{n_i}$$

where $N = N_1 + N_2 + \cdots + N_k$ and $p_i = \frac{N_i}{N}$

$$E(\bar{y}_w) = E\left\{\sum_{i=1}^{k} p_i\, \bar{y}_{n_i}\right\}$$
$$= \sum_{i=1}^{k} p_i\, E(\bar{y}_{n_i})$$

Since $\bar{y}_{n_i}$ is the sample mean of the simple random sample from i^{th} stratum and sample mean in simple random sampling is an unbiased estimate of population mean.

i.e. $$E(\bar{y}_{n_i}) = \bar{y}_{N_i}$$

Thus $$E(\bar{y}_w) = \sum_{i=1}^{k} p_i\, \bar{y}_{N_i}$$
$$= \bar{y}_N \text{ (Population mean)}$$

Hence, $\bar{y}_w$ is an unbiased estimate of population mean.

4.8 Variance of $\bar{y}_w$

For the population mean per unit, the estimate used in stratified sampling is $\bar{y}_w$, where

$$\bar{y}_w = \frac{1}{N}\sum_{i=1}^{k} N_i\, \bar{y}_{n_i} = \sum_{i=1}^{k} p_i\ \bar{y}_{n_i}$$

$$V(\bar{y}_w) = V\left\{\sum_{i=1}^{k} p_i \bar{y}_{n_i}\right\} = \sum_{i=1}^{k} p_i^2 V(\bar{y}_{n_i}) \qquad (1.8.1)$$

Since samples are drawn independently in different strata, therefore all covariance terms vanish.

Since in simple random sampling

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2, \text{ where } S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$$

So, let us define S_i^2 for i^{th} stratum, where

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(y_{ij} - \bar{y}_{N_i}\right)^2$$

Therefore, $V(\bar{y}_{n_i}) = \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$

(1.8.2)

Thus, with the help of equation (1.8.2), we can write the equation (1.8.1) as

$$V(\bar{y}_w) = \sum_{i=1}^{k} p_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S_i^2$$

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 \qquad (1.7.3)$$

4.9 Unbiased Estimate of $V(\bar{y}_w)$

We know that that in simple random sampling without replacement

$$E(s_i^2) = S_i^2 \qquad (1.8.1)$$

where $s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{n_i}\right)^2$

Hence $E\left\{\sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 s_i^2\right\} = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 E(s_i^2)$

$= \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$ (From (1.8.1))

$= V(\bar{y}_w)$

Hence, an unbiased estimate of $V(\bar{y}_w)$ is given by

$$Est.V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 s_i^2 \qquad (1.8.2)$$

4.10 Allocation of Sample Size in Different Strata

The allocation of the sample to different strata is done by the consideration of three factors:

i. the total number of units in the stratum i.e. stratum size
ii. the variability within the stratum and
iii. the cost in taking observations per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources. There are four methods of allocations of sample sizes to different strata in a stratified sampling procedure, which are as follows:

a) Equal allocation
b) Proportional allocations
c) Neyman allocation
d) Optimum allocation

a) Equal Samples from each Stratum- This is a situation of considerable practical interest for reasons of administrative or field work convenience. In this method, the total sample size n is divided equally among all the strata i.e. for i^{th} stratum

$$n_i = n/k$$

b) Proportional Allocation- In this allocation the sample size is proportional to the stratum size in each stratum i.e. $n_i \alpha N_i$

$$n_i = AN_i$$

where A is any constant.

$$\sum_{i=1}^{k} n_i = A \sum_{i=1}^{k} N_i$$

or $$n = A\,N$$

or $$A = \frac{n}{N}$$

Therefore $$n_i = \frac{n}{N} N_i$$

or $$n_i = np_i \qquad (1.9.1)$$

Now, variance of $\bar{y}_w$ in proportional allocation is given by

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 \qquad (1.9.2)$$

From equations (1.9.1) and (1.9.2), we have

$$V(\bar{y}_w)_p = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

or $$V(\bar{y}_w)_p = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

or $$V(\bar{y}_w)_p = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 \qquad (1.9.3)$$

b) Optimum Allocation- In this method of allocation the sample sizes n_i in the respective strata are determined with a view to minimize $V(\bar{y}_w)$ for a specified cost of conducting the sample survey or to minimize the cost for a specified value of $V(\bar{y}_w)$.

Let c_i be the cost of survey per unit from i^{th} stratum. Then total cost C is given by

$$C = \sum_{i=1}^{k} n_i c_i \qquad (1.9.4)$$

Let us define $$\emptyset = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 + \mu \sum_{i=1}^{k} n_i c_i \qquad (1.9.5)$$

Differentiating the above equation with respect to n_i, we get

$$\frac{\partial \emptyset}{\partial n_i} = -\frac{p_i^2 S_i^2}{n_i^2} + \mu c_i$$

For optimum value, we have

$$\frac{\partial \emptyset}{\partial n_i} = 0$$

$\Rightarrow$ $$n_i = \frac{p_i S_i}{\sqrt{\mu c_i}}$$

$\Rightarrow$ $$\begin{cases} n_i \alpha N_i \\ n_i \alpha S_i \\ n_i \alpha\, 1/\sqrt{c_i} \end{cases} \qquad (1.9.6)$$

$\Rightarrow$ $$n_i \alpha \frac{N_i S_i}{\sqrt{c_i}} \qquad (1.9.7)$$

This shows that under optimum allocation, sample size from a stratum should be proportional to population size of that stratum, standard

deviation of that stratum and is inversely proportional to the square root of per unit cost of that stratum.

Case I: (Allocation of Sample for Fixed Cost):

Let us have a fixed total cost for the survey be C_0. Then total cost is given by

$$C_0 = \sum_{i=1}^{k} c_i n_i \qquad (1.9.8)$$

We know that $n_i = \frac{p_i S_i}{\sqrt{\mu c_i}} = \frac{p_i S_i}{\sqrt{c_i}} \cdot \frac{1}{\sqrt{\mu}}$

$\therefore$ $C_0 = \frac{1}{\sqrt{\mu}} \sum_{i=1}^{k} p_i S_i \sqrt{c_i}$

or $\frac{1}{\sqrt{\mu}} = \frac{C_0}{\sum_{i=1}^{k} p_i S_i \sqrt{c_i}}$

$\therefore$ $n_i = \frac{p_i S_i}{\sqrt{c_i}} \cdot \frac{C_0}{\sum_{i=1}^{k} p_i S_i \sqrt{c_i}} \qquad (1.9.9)$

So, the total sample size required for estimating the population mean with maximum precision for a fixed cost C_0 is given by

$$n = \frac{C_0 \sum_{i=1}^{k} (p_i S_i / \sqrt{c_i})}{\sum_{i=1}^{k} p_i S_i \sqrt{c_i}} \qquad (1.9.10)$$

The allocation of the sample according to (1.9.10) is known as optimum allocation. If $c_i = C$ i.e. the cost of survey per unit is same for each stratum. Then we have

$$C_0 = \sum_{i=1}^{k} n_i C$$

or $C_0 = nC$

or $n = \frac{C_0}{C}$

Hence $n_i = \frac{p_i S_i}{\sqrt{C}} \cdot \frac{C_0}{\sum_{i=1}^{k} p_i S_i \sqrt{C}}$ (From equation (1.9.9))

$$= \frac{C_0}{C} \cdot \frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i} = n \frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i}$$

or $n_i = n \frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i} \qquad (1.9.11)$

This allocation is sometimes called ***Neyman Allocation***.

Now, the minimum variance of $\bar{y}_w$ is given as

$$V(\bar{y}_w)_0 = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

When n_i are taken according to (1.9.9), then

$$V(\bar{y}_w)_0 = \sum_{i=1}^{k} \left[\frac{\left(\sum_{i=1}^{k} p_i S_i \sqrt{c_i} \right)}{C_0 / \sqrt{c_i}} p_i S_i \right] - \sum_{i=1}^{k} \frac{1}{N_i} p_i^2 S_i^2$$

Or $V(\bar{y}_w)_0 = \frac{\left(\sum_{i=1}^{k} p_i S_i \sqrt{c_i} \right)^2}{C_0} - \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2 \qquad (1.9.12)$

And under Neyman allocation, where n_i are taken according to (1.9.11), then

$$V(\bar{y}_w)_N = \frac{C_0}{C} \left(\sum_{i=1}^{k} p_i S_i \right)^2 - \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2$$

Or $$V(\bar{y}_w)_N = \frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{n} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 \qquad (1.9.13)$$

Case II: (Minimum Cost for fixed Precision):

Let V_0 be the fixed variance and is given by

$$V_0 = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 = \sum_{i=1}^{k} \frac{p_i^2 S_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

For optimum allocation, $$n_i = \frac{p_i S_i}{\sqrt{\mu c_i}} \qquad (1.9.14)$$

Therefore, $$V_0 = \sqrt{\mu}\sum_{i=1}^{k} p_i S_i \sqrt{c_i} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2$$

or $$\sqrt{\mu} = \frac{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}{\sum_{i=1}^{k} p_i S_i \sqrt{c_i}}$$

Putting the value of $\sqrt{\mu}$ in equation (1.9.14), we have

$$n_i = \frac{p_i S_i}{\sqrt{c_i}} \cdot \frac{\sum_{i=1}^{k} p_i S_i \sqrt{c_i}}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2},$$ under optimum allocation.

If $c_i = c$ then, $n_i = p_i S_i \frac{\sum_{i=1}^{k} p_i S_i}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}$, under Neyman Allocation.

So, the minimum sample size required for estimating the mean with fixed variance V_0 under optimum allocation is given by

$$n = \frac{\sum_{i=1}^{k}(p_i S_i/\sqrt{c_i})\sum_{i=1}^{k} p_i S_i \sqrt{c_i}}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}$$

If $c_i = c$ then minimum sample size required for estimating the mean with fixed variance V_0 under Neyman allocation is given by

$$n = \frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}.$$

Hence, minimum cost under optimum allocation and Neyman allocation are given by

$$(Min.\ cost)_{opt.} = \frac{\left(\sum_{i=1}^{k} p_i S_i \sqrt{c_i}\right)^2}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}$$

and $$(Min.\ cost)_{Ney.} = \frac{c\left\{\sum_{i=1}^{k} p_i S_i\right\}^2}{V_0 + \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2}$$

4.11 Comparison between Simple Random Sampling, Proportional Allocation and Optimum Allocation

We know that in simple random sampling without replacement (SRSWOR) method

$$V_R \to V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

and in proportional allocation

$$V_p \to V(\bar{y}_w)_p = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i S_i^2$$

and in Optimum allocation

$$V_0 \to V(\bar{y}_w)_0 = \frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{n} - \frac{1}{N} \sum_{i=1}^{k} p_i S_i^2$$

Now, $S^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(y_{ij} - \bar{y}_N\right)^2$

$$(N-1)S^2 = \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left(y_{ij} - \bar{y}_N\right)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left\{\left(y_{ij} - \bar{y}_{N_i}\right) + \left(\bar{y}_{N_i} - \bar{y}_N\right)\right\}^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{N_i} \left\{\left(y_{ij} - \bar{y}_{N_i}\right)^2 + \left(\bar{y}_{N_i} - \bar{y}_N\right)^2 + 2\left(y_{ij} - \bar{y}_{N_i}\right)\left(\bar{y}_{N_i} - \bar{y}_N\right)\right\}$$

By solving it, we get

$$(N-1)S^2 = \sum_{i=1}^{k} (N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2 + 2 \sum_{i=1}^{k} \left(\bar{y}_{N_i} - \bar{y}_N\right) \left\{\sum_{j=1}^{N_i} y_{ij} - N_i \bar{y}_{N_i}\right\} \quad (1.10.1)$$

But $\bar{y}_{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} \Rightarrow \sum_{j=1}^{N_i} y_{ij} = N_i \bar{y}_{N_i}$

$\therefore$ $(N-1)S^2 = \sum_{i=1}^{k} (N_i - 1)S_i^2 + \sum_{i=1}^{k} N_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2$

Now divided both sides by N and removing finite population correction term $\left(\frac{N-1}{N}\right)$, we get

$$\frac{N-1}{N} S^2 = \sum_{i=1}^{k} \frac{(N_i - 1)}{N_i} \cdot \frac{N_i}{N} S_i^2 + \sum_{i=1}^{k} \frac{N_i}{N} \left(\bar{y}_{N_i} - \bar{y}_N\right)^2$$

Or $S^2 = \sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2 \quad \left\{\frac{N_i - 1}{N_i} \simeq 1 \; and \; \frac{N-1}{N} \simeq 1\right\}$

Therefore,

$$V_R - V_P = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 - \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i S_i^2$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2\right\} - \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i S_i^2$$

or $V_R - V_P = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2$

Hence $V_R \geq V_P$ (1.10.2)

Again,

$$V_P - V_0 = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 - \left\{\frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{n} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2\right\}$$

Let $\sum_{i=1}^{k} p_i S_i = \bar{S}_w$, $\bar{S}_w$ →weight mean

Then
$$V_P - V_0 = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 - \frac{\bar{S}_w^2}{n}$$
$$= \frac{1}{n}\left\{\sum_{i=1}^{k} p_i S_i^2 - \bar{S}_w^2\right\}$$
$$= \frac{1}{n}\left\{\sum_{i=1}^{k} p_i S_i^2 - 2\bar{S}_w^2 + \bar{S}_w^2\right\}$$
$$= \frac{1}{n}\left\{\sum_{i=1}^{k} p_i S_i^2 - 2\sum_{i=1}^{k} p_i S_i \bar{S}_w + \sum_{i=1}^{k} p_i \bar{S}_w^2\right\}, \qquad \left[\because \sum_{i=1}^{k} p_i = 1\right]$$
$$= \frac{1}{n}\left\{\sum_{i=1}^{k} p_i \left(S_i^2 - 2S_i\bar{S}_w + \bar{S}_w^2\right)\right\}$$
$$= \frac{1}{n}\sum_{i=1}^{k} p_i (S_i - \bar{S}_w)^2 \geq 0$$

Hence $V_P \geq V_0$ (1.10.3)

Hence from (1.10.2) and (1.10.3), we have

$$V_0 \leq V_p \leq V_R$$

4.12 Comparison between Simple Random Sampling, Stratified Random Sampling under Optimum and Proportional Allocation

We know that, in simple random sampling without replacement

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

and in stratified random sampling

$$V(\bar{y}_w) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$

∴ Gain due to stratification is given by

$$V(\bar{y}_n) - V(\bar{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 - \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\sum_{i=1}^{k} p_i S_i^2 + \sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2\right\} - \left\{\sum_{i=1}^{k} \frac{1}{n_i} p_i^2 S_i^2 - \sum_{i=1}^{k} \frac{1}{N} p_i S_i^2\right\}$$
$$= \sum_{i=1}^{k} \left(\frac{1}{n} - \frac{p_i}{n_i}\right) p_i S_i^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i \left(\bar{y}_{N_i} - \bar{y}_N\right)^2 \qquad (1.11.1)$$

From the above expression, it is clear that the second term of the gain is always positive. First term of the gain may or may not be positive. It depends upon the allocation of the sample size in the strata or we can say that it depends on the value of n_i .

In case of proportional allocation $n_i = np_i$, due to which first term become zero and their gain becomes positive.

But in Optimum allocation $n_i = n\frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i}$ due to which the first term becomes negative or positive. If the first term becomes negative but its modulus value is smaller than the second term, then gain due to stratification will be positive. But if the first term is negative and its modulus value is greater

than second term, the gain due to stratification becomes negative and in such a case variance in stratified random sampling becomes greater than variance in simple random sampling.

Example 1.1: The following table shows the data of 2000 holdings (N_i), mean area under wheat per holding $(\bar{y}_{N_i})$ and standard deviation (s. d.) of area under wheat per holding for each stratum.

Stratum No.	1	2	3	4	5	6	7
Number of holdings (N_i)	394	461	381	334	169	113	148
mean area under wheat per holding $(\bar{y}_{N_i})$	5.4	16.3	24.3	34.5	42.1	50.1	63.8
s. d. of area under wheat per holding (S_i)	8.3	13.3	15.1	19.8	24.5	26.0	35.2

Compute the sample size in each stratum under proportional and optimum allocations for a sample of 200 holdings. Calculate the sampling variance of the estimated area under wheat from the sample

(i) If the holdings are selected under proportional allocation by with and without replacement methods,

(ii) If the holdings are selected under Neyman's allocation by with and without replacement methods

Also compute the gain in efficiency from these procedures as compared to simple random sampling.

Solution: The necessary calculations are shown in the table

Stratum no. (1)	N_i (2)	$\bar{y}_{N_i}$ (3)	S_i (4)	p_i (5)	$n\,p_i$ (6)	$p_i S_i$ (7)	$\frac{n\,p_i S_i}{\sum p_i S_i}$ (8)	$p_i\,\bar{y}_{N_i}$ (9)	$p_i\,\bar{y}_{N_i}^2$ (10)	$p_i\,S_i^2$ (11)
1	394	5.4	8.3	0.1970	40	1.64	19	1.06	5.72	13..612
2	461	16.3	13.3	0.2305	46	3.07	36	3.76	61.29	40.831
3	381	24.3	15.1	0.1905	38	2.88	34	4.63	112.51	43.488
4	334	34.5	19.8	0.1670	33	3.31	39	5.76	198.72	65.538
5	169	42.1	24.5	0.0845	17	2.07	24	3.56	149.88	50.715
6	113	50.1	26.0	0.0565	11	1.47	17	2.83	141.78	38.220
7	148	63.8	35.2	0.0740	15	2.61	31	4.72	301.14	91.872
Total	2000			1.0000	200	17.05	200	26.32	971.04	344.276

Estimation under proportional allocation:

We know that $n_i \alpha N_i$ or $n_i = AN_i$

The numbers of holdings to be selected from strata are given in column (6) of the table and are 40, 46, 38, 33, 17, 11 and 15, respectively.

$$\left[V(\bar{y}_w)_p\right]_{WOR} = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{i=1}^{k} p_i S_i^2 = 1.5492$$

If fpc is neglected, then we have

$$\left[V(\bar{y}_w)_p\right]_{WR} = \frac{1}{n}\sum_{i=1}^{k} p_i S_i^2 = 1.7214$$

Estimation under proportional allocation:

In this case $n_i \; \alpha \; p_i S_i$ or $n_i = n\frac{p_i S_i}{\sum_{i=1}^{k} p_i S_i}$

The numbers of holdings to be selected from strata are given in column (8) of the table and are 19, 36, 34, 39, 24, 17 and 31, respectively.

$$[V(\bar{y}_w)_N]_{WOR} = \frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{n} - \frac{1}{N}\sum_{i=1}^{k} p_i S_i^2 = 1.2813$$

If fpc is neglected, then we have

$$[V(\bar{y}_w)_N]_{WR} = \frac{\left(\sum_{i=1}^{k} p_i S_i\right)^2}{n} = 1.4535$$

Estimation under simple random sampling:

$$[V\,(\bar{y}_{SRS})]_{WOR} = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$$

$$= [V(\bar{y}_w)_N]_{WOR} + \frac{(N-n)}{n\,(N-1)}\left[\sum p_i \bar{y}_{N_i}^2 - \sum\left(p_i \bar{y}_{N_i}\right)^2 + \sum p_i S_i^2 - \sum (p_i S_i)^2\right] = 3.0420$$

If fpc is neglected, then we have

$$[V\,(\bar{y}_{SRS})]_{WR} = \left[V(\bar{y}_w)_p\right]_{WR} + \frac{1}{n}\sum\left(\bar{y}_{N_i} - \bar{y}_N\right)^2 = 3.1129$$

(i) The relative precision of proportional allocation is given by

(a) Without replacement method $[V\,(\bar{y}_{SRS})]_{WOR} / \left[V(\bar{y}_w)_p\right]_{WOR} = 1.9636$

(b) With replacement method $[V\,(\bar{y}_{SRS})]_{WR} / \left[V(\bar{y}_w)_p\right]_{WR} = 1.8084$

(ii) The relative precision of optimum allocation is given by

(a) Without replacement method $[V\,(\bar{y}_{SRS})]_{WOR} / [V(\bar{y}_w)_N]_{WOR} = 2.3742$

(b) With replacement method $[V\,(\bar{y}_{SRS})]_{WR} / [V(\bar{y}_w)_N]_{WR} = 2.1416$

4.13 Post Stratification

In stratified random sampling, it is presupposed that the knowledge of the stratum sizes as well as the sampling frame in each stratum are available. However, there are instances where the latter is not available, e.g. from the voter list of a given locality, the age of an individual voter is available although the lists of voters belonging to different age groups are not. With some characteristic, which is suitable for stratification, it is not possible to know in advance to which stratum a sampling unit belongs until the sample is selected. So the technique of post-stratification consists in classifying the population and the selected sample into a given number of strata after selection of sample. The problem of post-stratification has been discussed by Hansen, Hurwitz and Madow (1953). Here we have discussed the gain in precision due to such post-stratification.

If the sample is to be treated as if it were a stratified sample, then the weighted mean $\bar{y}_\omega$ would be the appropriate estimate of the population mean. This is easily seen to be an unbiased estimate of the population mean, since, for each i

$$E(\bar{y}_{n_i}) = E\{E(\bar{y}_{n_i}|n_i)\} = E(\bar{y}_{N_i}) = \bar{y}_{N_i} \qquad (1.12.1)$$

Hence, $E(\bar{y}_w) = \sum_{i=1}^{k} p_i E(\bar{y}_{n_i}) = \sum_{i=1}^{k} p_i \bar{y}_{N_i} = \bar{y}_N$ (1.12.2)

For fixed $n_1, n_2, \dots, n_k$, the variance of $\bar{y}_w$ is given by (1.7.3). Thus

$$V(\bar{y}_w|n_1, n_2, \dots, n_k) = \sum_{i=1}^{k} \left(\frac{1}{n_i} - \frac{1}{N_i}\right) p_i^2 S_i^2 \qquad (1.12.3)$$

In order to examine the gain in precision from post-stratification, we must find the unconditional variance of $\bar{y}_w$ to make it comparable to the variance of $\bar{y}_n$, the mean of a simple random sample. Now

$$\begin{aligned} V(\bar{y}_w) &= E[V(\bar{y}_w|n_1, n_2, \dots, n_k)] + V[E(\bar{y}_w|n_1, n_2, \dots, n_k)] \\ &= E[V(\bar{y}_w|n_1, n_2, \dots, n_k)] \end{aligned} \qquad (1.12.4)$$

since $E(\bar{y}_w|n_1, n_2, \dots, n_k)$ is a constant independent of n_i. From (1.12.3) and (1.12.4), we see that $V(\bar{y}_w) = \sum_{i=1}^{k} \left[E\left(\frac{1}{n_i}\right) - \frac{1}{N_i}\right] p_i^2 S_i^2$ (1.12.5)

An exact expression for (1.12.5) cannot be obtained. However, for large values of n and N, we may use the result $E\left(\frac{1}{n_i}\right) \cong \frac{1}{np_i} + \frac{1-p_i}{n^2 p_i^2}$ (1.12.6)

Substituting from (1.12.6) in (1.12.5), we have

$$\begin{aligned} V(\bar{y}_w) &\cong \sum_{i=1}^{k} \left[\frac{1}{np_i} + \frac{1-p_i}{n^2 p_i^2} - \frac{1}{Np_i}\right] p_i^2 S_i^2 \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i=1}^{k} p_i S_i^2 + \frac{1}{n^2} \sum_{i=1}^{k} (1 - p_i) S_i^2 \end{aligned} \qquad (1.12.7)$$

The first term in relation (1.12.7) is the value of variance of the mean of a stratified sample taken with proportional allocation. The second term represents the adjustment in the variance due to post-stratification, which will

be small in comparison to the first term if n is large. Hence post-stratification is almost as precise as proportional stratified sampling for large samples.

4.14 Deep Stratification

Let us consider the situation for the two alternative criteria of stratification. For example-village may be classified by regions or by the size of the agricultural area of the village. The question then naturally arises as to which of the two systems of stratification is to be preferred as their relative merits are bound to be different depending upon the characteristic under study. An obvious choice is to opt for two-way stratification which yields k rows and k' columns on the basis of those two stratification variables. There shall be kk' strata and thus a sample of size $n = kk'$ is required to estimate the population mean. If it is required to estimate the variance of the estimated mean, at least two observations should be taken from each stratum so the minimum sample size must be $2kk'$. A problem arises when $n < kk'$ and it is also desired to give proportional allocation to each criterion of stratification. Bryant, Hartley and Jessen (1960) have given an interesting and simple solution to this problem. The steps involved in this procedure are as follows:

(a) construct a square of n^2 cells with n rows and n columns,
(b) select n of the cells with equal probability such that no two selected cells belong to the same column or rows,
(c) combine the n rows to form k strata such that the i^{th} stratum has an allocation of n_i units, and
(d) combine the n columns to form k' strata such that the j^{th} stratum has an allocation of n_j units.

Let n_{ij} be the number of cells from the deep ij^{th} stratum selected in the sample. An unbiased estimator of the population mean is given by

$$\bar{y}_\omega = \frac{1}{n}\sum_i \sum_j G_{ij} n_{ij} \bar{y}_{ij} \qquad (1.13.1)$$

where $\bar{y}_{ij}$ is the sample mean in the ij^{th} stratum. G_{ij} is the weighting factor defined by $G_{ij} = \frac{n^2 p_{ij}}{n_i n_j}$ and $p_{ij} = N_{ij}/N$.

In case of proportional allocation when $p_{ij} = p_{i.}\, p_{.j}$, the variance of the estimator in relation (1.13.1) can be written as

$$V(\bar{y}_\omega) \cong \frac{1}{n}\sum_i \sum_j p_{ij} S_{ij}^2 - \frac{1}{(n-1)}\left[\sum_i \sum_j p_{ij}(y_{ij} - \bar{y}_{i.})^2 - \sum_j p_{.j}(\bar{y}_{.j} - \bar{y}_{..})^2\right],$$

where $\bar{y}_{i.} = \frac{\sum_j p_{ij}\bar{y}_{ij}}{p_{i.}}$, $\bar{y}_{.j} = \frac{\sum_i p_{ij}\bar{y}_{ij}}{p_{.j}}$, $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$.

4.15 Summery

The present unit deals with the problem of reducing the error occurred in simple random sampling without increasing the sample size. To reduce this error, stratified random sampling has been elaborately explained with its objectives, principles and advantages. The mathematical derivations of execrated mean and variance of stratified random sample have been given. Unbiased estimate of variance of stratified random sample along with the allocation of sample size in different strata has been explained. Comparisons of stratified and simple random sampling have been made. Post and deep stratifications are elaborately explained.

4.16 Check Yourself (Questions)

1. Explain Stratified random sampling. What are the various reasons for stratification? Under what situations, the stratified random sampling with Neyman allocation will be just as good as the simple random sampling?
2. A population of 900 units was divided into three strata whose sizes N_i and standard deviations S_i are given below:

Strata	Sizes (N_i)	Standard deviations (S_i)
1	200	9
2	350	15
3	350	27

   A stratified random sample of size $n = 126$ is drawn from the population. Determine the size of the samples drawn from three strata in case of (i) equal allocation, (ii) proportional allocation and (iii) Neyman allocation.
3. Compare simple random sampling with Proportional allocation and Neyman allocation with precision point of view.
4. Explain the concept of post and deep stratifications and their properties.

4.17 References

- Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): Sample survey methods and theory, vol. I and II, New York: John Wiley and Sons, Inc.
- Bryant, E. C., Hartley, H. O. and Jessen, R. J. (1960): Design and estimation in two-way stratification, J. Amer. Stat. Assoc., 55, 105-124.

4.18 Further Readings

- **Cochran, W. G. (1977):** Sampling Techniques, New York: Wiley.
- **Mukhopadhyay, P (2009):** Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- **Singh, D. and Chaudhary, F. S. (1986):** Theory and Analysis of Sample Surveys Designs, Wiley Eastern Ltd., New Age International Ltd.
- **Sukhatme, P. V. and Sukhatme, B. V. (1997):** Sampling Theory of Surveys with Applications, The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

Unit-5 Ratio and Regression Methods of Estimation

Structure

5.1 Introduction

In the conventional estimation of parameter like mean, simple arithmetic mean based on the observed values in the sample of the character under study are used. But sometimes, additional information on so called auxiliary character which is correlated with the study character helps to improve the precision of the estimate without increasing the sample size. In this context, the two main methods for estimating the mean using auxiliary character are (i) ratio method of estimation and (ii) regression method of estimation.

5.2 Objectives

After this unit, the learner would be able to understand about:

- Ratio and regression methods of estimation for estimating the population mean with known population mean of auxiliary character.
- Extension of ratio and regression methods for estimating the population mean under double sampling when the population mean of auxiliary is unknown.
- Application of ratio and regression methods for estimating the unknown parameters under different situations.

5.3 Ratio Method of Estimation

In many Surveys, information on an auxiliary variate which is highly correlated with the variable under study is readily available and can be used for improving sampling design. In ratio method of estimation, the aggregated data on auxiliary variate be used at the time of estimation of the parameters under study provided the data on auxiliary variate for the sampled units be easily obtained while recording the values of the study variate.

5.4 Ratio Estimator

Let y_i and x_i are the values of the characteristics under study and auxiliary for the i^{th} unit of the population. Let us assume that based on 'n' pairs of observations. Let $\bar{y}_n$ and $\bar{x}_n$ are the sample means of the characteristics y and x respectively. and the population mean $\bar{x}_N$ is known (or population total x_N is known). The ratio estimators of the population ratio $y_N/x_N = R_N$, the total y_N and the mean $\bar{y}_N$ may be defined as

$$R_n = \frac{y_n}{x_n} = \frac{\bar{y}_n}{\bar{x}_n}$$

$$y_R = \frac{y_n}{x_n}\, x_N = \frac{\bar{y}_n}{\bar{x}_n} x_N$$

and $$\bar{y}_R = \frac{y_n}{x_n}\, \bar{x}_N = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_N$$ respectively.

where y_n and x_n are the sample totals for y and x respectively.

Notations:

y_i- value of the characteristic under study for the i^{th} unit of the population.

x_i- value of the auxiliary character for the i^{th} unit of the population.

$\gamma_i = \frac{y_i}{x_i}$- ratio of y to x for i^{th} unit

$\bar{r}_n = \frac{1}{n}\sum^{n} r_i$- the simple mean of the ratios for the units in the sample.

$\bar{r}_N = \frac{1}{N}\sum_{i=1}^{N} r_i$- the simple mean of the ratios for the units in the population

$R_n = \frac{\bar{y}_n}{\bar{x}_n}$-the ratio of the sample mean of y to the sample mean of x.

$R_N = \frac{\bar{y}_N}{\bar{x}_N}$-the ratio of the population mean of y to the population mean of x.

5.5 Bias of Ratio Estimator (R_n)

The ratio estimator R_n of population ratio R_N is given by

$$R_n = \frac{\bar{y}_n}{\bar{x}_n}$$

Since $\bar{y}_n$ and $\bar{x}_n$ both are the unbiased estimates of $\bar{y}_N$ and $\bar{x}_N$. Therefore,

$$R_N = \frac{\bar{y}_N}{\bar{x}_N} = \frac{E(\bar{y}_n)}{E(\bar{x}_n)} = \frac{E(R_n\bar{x}_n)}{E(\bar{x}_n)} = -\frac{Cov(R_n\bar{x}_n)}{E(\bar{x}_n)}$$

So, the upper bound to the bias in ratio estimate R_n is given by

$$|Bias(R_n)| \leq \frac{\sigma_{R_n}.\sigma_{\bar{x}_n}}{\bar{x}_N} = \sigma_{R_n}\frac{\sqrt{\frac{N-n}{Nn}S_x^2}}{\bar{x}_N} = \sigma_{R_n}\sqrt{\frac{N-n}{Nn}}\left(\frac{S_x}{\bar{x}_N}\right) =$$

$$\sigma_{R_n}\sqrt{\frac{N-n}{Nn}}C_x$$

where $C_x = \frac{S_x}{\bar{x}_N}$, coefficient of variation of x.

For sufficiently large value of n, we see that the bias in the ratio estimate R_n is negligible as compared to its standard deviation.

5.6 Expected Value of Ratio Estimator for Population Mean ($\bar{y}_N$)

The ratio estimator of the population mean is defined as

$$\bar{y}_R = R_n\bar{x}_N \text{ where } R_n = \frac{\bar{y}_n}{\bar{x}_n}$$

$$\therefore \quad E(\bar{y}_R) = E(R_n\bar{x}_N) = E(R_n)^*\,\bar{x}_N \tag{2.6.1}$$

Let $y_i = \bar{y}_N + \epsilon_i$ and $x_i = \bar{x}_N + \epsilon_i'$

or $\bar{y}_n = \bar{y}_N + \bar{\epsilon}_n \quad \Rightarrow E(\bar{\epsilon}_n) = 0$

Similarly, $\bar{x}_n = \bar{x}_N + \bar{\epsilon}_n' \Rightarrow E(\bar{\epsilon}_n') = 0$

$$R_n = \frac{\bar{y}_n}{\bar{x}_n} = \frac{\bar{y}_N + \bar{\epsilon}_n}{\bar{x}_N + \bar{\epsilon}_n'}$$

$$= R_N\left(1 + \frac{\bar{\epsilon}_n}{\bar{y}_N}\right)\left(1 + \frac{\bar{\epsilon}_n'}{\bar{x}_N}\right)^{-1}$$

$$= R_N\left(1 + \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}_n'}{\bar{x}_N} - \frac{\bar{\epsilon}_n\bar{\epsilon}_n'}{\bar{y}_N\bar{x}_N} + \frac{\bar{\epsilon}_n'^2}{\bar{x}_N^2} + \frac{\bar{\epsilon}_n\bar{\epsilon}_n'^2}{\bar{y}_N\bar{x}_N^2} \ldots .\right)$$

By ignoring the terms of the second and higher orders, we have

$$E(R_n) = R_N\left(1 + \frac{V(\bar{x}_N)}{\bar{x}_N^2} - \frac{E(\bar{\epsilon}_n\bar{\epsilon}_n')}{\bar{y}_N\bar{x}_N}\right) \tag{2.6.2}$$

Now, $$E(\bar{\epsilon}_n\bar{\epsilon}'_n) = E\left[\left(\frac{1}{n}\sum_i^n \epsilon_i\right)\left(\frac{1}{n}\sum_i^n \epsilon'_i\right)\right]$$

$$= \frac{1}{n^2}\left[\sum_i^n E(\epsilon_i\epsilon'_i) + \sum_{i\neq j}^n E(\epsilon_i\epsilon'_i)\right]$$

Or $$E(\bar{\epsilon}_n\bar{\epsilon}'_n) = \frac{N-n}{Nn}\left\{\frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{y}_N)(x_i - \bar{x}_N)\right\}$$

$$= \frac{N-n}{Nn}S_{yx} = \frac{N-n}{Nn}\rho S_y S_x \qquad (2.6.3)$$

Where ρ is the correlation coefficient between y and x, which is given by

$$\rho = \frac{S_{yx}}{S_y S_x} = \frac{E(y_i-\bar{y}_N)(x_i-\bar{x}_N)}{\sqrt{E(y_i-\bar{y}_N)^2 E(x_i-\bar{x}_N)^2}}$$

Therefore, from (2.6.2) and (2.6.3), we have

$$E(R_n) = R_N\left[1 + \left(\frac{N-n}{Nn}\right)\frac{S_x^2}{\bar{x}_N^2} - \left(\frac{N-n}{Nn}\right)\rho\frac{S_y}{\bar{y}_N}\frac{S_x}{\bar{x}_N}\right],$$

$$E(R_n) = R_N\left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_y C_x\right)\right] \qquad (2.6.4)$$

where $C_x = \frac{S_x}{\bar{x}_N}$, $C_y = \frac{S_y}{\bar{y}_N}$.

From (2.6.1) and (2.6.4), we get

$$E(\bar{y}_R) = \bar{y}_N\left[1 + \frac{N-n}{Nn}\left(C_x^2 - \rho C_y C_x\right)\right] \qquad (2.6.5)$$

From (2.6.4), the first approximation to the relative bias of the ratio estimator (R_n) is given by

$$R.B.(R_n) = \frac{E(R_n)-R_N}{R_N}$$

or $$R.B.(R_n) = \frac{N-n}{N}.\frac{1}{n}\left(C_x^2 - \rho C_y C_x\right) \qquad (2.6.6)$$

From (2.6.5), Bias of $\bar{y}_R$ is given by

$$Bias(\bar{y}_R) = E(\bar{y}_R) - \bar{y}_N$$

$$Bias(\bar{y}_R) = \frac{N-n}{N}.\frac{1}{n}\bar{y}_N\left(C_x^2 - \rho C_y C_x\right) \qquad (2.6.7)$$

From (2.6.6) and (2.6.7), we observe that for large and sufficiently high value of ρ, the bias will be negligible. It should be noted that the bias in the ratio estimator becomes zero, when

$$C_x^2 - \rho C_y C_x = 0$$

or $$\rho = R_N\frac{S_x}{S_y}$$

or $$\bar{y}_N = \rho\frac{S_x}{S_y}\bar{x}_N$$

which is satisfied only if the line of regression y on x is a straight line and passes through origin.

5.7 Variance of Ratio Estimate of Population Mean $\bar{y}_R$

The ratio estimate of population mean $\bar{y}_N$ is define as

$$\bar{y}_R = R_n\bar{x}_N$$

Therefore, $V(\bar{y}_R) = V(R_n \bar{x}_N) = \bar{x}_N^2 \, V(R_n)$

(2.7.1)

Now
$$V(R_n) = E[R_n - E(R_n)]^2$$
$$= E(R_n^2) - [E(R_n)]^2$$
$$\simeq E(R_n - R_N)^2 \qquad (2.7.2)$$

we know that $R_n = R_N + R_N \left\{ \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}'_n}{\bar{x}_N} + \cdots \right\}$

or $R_n - R_N = R_N \left\{ \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}'_n}{\bar{x}_N} + \cdots \right\}$

Now squaring both sides and neglecting the higher order terms, we have

$$(R_n - R_N)^2 = R_N^2 \left[\frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}'_n}{\bar{x}_N} \right]^2 \qquad (2.7.3)$$

From (2.7.2) and (2.7.3), we have

$$V(R_n) = E\left[R_N^2 \left\{ \frac{\bar{\epsilon}_n}{\bar{y}_N} - \frac{\bar{\epsilon}'_n}{\bar{x}_N} \right\}^2 \right]$$

or $$V(R_n) = R_N^2 \frac{N-n}{Nn} \left\{ \frac{S_y^2}{\bar{y}_N^2} + \frac{S_x^2}{\bar{x}_N^2} - \frac{2S_{yx}}{\bar{y}_N \bar{x}_N} \right\} \qquad (2.7.4)$$

Since $\frac{S_{yx}}{\bar{y}_N \bar{x}_N} = \rho C_y C_x$, where ρ is the correlation coefficient between y and x.

Therefore,

$$V(R_n) = R_N^2 \frac{N-n}{Nn} \{ C_y^2 + C_x^2 - 2\rho C_y C_x \} \qquad (2.7.5)$$

Hence, substituting the above value of $V(R_n)$ in (2.7.1), we get

$$V(\bar{y}_R) = \bar{y}_N^2 \frac{N-n}{Nn} \{ C_y^2 + C_x^2 - 2\rho C_y C_x \}$$

or $$V(\bar{y}_R) = \frac{N-n}{Nn} \{ S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x \}$$

Also, since $\frac{S_y^2}{\bar{y}_N^2} + \frac{S_x^2}{\bar{x}_N^2} - 2\rho \frac{S_y S_x}{\bar{y}_N \bar{x}_N} = \frac{1}{(N-1)\bar{y}_N^2} \sum_{i=1}^{N} (y_i - R_N x_i)^2$

Therefore, $V(\bar{y}_R) = \frac{N-n}{Nn} . \frac{1}{N-1} \sum_{i=1}^{N} (y_i - R_N x_i)^2$

5.8 Relationship between Biased and Variance

Relation between biased and variance is given by

$$R.B.(R_n) = \frac{E(R_n) - R_N}{R_N} = \frac{N-n}{Nn} \left(C_x^2 - \rho C_y C_x \right)$$
$$\simeq \frac{C^2}{n} (1 - \rho) \text{ if } C_x = C_y = C \text{ and } N \text{ is large}$$

and relative variance is given by

$$V\left(\frac{R_n}{R_N} \right) = \frac{N-n}{Nn} \left(C_y^2 + C_x^2 - 2\rho C_y C_x \right)$$

If $C_x = C_y = C$ and N is large then $\frac{N-n}{N} \simeq 1$.

Therefore, $V\left(\frac{R_n}{R_N} \right) = \frac{2C^2}{n} (1 - \rho)$

Therefore, Relative variance $= 2$ Relaitve bias.

5.9 Estimate of Variance of (R_n)

The variance of ratio estimate is given by

$$V(R_n) = \frac{N-n}{Nn}.\frac{1}{\bar{x}_N^2}\left[S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x\right]$$

As we know that s_y^2, $\bar{y}_n$, s_x^2 and $\bar{x}_n$ are an unbiased estimates of the corresponding population values. Similarly,

$$s_{yx} = \frac{1}{n-1}\sum_i^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)$$

provides an unbiased estimate of the corresponding population value $S_{yx} = \rho S_y S_x$ and if r is the estimate of correlation coefficient from sample values, then

$$Est.\left(V(R_n)\right) = \frac{N-n}{Nn}.\frac{1}{\bar{x}_N^2}\left[s_y^2 + R_n^2 s_x^2 - 2R_n s_{yx}\right]$$

or $$Est.\left(V(R_n)\right) = \frac{N-n}{Nn}.\frac{1}{\bar{x}_N^2}\left[s_y^2 + R_n^2 s_x^2 - 2rR_n s_y s_x\right]$$

In the same way, the estimates of the variance of R_n and $\bar{y}_R$ are given by

$$Est.\left(V(R_n)\right) = \frac{N-n}{Nn}.\frac{1}{\bar{x}_N^2}\frac{1}{n-1}\sum^n (y_i - R_n x_i)^2$$

and $$Est.\left(V(\bar{y}_R)\right) = \frac{N-n}{Nn}.\frac{1}{n-1}\sum^n (y_i - R_n x_i)^2.$$

5.10 Comparison of Ratio Estimator with Mean per Unit

The conditions under which the ratio estimator is superior to the mean per unit can be work out with a comparison of their variances.

In SRSWOR, the variance of the mean per unit is given by

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

Also, the variance of the mean based on the ratio method to the first order approximation is given by $$V(\bar{y}_R) = \frac{N-n}{Nn}\left\{S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x\right\}$$

Hence, the ratio estimate will have smaller variance if

$$S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x < S_y^2$$

or $$\rho > \frac{R_N S_x}{2S_y} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right) = \frac{C_x}{2C_y}$$

i.e. $$\rho > \frac{1}{2}\frac{C_x}{C_y}.$$

Thus, it depends on the value of correlation between y and x. If $C_x = C_y = C$ i.e the values of coefficient of variation of x and y are the same, the ratio estimator is superior if ρ exceeds 0.5. The variability of the auxiliary variate x is an important factor. If C.V. of x is more than twice that of y, the ratio estimate is always less precise.

5.11 Double Sampling in Ratio Method of Sampling

In most of the time of sample surveys, it happens that the population mean $\bar{x}_N$ is unknown and in this case the ratio method of estimation cannot be used to estimate the population mean $\bar{y}_N$. The usual procedure in such case is to use the technique known as two-phase or double sampling. The technique consists in taking a first phase large sample of size n' to estimate the population mean $\bar{x}_N$ while a sub-sample of size n at the second phase is drawn from n' to observe characteristic under study. Several estimates of the population mean $\bar{y}_N$ can be formed. The simplest is the usual biased ratio estimate $\bar{y}_{Rd}$, with $\bar{x}_N$ replaced by its estimate $\bar{x}_{n'}$, based on a sample of size n', given by

$$\bar{y}_{Rd} = \frac{\bar{y}_n}{\bar{x}_n}.\bar{x}_{n'}$$

To find the expectation and variance of the ratio estimate, we write

$$\bar{y}_n = \bar{y}_N + \epsilon_0, \bar{x}_n = \bar{x}_N + \epsilon_1, \bar{x}_{n'} = \bar{x}_N + \epsilon_1'$$

such that $E(\epsilon_0) = E(\epsilon_1) = E(\epsilon_1') = 0.$

Then,
$$\bar{y}_{Rd} = \left(\frac{\bar{y}_N+\epsilon_0}{\bar{x}_N+\epsilon_1}\right).(\bar{x}_N + \epsilon_1')$$
$$= \bar{y}_N\left[1+\frac{\epsilon_0}{\bar{y}_N}\right]\left[1+\frac{\epsilon_1'}{\bar{x}_N}\right]\left[1+\frac{\epsilon_1}{\bar{x}_N}\right]^{-1}$$

Assuming that $\left|\frac{\epsilon_1}{\bar{x}_N}\right| < 1$, so that the expansion of $\left[1+\frac{\epsilon_1}{\bar{x}_N}\right]^{-1}$ is valid and ignoring terms of order higher than two, we have

$$\bar{y}_{Rd} = \bar{y}_N\left[1+\frac{\epsilon_0}{\bar{y}_N}+\frac{\epsilon_1'}{\bar{x}_N}-\frac{\epsilon_1}{\bar{x}_N}+\frac{\epsilon_0\epsilon_1'}{\bar{y}_N\bar{x}_N}-\frac{\epsilon_0\epsilon_1}{\bar{y}_N\bar{x}_N}-\frac{\epsilon_1\epsilon_1'}{\bar{x}_N^2}+\frac{\epsilon_1^2}{\bar{x}_N^2}\right]$$

Now
$$E(\epsilon_0\epsilon_1') = Cov(\bar{y}_n, \bar{x}_{n'})$$
$$= Cov[E(\bar{y}_n|n'), E(\bar{x}_{n'}|n')] + E[Cov(\bar{y}_n, \bar{x}_{n'}|n')]$$
$$= Cov(\bar{y}_n, \bar{x}_{n'})$$
$$= \left(\frac{1}{n'}-\frac{1}{N}\right)S_{yx}$$

$$E(\epsilon_0\epsilon_1) = Cov(\bar{y}_n, \bar{x}_n) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{yx}$$

Hence, ignoring the terms in ϵ of higher than two and taking expectation term by term, we have

$$E(\bar{y}_{Rd}) \simeq \bar{y}_N\left[1+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_x^2-\rho_{yx}C_xC_y\right)\right]$$

The relative bias of $\bar{y}_{Rd}$ is thus given by

$$RB(\bar{y}_{Rd}) = \left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_x^2-\rho_{yx}C_xC_y\right)$$

which will be negligible if the sample size n is sufficiently large. If the regression of y on x is linear and passes through the origin, it will be zero to the first degree of approximation.

Again, to find the variance, we have to take a first approximation

$$V_1(\bar{y}_{Rd}) = E(\bar{y}_{Rd} - \bar{y}_N)^2 = \bar{y}_N^2 E\left[\frac{\epsilon_0}{\bar{y}_N} + \frac{\epsilon_1'}{\bar{x}_N} - \frac{\epsilon_1}{\bar{x}_N}\right]^2$$

Expanding and taking expectation term by term, we have

$$V_1(\bar{y}_{Rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(S_y^2 + R_N^2 S_x^2 - 2R_N S_{yx}\right) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2$$

It follows that the estimate $\bar{y}_{Rd}$ based on double sampling is more efficient than the estimate $\bar{y}_n$ based on simple random sampling when no auxiliary variable is used, if

$$R_N^2 S_x^2 - 2R_N S_{yx} < 0$$

i.e., if $\rho_{yx}.\frac{C_y}{C_x} > \frac{1}{2}$

The comparison indicates that if the ratio estimate using some auxiliary characteristic is more efficient than the one based on simple random sampling, the ratio estimate based on double sampling will also be more efficient.

Example 2.1: Estimate the total number of literate persons in 117 villages of a tahsil on the basis of following survey and census data by

(i) ratio method and

(ii) simple mean per unit method

S. No. of village	1	2	3	4	5	6	7	8	9
No. of literate persons in survey (y)	1129	1144	1125	1138	1137	1127	1163	1153	1164
No. of literate persons in census (x)	1141	1144	1127	1153	1117	1140	1153	1146	1189
S. No. of village	10	11	12	13	14	15	16	17	
No. of literate persons in survey (y)	1130	1153	1125	1116	1115	1112	1112	1123	
No. of literate persons in census (x)	1137	1170	1115	1130	1118	1122	1113	1166	

Also compare its precision, given the total number of literate persons in the census is 143968.

Solution: The relevant values for the calculation are as follows:

$N = 117$, $n = 17$, $x_N = 143968$, $\sum x_i = 19381$, $\sum y_i = 19266$, $\bar{x}_n = 1140.06$,

$\bar{y}_n = 1133.29$, $R_n = 0.994$, $s_y^2 = 287.85$, $s_x^2 = 458.56$, $s_{yx} = 262.8\ 6$

(i) The total number of literate persons by the ratio method of estimation is given by

$$y_R = \hat{R}x_N = \frac{y_n}{x_n}\, x_N = 0.994 \times 143968 = 143120$$

and the estimate of the variance of y_R is given by

$$V(y_R) = N^2 V(\bar{y}_R) = \frac{N(N-n)}{n}\{S_y^2 + R_N^2 S_x^2 - 2\rho R_N S_y S_x\}$$
$$= \frac{117\times100}{17}\{287.85 + 453.61 - 522.62\} = 1{,}50{,}304$$

(ii) The total number of literate persons by mean per unit estimate is given by

$$y_{SR} = N\,\bar{y}_n = 117 \times 1133.29 = 132595$$

and the estimate of the variance of y_{SR} is given by

$$V(y_{SR}) = N^2 V(\bar{y}_n) = \frac{N(N-n)}{n}\,S_y^2$$
$$= \frac{117\times100}{17} \times 287.85 = 1{,}98{,}108$$

The relative precision of ratio estimate is given by

$$R.P. = \left\{\frac{V(y_{SR})}{V(y_R)}\right\} \times 100 = 131.8\%$$

5.12 Regression Method of Estimation

Linear regression estimators also make use of auxiliary information for increasing precision. It has been seen that the ratio estimator provides a precise estimate of the population mean if regression is linear and the line passes through the origin. When regression is linear and the line does not go through the origin, it is better to use estimators based on linear regression.

In other words, if the study variate (y) is approximately a constant and a multiple of auxiliary variate, it is more precise to estimate the population mean or total by fitting a linear regression. Such an estimator is called a regression estimator.

5.13 Difference Estimator

Let y and x are correlated characteristics. If $\bar{y}_n$ and $\bar{x}_n$ are the unbiased estimators of population mean of study and auxiliary characters respectively based on simple random sample, then we can improve the estimator $\bar{y}_n$ by introducing a difference function.

Hence a generalized difference estimator for estimating the population mean $\bar{y}_N$ is defined as

$$\bar{y}_d = \bar{y}_n + C(\bar{x}_N - \bar{x}_n) \qquad (2.13.1)$$

where C and $\bar{x}_N$ are known quantities.

From (2.13.1), we find that $\bar{y}_d$ is an unbiased estimator of population mean $\bar{y}_N$ and its variance is given by

$$V(\bar{y}_d) = V(\bar{y}_n) + C^2 V(\bar{x}_n) - 2CCov(\bar{y}_n, \bar{x}_n) \qquad (2.13.2)$$

Now differentiate (2.13.2) with respect to C and equating the result to zero for minimum variance, we have

$$C = \frac{Cov(\bar{y}_n, \bar{x}_n)}{V(\bar{x}_n)} = \frac{S_{yx}}{S_x^2} \ (= \beta) \qquad (2.13.3)$$

Hence $V(\bar{y}_d)$will have its minimum value when C is β, the value of the regression coefficient of y on x.

Here $S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)(x_i - \bar{x}_N)$

$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2$

5.14 Regression Estimator

The difference estimator for population mean is given by

$$\bar{y}_d = \bar{y}_n + C(\bar{x}_N - \bar{x}_n)$$

The optimum value of C is given to equal to β, which is regression coefficient of y on x. Generally, β is not known in advance and its value is estimated from the sample. So a consistent estimator of β is given by

$$\hat{\beta} = \frac{s_{yx}}{s_x^2},$$

where $s_{yx} = \frac{1}{n-1}\sum_{i}^{n}(y_i - \bar{y}_n)(x_i - \bar{x}_n)$ and $s_x^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x}_n)^2$

Hence, linear regression estimator of the population mean $\bar{y}_N$ and population total y_N are given by

$$\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$$

or $y_l = N\bar{y}_l$

5.15 Bias or Expected Value of Simple Regression Estimator

The linear regression estimator of the population mean $\bar{y}_N$ is given by

$$\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$$

(2.15.1)

where $\hat{\beta} = \frac{s_{yx}}{s_x^2}$

Let $\bar{x}_n = \bar{x}_N + \epsilon_1$, $s_{yx} = S_{yx} + \epsilon_2$, $s_x^2 = S_x^2 + \epsilon_3$ such that $E(\epsilon_1) = E(\epsilon_2) = E(\epsilon_3) = 0$

Then, $\bar{y}_l = \bar{y}_n + \frac{s_{yx}}{s_x^2}(\bar{x}_N - \bar{x}_n)$

$= \bar{y}_n + \frac{S_{yx}+\epsilon_2}{S_x^2+\epsilon_3}(-\epsilon_1)$

or $\bar{y}_l = \bar{y}_n - \epsilon_1 \frac{S_{yx}}{S_x^2}\left(1 + \frac{\epsilon_2}{S_{yx}}\right)\left(1 + \frac{\epsilon_3}{S_x^2}\right)^{-1}$

$= \bar{y}_n - \epsilon_1 \beta\left(1 + \frac{\epsilon_2}{S_{yx}}\right)\left(1 + \frac{\epsilon_3}{S_x^2}\right)^{-1},$

where β is the regression coefficient of y on x. Since $\left|\frac{\epsilon_3}{S_x^2}\right| < 1$, the expansion of $\left(1+\frac{\epsilon_3}{S_x^2}\right)^{-1}$ is valid. Expanding and ignoring the terms of order higher than two, we have

$$\bar{y}_l = \bar{y}_n - \epsilon_1 \beta\left(1+\frac{\epsilon_2}{S_{yx}}\right)\left(1-\frac{\epsilon_3}{S_x^2}\right)$$

$$= \bar{y}_n - \beta\left(\epsilon_1 + \frac{\epsilon_1\epsilon_2}{S_{yx}} - \frac{\epsilon_1\epsilon_3}{S_x^2}\right)$$

Therefore, $E(\bar{y}_l) = E(\bar{y}_n) - \beta\left[\frac{E(\epsilon_1\epsilon_2)}{S_{yx}} - \frac{E(\epsilon_1\epsilon_3)}{S_x^2}\right]$

or $E(\bar{y}_l) = \bar{y}_N - \beta\left[\frac{Cov(\bar{x}_n, s_{yx})}{S_{yx}} - \frac{Cov(\bar{x}_n, s_x^2)}{S_x^2}\right]$

(2.15.2)

If $Cov(\bar{x}_n, s_{yx}) \simeq \frac{N-n}{Nn}\mu_{21}$, where $\mu_{21} = E\{(x-\bar{x}_N)^2(y-\bar{y}_N)\}$ and $Cov(\bar{x}_n, s_x^2) \simeq \frac{N-n}{Nn}\mu_{30}$, where $\mu_{30} = E\{(x-\bar{x}_N)^3(y-\bar{y}_N)^0\}$

The equation (2.15.2) can be rewritten as

$$E(\bar{y}_l) = \bar{y}_N - \beta\left[\frac{\mu_{21}}{S_{yx}} - \frac{\mu_{30}}{S_x^2}\right]$$

Hence, regression estimator is a biased estimator and bias of regression estimator is given by

$$Bias(\bar{y}_l) = -\frac{N-n}{Nn}\beta\left[\frac{\mu_{21}}{S_{yx}} - \frac{\mu_{30}}{S_x^2}\right]$$

The above expression shows that the bias will be zero for sufficiently large value of n or if the joint distribution of x and y is bivariate normal.

Another expression for bias:

Since $\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$

Therefore, $E(\bar{y}_l) = E(\bar{y}_n) + E\left[\hat{\beta}(\bar{x}_N - \bar{x}_n)\right]$

$$= \bar{y}_N + E\{\hat{\beta}(\bar{x}_N)\} - E\{\hat{\beta}(\bar{x}_n)\}$$

$$= \bar{y}_N + E\{\hat{\beta}, \bar{x}_N\} - E\{\hat{\beta}, \bar{x}_n\}$$

$$= \bar{y}_N - Cov(\hat{\beta}, \bar{x}_n)$$

Therefore, $Bias(\bar{y}_l) \simeq -Cov(\hat{\beta}, \bar{x}_n)$.

5.16 Variance of Simple Regression Estimate or $(\bar{y}_l)$

The simple regression estimator $\bar{y}_l$ for population mean is given by

$$\bar{y}_l = \bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)$$

where $\hat{\beta} = \frac{s_{yx}}{s_x^2}$.

Therefore, $V(\bar{y}_l) = E[\bar{y}_l - E(\bar{y}_l)]^2$ (2.16.1)

Let $\bar{x}_n = \bar{x}_N + \epsilon_1$, $s_{yx} = S_{yx} + \epsilon_2$, $s_x^2 = S_x^2 + \epsilon_3$ such that $E(\epsilon_1) = E(\epsilon_2) = E(\epsilon_3) = 0$

$$\bar{y}_l = \bar{y}_n - \beta \in_1 - \beta\left[\frac{\in_1 \in_2}{S_{yx}} - \frac{\in_1 \in_3}{S_x^2}\right] + \cdots \qquad (2.16.2)$$

and $$E(\bar{y}_l) \simeq \bar{y}_N - \beta\left[\frac{Cov(\bar{x}_n, s_{yx})}{S_{yx}} - \frac{Cov(\bar{x}_n, s_x^2)}{S_x^2}\right] \qquad (2.16.3)$$

Now substituting from (2.16.2) and (2.16.3) in (2.16.1) and ignoring the terms of order greater than two, we have

$$V(\bar{y}_l) = E[\bar{y}_n - \beta \in_1 - \bar{y}_N)]^2$$
$$= E[(\bar{y}_n - \bar{y}_N) - \beta \in_1]^2$$
$$= E(\bar{y}_n - \bar{y}_N)^2 + \beta^2 E(\bar{x}_n - \bar{x}_N)^2 - 2\beta E\{(\bar{y}_n - \bar{y}_N)(\bar{x}_n - \bar{x}_N)\}$$
$$= \frac{N-n}{Nn} S_y^2 + \beta^2 \frac{N-n}{Nn} S_x^2 - 2\beta \frac{N-n}{Nn} S_{yx}$$

or $$V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + \beta^2 S_x^2 - 2\beta S_{yx}\right)$$

Let ρ is the correlation coefficient between y and x, then $\rho = \frac{S_{yx}}{S_y S_x}$ and $\beta = \frac{S_{yx}}{S_x^2}$.

Therefore, $$V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 - \frac{S_{yx}^2}{S_x^2}\right)$$

or $$V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 (1 - \rho^2)$$

This shows that if regression is linear and $\hat{\beta}$ is the least square of β then the regression estimator $(\bar{y}_l)$ is more precise than the difference estimator $\bar{y}_d$. If the regression of y on x is perfectly linear i.e. $|\rho| = 1$, the variance of the regression estimator becomes zero. Also, if y and x are uncorrelated, the variance of $\bar{y}_l$ reduces to the of the mean per unit estimator $\bar{y}$.

If we take N is sufficiently large, then we can ignore the finite correction factor and hence

$$V(\bar{y}_l) \simeq \sigma_y^2 \frac{(1-\rho^2)}{n}.$$

5.17 Estimate of Variance of the Simple Regression Estimator

Since s_y^2, s_x^2 and s_{yx} are unbiased estimate of S_y^2, S_x^2 and S_{yx} respectively a consistent estimate of the variance of the regression estimate is given by

$$Est.V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(s_y^2 + \hat{\beta}^2 s_x^2 - 2\hat{\beta} R_{yx}\right)$$

which on simplification, put in the form

$Est.V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 (1 - r^2)$,

where $r = s_{yx}/s_y s_x$ is the sample regression coefficient.

5.18 Comparison between Variance of Sample Mean in SRS $[V(\bar{y}_n)]$, Ratio Estimate $[V(\bar{y}_R)]$ and Simple Regression Estimate $[V(\bar{y}_l)]$

We know that

a) Variance of sample mean in SRS

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 \tag{2.18.1}$$

b) Variance of ratio estimate

$$V(\bar{y}_R) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 - 2R_N\rho S_y S_x + R_N^2 S_x^2\right) \tag{2.18.2}$$

c) Variance of simple regression estimate

$$V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(1-\rho^2) \tag{2.18.3}$$

From equation (2.18.1) and (2.18.3)

$$V(\bar{y}_n) - V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2\rho^2 \tag{2.18.4}$$

From (2.18.4), we observe that the variance of regression estimator is always smaller unless $\rho = 0$. In case, $\rho = 0$, the variances for both are equal. The reduction in variance is large when ρ is high and small when ρ is low.

Now from (2.18.2) and (2.18.3), we have

$$V(\bar{y}_R) - V(\bar{y}_l) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(\rho^2 S_y^2 - 2R_N\rho S_y S_x + R_N^2 S_x^2\right) \tag{2.18.5}$$

By comparing the ratio estimator, it has been observed that variance of regression estimator is less than that of the ratio estimator if

$$\left(\rho S_y - R_N S_x\right)^2 > 0$$

or $$(\beta - R_N)^2 > 0$$

which is always true except $\beta = R_N$. In this situation, both estimates will have the same variance and this occurs only when regression of y on x is straight line through the origin.

5.19 Double Sampling in Regression Method of Estimation

The regression estimate of the population mean of study character y pre-assumes that the population mean of auxiliary character x, namely $\bar{x}_N$, is known. However, $\bar{x}_N$ is not always known, although in many situations, it is possible to estimate it from a second sample of the population without appreciably adding to the cost of the inquiry. Here, we shall assume that a larger first phase sample of size n' is drawn with equal probability and without replacement to estimate the population mean $\bar{x}_N$ of the auxiliary character x while a second phase sub-sample of size n is drawn from n' to observe the characteristic under study. Since $\bar{x}_{n'}$ is an unbiased estimate of the population mean $\bar{x}_N$, so an estimator for estimating $\bar{y}_N$ is given by

$$\bar{y}_{ld} = \bar{y}_n + \hat{\beta}(\bar{x}_{n'} - \bar{x}_n)$$

Let us consider, $\bar{x}_n = \bar{x}_N + \epsilon_1$, $\bar{x}_{n'} = \bar{x}_N + \epsilon_1'$, $s_{yx} = S_{yx} + \epsilon_2$, $s_x^2 = S_x^2 + \epsilon_3$ such that $E(\epsilon_1) = E(\epsilon_1') = E(\epsilon_2) = E(\epsilon_3) = 0$.

Then, we have

$$\bar{y}_{ld} = \bar{y}_n + \frac{S_{yx}+\epsilon_2}{S_x^2+\epsilon_3}.(\epsilon_1'-\epsilon_1)$$

$$= \bar{y}_n + \beta\left[1 + \frac{\epsilon_2}{S_{yx}}\right]\left[1 + \frac{\epsilon_3}{S_x^2}\right]^{-1}(\epsilon_1'-\epsilon_1)$$

Assuming that $\left|\frac{\epsilon_3}{S_x^2}\right| < 1$ so that the expansion of $\left[1 + \frac{\epsilon_3}{S_x^2}\right]^{-1}$ is valid, expanding and ignoring terms in ϵ of order higher than two and taking expectation, we obtain

$$E(\bar{y}_{ld}) \simeq \bar{y}_N + \beta\left[\frac{E(\epsilon_2\epsilon_1')-E(\epsilon_2\epsilon_1)}{S_{yx}} - \frac{E(\epsilon_3\epsilon_1')-E(\epsilon_3\epsilon_1)}{S_x^2}\right]$$

$$= \bar{y}_N + \beta\left[\frac{Cov(s_{yx},\bar{x}_{n'})-Cov(s_{yx},\bar{x}_n)}{S_{yx}} - \frac{Cov(s_x^2,\bar{x}_{n'})-Cov(s_x^2,\bar{x}_n)}{S_x^2}\right]$$

$$\simeq \bar{y}_N - \beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left[\frac{\mu_{21}}{S_{yx}} - \frac{\mu_{30}}{S_x^2}\right]$$

The mean square error of $\bar{y}_{ld}$ up to the same order of approximation is given by

$$M.S.E.(\bar{y}_{ld}) = E(\bar{y}_{ld} - \bar{y}_N)^2$$

or $$M.S.E.(\bar{y}_{ld}) \simeq E[\bar{y}_n - \bar{y}_N + \beta(\epsilon_1'-\epsilon_1)]^2$$

$$= V(\bar{y}_n) + \beta^2 E(\epsilon_1'-\epsilon_1)^2 + 2\beta E[(\epsilon_1'-\epsilon_1)(\bar{y}_n - \bar{y}_N)],$$

where $E(\epsilon_1\epsilon_1') = \left(\frac{1}{n'} - \frac{1}{N}\right)S_x^2$ and $E[(\bar{y}_n - \bar{y}_N)\,\epsilon_1'] = \left(\frac{1}{n'} - \frac{1}{N}\right)S_{yx}$

After simplification, we have

$$M.S.E.(\bar{y}_{ld}) = \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2(1-\rho^2) + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2$$

$$\simeq \frac{S_y^2(1-\rho^2)}{n} + \frac{S_y^2\rho^2}{n'}$$

It is clear that the regression estimate $\bar{y}_{ld}$ in the case of double sampling is always more efficient than the estimate $\bar{y}_n$ based on simple random sampling when no auxiliary variable is used.

Example 2.2: Estimate the total number of literate persons in 117 villages of a tahsil using the data given in Example 2.1 by regression method of estimation and compare its precision with ratio estimate and mean per unit estimate.

Solution: From Example 1.2, we have

$N = 117$, $n = 17$, $x_N = 143968$, $\sum x_i = 19381$, $\sum y_i = 19266$, $\bar{x}_n = 1140.06$,

$\bar{y}_n = 1133.29$, $R_n = 0.994$, $s_y^2 = 287.85$, $s_x^2 = 458.56$, $s_{yx} = 262.86$, $\hat{\beta} = \frac{s_{yx}}{s_x^2} = 0.573$,

$r = s_{yx}/s_y s_x = 0.723$

The total number of literate persons by the regression method of estimation is given by

$$y_l = N(\bar{y}_l) = N\{\bar{y}_n + \hat{\beta}(\bar{x}_N - \bar{x}_n)\}$$
$$= 117\{1133.29 + 0.573\left(\frac{143968}{117} - 1140.06\right) = 138658$$

and the estimate of the variance of y_R is given by

$$V(y_l) = N^2 V(\bar{y}_l) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)s_y^2(1 - r^2)$$
$$= \frac{N(N-n)}{n}(1 - r^2)s_y^2$$
$$= \frac{117(117- \quad)}{17}(1 - 0.723^2)85 = 94551$$

From Example 2.1, we have $V(y_R) = 1{,}50{,}304$, $V(y_{SR}) = 1{,}98{,}108$
Therefore, the relative precisions of regression estimate over mean per unit and ratio estimate are respectively given by

$$R.P. = \left\{\frac{V(y_{SR})}{V(y_l)}\right\} \times 100 = 209.5\%$$
$$R.P. = \left\{\frac{V(y_R)}{V(y_l)}\right\} \times 100 = 158.9\%$$

5.20 Summary

The present unit provides a brief idea about ratio and regression methods of estimation for estimating the population mean of study character. This unit explains the application of auxiliary character in ratio and regression methods of estimation and extended over double sampling to estimate the parameter of auxiliary character. Theoretical comparisons for the efficiency of the estimate obtained by these methods have done with examples. For further study and exercise, readers are required to study the referred text books of sampling techniques.

5.21 Check Yourself (Questions)

1. Define ratio estimator and give justification for its use on place of sample mean $(\bar{y}_n)$ for estimating the population mean.
2. Write the first order approximation formula to work out the bias and mean square error of the ratio estimator $\bar{y}_R$. Obtained the expression to estimate the sampling variance of ratio estimator $\bar{y}_R$ and explain each term used in the expression.
3. Define regression estimators $(\bar{y}_l)$ and give justification for its use in place of sample estimator $(\bar{y}_n)$ based on mean per unit. Obtain the expression for the sampling variance of regression estimator.
4. A random sample of size $N = 15$ was drawn from a bi-variate population using SRSWOR and reported as

y:	7	8	9	10	11	2	3	4	5	6	12	13	14	15	16
x:	6	7	8	9	10	1	2	3	4	5	11	12	13	14	15

Given that the population size $N = 75$ and population mean of x variate $\bar{x}_N = 13.0$.

Obtain the estimate of population mean $\bar{y}_N$ using ratio and regression methods of estimation. Also find the estimates of sampling variance of both the considered estimators.

5.22 Further Readings

- Cochran, W. G. (1977): Sampling Techniques, New York: Wiley.
- Mukhopadhyay, P (2009): Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- Singh, D. and Chaudhary, F. S. (1986): Theory and Analysis of Sample Surveys Designs, Wiley Eastern Ltd., New Age International Ltd.
- Sukhatme, P. V. and Sukhatme, B. V. (1997): Sampling Theory of Surveys with Applications, The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

Unit-6 Cluster and Multi-Stage Sampling

Structure

6.1 Introduction

The group of smallest unit into which the population can be divided is called cluster and the procedure of sampling in which the sampling unit is a cluster is called cluster sampling.

For a given number of sampling units, cluster sampling is more convenient and cost efficient. Some advantages of cluster sampling are as follows:

i. It is less costly than simple random sampling due to the saving of time in journeys, identification, contacts etc.,
ii. Collection of data for neighboring elements is easier, cheaper, faster and operationally more convenient than observing units over a region,
iii. When the sampling frame of elements may not be readily available.

The cluster sampling is generally less efficient than simple random sampling due to the usual tendency of units in a cluster to be similar. In fact, the efficiency of cluster sampling is likely to decrease with increase in cluster size. In most practical situations, the loss in efficiency may be balanced by reduction in cost. Therefore, the efficiency per unit cost may be more in cluster sampling than in simple random sampling.

Example-In a city, a list of all houses may be available but a list of all the persons is generally not available. Cluster sampling in such case is mostly preferred.

6.2 Objectives

After this unit, the learner should able to understand about:

- cluster sampling with equal as well as varying size of clusters and their properties
- sub sampling and its properties
- extension of sub sampling to multi-stage sampling and its advantages

6.3 Cluster Sampling with Equal Clusters

In this section, cluster sampling with equal clusters is discussed along with its properties. The notations used in this sampling are given as below:

Notations: (Equal Cluster)

N-Number of clusters in the population

M- Number of elements in each cluster

n- Number of clusters selected in the sample

y_{ij}-value of characteristic under study for j^{th} element $(j = 1,2, \dots, M)$ in the i^{th} cluster $(i = 1,2, \dots, N)$

Mean per element in the i^{th} cluster

$$\bar{y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$$

Mean of cluster means in the population

$$\bar{\bar{y}}_{N.} = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_{i.}$$

Mean per element in the population

$$\bar{y}_{..} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$$

Therefore, $\bar{y}_{..} = \bar{\bar{y}}_{N.}$, because $\bar{y}_{..} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$.

Now let a sample of 'n' clusters is selected out of N clusters according to simple random sampling (SRS), then we define mean of cluster means in a simple random sample of 'n' clusters

$$\bar{\bar{y}}_{n.} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_{i.}$$

Mean square between elements in the i^{th} cluster

$$S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

Mean square within cluster

$$\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2$$

Mean square between clusters means in the population

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{\bar{y}}_{N.})^2$$

and mean square between elements in the population

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{\bar{y}}_{..}\right)^2$$

6.4 Expectation of $\bar{\bar{y}}_{n.}$ or Unbiased Estimate of Population Mean

Suppose we draw a sample of size n from the population having N clusters and there are M elements in each cluster,
According to SRS

$$\bar{\bar{y}}_{n.} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$$

Therefore, $E(\bar{\bar{y}}_{n.}) = E\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right\} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i.} = \bar{\bar{y}}_{N.} = \bar{y}_{..}$ = population mean
Hence, $\bar{\bar{y}}_{n.}$ is an unbiased estimate of population mean.

6.5 Variance of the Estimate of Population Mean

Let a sample of 'n' clusters is drawn from the population having N clusters and there are M elements in each cluster, According to SRS

$$\bar{\bar{y}}_{n.} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}$$

Therefore, $V(\bar{\bar{y}}_{n.}) = V\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.}\right)$
In SRS $V\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$, where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$
when we replace y_i to $\bar{y}_{i.}$, then let us have

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{\bar{y}}_{N.})^2$$

Therefore, $V(\bar{\bar{y}}_{n.}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2$.
To compare or to get relative efficiency, if we consider that a simple random sample of nM units are selected from NM units, therefore

$$V(\bar{y}_{nM}) = \left(\frac{1}{nM} - \frac{1}{NM}\right)S^2,$$

where $S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{N.}\right)^2$
Then relative efficiency (R.E.) is given by

$$R.E. = \frac{V(\bar{y}_{nM})}{V(\bar{\bar{y}}_{n.})} = \frac{S^2}{MS_b^2}.$$

This shows that $R.E.$ is to be equal to the ratio of the overall mean square between elements to that between clusters in the population.
For large N, an estimate of the relative efficiency of cluster sampling can be given by

$$Est.R.E.(e) = \frac{s^2}{Ms_b^2},$$

where, $Ms_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{\bar{y}}_{n.})^2$,

$\bar{s}_w^2 = \frac{1}{n(M-1)}\sum_{i}^{n}\sum_{j=1}^{M}\left(y_{ij} - \bar{y}_{i.}\right)^2$ and $s^2 = Ms_b^2 + \bar{s}_w^2$.

6.6 Comparison of SRS and Cluster Sampling in Terms of Intra-Class Correlation Coefficient

Let the intra-class correlation coefficient between elements of a cluster is ρ, which is given by

$$\rho = \frac{E(y_{ij}-\bar{\bar{y}}_{N.})(y_{ik}-\bar{\bar{y}}_{N.})}{E(y_{ij}-\bar{\bar{y}}_{N.})^2} \quad (3.6.1)$$

Now, $E(y_{ij}-\bar{\bar{y}}_{N.})^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{y}}_{N.})^2 = \frac{NM-1}{NM}S^2$

and $E(y_{ij}-\bar{\bar{y}}_{N.})(y_{ik}-\bar{\bar{y}}_{N.}) = E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} + E(\bar{y}_{i.}-\bar{\bar{y}}_{N.})^2$ (3.6.2)

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} = E\left[\sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\frac{1}{M(M-1)}\right] \quad (3.6.3)$$

Since $\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.}) = 0$ (3.6.4)

Therefore,

$$\left[\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})\right]^2 = \sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 + \sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})$$

or $$0 = \sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2 + \sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})$$

$$\Rightarrow \sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.}) = -\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2$$

or $$\frac{1}{M(M-1)}\sum_{j\neq k=1}^{M}(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.}) = \frac{1}{M(M-1)}\left[-\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2\right]$$

$$= -\frac{1}{M}\left[\frac{1}{M-1}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i.})^2\right] = -\frac{S_i^2}{M} \quad (3.6.5)$$

Therefore, from (3.6.3) and (3.6.5), we have

$$E\{(y_{ij}-\bar{y}_{i.})(y_{ik}-\bar{y}_{i.})\} = E\left[-\frac{S_i^2}{M}\right] = -\frac{1}{M}E(S_i^2)$$

$$= -\frac{1}{M}\frac{1}{N}\sum_{i=1}^{N}S_i^2 = -\frac{\bar{S}_w^2}{M} \quad (3.6.6)$$

and $$E(y_{ij}-\bar{\bar{y}}_{N.})^2 = \frac{1}{N}\sum_{i=1}^{N}(y_{ij}-\bar{\bar{y}}_{N.})^2$$

$$= \frac{N-1}{N}\cdot\frac{1}{N-1}\sum_{i=1}^{N}(y_{ij}-\bar{\bar{y}}_{N.})^2 = \frac{N-1}{N}S_b^2 \quad (3.6.7)$$

From (3.6.2), (3.6.6) and (3.6.7), we get

$$E(y_{ij}-\bar{\bar{y}}_{N.})(y_{ik}-\bar{\bar{y}}_{N.}) = -\frac{\bar{S}_w^2}{M} + \frac{N-1}{N}S_b^2 \quad (3.6.8)$$

and $$E(y_{ij}-\bar{\bar{y}}_{N.})^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{y}}_{N.})^2$$

$$= \frac{NM-1}{NM}\cdot\frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{y}}_{N.})^2$$

$$= \frac{NM-1}{NM}S^2 \quad (3.6.9)$$

Substituting the values from (3.6.8) and (3.6.9) in (3.6.1), we get

$$\rho = \frac{\frac{N-1}{N}S_b^2 - \frac{\bar{S}_w^2}{M}}{\frac{NM-1}{NM}S^2} \quad (3.6.10)$$

Now,

$$(NM-1)S^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{\bar{y}}_{N.}\right)^2$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{\bar{y}}_{N.}\right)^2$$
$$= (M-1)\sum_{i=1}^{N}S_i^2 + M\left(y_{ij}-\bar{\bar{y}}_{N.}\right)^2$$
$$= N(M-1)\bar{S}_w^2 + M(N-1)S_b^2 \qquad (3.6.11)$$

But from (3.6.10), we have

$$\rho(NM-1)S^2 = M(N-1)S_b^2 - N\bar{S}_w^2 \qquad (3.6.12)$$

or $$\rho(M-1)(NM-1)S^2 = M(M-1)(N-1)S_b^2 - N(M-1)\bar{S}_w^2 \qquad (3.6.13)$$

Adding (3.6.11) and (3.6.13), we get

$$(NM-1)S^2\{1+\rho(M-1)\} = M^2(N-1)S_b^2$$

or $$S_b^2 = \frac{NM-1}{M^2(N-1)}S^2\{1+\rho(M-1)\} \qquad (3.6.14)$$

Subtract (3.6.12) from (3.6.11), we get

$$\bar{S}_w^2 = \frac{NM-1}{NM}S^2(1-\rho) \qquad (3.6.15)$$

For comparing simple random sampling, we may consider population of size NM elements and a sample of equivalent of size nM elements is drawn

Therefore, variance of the estimate of population mean

$$V(\bar{y}_{nM}) = \left(\frac{1}{nM}-\frac{1}{NM}\right)S^2,$$

where $S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}-\bar{y}_{..}\right)^2$.

Therefore, $$V(\bar{y}_{NM}) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{S^2}{M} \qquad (3.6.16)$$

and in cluster sampling, variance of the estimate of the population mean

$$V(\bar{\bar{y}}_{n.}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$
$$V(\bar{\bar{y}}_{n.}) = \left(\frac{1}{n}-\frac{1}{N}\right)\frac{NM-1}{M^2(N-1)}\{1+\rho(M-1)\} \qquad (3.6.17)$$

Therefore, Relative efficiency $(R.E.) = \frac{V(\bar{y}_{nM})}{V(\bar{\bar{y}}_{n.})} = \frac{1}{\frac{(NM-1)N}{NM(N-1)}\{1+\rho(M-1)\}}$

But for large value of N, we can consider

$$\frac{NM-1}{NM}.\frac{N}{N-1} \simeq 1$$

So, $R.E.$ of cluster sampling in terms of intra-class correlation ρ is given by

$$R.E. \simeq \frac{1}{1+(M-1)\rho} = \{1+\rho(M-1)\}^{-1} \qquad (3.6.18)$$

From (3.6.18), we observe that

If $M = 1$ or $\rho = 0$, in both the cases SRS and cluster sampling are equally efficient.

If $\rho = -\frac{1}{M-1}$ then $R.E. \to \infty, V(\bar{\bar{y}}_{n.}) = 0$.

i.e. $R.E.$ increases as ρ decreases.

For the large N, an estimate of intra-class correlation $(\hat{\rho})$ can be given by

$$\hat{\rho} = \frac{(1-e)}{(M-1)e}$$

Since multi-stage sampling is not given in this section, so readers may extend this sampling technique to the multi-stage sampling by referring the text books whose references are given in the end of this study material.

Example 3.1: For studying the cultivation practice and yield of apple, a sample surveys was conducted in a district of H P. The yields (kg) of 15 clusters each of 4 trees selected at random out of 412 bearing trees in a village are as follows:

Cluster	Tree			
	1	2	3	4
1	5.53	4.84	0.69	15.79
2	26.11	10.93	19.08	11.18
3	11.08	0.65	4.21	7.56
4	12.66	32.52	16.92	37.02
5	0.87	3.56	4.81	57.54
6	6.40	11.68	40.05	5.15
7	54.21	34.63	52.55	37.96
8	1.94	35.97	29.54	25.98
9	37.94	47.07	16.94	28.11
10	56.92	17.69	26.24	6.77
11	27.59	38.10	24.76	6.53
12	45.98	5.17	1.17	6.53
13	7.13	34.35	12.18	9.86
14	14.23	16.89	28.93	21.70
15	3.53	40.76	5.15	1.25

(i) Estimate the average yield per tree as well as the production of apple in the village and their standard error.

(ii) Estimate the efficiency of cluster sampling as compared to the simple random sampling.

(iii) Estimate the intra-class (intra-cluster) correlation coefficient between trees within cluster.

Solution: We have from this data $M = 4,\ N = 103,\ NM = 412$ and $n = 15$.

The calculations of cluster means and their variances are given in the Table below:

Cluster	Mean $(\bar{y}_{i.})$	$\sum_i y_{ij}^2$	$My_{i.}^2$	s_i^2
1	6.71	303.8067	180.0954	41.237
2	14.58	1027.7958	854.3056	59.163
3	5.88	198.0666	138.2976	19.923

4	24.78	2874.5927	2456.1936	139.466
5	7.20	495.0185	338.5600	152.157
6	15.81	1807.5993	999.8244	269.258
7	44.84	834.4251	8042.5024	99.264
8	23.36	2845.1765	2182.7584	220.806
9	33.19	4830.9302	4406.3044	148.542
10	26.91	4287.1930	2896.5924	463.533
11	23.08	2828.4957	2130.7456	232.533
12	14.71	2184.8911	865.5364	439.787
13	15.88	1476.3314	1008.6976	155.878
14	20.44	1795.5999	1676.1744	41.475
15	12.67	1701.9235	642.1156	353.269
Total	292.04			2829.341

Let y_{ij} be the yield of the j^{th} tree $(j = 1,2,\dots,M)$ in the i^{th} cluster $(i = 1,2,\dots,N)$

(i) An estimate of the average yield per tree of apple in the village is given by

$$\bar{\bar{y}}_{n.} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{i.} = \frac{292.04}{15} = 19.47$$

and estimated variance of $\bar{\bar{y}}_{n.}$ is given by

$$Est.V(\bar{\bar{y}}_{n.}) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_{i.} - \bar{\bar{y}}_{n})^2$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left(\sum_{i=1}^{n}\bar{y}_{i.}^2 - n\,\bar{\bar{y}}_{n}^2\right)$$
$$= \left(\frac{1}{15} - \frac{1}{103}\right)\frac{1}{14}(7202.4262 - 15 \times 379.0809) = 6.1686$$

Thus, the estimate of standard error of $\bar{\bar{y}}_{n.}$ is given by

$$Est.S.E.(\bar{\bar{y}}_{n.}) = \sqrt{Est.V(\bar{\bar{y}}_{n.})} = \sqrt{6.1686} = 2.48$$

To find the estimate of the efficiency of cluster sampling over simple random sampling, the analysis of variance table is required, which is as follows:

Source of variation	d. f.	Mean square
Between clusters	14	108.3 $(= Ms_b^2)$
Within clusters	397	188.6 $(=\bar{s}_w^2)$
Total	411	249.3$(= \hat{S}^2)$

(ii) The estimate of the relative efficiency of cluster sampling as compared to the simple random sampling is given by

$$R.E. = \frac{249.32}{4 \times 108.3} = 0.556$$

(iii) The estimate of the intra-class correlation coefficient between trees within cluster is given by

$$\hat{\rho} = \frac{(1-e)}{(M-1)e} = \frac{(1-0.576)}{(4-1)\times 0.576} = 0.245$$

6.7 Cluster Sampling with Varying Size of Clusters (Unequal Cluster Sampling): Estimators of Mean and their Variances

In the previous section, we have discussed the case of cluster sampling with equal size of clusters. But mostly in many practical circumstances, cluster sizes vary. Here, estimator of Mean and its Variances under Cluster Sampling with Varying Size of Clusters has been explained.

Let us consider a population of N clusters in which i^{th} cluster consist with M_i elements; $(i = 1, 2, \dots, N)$ and $\sum_i^N M_i = M_0$. The population mean per element $\bar{Y}$ is defined by-

$$\bar{Y} = \frac{1}{\sum_i^N M_i}\left\{\sum_i^N \sum_j^{M_i} y_{ij}\right\} = \frac{1}{M_0}\{\sum_i^N M_i \bar{y}_{i.}\},$$

where $\bar{y}_{i.} = \frac{1}{M_i}\sum_j^{M_i} y_{ij}$ is the mean per element of the i^{th} cluster.

The pooled mean of the cluster means is defined as

$$\bar{Y}_N = \frac{1}{N}\sum_i^N \bar{y}_{i.}$$

Let a random sample of size n clusters have been drawn by simple random sampling without replacement method all elements of the clusters surveyed. The three different estimators of $\bar{Y}$ may be considered as follows:

(i) $\bar{y}_n = \sum_i^n \frac{\bar{y}_{i.}}{n}$ (3.7.1)

(ii) $\bar{y}_n' = \frac{\sum_i^n M_i \bar{y}_{i.}}{\sum_i^n M_i}$ (3.7.2)

and (iii) $\bar{y}_n^* = \frac{N}{nM_0}\sum_i^n M_i \bar{y}_{i.} = \frac{\sum_i^n M_i \bar{y}_{i.}}{n\,\bar{M}}$ (3.7.3)

where $\bar{M} = \frac{\sum_i^n M_i}{N} = \frac{M_0}{N}$.

For (i): The expectation of $\bar{y}_n$ is given by

$$E(\bar{y}_n) = E\left[\frac{1}{n}\sum_i^n \bar{y}_{i.}\right] = \frac{1}{n}\sum_i^n E(\bar{y}_{i.})$$
$$= \sum_i^N \frac{\bar{y}_{i.}}{N} = \bar{Y}_N \neq \bar{Y}$$

Therefore, $\bar{y}_n$ is a not an unbiased estimator of the population mean $\bar{Y}$. The bias of the estimator is given by

$$B(\bar{y}_n) = E(\bar{y}_n) - \bar{Y} = \frac{\sum_i^N \bar{y}_{i.}}{N} - \frac{\sum_i^N M_i \bar{y}_{i.}}{N\,\bar{M}}$$
$$= \frac{\sum_i^N (\bar{M} - M_i)\bar{y}_{i.}}{N\,\bar{M}} = -\frac{Cov(\bar{y}_{i.}, M_i)}{\bar{M}}.$$

For such type of population in which M_i's do not considerably vary from one cluster to another, the bias may not be materially significant. For uncorrelated M_i and $\bar{y}_{i.}$, the bias will be zero and then $\bar{y}_n$ would an unbiased estimator of $\bar{Y}$. So, if M_i and $\bar{y}_{i.}$ are not highly correlated then $\bar{y}_n$ may be advisable to use.

The sampling variance of $\bar{y}_n$ is given by

$$V(\bar{y}_n) = E[\bar{y}_n - E(\bar{y}_n)]^2 = E(\bar{y}_n - \bar{Y}_N)^2$$
$$= \frac{(1-f)}{n} \frac{\sum_i^N (\bar{y}_{i.} - \bar{Y}_N)^2}{N-1} = \frac{(1-f)}{n} S_b^2.$$

An unbiased estimator of $V(\bar{y}_n)$ is given by

$$Est.[V(\bar{y}_n)] = \frac{(1-f)}{n} s_b^2, \qquad (3.7.4)$$

where $s_b^2 = \frac{\sum_i^N (\bar{y}_{i.} - \bar{y}_n)^2}{n-1}$ and $f = \frac{n}{N}$.

For (ii): As ratio estimator is a biased estimator so if we replace x_i by M_i and y_i by $M_i \bar{y}_{i.}$, it can be seen that the weighted mean of the cluster means $\bar{y}_n'$ is a biased estimator but consistent. Its sampling variance up to the first order of approximation is given by

$$V(\bar{y}_n') = \frac{(1-f)}{n} {S_b'}^2, \qquad (3.7.5)$$

where ${S_b'}^2 = \frac{\sum_i^N M_i^2 (\bar{y}_{i.} - \bar{Y})^2}{\bar{M}^2 (N-1)}$.

An unbiased estimator of $V(\bar{y}_n')$ is given by

$$Est.[V(\bar{y}_n')] = \frac{(1-f)}{n} {s_b'}^2$$

(3.7.6)

where ${S_b'}^2 = \frac{\sum_i^N M_i^2 (\bar{y}_{i.} - \bar{Y})^2}{\bar{M}'^2 (N-1)}$

and $\bar{M}' = \frac{\sum_i^n M_i}{N}$.

For (ii): The expectation of the estimator $\bar{y}_n^* = \frac{\sum_i^n M_i \bar{y}_{i.}}{n \bar{M}}$ is given as

$$E(\bar{y}_n^*) = E\left(\frac{\sum_i^n M_i \bar{y}_{i.}}{n \bar{M}}\right) = \frac{\sum_i^n E(M_i \bar{y}_{i.})}{n \bar{M}} = \bar{Y}$$

Therefore, $\bar{y}_n^*$ is an unbiased estimator.

The sampling variance of the estimator is given by

$$V(\bar{y}_n^*) = \frac{(1-f)}{n} \frac{1}{N-1} \sum_i^N \left(\frac{M_i}{\bar{M}} \bar{y}_{i.} - \bar{Y}\right)^2$$
$$= \frac{(1-f)}{n} {S_b^*}^2$$

where ${S_b^*}^2 = \frac{1}{N-1} \sum_i^N \left(\frac{M_i}{\bar{M}} \bar{y}_{i.} - \bar{Y}\right)^2$.

It is pertinent to be mention that the estimator $\bar{y}_n^*$ will often be less precise. This happens because the variance depends upon the variation of the product $M_i \bar{y}_{i.}$ and is likely to be larger than $\bar{y}_n$ unless $\bar{y}_{i.}$ and M_i vary in such a way that their product is almost constant.

An unbiased estimator of $V(\bar{y}_n^*)$ is given by

$$Est.[V(\bar{y}_n^*)] = \frac{(1-f)}{n} {s_b^*}^2, \qquad (3.7.7)$$

where, ${s_b^*}^2 = \frac{1}{n-1} \sum_i^N \left(\frac{M_i}{\bar{M}} \bar{y}_{i.} - \bar{y}_n^*\right)^2$.

6.8 Sub Sampling: Two Stage Sampling

The sampling procedure which consists in first selecting the groups of elements (say- clusters) and then selecting a specified number of elements from each selected groups is known as sub-sampling or two stage sampling. In each sampling design, clusters/groups which form the units of sampling at the first stage are called the first stage units $(f.s.u.)$ or primary sampling units $(p.s.u.)$ and the elements within clusters are called second stage units $(s.s.u.)$.

The main advantage of this sampling procedure is that the frame of $f.s.u.$ is required only at the first stage, which can be easily prepared. At the second stage, the frame of $s.s.u's$ is required only for the selected $f.s.u.'s$. This design is very flexible as it allows the use of different sampling procedures at different stages. This sampling design may be advisable in a number of practical situations where a satisfactory sampling frame of ultimate stage unit is not readily available and the cost of obtaining such a frame is too large.

Notations:

Let us assume that the population consists of NM elements grouped into $'N'$ $f.s.u.'s$ and $'M'$ $s.s.u.'s$ each. Let $'n'$ be the number of first stage units in the sample and $'m'$ be the number of second stage units to be selected from each sampled first stage unit. Here, equal cluster-equal first stage unit case has been considered. Also we assume that the units at each stage are selected with equal probability.

Let

y_{ij} −value of the characteristic under study on j^{th} second-stage unit from the i^{th} firststage unit $(j = 1,2,\dots,M;\ i = 1,2,\dots,N)$,

$\bar{y}_{i.} = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ −mean per second stage unit in the i^{th} first-stage unit $(i = 1,2,\dots,N)$,

$\bar{y}_{..} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$ −mean per second stage unit in the population,

$\bar{y}_{im} = \frac{1}{m}\sum_{j}^{m} y_{ij}$ −mean per secondstage unit of the ith first stage unitin the sample

and

$\bar{y}_{nm} = \frac{1}{nm}\sum_{i}^{n}\sum_{j}^{m} y_{ij}\frac{1}{n}\sum_{i}^{n}\bar{y}_{im}$ −mean per second stage unit in the sample.

$$= \frac{1}{n}\sum_{i}^{n}\frac{1}{m}\sum_{j}^{m} y_{ij}$$

$$= \frac{1}{n}\sum_{i}^{n}\bar{y}_{im}$$

To obtain the expectation of sample mean $\bar{y}_{nm}$, consider

$$\bar{y}_{nm} = \frac{1}{n}\sum_i^n \bar{y}_{im}$$

or $$E(\bar{y}_{nm}) = E\left(\frac{1}{n}\sum_i^n \bar{y}_{im}\right)$$

$$= E\left\{E\left(\frac{1}{n}\sum_i^n \bar{y}_{im}|i\right)\right\}$$

$$= E\left\{\frac{1}{n}\sum_i^n \bar{y}_{i.}\right\} = \bar{y}_{..}$$

Hence, $\bar{y}_{nm}$ is an unbiased estimate of population mean $(\bar{y}_{..})$.

6.8.1 Variance of Sample Mean $(\bar{y}_{nm})$

$$V(\bar{y}_{nm}) = V[E(\bar{y}_{nm}|n)] = E[V(\bar{y}_{nm}|n)]$$

$$= V\left(\frac{1}{n}\sum_i^n \bar{y}_{i.}\right) + E\left\{\frac{1}{n^2}\sum_i^n V(\bar{y}_{im}|i)\right\}$$

$$= V\left(\frac{1}{n}\sum_i^n \bar{y}_{i.}\right) + \frac{1}{n}.\frac{1}{N}\sum_{i=1}^N \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2 \quad \text{as} \quad \frac{1}{N}\sum_{i=1}^N S_i^2 = \bar{S}_w^2$$

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \quad (3.8.1)$$

Where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^N (\bar{y}_{i.} - \bar{y}_{..})^2$ and $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^N S_i^2 = \frac{1}{N}\sum_{i=1}^N \left(\frac{1}{M-1}\right)\sum_{j=1}^M (y_{ij} - \bar{y}_{i.})^2$.

The variance obtained in (3.8.1) is consists with two components, first component comes from the variability of $s.s.u.'s$ within $f.s.u.'s$ and the second one takes place from the variance of $f.s.u.'s$. If the selected $f.s.u.'s$ are completely enumerate or in other words if $m = M$, the variance of the sample mean will be given by the first component only and the situation has been given in cluster sampling. If $n = N$ or in other words, every $f.s.u.$ in the population is included in the sample, then this case corresponds to stratified random sampling with $f.s.u.'s$ as strata and a simple random sample of m $s.s.u.'s$ is drawn from each of the strata. The three following cases for approximation point of view are as follows:

Case-(i)- When $N >> n$, so we can take $\frac{N-n}{N} \simeq 1$, then

$$V(\bar{y}_{nm}) = \frac{S_b^2}{n} + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \quad (3.8.2)$$

Case-(ii)- When $M >> m$, so we can take $\frac{M-m}{M} \simeq 1$, Hence in this case

$$V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{\bar{S}_w^2}{nm} \quad (3.8.3)$$

Case-(iii)- When finite multipliers at both stages are considered as unity. i. e. $\frac{N-n}{N} \simeq 1$ and $\frac{M-m}{M} \simeq 1$, then

$$V(\bar{y}_{nm}) = \frac{S_b^2}{n} + \frac{\bar{S}_w^2}{nm} \quad (3.8.4)$$

To obtained the unbiased estimate of $V(\bar{y}_{nm})$, suppose s_b^2 denote the mean square between first stage units means in the sample which is defined as

$$s_b^2 = \frac{1}{n-1}\sum_i^n(\bar{y}_{im} - \bar{y}_{nm})^2 \qquad (3.8.5)$$

and let s_i^2 is the mean square between second stage units drawn from the ith first stage unit in the sample, therefore

$$s_i^2 = \frac{1}{m-1}\sum_j^m\left(\bar{y}_{ij} - \bar{y}_{im}\right)^2 \qquad (3.8.6)$$

Here, in this case, we know that for varying i

$$E\left(\frac{1}{n}\sum_i^n s_i^2\right) = \frac{1}{N}\sum_{i=1}^N S_i^2 = \bar{S}_w^2$$

and for fixed i; $E(s_i^2) = S_i^2$.

So, $E\left(\frac{1}{n}\sum_i^n s_i^2\right) = \bar{S}_w^2$.

Thus, an unbiased estimate of $\bar{S}_w^2$ is given by

$Est.\bar{S}_w^2 = \bar{s}_w^2$, where $\bar{s}_w^2 = \frac{1}{n}\sum_i^n s_i^2$.

and an unbiased estimate of S_b^2 is given by

$$Est.S_b^2 = s_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right)\bar{s}_w^2 .$$

When we compare the efficiency of the estimator of mean obtained by two-stage sampling with SRS (one stage sampling), we assume

population size- $=NM$, sample size- nm

and $S^2 = \frac{1}{NM-1}\sum_{i=1}^N\sum_{j=1}^M\left(y_{ij} - \bar{y}_{..}\right)^2$

In simple random sampling and two stage sampling, the variance of $\bar{y}_{nm}$ are respectively given by

$$V(\bar{y}_{nm})_R = \left(\frac{1}{nm} - \frac{1}{NM}\right)S^2 = \frac{NM-nm}{nmNM}S^2$$

and in two stage sampling

$$V(\bar{y}_{nm})_T = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2$$

Since $S_b^2 = \frac{NM-1}{M(N-1)}\frac{S^2}{M}\{1 + (M-1)\rho\}$

and $\bar{S}_w^2 = \frac{NM-1}{NM}S^2(1-\rho)$, where ρ is the intra class correlation.

Therefore, the relative efficiency (R.E.) of two stage is given by

$$R.E. = \frac{V(\bar{y}_{nm})_R}{V(\bar{y}_{nm})_T}$$

$$= \frac{1}{1+\rho\left\{\left(\frac{N-n}{N-1}\right)m-1\right\}}$$

On the basis of above result, it is concluded that as ρ increases, relative efficiency of two stage sampling compared to simple random sampling will decrease and R.E. will also decrease when sample size of second stage unit m will increase.

6.8.2 Allocation of Sample Size in Two Stages Sampling

Let C −cost of survey proportional to the size of sample, i.e. $C \propto n.m$, Then

Total cost $C = c.n.m,$ (3.8.7)

where c −constant

When cost is fixed- Let C_0 be the given fixed cost, then

$C_0 = c.n.m$, where c is a constant.

so $m = \frac{C_0}{nc}$ (3.8.8)

Thus $V(\bar{y}_{nm}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2$

$= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \left(\frac{nc}{C_0} - \frac{1}{M}\right) \frac{\bar{S}_w^2}{n}$

$= \frac{1}{n}\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) - \frac{S_b^2}{N} + \frac{\bar{S}_w^2}{C_0}.c$ (3.8.9)

which is monotonic decreasing function of n, if $S_b^2 - \frac{\bar{S}_w^2}{M} > 0$, reaching its minimum value when n assumes the maximum value, namely $\hat{n} = {C_0}/{c}$, corresponding to $\hat{m} = 1$

If $S_b^2 - \frac{\bar{S}_w^2}{M} < 0$, which for large N is equivalent to stating that the intra-class correlation is negative, then the R.H.S. of (3.8.9) becomes a monotonic increasing function of $'n'$ is minimum, given by $\hat{n} = \frac{C_0}{cM}$ or in other words, when there is no sub-sampling.

6.9 Multistage Sampling: Three-Stage Sampling-Equal First-Stage and Second-Stage Units: Sample Mean and its Variance

The generalization of two stage sampling in to three of more stages is termed as multi-stage sampling. For example- in crop surveys for estimating yield of a crop in a district, a block may be considered as a primary sampling unit, the villages the second stage units, the crop fields the third stage units and a plot of fixed size the ultimate unit of sampling. To understand multi-stage sampling, three-stage sampling with equal first-stage and second-stage units has been considered in this section

Let

$N =$ the number of first-stage units in the population,

$M =$ the number of second-stage units in each of N first-stage units,

$L =$ the number of third-stage units in each of NM second-stage units in the population,

and n, m and l the corresponding values in the sample.

Further, let y_{ijk} be the value obtained for the $k-th$ third stage unit in the $j-th$ second stage unit of the $i-th$ first stage unit. The relevant population means per element are as follows:

$\bar{y}_{ij.} = \frac{1}{L}\sum_{k=1}^{L} y_{ijk}$ (3.9.1)

$$\bar{y}_{i..} = \frac{1}{ML}\sum_{j=1}^{M}\sum_{k=1}^{L} y_{ijk} \quad (3.9.2)$$

$$\bar{y}_{...} = \frac{1}{NML}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{L} y_{ijk} \quad (3.9.3)$$

and $\bar{y}_{ijl}$, $\bar{y}_{iml}$, $\bar{y}_{nml}$ denote the corresponding values for the sample. Here it has been assumed that the units at each stage are selected with simple random sampling without replacement.

Now it can be easily shown as in two-stage sampling that the sample mean $\bar{y}_{nmp}$ provides an unbiased estimate of the population mean $\bar{y}_{...}$. Hence

$$E(\bar{y}_{nml}) = E\left\{\frac{1}{n}\sum_{i}^{n} E\left(\bar{y}_{imp}|i\right)\right\} \quad (3.9.4)$$

Since mp is a sample two-stage sample from the $i-th$ first stage unit, therefore

$$E\left(\bar{y}_{nmp}\right) = E\left\{\frac{1}{n}\sum_{i}^{n}\bar{y}_{i..}\right\} = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_{i..} = \bar{y}_{...} \quad (3.9.5)$$

To obtain the variance of $\bar{y}_{nmp}$, consider

$$\begin{aligned} V(\bar{y}_{nml}) &= V\left(\frac{1}{n}\sum_{i}^{n}\bar{y}_{iml}\right) \\ &= V\left[E\left\{\frac{1}{n}\sum_{i}^{n}(\bar{y}_{iml}|n)\right\}\right] + E\left[V\left\{\frac{1}{n}\sum_{i}^{n}(\bar{y}_{iml}|n)\right\}\right] \\ &= V\left[\frac{1}{n}\sum_{i}^{n}\bar{y}_{i..}\right] + E\left[\frac{1}{n^2}\sum_{i}^{n}V(\bar{y}_{iml}|i)\right] \\ = V\left[\frac{1}{n}\sum_{i}^{n}\bar{y}_{i..}\right] &+ E\left[\frac{1}{n^2}\sum_{i}^{n}\left\{\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \frac{1}{m}\left(\frac{1}{l}-\frac{1}{L}\right).\frac{1}{M}\sum_{j=1}^{M}S_{ij}^2\right\}\right] \end{aligned} \quad (3.9.6)$$

where $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(\bar{y}_{ij.} - \bar{y}_{i..}\right)^2$

and $S_{ij}^2 = \frac{1}{L-1}\sum_{k=1}^{L}\left(\bar{y}_{ijk} - \bar{y}_{ij.}\right)^2$ (3.9.7)

By taking expectations over samples of size n, we get

$$\begin{aligned} V(\bar{y}_{nml}) &= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{nN}\sum_{i=1}^{N}\left\{\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \frac{1}{m}\left(\frac{1}{l}-\frac{1}{L}\right).\frac{1}{M}\sum_{j=1}^{M}S_{ij}^2\right\} \\ &= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{1}{nm}\left(\frac{1}{l}-\frac{1}{L}\right)\bar{\bar{S}}_l^2 \end{aligned} \quad (3.9.8)$$

where

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_{i..} - \bar{y}_{...})^2, \quad (3.9.9)$$

$$\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2 \quad \text{and} \quad \bar{\bar{S}}_l^2 = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}S_{ij}^2 \quad (3.9.10)$$

It can be seen that the variance of the sample mean is comprised with three components of three stages of sampling. If each of the nm selected second-stage units are completely enumerated, i.e. if $l = L$, the variance would be obtained by the first two terms similar as two-stage sampling design. If each of the n first-stage units in the sample are completely enumerated, i.e. if $m = M$ and $l = L$, we will have the first term only representing the variance similar to one-stage sampling.

When $n = N$ and from each of the N units a two-stage sample is drawn, we will have the second and third terms to represent the variance of the mean and the case will correspond to a stratified two-stage sampling design with the first-stage units in the population constituting the strata.

Finally, when finite population corrections are ignored, we get a simple expression for the variance as

$$V(\bar{y}_{nmp}) = \frac{S_b^2}{n} + \frac{\bar{S}_w^2}{nm} + \frac{\bar{\bar{S}}_l^2}{nml} \qquad (3.9.11)$$

6.9 Summery

This unit provides a brief idea about cluster sampling with equal as well as varying size of clusters. The unbiased estimate of population mean and its variance under this sampling design is obtained. A comparison of simple random sampling and Cluster sampling in terms of intra-class correlation coefficient is described. Sub-sampling and its extension to multi stage sampling are described with its properties and applications. For further study and exercise, readers are required to study the referred text books of sampling techniques.

6.10 Check yourself (Questions)

1. Explain cluster sampling with its advantages and disadvantages. Obtained an unbiased estimate of population mean and its standard error in cluster sampling?
2. If NM elements in a population are grouped at random to form N clusters of M elements each, show that a random sample selected through without replacement of n clusters would have the same efficiency as sampling nM elements in random sample using without replacement scheme.
3. Describe two stage sampling. Obtained an unbiased estimate of population mean and its variance up-to first degree of approximation.
4. Explain multi stage sampling and its advantages in estimation of parameters. Suggest an unbiased estimator for estimating population mean under three stage sampling. Also obtained its variance

6.11 Further Readings

- Cochran, W. G. (1977). Sampling Techniques, New York: Wiley.
- Mukhopadhyay, P (2009). Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- Singh, D. and Chaudhary, F. S. (1986). Theory and Analysis of Sample Surveys Designs, Wiley Eastern Ltd., New Age International Ltd.
- Sukhatme, P. V. and Sukhatme, B. V. (1997). Sampling Theory of Surveys with Applications, The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

Unit-7 Response and Non Response Sampling

Structure

7.1 Introduction

While conducting any type of survey, researchers often face the problem of response and non response bias and these biases must be avoid to get significant and precious results. Despite of many efforts by the researchers to maximize response in the surveys, there is another big issue of non-response.

Non-response is an observable fact that may affect the quality of the survey results. It occurs when the people who are selected as eligible for the sample do not provide the requested information, or when the provided information is not usable. Non response causes estimators of the population parameters to be biased. This occurs when specific groups are over or under represented and these groups may behave differently with respect to the surveys variables. Non-response is mainly due to noncontact, refusal to answer and not being able to answer. It is important to distinguish among these different causes because they may have different impacts on estimates.

Response can be related to all the data obtained either directly from respondents or from administrative data. This extensive definition of response

is crucial to reflect the increased use of different collection strategies in the same survey. Further it is pertinent to mention that administrative data is also not exempted from either partial or total non-response as simply happens in survey data.

Therefore, in this unit response and non-response sampling has been introduced with different methodologies and strategies to handle the problem happened from the incompleteness of data due to non-response.

7.2 Objectives

After this unit, the learner should able to understand about:

- Non-sampling errors with non-response technique
- Randomized response techniques to collect information on sensitive issues and Warner's model to estimate the proportion
- Rank set sampling and controlled sampling

7.3 Non-Sampling Errors

Non-sampling errors are mainly associated to data collection and processing procedures. Non-sampling error can occur due to many reasons, such as

- the complexity of survey processes,
- inconsistencies in procedures,
- lack of understanding of issues by staff,
- poor questionnaire design leading to respondent misinterpretation,
- poor systems leading to processing errors and data adjustments, or even user misunderstanding, etc.

These problems can occur at one or more places in the survey process. However, the non-sampling errors may occur even if the whole population is investigated. Non-sampling error can arise from both observation and non-observation errors. Non-observation errors are the errors made when the intended measurements are not obtained.

Under coverage occurs when elements of the target population do not have a corresponding entry in the sampling frame. The population members cannot ever be contacted. Another type of non-observation error is non-response. This phenomenon occurs if the sampled person does not provide the required information. More precisely, non-response denotes the situation where a member of the target population (and thus eligible for the survey) does not submit the required information.

7.4 Randomized Response Techniques

Randomized response is a technique basically used in collection of sensitive information from individuals in such a manner that survey interviewers and the

persons who process the data don't know the answered two alternative questions that respondent has made. Data collected using option arrangement in randomized response contain errors due to the reason of untruthful answer given by the respondents who are forced to do like this. Different statistical methods exist to rectify this type of errors and obtained unbiased estimates to examine the sensitive behavior or attitude.

In other words, randomized response may be defined as a research method to assess the behavior, opinions or attitudes of respondents towards any sensitive topics like addiction of drugs or alcohol, breaking the law and order, favoring of dowry, etc. In research of such topics, the respondents often have a phenomenon to give popular answers and therefore randomized response technique is designed in that way to obtain more valid responses. It is a truly a public health oriented research technique.

To understand this mechanism with a coin, let the respondents allow to throw a coin and ask to answer the question according to the faces come up in the toss. Before the respondents answer, they are then instructed to answer 'yes' if the coin comes up tails, and truthfully, if it comes up heads. Since only the respondents can see the results of the coin, so the anonymity of the respondent is protected as nobody knows the real answer while the other half will answer truthfully according to their experience.

Let in a response of a question, half of the respondents get tails and other half get heads, then half of the respondents will answer 'yes' despite of whether they have involved in it. Hence, whatsoever proportion of the group of respondents said 'no', the factual number who did involved in that, based on the supposition that the two halves are probably close to the same since it is a large randomized sampling. For example, if 10% of the surveyed group said 'no', the factual proportion who didn't involve will be 20%.

7.5 Warner's Model

Warner (1965) introduced a randomized response technique to draw the information related to delicate questions in a survey and suggested estimation procedure for the proportion π of the population belongs to the sensitive category$A$. This technique is based on a spinner with two outcomes A or A^c with probabilities p and $\bar{p} = 1 - p$ respectively to each respondent. The respondent spins the spinner unseen by the survey interviewer and answers 'yes' according to the characteristic indicated by the pointer and no otherwise. Therefore the probability of a 'yes' response (say λ) is given by

$$\begin{aligned}\lambda &= p\pi + \bar{p}(1-\pi)\\ &= \bar{p} + (2p-1)\pi.\end{aligned}$$

Let the number of truthful 'yes' responses among a total of n interviews is r then Warner stated that the maximum likelihood estimator $(m.l.e.)$ of λ is $\hat{\lambda} = {}^{r}/_{n}$ and accordingly π is given by

$$\hat{\pi} = \frac{\left[\frac{r}{n} - (1-p)\right]}{(2p-1)}$$

provided $p \neq 0.5$.

Suppose that a sample of size n has been drawn with or without replacement method from a large enough and results in r 'Yes' answers. Let us assume without any loss of generality that $p > {}^{1}/_{2}$. It is pertinent to be mention that λ is restricted to be in the interval $(\bar{p},\ p)$, since π is in $(0, 1)$. Therefore the $m.l.e.$ of λ is

$$\tilde{\lambda} = \begin{cases} {}^{r}/_{n}, & if\ \bar{p} < {}^{r}/_{n} < p \\ p, & if\ \bar{p} \geq {}^{r}/_{n} \\ p, & if\ {}^{r}/_{n} \geq p \end{cases}$$

Since the sampling process is Bernoulli in X and let $b(x: n, \lambda)$ be the binomial mass function of r, therefore we can write

$$E(\tilde{\lambda}) = \lambda + B$$

where the bias B is

$$B = \frac{1}{n}\left[\sum_{x<n\bar{p}}(n\bar{p} - x)b(x: n, \lambda) + \sum_{x>np}(np - x)b(x: n, \lambda)\right].$$

As the values of the first and the second summations come up positive and negative respectively, therefore the overall magnitude and the sign of B depends upon the values of the parameters n, p, and π. For fixed values of n and p, it can be obtained that the bias evaluated at $\pi$is equal in magnitude to that at $(1 - \pi)$ but opposite in sign. Further the magnitude of bias for specific choices of the parameters has been found to be very small after computing and hence it can be verified that λ is consistent.

To show the efficiency of $\tilde{\lambda}$ over $\hat{\lambda}$ in terms of squared error loss, consider

$$E(\tilde{\lambda} - \lambda)^2 - E(\hat{\lambda} - \lambda)^2 = \sum_{x\leq n\bar{p}}\left\{(\bar{p} - \lambda)^2 - \left(\frac{x}{n} - \lambda\right)^2\right\} b(x: n, \lambda) + \sum_{x\geq np}\left\{(p - \lambda)^2 - \left(\frac{x}{n} - \lambda\right)^2\right\} b(x: n, \lambda)$$

It can be easily seen that ${}^{x}/_{n} \leq \bar{p}$ in the first summation this implies ${}^{x}/_{n} - \lambda \leq \bar{p} - \theta < 0$, therefore $(\bar{p} - \lambda)^2 - ({}^{x}/_{n} - \lambda)^2 < 0$. It proves the value of the first summation to be negative. The second summation is also negative by the similar argument. Thus $\hat{\lambda}$ is not admissible.

Since π is linearly related to λ, therefore the $m.l.e.$ of π can be obtained from $\tilde{\lambda}$ as

$$\tilde{\pi} = \begin{cases} \dfrac{\left[\frac{r}{n} - (1-p)\right]}{(2p-1)}, & if\ \bar{p} < {}^{r}/_{n} < p, \quad p \neq \dfrac{1}{2} \\ 0, & if\ \bar{p} \geq {}^{r}/_{n} \\ 1, & if\ {}^{r}/_{n} \geq p \end{cases}$$

Of course, statements which are true for $\tilde{\theta}$ relative to $\hat{\theta}$ continue to be true for $\tilde{\pi}$ relative to $\hat{\pi}$. Asymptotically, on the basis of the results for $\tilde{\lambda}$ relative to $\hat{\lambda}$ both $\tilde{\pi}$ and $\hat{\pi}$ are found to be equally efficient however for small surveys $\tilde{\pi}$ is highly efficient.

In conclusion, it has been found that Warner's estimator is undeniably neither the maximum likelihood estimator nor even admissible. The maximum likelihood estimator discussed in this section is uniformly efficient in terms of squared error loss than that of Warner's estimator.

7.6 Rank Set Sampling

McIntyre (1952) first proposed a concept of sampling known as ranked set sampling (RSS) in the framework of obtaining reliable estimates of farm yields based on sampling of pastures and crop plots. He provided a clear and insightful introduction of ranked set sampling for the situations where the actual measurements for sample observations is difficult (*e.g.*, costly, destructively, time-consuming), and shown that how this easy and reliable mechanisms can lead to improved estimation relative to simple random sampling. On the contrary to the classical approaches of sampling that presume stratification of a population, it works at stratification of samples. The implementation of this *RSS* technique depends only the ranking of the randomly selected units, which does not depend upon the method employed for determining the ranks. This procedure provides unbiased and more efficient estimators as compared to simple random sampling, of various population parameters. Apart from McIntyre's method of ranked set sampling (MRSS) there are some more RSS methods like that of Takahasi and Wakimoto(1968).

7.6.1 McIntyre's RSS Method

For obtaining McIntyre's RSS first of all *m* random samples with *m* units in each sample are selected from a population with mean μ and a finite variance σ^2. This is the same as drawing m^2units randomly and then these are partitioned into *m* equal samples. The *m* units of each subset are ranked with respect to the variable of interest without using the exact measurements. For this purpose some outside information like visual perception, past experience, etc. are used. Using this ranking information the unit with the smallest rank is measured from the first subset; the unit with the second smallest rank is measured from the second subset, and this process of quantification is continued until the unit with the m^{th} rank is measured from the m^{th} subset. This yields *m* measurements with each of the first *m* ranks, and these constitute an MRSS of size *m*. For obtaining a larger sample of size *mr* the whole procedure is repeated *r* times. Here *m* is referred to as the set size while *r* is called as the number of cycles. In terms of usual notations we get the sample of size $n = mr$

from the population with its size $N \geq m^2 r$. This procedure is called as MRSS with balanced (equal) allocation because each rank order consists of size r observations. Further, one could utilize the prior information, if available, about the population for obtaining a more efficient estimator of the population parameter of interest. For a skewed (*i.e.* a non-symmetrical) population the number of quantifications (*i.e.* measurements) of the i^{th} rank r_i is taken as proportional to the standard deviation of the rank order, $\sigma_{(i:m)}$. In this case there would be unequal number of quantifications of different ranks. Hence, this allocation is termed as MRSS with unbalanced allocation (MRSSUA).

Example 4.1: For illustration of this procedure we draw a random sample of 36 units from the data set showing (in millions of kroner) earned by 40 municipalities in Sweden. The drawn sample in a set of three observations is given as follows:

24	386	74	163	37	97	134	128	96	199	290	626
422	101	536	412	250	155	249	240	196	144	249	63
1277	230	467	241	159	111	55	623	288	277	720	488

Draw an MRSS with equal allocation of size, n =12 while keeping the set size, m = 3 and the number of cycles, r = 4.

Solution: For drawing the sample, the ranking is carried out vertically (visual perception) in each column and the lowest value is taken from the first column, the value having the second rank is selected from the second column and the highest value is drawn from the third column. This process is repeated for each set of three columns and the values so selected are given below.

Rank		
1	2	3
24	230	536
163	159	155
55	240	288
144	290	626

Example 4.2: From the data of *Example1* draw an MRSS with unequal allocation with r_1 = 2, r_2 = 2 and r_3 = 8.

Solution: Here r_1 = 2, r_2 = 2 and r_3 = 8 means we should have two observations with rank 1, two with rank 2 and eight observations with rank 3. This can be achieved by taking the first two sets of three columns and finding the observations with lowest rank from the first column of the set, the second rank from the third column of each set and the observation with highest rank from the third column of each set. Then from the rest of the six columns find the observation with the highest rank. Thus the selected values are as follows.

Rank	Value							
1	24	163						
2	230	159						
3	536	155	249	623	288	277	720	626

7.6.2 McIntyre's Estimators

Let us consider only one cycle and suppose that $X_{(i:m)}$ denote the $i:m^{\text{th}}$ order statistic from the population, *i.e.* the i^{th} ranked observation of a cycle consisting of subsets of size m. The parentheses are used to surround the subscript to show that $X_{(i:m)}$ are independent unlike the usual i^{th} order statistic from a sample of size m denoted by $X_{i:m}$. Thus the mean of the ranked set sample denoted by $\bar{X}_{(m)}$ is obtained by

$$\bar{X}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(i:m)},$$

while the sample mean $\bar{X}$ based on the same number of independent and identically distributed (*iid*) quantifications is computed from

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i = \frac{1}{m}\sum_{i=1}^{m} X_{i:m}.$$

Here $X_{(i:m)}$ and $X_{i:m}$ have the same marginal distribution and hence equal variances. As the former are independent while the latter are positively correlated, it follows that $\bar{X}_{(m)}$ is more efficient than $\bar{X}$ for estimating the population mean since

$$V\left(\bar{X}_{(m)}\right) = \frac{1}{m^2}\sum_{i=1}^{m} V\left(X_{(i:m)}\right) = \frac{1}{m^2}\sum_{i=1}^{m} \sigma_{ii:m}$$

and

$$V(\bar{X}) = \frac{1}{m^2}\left[\sum_{i=1}^{m} \sigma_{ii:m} + \sum_{i=1}^{m}\sum_{\substack{j=1 \\ i\neq j}}^{m} \sigma_{ij:m}\right],$$

where $\sigma_{ii:m} = V\left(X_{(i:m)}\right) = V\left(X_{i:m}\right)$ and $\sigma_{ij:m} = Cov\left(X_{i:m}, X_{j:m}\right); i, j = 1,2,\cdots,m.$

As $X_{i:m}$ are positively correlated, $\sigma_{ij:m} > 0$ and hence, $V\left(\bar{X}_{(m)}\right) < V(\bar{X})$.

In general suppose that $X_{(i:m)j}$ denotes the i^{th} order statistic based on perfect ranking in the j^{th} cycle, for $i = 1,2,\cdots,m$ and $j = 1,2,\cdots,r$. Note that these are not *iid* in general, but for a given value of i these are so with $E\left(X_{(i:m)j}\right) = \mu_{(i:m)}$ and $V\left(X_{(i:m)j}\right) = \sigma^2_{(i:m)}$. The McIntyre's estimator $\hat{\mu}_{MRSS}$ of the population mean μ is defined as follows.

$$\hat{\mu}_{MRSS} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} X_{(i:m)j}\ . \qquad (4.6.1)$$

If $\hat{\mu}_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r} X_{(i:m)j}$ then $\hat{\mu}_{MRSS} = \frac{1}{m}\sum_{i=1}^{m} \hat{\mu}_{(i:m)}$.

Here $E\left(\hat{\mu}_{(i:m)}\right) = \mu_{(i:m)}$; $E(\hat{\mu}_{MRSS}) = \mu$ and $V\left(\hat{\mu}_{(i:m)}\right) = \sigma^2_{(i:m)}/r$. Thus, we have

$V(\hat{\mu}_{MRSS}) = V\left\{\frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_{(i:m)}\right\} = \frac{1}{m^2}V\left(\hat{\mu}_{(i:m)}\right) = \frac{1}{m^2 r}\sum_{i=1}^{m}\sigma^2_{(i:m)}$. (4.6.2)

Also, $V(\hat{\mu}_{MRSS}) = V\left(\frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r}X_{(i:m)j}\right)$

$= \frac{1}{m^2r^2}\sum_{i=1}^{m}\sum_{j=1}^{r}V(X_{(i:m)j})$

[since for any given i, $X_{(i:m)j}$ are independently distributed]

$= \frac{1}{m^2r^2}\sum_{i=1}^{m}\sum_{j=1}^{r}E[X_{(i:m)j} - E(X_{(i:m)j})]^2$

$= \frac{1}{mr}\left[\sigma^2 - \frac{1}{m}\sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2\right]$. (4.6.3)

Expression (4.6.3) is useful when the variance of the i^{th} order statistic is not available. If we denote the SRS estimator of the population mean with the same sample size $n = mr$ by $\hat{\mu}_{SRS}$, then

$$V(\hat{\mu}_{SRS}) = \frac{\sigma^2}{n} = \frac{\sigma^2}{mr}.$$

Comparing this with $V(\hat{\mu}_{MRSS})$ given by (4.6.3), we see that

$$V(\hat{\mu}_{MRSS}) < V(\hat{\mu}_{SRS}) \text{ as } \sum_{i=1}^{m}(\mu_{(i:m)} - \mu)^2 > 0.$$

Hence the estimate based on ranked set sampling is more efficient than that of a simple random sampling.

Example 4.3: From the data of *Example1* obtain the value of $\hat{\mu}_{MRSS}$, an estimate of $V(\hat{\mu}_{MRSS})$,the estimate of the population variance based on MRSS and SRS. Also obtain an estimate of *RP*.

Solution: We have the set size, $m = 3$ and the number of cycles, $r = 4$.

Cycle J	Rank i			**Total** $\sum_{i=1}^{m} X_{(i:m)j}$
	1	**2**	**3**	
1	24	230	536	790
2	163	159	155	477
3	55	240	288	583
4	144	290	626	1060
Total $\sum_{j=1}^{r} X_{(i:m)j}$	386	919	1605	2910

As $\hat{\mu}_{(i:m)} = \frac{1}{r}\sum_{j=1}^{r}X_{(i:m)j}$, $\hat{\mu}_{(1:m)} = \frac{386}{4} = 96.5$, $\hat{\mu}_{(2:m)} = \frac{919}{4} = 229.75$,

$\hat{\mu}_{(3:m)} = \frac{1605}{4} = 401.25$;

$\hat{\mu}_{MRSS} = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r}X_{(i:m)j} = \frac{1}{12} \times 2910 = 242.5$

$= \frac{1}{3}[96.5 + 229.75 + 401.25] = \frac{1}{m}\sum_{i=1}^{m}\hat{\mu}_{(i:m)}$.

$\hat{\sigma}^2_{(i:m)} = \frac{1}{r-1}\sum_{j=1}^{r}\left(X_{(i:m)j} - \hat{\mu}_{(i:m)}\right)^2$; $\hat{\sigma}^2_{(1:m)} = 4552.333$, $\hat{\sigma}^2_{(2:m)} = 2913.583$, $\hat{\sigma}^2_{(3:m)} = 47378.25$

$Est.V(\hat{\mu}_{MRSS}) = \frac{1}{m^2 r}\sum_{i=1}^{m}\hat{\sigma}^2_{(1:m)} = \frac{1}{9\times4}[4552.333 + 2913.583 + 47378.25]$

$$= \frac{54844.17}{36} = 1523.449.$$

Unbiased estimate of the population variance based on MRSS is given by

$$\hat{\sigma}^2_{MRSS} = \left\{\frac{mr-m+1}{m^2r(r-1)}\right\}\sum_{i=1}^{m}\sum_{j=1}^{r}\left(X_{(i:m)j} - \hat{\mu}_{(i:m)}\right)^2 + \left(\frac{1}{m}\right)\sum_{i=1}^{m}(\hat{\mu}_{(i:m)} - \mu_{MRSS})^2$$

$$= \frac{10}{108} \times 160110.3 + \frac{1}{3} \times 46680.13 = 30385.06.$$

Hence

$$V(\hat{\mu}_{SRS}) = \hat{\sigma}^2_{MRSS}/mr = 30385.06/12 = 2532.089.$$

An estimator of *RP*, based on unbiased estimators is given by

$$\widehat{RP} = \frac{Est.V(\hat{\mu}_{SRS})}{Est.V(\hat{\mu}_{MRSS})} = \frac{2532.089}{1523.449} = 1.662076.0$$

7.7 Controlled Sampling

The controlled sampling introduced by Goodman and Kish (1950) is a technique of sample selection that reduces the probability of selection of undesirable samples while retaining certain properties of an associated uncontrolled plan. The scope of controlled sampling is that it may include suitable distribution of sampling units over different subgroups of the population to obtain consistent estimates from each of the subgroups. In multi-variable surveys, controlled sampling is also used to increase the efficiency of the estimators for estimating the unknown parameters. Stratified sampling can be considered as one of the example of controlled sampling in such a way that that the chance of selecting a non-preferred sample has been reduced to zero. It is pertinent to be mention here that even after fully exploiting the mechanism of stratification it may still be necessary to further control the selection of the sample within each stratum. Waterton (1983) showed that controlled sampling provides more efficient estimates than multi proportionate stratified sampling. Five different approaches of controlled sampling available in the literature are (i) experimental design configurations, (ii) linear programming, (iii) nonlinear programming, (iv) nearest proportional to size design and (v) co-ordination of samples over time.

7.8 Non Sampling Errors with Non-Response Techniques

The errors occurred in collection, processing and analysis of the data in a sample survey are classified in sampling and non-sampling errors. The data collected through complete enumeration or census may be free from sampling error but would not be free from non-sampling errors. The non-sampling errors are inescapable in census as well as surveys however the data collected by sample surveys involved both sampling and non-sampling errors.

A non-sampling error may be defined as a statistical term which refers to an error that results bias during the data collection and causing the data to differ

and it has been observed that n_1 units respond and n_2 units do not respond. Here it is assumed that the whole population is divided into two unknown non overlapping strata of N_1 responding and N_2 non-responding units such that $N_1 + N_2 = N$, though the stratum weights $W_1 \left(= \frac{N_1}{N}\right)$ and $W_2 \left(= \frac{N_2}{N}\right)$ are estimated by $\widehat{W}_1 \left(= \frac{n_1}{n}\right)$ and $\widehat{W}_2 \left(= \frac{n_2}{n}\right)$ respectively. Further a subsample of size $r(= n_2/k\,, k > 1)$ from n_2 non-responding units has been drawn by using SRSWOR method and necessary information is collected by interviewed the units.

Clearly, N_1 and N_2 cannot be known and can only be estimated from the sample. We have, for unbiased estimates of N_1 and N_2,

$$\widehat{N}_1 = \frac{n_1 \mathrm{N}}{n} \quad \text{and} \quad \widehat{N}_1 = \frac{n_2 \mathrm{N}}{n} \tag{4.9.1}$$

Clearly, an unbiased estimate of the population mean is given by

$$\overline{y}^{*} = \frac{n_1}{n}\overline{y}_{(1)} + \frac{n_2}{n}\overline{y}_{(2)}^{'} \tag{4.9.2}$$

Here, $\overline{y}_{(1)}$ and $\overline{y}_{(2)}^{'}$ are the sample means based on n_1 and r units respectively.

$$E\left(\overline{y}^{*}\right) = E\left[E\left\{\overline{y}^{*} | n_1,\, n_2\right\}\right] = E\left(\frac{n_1}{n}\overline{y}_{(1)} + \frac{n_2}{n}\overline{y}_{(2)}\right)$$
$$= E(\bar{y}_n) = \bar{y}_N \tag{4.9.3}$$

where the expectation within bracket is taken over the set of values corresponding to the samples of sizes n_1 and n_2 drawn from the responding and non-responding group of population respectively.

To obtain the variance, consider

$$V\left(\overline{y}^{*}\right) = V\left[E\left\{\overline{y}^{*} | n_1,\, n_2\right\}\right] + E\left[V\left\{\overline{y}^{*} | n_1,\, n_2\right\}\right]$$

or
$$= V(\bar{y}_n) + E\left[\frac{n_2^2}{n^2}\left(\frac{1}{r} - \frac{1}{n_2}\right)s_2^2\right] \tag{4.9.4}$$

where s_2^2 is the mean square based on n_2 units. Putting $n_2 = rk$, we get

$$E\left[\frac{n_2^2}{n^2}\left(\frac{1}{r} - \frac{1}{n_2}\right)s_2^2\right] = E\left[\frac{n_2}{n^2}(k-1)s_2^2\right]$$
$$= \frac{(k-1)}{n} E\left[E\left(\frac{n_2}{n}s_2^2 | n_2\right)\right]$$
$$= \frac{(f-1)}{n} S_2^2\, E\left(\frac{n_2}{n}\right)$$
$$= \frac{(k-1)}{n}\frac{N_2}{N} S_2^2\,. \tag{4.9.5}$$

From (4.9.4) and (4.9.5) we have

$$V\left(\overline{y}^{*}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2 + \frac{W_2(k-1)}{n} S_2^2, \tag{4.9.6}$$

where S^2 and S_2^2 are the population mean square of the entire and non-response group of the population. If we assume $k = 1$, the second term of (4.9.6) will vanish and we would have the variance of a simple random sample on n units. That implies the second term is an increase in variance arising from sub-sampling r out of n_2 units.

from the true values. In general, errors which are due to the factors other than sampling are defined as non-sampling errors. A non-sampling error differs from a sampling error. A sampling error is limited to any differences between sample values and true values that arise because the sample size was limited.

Non-sampling errors refer to either random or systematic errors, and these errors can be challenging to mark in a survey, sample, or census. Systematic non-sampling errors are not as good as than random non-sampling errors because study, survey or census can have to be scrapped due to systematic errors.

Non sampling errors can take place at every phase of planning and execution of survey or census. The key sources of the non-sampling errors are lack of proper knowledge/specification of the area of variable under study and scope of the investigation, incomplete coverage of the population or sample, faulty definition, defective methods of data collection and tabulation errors.

When non-sampling errors occur, the rate of bias in a study or survey goes up. The higher the number of errors, the less reliable the information. It is pertinent to be mention here that increase in sample size may help to reduce the sampling errors but would not help to reduce non-sampling errors. This is because non-sampling errors are often difficult to detect, and it is virtually impossible to eliminate them.

7.9 Non-Response Techniques

Non-response is a phenomenon that may affect the quality of the survey outcomes. It occurs when the people who are selected as eligible for the sample do not provide the requested information, or when the provided information is not usable. Non response can cause estimators of the population characteristics to be biased. This occurs when specific groups are over or underrepresented and these groups may behave differently with respect to the surveys variables. Non-response is mainly due to noncontact, refusal to answer and not being able to answer. It is important to distinguish among these different causes because they may have different impacts on estimates. Basically there are two types of non-response. First one is called "unit non-response" in which a sampled unit that is contacted may fail to respond and the second one is known as "item non-response" where the unit may respond to the questionnaire incompletely.

Hansen and Hurwitz (1946) considered the problem of estimating the population mean under non-response by taking a sub-sample from the non-respondent group with the help of some extra efforts and suggested an unbiased estimator by combining the information available from responding and non-responding groups.

Let n be the size of the sample drawn from the population of size N by using simple random sampling without replacement (SRSWOR) method of sampling

7.10 Summery

The present unit deals with the consequences and effect of non- sampling errors in sample surveys. Warner's model for estimating the proportion has been explained under randomized response technique. Rank set sampling is given with its application and McIntyre's RSS method with estimator for mean has been explained through examples. A brief note on controlled sampling is provided in this unit and Hansen and Hurwitz (1946) technique to tackle the problem of non-response has also been explained. For further study and exercise, learners are also required to study the referred text books of sampling techniques.

7.11 Check Yourself (Questions)

1 What are non-sampling errors? How do you distinguish between sampling and non-sampling errors?

2 Explain randomized response technique and Warner's model to estimate the proportion of the population.

8 Discuss rank set sampling and McIntyre's RSS Method. Show that the estimate of population mean based on ranked set sampling is more efficient than that of a simple random sampling.

9 Give a brief note on controlled sampling and its importance.

10 For the following data. Obtain the value of $\hat{\mu}_{MRSS}$, an estimate of $V(\hat{\mu}_{MRSS})$, the estimate of the population variance based on MRSS and SRS. Also obtain an estimate of *RP*.

Cycle j	Rank i			**Total** $\sum_{i=1}^{m} X_{(i:m)j}$
	1	**2**	**3**	
1	28	234	540	802
2	167	163	159	489
3	57	242	290	589
4	146	292	628	1066
Total $\sum_{j=1}^{r} X_{(i:m)j}$	398	931	1617	2946

11 Explain the reasons and consequences of non-response in sample surveys. Describe Hansen and Hurwitz unbiased estimate of population mean to eradicate non-response and obtain its variance.

7.12 References

- Goodman, R. and Kish, L. (1950): Controlled selection- a technique in probability sampling. Jour. Amer. Statist. Assoc., 45, 350-372.
- Hansen, M.H. and Hurwitz, W. N. (1946): The problem of non-response in sample surveys, Jour. Amer. Stat. Assoc. 41, 517-529.
- McIntyre, G. A. (1952): A method for unbiased selective sampling, using ranked sets, Australian Journal of Agricultural Research, 3, 385–390.
- Takahasi, K. and Wakimoto, K. (1968): On unbiased estimates of the population mean based on the sample stratified by means of ordering, Annals of the Institute of Statistical Mathematics, 20(1), 1–31.
- Warner, S. L. (1965): Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias, Jour. Amer. Stat. Assoc., 60, 63 -69.
- Waterton, J.J. (1983): A exercise in controlled selections, Appl. Statist., 32, 150-164.

7.13 Further Readings

- Cochran, W. G. (1977). Sampling Techniques, New York: Wiley.
- Mukhopadhyay, P (2009). Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- Singh, D. and Chaudhary, F. S. (1986). Theory and Analysis of Sample Surveys Designs, Wiley Eastern Ltd., New Age International Ltd.
- Sukhatme, P. V. and Sukhatme, B. V. (1997). Sampling Theory of Surveys with Applications, The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

U.P. Rajarshi Tandon Open University, Prayagraj

PGSTAT – 103/ MASTAT – 103 Survey Sampling

Block: 3 Varying Probability Sampling

Course Design Committee

Course Preparation Committee

Block: 1 ***Random Sampling Procedures - I***

PGSTAT – 103/ MASTAT – 103 **SURVEY SAMPLING**

First Edition: ***July 2021***
ISBN : : 978-93-94487-03-1

Printed By : K.C. Printing & Allied Works, Panchwati, Mathura - 281003

Block & Units Introduction

The *Block - 3 – Varying Probability Sampling* has three units. it is consists of varying probability sampling in sixth and seventh units. The first unit of present block explains the procedure of selecting a sample and estimation of population mean, under probability proportional to size, with and without replacement. Des- Raj's estimator in ordered estimator is discussed in the second part of this unit. This block is also deals with the estimation of unordered estimators. Horwitz-Thompson Estimator in this class of unordered estimators is explained to estimate the population mean. Midzuno System and Narain Method of sampling are also given with examples.

Unit – 8 – Sampling on Probability Proportional to Size, comprises the methods of selection and related theorems.

In *Unit – 9 – Ordered Estimators*, ordered estimators with their requirements and properties are explained.

In *Unit – 10 – Unordered Estimators*, Horvitz- Thompson estimators along with Midzuno system and Narain method of sampling have been given with suitable example to understand the unordered estimators.

At the end of every block/unit the summary, self assessment questions and further readings are given.

Unit-8 Sampling on Probability Proportional to Size

Structure

8.1 Introduction

The simplest method of probability sampling is known as simple random sampling, but this method is effective when the selection probabilities for all units are equal in the population. Practically, we often find that the units vary considerably in their size, so in this case simple random sampling is not considered as an appropriate procedure, since no importance is given to the size of the unit. There are many ways of utilizing such ancillary information about the size of the units in selecting the sample to obtain more efficient estimators of the population parameters. Therefore, when units vary in size and the variate under study is highly correlated with the size of the units, the probabilities of selection may be assigned in proportion to the size of the unit. For example- in the selection of orchards having varying numbers of fruit trees, orchards are selected with probabilities proportional to the number of trees in the orchards. Such type of sampling procedure in which the units are selected

with probabilities proportional to some measure of their size is known as probability proportional to size sampling, abbreviated as *pps* sampling.

The basic difference between simple random sampling and *pps* sampling is that in simple ransom sampling, the probability of drawing any specified unit at any given drawn is same, while in *pps* sampling, it differs from draw to draw.

There are two methods of selecting a sample in probability proportional to size sampling, which are

1. probability proportional to size with replacement (ppswr)
2. probability proportional to size without replacement (ppswor)

8.2 Objectives:

After this unit, the learner should able to understand about:

- Procedure of selecting a sample under probability proportional to size with replacement (ppswr)
- Estimation of population mean and its variance under probability proportional to size with replacement
- Method of selecting a sample under probability proportional to size without replacement (ppswor)

8.3 Procedure of Selecting a Sample in Probability Proportional to Size with Replacement (ppswr):

There are two methods of selecting a sample in ppswr sampling:

(i) Cumulative Total Method

(ii) Lahiri's Method

8.3.1 Cumulative Total Method:

In case of varying probabilities, the procedures of selecting a sample, consists in association with each unit a set of consecutive natural numbers, the size of the set being proportional to the desired probability. So, if the size of the i^{th} unit be y_i $(i = 1.2, \dots, N)$, we associate the numbers 1 to y_1 with the first unit,

the numbers $(y_1 + 1)$ to $(y_1 + y_2)$ with the second unit, and so on. A number k is chosen at random from 1to $Y = \sum_{i=1}^{N} y_i$and selected the i^{th} unit in the population for which $\sum_{i=0}^{i-1} y_i < k \leq \sum_{i=0}^{i} y_i$, where y_0 is to be interpreted as zero. Clearly, the i^{th} unit in the population is being selected with a probability proportional y_i . The procedure is to be repeated n times with replacement of the unit selected, if a sample of size n is required and this procedure of selection is known as the cumulative total method.

The main drawback of this method is that it requires complete successive cumulative totals, which is time consuming, tedious and costly, especially when the population size is large.

8.3.2 Lahiri's Method:

To avoid the cumulation of the size of the units, Lahiri (1951) suggested an alternative method for selecting a sample in case of varying probabilities. The 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 size of the N units in the population. If y_i, the size of the i^{th} unit selected is greater than or equal to j then i^{th} unit is selected otherwise it is rejected. The entire procedure is repeated until a unit is finally selected. To select a sample of n units with probability proportional to size and with replacement, the procedure is to be repeated until n units are selected. It can be seen that the procedure leads to the required probabilities of selection.

The main drawback of this method is the wastage of time and efforts if units get rejected.

Example 1.1: A village has 10 orchards containing 51, 31, 46, 26, 41, 27, 45, 36, 29 and 28 trees respectively. Select a number of four orchards with replacement and with probability proportional to the number of trees in the orchard using cumulative total method and Lahiri's method.

Solution: Selection of four orchards by cumulative total method: The first step in the selection of orchards is to form cumulative totals as shown below-

S. No. of the orchards	Size (y_i)	Cumulative size	Number associated
1	51	51	1-51
2	31	82	52-82
3	46	128	83-128
4	26	154	129-154
5	41	195	155-195
6	27	222	196-222
7	45	267	223-267
8	36	303	268-303
9	29	332	304-332
10	28	360	333-360

To select a orchard, a random number not exceeding 360 is drawn with the help of random number table. Suppose this random number is 273. It can be seen from the successive cumulative totals that the number is associated with the group 268-303 of 8^{th} orchard. Similarly suppose the next three more random numbers are 347, 168 and 096 which associate the orchards 10^{th}, 5^{th} and 3^{rd} respectively. Therefore, the 8^{th}, 10^{th}, 5^{th} and 3^{rd} are the four orchards selected with probability proportional to size.

Selection of four orchards by Lahiri's method: Here we have total numbers of orchards (N) is 10 and the maximum size (M) of the orchard is 51. In this method, first we have to select a random number which is not greater than 10 and a second number which is not greater than 51. Using the random number table, let the pair is (10, 14). Hence, the 10^{th} orchard is selected in the sample. Similarly suppose the next three pairs are (4, 28), (5, 34) and (7, 25). The pair (4, 28) is rejected as 28 is greater than the size value (26), so another pair is required to drawn and let it be (8, 21). Therefore, the 10^{th}, 5^{th}, 7^{th} and 8^{th} are the four orchards selected with probability proportional to size.

8.4 Estimation of Population Mean and its Variance in Probability Proportional to Size with Replacement (ppswr)

Let $y_1, y_2, \dots, y_n$ are the values of the study character of size N. Let $p_i (i = 1,2, \dots, N)$ denotes the probability of selecting the i^{th} unit of the population at any given draw in the case of with replacement sampling. Obviously $\sum_{i=1}^{N} p_i = 1$.

Let us define a variate τ having the value $\tau_i = \frac{y_i}{Np_i}; i = 1,2, \dots, N.$ (1.4.1)

The sample arithmetic mean based on n of variate τ is given by

$$\bar{\tau}_n = \frac{1}{n}\sum^{n} \tau_i \qquad (1.4.2)$$

Since the variatesτ_i $(i = 1,2, \dots, N)$ are independently and identically distributed. Therefore

$$E(\tau_i) = \sum_{i=1}^{N} p_i \tau_i = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}_N \qquad (1.4.3)$$

or $$E(\bar{y}_{pps}) = E(\bar{\tau}_n) = \frac{1}{n}\sum^{n} E(\tau_i) = \bar{y}_N = \bar{\tau}. \qquad (1.4.4)$$

Therefore $\bar{y}_{pps}$ is an unbiased estimate of population mean $(\bar{y}_N)$.

To obtain the sampling variance of $\bar{y}_{pps}$, we have

$$\begin{aligned} V(\bar{y}_{pps}) &= V(\bar{\tau}_n) = E(\bar{\tau}_n^2) - \{E(\bar{\tau}_n)\}^2 \\ &= E\left\{\frac{1}{n}\sum^{n} \tau_i\right\}^2 - \bar{\tau}^2 \\ &= \frac{1}{n^2} E\left\{\sum^{n} \tau_i^2 + \sum_{i\neq j}^{n} \tau_i \tau_j\right\} - \bar{\tau}^2 \\ &= \frac{1}{n^2}\left\{\sum^{n} E(\tau_i^2) + \sum_{i\neq j}^{n} E(\tau_i \tau_j)\right\} - \bar{\tau}^2. \end{aligned} \qquad (1.4.5)$$

Since draws are made with replacement, so

$$E(\tau_i \tau_j) = E(\tau_i)E(\tau_j) = \bar{\tau}^2 \qquad (1.4.6)$$

and $$E(\tau_i^2) = \sum_{i=1}^{N} p_i \tau_i^2. \qquad (1.4.7)$$

From equations (1.4.5), (1.4.6) and(1.4.7), we have

$$V(\bar{\tau}_n) = \frac{1}{n}\{\sum_{i=1}^{N} p_i \tau_i^2 - \bar{\tau}^2\} = \frac{\sigma_\tau^2}{n},$$

where $$\sigma_\tau^2 = \sum_{i=1}^{N} p_i (\tau_i - \bar{\tau})^2.$$

Here it is important to notice that the finite multiplier $\left(\frac{N-n}{N}\right)$ does not play any role into the expression for the variance of the estimate if sample is based on with replacement.

Now, $\bar{\tau}_n = \frac{1}{n}\sum^n \frac{y_i}{N\,p_i}$

and $V(\bar{\tau}_n) = \frac{1}{n}\sum_{i=1}^N p_i \left(\frac{y_i}{N\,P_i} - \bar{y}_N\right)^2 = V(\bar{y}_{pps})$.

So, in case of sampling with equal probabilities i.e. $p_i = {}^1/_N$, we have

$$V(\bar{y}_{pps}) = \frac{1}{n\,N}\sum_{i=1}^N (y_i - \bar{y}_N)^2$$
$$= \frac{\sigma_y^2}{n} = \frac{N-1}{N}S^2,$$

where $S^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{y}_N)^2$.

Here it is remarkable that if the selection of unit is based on probability proportional to the value of the variate i.e. $p_i \propto y_i$, the variate τ assumes the constant value $\tau_i = \bar{y}_N \;\forall\, i$ and then σ_τ^2 will become zero in this case. Since in practice, the values of the variate is not known in advance, but the values of another variate which is correlated with the study variate may be known. Therefore, using the p_i proportional to some measure of size of y_i, we may get significantly more efficient estimate than that based on simple random sampling.

To estimate of sampling variance, let us consider the sample mean square, defined as

$$s_\tau^2 = \frac{1}{n-1}\sum^n (\tau_i - \bar{\tau}_n)^2$$
$$= \frac{1}{n-1}\{\sum^n \tau_i^2 - n\bar{\tau}_n^2\} \qquad (1.4.8)$$

Taking expectation, we have

$$E(s_\tau^2) = \frac{1}{n-1}\{\sum^n E(\tau_i^2) - nE(\bar{\tau}_n^2)\} \qquad (1.4.9)$$

But, by definition $V(\bar{\tau}_n) = E(\bar{\tau}_n^2) - \bar{\tau}^2$

or $$E(\bar{\tau}_n^2) = V(\bar{\tau}_n) + \bar{\tau}^2 = \frac{\sigma_\tau^2}{n} + \bar{\tau}^2 \qquad (1.4.10)$$

Substituting (1.4.7) and (1.4.10) in (1.4.9), we have

$$E(s_\tau^2) = \frac{1}{n-1}\left[\sum^n \{\sum_{i=1}^N p_i\,\tau_i^2\} - n\left\{\frac{\sigma_\tau^2}{n} + \bar{\tau}^2\right\}\right]$$
$$= \frac{1}{n-1}[\sum^n \{\sum_{i=1}^N p_i\,\tau_i^2 - \bar{\tau}^2\} - \sigma_\tau^2]$$
$$= \frac{1}{n-1}[n\sigma_\tau^2 - \sigma_\tau^2]$$
$$= \sigma_\tau^2 \qquad (1.4.11)$$

Hence, s_τ^2 is an unbiased estimate of σ_τ^2. Therefore an unbiased estimate of $V(\bar{\tau}_n)$ is given by

$$Est.V(\bar{\tau}_n) = \frac{s_\tau^2}{n}$$
$$= \frac{1}{n(n-1)}\sum^n \left(\frac{y_i}{N\,p_i} - \frac{y_n}{N}\right)^2$$
$$= \frac{1}{n(n-1)}\frac{1}{N^2}\left\{\sum^n \left(\frac{y_i}{p_i}\right)^2 - n\,y_n^2\right\}$$

where $y_n = N\,\bar{y}_n$ is an unbiased estimate of population total $N\,\bar{y}_N$.

Example 1.2: The following table shows the cultivated area (a_i) and area under rice (y_i) for a sample of 25 villages of a tehsil. The sample was selected with replacement and with probability proportional to cultivated area. The total cultivated area in the tehsil was 5,68,565 acres.

S. No. of village	1	2	3	4	5	6	7	8	9	10	11	12	13
Total cultivated area of the village (a_i) in acres	1,232	327	1,346	1,285	428	871	1,042	1,262	497	1,016	651	1,170	2,630
Area under rice (y_i) in acres	688	231	768	898	417	697	785	1,190	338	745	392	1,055	2,400

S. No. of village	14	15	16	17	18	19	20	21	22	23	24	25
Total cultivated area of the village (a_i) in acres	515	895	1,055	2,110	979	671	120	541	1,331	842	162	206
Area under rice (y_i) in acres	330	810	1,026	1,666	929	565	101	516	1,036	568	137	107

Estimate the area under rice in the tehsil with its standard error$(S.E.)$. Give an estimate of the variance of the estimated area under rice if the villages were selected with equal probability and without replacement.

Solution: An unbiased estimate of the total area under rice is given by

$$y_N = N\bar{\tau}_n = N\left(\frac{1}{n}\sum^n \frac{y_i}{N\,p_i}\right) = \frac{1}{n}\sum^n \frac{y_i}{p_i}$$

We know that $p_i \propto a_i \Rightarrow p_i = k\,a_i \Rightarrow k = \sum p_i / \sum a_i = 1/\sum a_i$,

After calculating k and using this value of k we find $\sum^n \frac{y_i}{p_i} = 11{,}112 \times 10^3$

Now $\hat{y}_N = \frac{1}{n}\sum^n \frac{y_i}{p_i} = \frac{11{,}112\times10^3}{25} = 4{,}44{,}480$ acres

An unbiased estimate of the variance of the estimate $\hat{y}_N$ is obtained from

$$Est.V(\hat{y}_N) = \frac{1}{n(n-1)}\left\{\sum^n \left(\frac{y_i}{p_i}\right)^2 - n\,\hat{y}_N^2\right\}$$

$$= \frac{1}{25\times24}[50{,}908 \times 10^8 - 25 \times (4{,}44{,}480)^2]$$

$$= 253 \times 10^6 \text{ (acres)}^2$$

Therefore, $S.E.(\hat{y}_N) = 15{,}900$ acres

and $\%\, S.E.(\hat{y}_N) = \frac{15{,}900}{4{,}44{,}480} \times 100 = 3.58$

To estimate the variance of the estimated total area under rice if the villages were selected with equal probability and without replacement, we know that

$$V(N\,\bar{y}_n) = N^2\left(\frac{N-n}{Nn}\right)S^2$$

$$= \frac{N(N-n)}{n(N-1)}\left[\sum_{i=1}^{N} y_i^2 - N\bar{y}_N^2\right]$$

An unbiased estimate of $\sum_{i=1}^{N} y_i^2$ is given by $\frac{1}{n}\sum^{n} \frac{y_i^2}{p_i}$ and unbiased estimate of $N^2\bar{y}_N^2$ is obtained from $[\hat{y}_N^2 - Est.V(\hat{y}_N)]$. Therefore substituting these estimates, we have

$$Est.V(N\,\bar{y}_n) = \frac{N(N-n)}{n(N-1)}\left[\frac{1}{n}\sum^{n} \frac{y_i^2}{p_i} - \frac{1}{N}\{\hat{y}_N^2 - Est.V(\hat{y}_N)\}\right]$$

$$= \frac{892\times867}{25\times891}\left[\frac{8{,}52{,}548\times10^4}{25} - \frac{1}{892}\{(4{,}44{,}480)^2 - 253\times10^6\}\right]$$

$$= 4{,}16{,}010 \times 10^4 \text{ (acres)}^2.$$

From the above result, it can be easily seen that the estimated variance in the case of simple random sampling is very much larger than in sampling with probability proportional to size.

An estimate of the percent gain in efficiency due to the latter procedure is given by

$$\frac{Est.V(N\,\bar{y}_n) - Est.V(\hat{y}_N)}{Est.V(\hat{y}_N)} \times 100 = 1{,}540.$$

The larger gain in efficiency is due to the high correlation between the cultivated area and the area under rice.

8.5 Procedure of Selecting a Sample in Probability Proportional to Size without Replacement (ppswor):

There are several procedures for selecting a sample with unequal probabilities without replacement. Here, we have discussed the general selection procedure of selecting a sample with probability proportional to size without replacement (ppswor) method.

In this procedure, we first select a *pps* sample of size unity and remove the selected unit from the population. From the remaining units, another *pps* sample of size one is taken as before and remove this selected unit from the population. This process is repeated until 'n' selections are made.

Suppose 'n' units are selected one by one, with probability proportional to size at each draw, without replacing the units selected in the previous draw.

Now, the probability of selecting the i^{th} unit at the first draw is given by

$$p_i = \frac{y_i}{y}, \quad i = 1,2, \dots \dots N.$$

Where$y = \sum_{i=1}^{N} y_i$ and the probability of selecting the j^{th} unit at the second draw, when i^{th} unit is selected at the first draw and it has been replaced from the population, is given by

$$p_{j|i} = p_j/(1 - p_i); \quad i \neq j$$

and so on. Therefore, in this way of sampling, we have an ordered set of sample value$(y_i; i = 1,2, \dots ., n)$ with probabilities $(p_i; i = 1,2, \dots ., n)$.

To show that the expected value of the variate under ppswor changes with successive draws, let $p_{i_r}$is the probability of selecting y_i at the r^{th} draw$(r = 1,2, \dots, n)$. Since in without replacement sampling, we assume that at any subsequent draw, the probability of selecting a unit among the available units at that draw is proportional to the probability of selecting it at the first draw.

i.e. $p_{i_1} = p_i \quad (i = 1,2, \dots, N)$ (1.5.1)

and $p_{i_2} = P\{y_i \text{is not selected at the first draw}\}$

$\times P\{y_i \text{ is selected at the second draw } | \, y_i \text{ is not selected at the first draw }\}$

$= \sum_{j(\neq i)=1}^{N} P\{y_i \text{is not selected at the first draw}\}$

$\times$ P$\{y_i \text{ is selected at the second draw } | \, y_i \text{ is not selected at the first draw}$

$= \sum_{j(\neq i)=1}^{N} \frac{p_j p_i}{1-p_j}$

$= \left[\sum_{j=1}^{N} \frac{p_j}{1-p_j} - \frac{p_i}{1-p_i}\right] p_i$

$= \left[S - \frac{p_i}{1-p_i}\right] p_i$, where $S = \sum_{j=1}^{N} \frac{p_j}{1-p_j}$. (1.5.2)

Hence, it means $p_{i_2} \neq p_{i_1} \; \forall \; i = 1,2, \dots, N$ unless $p_i = 1/N$ and it follows that the expected value of the variate under this case will change with successive

draws. It makes the theory of sampling with varying probabilities and without replacement necessarily complex and not easily applicable in practices.

Example 1.3: In a village, there are orchards with 50, 30, 25, 40, 26, 44, 20 and 35 trees, respectively. Select a sample of 2 orchards with probability proportional to the number of trees in the orchard and without replacement.

Solution:

(i) For selecting a unit by probability proportional to the number of trees, using Lahiri's method of selection, consider the following arrangement:

Orchard number	1	2	3	4	5	6	7	8
Number of trees	50	30	25	40	26	44	20	35

Selecting a pair of random numbers $(i, j)/(i \leq 8,\ j \leq 50)$, using the random number table, we get the pair (5,17). Since the number of trees X_i for orchard number 5 is greater than the second number (17) of the selected random pair, the 5th orchard is selected in the sample.

(ii) For selecting the second unit by probability proportional to the number of trees in the orchard, we prepare the following arrangement once again after deleting the 5th orchard:

Orchard number	1	2	3	4	5	6	7
Number of trees	50	30	25	40	44	20	35

As in (i), a pair of random numbers $(i, j)/(i \leq 7, j \leq 50)$ has to selected, using the random number table. Referring to the table of random numbers, the pair selected is (6,18). As the size of the 6th unit in the above arrangement is greater than the second number of the random pair selected, orchard 6 is selected into the sample. Thus, the sample selected consists of the units at serial numbers 5 and 7 of the original list with the number of trees being 26 and 20, respectively. However, we know that sampling without replacement is more efficient, than sampling with replacement. This rule also applies to *pps* sampling. So we find that there are enumerable sampling procedures for selecting a sample with varying probabilities without replacement. Since the probability of inclusion changes by draws or the order of selected units, therefore we have discussed ordered estimators in unit 2 and unordered estimators in unit 3.

8.6 Summery

This unit provides a brief idea about varying probability sampling. In this unit, procedures of selecting a sample in probability proportional to size with replacement and without replacement are explained. Estimation of population mean and its variance under probability proportional to size with replacement are given. Suitable examples are given to understand the methods and properties.

For further study and exercise, readers are required to study the referred text books of sampling techniques.

8.7 Check Your Self (Questions)

1. Describe the requirements and applications of PPS sampling. Explain the cumulative total and Lahiri's methods to select a sample under probability proportional to size with replacement.

2. Suggest unbiased estimators for estimating the population mean under the strategy of probability proportional to size with replacement and hence obtain its variance.

3 Let $y_1, y_2, \ldots, y_n$ and $P_1, P_2, \ldots, P_n$ be the values of the units in the order in which they are drawn and their initial probabilities of selection. Define

$$z_1 = y_1/NP_1$$

$$z_i = \frac{y_i(1-P_1)(1-P_1-P_2)\ldots\ldots(1-P_1-P_2-\cdots P_{i-1})}{N(N-1)(N-2)\ldots\ldots(N-i+1)P_1P_2\ldots\ldots P_i} (i = 2, \ldots, n)$$

Show that $\bar{z}_n = \frac{1}{n}\sum^n z_i$ is an unbiased estimate of the population mean.

Further, show that $E\left\{\frac{1}{nN}\sum^n y_i z_i + \frac{2(N-1)}{n(n-1)}\sum_{i<j}^n y_i z_i\right\} = \bar{y}_n^2$

Hence, or otherwise, obtain an unbiased estimate of the variance of $\bar{z}_n$.

4 A sample survey was conducted to study the yield of a crop in a State. A sample of 20 villages from a total of 100 was taken, with probability

proportional to area under the crop, with replacement method. The total area under the crop was 484.5 hectares. The area under crop(a_i) and yield (y_i) were respectively noted in hectares and quintals per hectares, which are as follows:

S. No. of village	1	2	3	4	5	6	7	8	9	10
Area under crop (a_i)	4.8	4.1	1.3	5.2	6.9	6.0	2.0	6.3	5.2	4.2
Yield of crop(y_i)	22	19	6	25	54	43	4	40	28	29

S. No. of village	11	12	13	14	15	16	17	18	19	20
Area under crop (a_i)	4.8	5.9	5.8	5.8	5.1	4.7	5.6	5.2	4.0	4.6
Yield of crop (y_i)	22	39	39	44	30	27	34	31	18	31

(i) Estimate the average yield per village along with its standard error for this sample.

(ii) Estimate the gain in efficiency due to pps sampling compared to simple random sampling with replacement.

8.8 References

- Lahiri, D. B. (1951) "A method of sample selection providing unbiased ratio estimates", *Bull. Inter. Stat. Inst.*, 33(2), 133-140.

8.9 Further Readings

- **Cochran, W. G. (1977)**. Sampling Techniques. New York: Wiley.
- **Mukhopadhyay, P (2009).**Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- **Singh, D. and Chaudhary, F. S. (1986).** Wiley Eastern Ltd. New Age International Ltd.
- **Sukhatme, P. V. and Sukhatme, B. V. (1997).** Sampling Theory of Surveys with Applications. The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

Unit-9 Ordered Estimators

Structure

9.1 Introduction

The estimators based on the order of units selected in the sample and do not require inclusion of probabilities are known as ordered estimators. The present unit describes the problem of estimation of population mean using ordered estimator suggested by Des Raj (1956) along with its properties.

9.2 Objectives

When learners go through this unit, they will able to understand about:

- The concept of ordered estimators
- Estimation of population mean using Des- Raj's ordered estimators

9.3 Concept of Ordered Estimators

In the previous few years of the sampling, the problem of selecting a sample with varying probabilities has received considerable attention. Estimates based on the strategies which take into account the order of the draw are called ordered estimates. Das (1951) and Des Raj (1956) have suggested some

ordered estimators which are based on the order of units selected in the sample and do not require inclusion of probabilities. In this unit, Des Raj's ordered estimators for the first two draws have been described and then generalization has been made.

9.4 Des- Raj's Ordered Estimators

In this section, we have considered the estimators proposed by Des Raj (1956) first for the case when $n = 2$ and then generalize the result.

(i) Case of Two Draws

Denote by y_1 and y_2 the values of the unit drawn at the first and second draw respectively, it being understood that y_1 and y_2 are not necessarily the values of the first two units in the population. Further, let p_1 and $p_2$are their probabilities of selection. Let

$$\tau_1 = \frac{y_1}{Np_1} \qquad (2.4.1)$$

and $$\tau_2 = \frac{1}{N}\left\{y_1 + y_2\frac{(1-p_1)}{p_2}\right\} \qquad (2.4.2)$$

Then, we have $$E(\tau_1) = \sum_{i=1}^{N}\frac{y_i}{Np_i}.p_i = \bar{y}_N \qquad (2.4.3)$$

and similarly $$E\left\{\frac{y_2(1-p_1)}{p_2}|y_1\right\} = \sum\frac{y_j(1-p_1)}{p_j}.\frac{p_j}{1-p_1} \qquad (2.4.4)$$

where the summation is taken over all the values of y except the value y_1 drawn at the first draw. It is easy simple to find

$$E\left\{\frac{y_2(1-p_1)}{p_2}|y_1\right\} = N\bar{y}_N - y_1 \qquad (2.4.5)$$

From the equations (2.4.2)and (2.4.5), we have

$$E\{\tau_2|y_1\} = \bar{y}_N \qquad (2.4.6)$$

Hence, we have $$E\{\tau_2\} = \bar{y}_N \qquad (2.4.7)$$

Therefore, the combined estimate $\bar{\tau}_{(n=2)}$ based on both the draws is given by

$$\bar{\tau}_{(n=2)} = \frac{\tau_1+\tau_2}{2} = \frac{1}{2N}\left\{(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right\} \qquad (2.4.8)$$

is an unbiased estimate of the population mean.

The variance of the estimate is given by

$$V\left(\bar{\tau}_{(n=2)}\right) = E\left[\frac{1}{2N}\left\{(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right\}\right]^2 - \bar{y}_N^2$$

$$= \frac{1}{4N^2}\sum_{i\neq j=1}^{N}\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}^2 \frac{p_i p_j}{1-p_i} - \bar{y}_N^2 \quad (2.4.9)$$

After some algebraic simplification and calculation, we obtain

$$V\left(\bar{\tau}_{(n=2)}\right) = \frac{1}{4N^2}\left\{\left(2 - \sum_{i=1}^{N} p_i^2\right)\sum_{i=1}^{N}\frac{y_i^2}{p_i} - \sum_{i=1}^{N} y_i^2 + 2\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} p_i y_i\right)\right\} - \frac{\bar{y}_N^2}{2} \quad (2.4.10)$$

To obtain an unbiased estimate of the variance, we have usually

$$Est.\left\{V\left(\bar{\tau}_{(n=2)}\right)\right\} = \left\{\bar{\tau}_{(n=2)}\right\}^2 - Est.[\bar{y}_N^2] \quad (2.4.11)$$

where the last term is an unbiased estimate of $\bar{y}_N^2$. Now, we have from the equations (2.4.11) and (2.4.6),

$$E\{\tau_1\tau_2\} = E[\tau_1\{E(\tau_2|y_1)\}]$$
$$= E\{\tau_1\bar{y}_N\} \quad = \bar{y}_N^2$$

and it follows that

$$Est.\left\{V\left(\bar{\tau}_{(n=2)}\right)\right\} = \left\{\bar{\tau}_{(n=2)}\right\}^2 - \tau_1\tau_2$$
$$= (\tau_1 - \tau_2)^2/4$$
$$= \frac{(1-p_1)^2}{4N^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2 \quad (2.4.12)$$

(ii) The General Case

Let $y_1, y_2, \ldots, y_n$ and $p_1, p_2, \ldots, p_n$ be the values of the units in the order in which they are drawn and their initial probabilities of selection.

Define τ_1 as in equation (2.4.1)

$$\tau_1 = \frac{1}{N}\left\{y_1 + y_2 + \cdots + y_{i-1} + y_i\frac{(1-p_1-p_2-\cdots-p_{i-1})}{p_i}\right\} \ (i = 2, \ldots, n) \quad (2.4.13)$$

Then it can written as before $E\{\tau_1\} = \bar{y}_N$ and

$$E\{\tau_1|y_1, y_2, \ldots, y_{i-1}\} = \bar{y}_N \ (i = 2, \ldots, n) \quad (2.4.14)$$

$$E(\tau_i) = \bar{y}_N \ (i = 2, \ldots, n) \quad (2.4.15)$$

It follows that

$$\bar{\tau}_n = \frac{1}{n}\sum^{n} \tau_i \qquad (2.4.16)$$

is an unbiased estimate of the population mean $\bar{y}_N$. Again, we observe that for i less than j,

$$E(\tau_i\tau_j) = E[\tau_i\{E(\tau_j|y_1, y_2, \ldots., y_{j-1})\}]$$
$$= E\{\tau_i\bar{y}_N\} = \bar{y}_N^2 \qquad (2.4.17)$$

Therefore, we have

$$E\left\{\frac{\sum_{i\neq j}^{n}\tau_i\tau_j}{n(n-1)}\right\} = \bar{y}_N^2 \qquad (2.4.18)$$

Since, by symmetry, (2.4.17)holds also for $i > j$.

Hence, an unbiased estimate of $\bar{y}_N^2$ is given by

$$Est.[\bar{y}_N^2] = \frac{1}{n(n-1)}\sum_{i\neq j}^{N}\tau_i\tau_j \qquad (2.4.19)$$

Although the expression for the variance $V(\bar{\tau}_n)$ is very much complex to write out explicitly, so an unbiased estimate of the variance takes a simple form as follows:

$$Est.\{V(\bar{\tau}_n)\} = \bar{\tau}_n^2 - Est.[\bar{y}_N^2]$$
$$= \bar{\tau}_n^2 - \frac{1}{n(n-1)}\sum_{i\neq j}^{N}\tau_i\tau_j$$
$$= \frac{1}{n(n-1)}\sum^{n}(\tau_i - \tau_j)^2 \qquad (2.4.20)$$

Example 2.1: Estimate the total production of the 8 orchards along with the standard error $(S.E.)$ using Des Raj ordered estimator, if the yields (in 10 kg) of the 8 orchards given in Example 1.3 are 60, 35, 30, 44, 30, 50, 22 and 40 respectively.

Solution: To estimate the total production and its standard error, the following table can be prepared from Example 1.3 as follows:

S. No.	1	2	3	4	**5**	6	**7**	8	Total
No. of trees (X_i)	50	30	25	40	**26**	44	**20**	35	290
Yield (Y_i)	60	35	30	44	**30**	50	**22**	40	311
$p_i = (X_i/X)$	0.185	0.111	0.093	0.148	**0.096**	0.163	**0.074**	0.130	1.000

Since the sample selected in the Example 1.3 includes the orchards at serial no. 5 and 7, the corresponding yield (in 10 kg) of two selected orchards are 30 and 22.

Hence we have $x_1 = 26, x_2 = 20, y_1 = 30, \quad y_2 = 22, \quad p_1 = 0.096$ and $p_2 = 0.074$.

The Des Raj ordered estimator for total production of orchards is given by the equation (1.19)

$$(\hat{y}_N)_{DR} = N(\bar{\tau}_{(n=2)}) = \frac{1}{2}\left\{(1+p_1)\frac{y_1}{p_1} + (1-p_1)\frac{y_2}{p_2}\right\}$$

$$= \frac{1}{2}\left\{(1+0.096)\frac{30}{0.096} + (1-0.096)\frac{22}{0.074}\right\}$$

$$= 305.625 \text{ (in 10 kg units)}$$

The estimate of $V[(\hat{y}_N)_{DR}]$ is given by the equation (1.23)

$$Est.V[(\hat{y}_N)_{DR}] = N^2[Est.\{V(\bar{\tau}_{(n=2)})\}]$$

$$= \frac{(1-p_1)^2}{4}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2$$

$$= \frac{(1-0.096)^2}{4}\left(\frac{30}{0.096} - \frac{22}{0.074}\right)^2 = 47.2645$$

and hence $S.E.[(\hat{y}_N)_{DR}] = \sqrt{47.2645} = 6.87.$

9.5 Summery

This unit provides an idea about ordered estimation. Estimation of population mean and its variance under probability proportional to size with replacement using Des Raj estimator are explained. Suitable examples and questions in exercise on these units are given for practice.

For further study and exercise, readers are required to study the referred text books of sampling techniques.

9.10 Check Your Self (Questions)

1. Define Des Raj's ordered estimator for population mean on the basis of a sample of size 2 and show that it is unbiased.
2. Select a sample of 2 villages from the data given below with probability proportional to the area under crop without replacement. On the basis of

these units selected in the sample, estimate total yield of 20 villages along with the standard error by using Des Raj ordered estimator.

S. No. of village	1	2	3	4	5	6	7	8	9	10
Area under crop (a_i)	4.8	4.1	1.3	5.2	6.9	6.0	2.0	6.3	5.2	4.2
Yield of crop(y_i)	22	19	6	25	54	43	4	40	28	29

S. No. of village	11	12	13	14	15	16	17	18	19	20
Area under crop (a_i)	4.8	5.9	5.8	5.8	5.1	4.7	5.6	5.2	4.0	4.6
Yield of crop (y_i)	22	39	39	44	30	27	34	31	18	31

9.11 References

- Das, A. C. (1951) "On two phase sampling and sampling with varying probabilities", *Bull. Inter. Stat. Inst.*, **33**, 105-112.
- Des Raj (1956) "Some estimators in sampling with varying probabilities", *J. Amer. Stat. Assoc.*, **61**, 384-390.

9.12 Further Readings

- Cochran, W. G. (1977). Sampling Techniques. New York: Wiley.
- Mukhopadhyay, P (2009).Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- Singh, D. and Chaudhary, F. S. (1986). Wiley Eastern Ltd. New Age International Ltd.
- Sukhatme, P. V. and Sukhatme, B. V. (1997). Sampling Theory of Surveys with Applications. The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.

Unit-10 Unordered Estimators

Structure

10.1 Introduction

In ordered estimator, the order in which the units are drawn is considered. Corresponding to any ordered estimator, Murthy (1957) and Basu (1958) have shown that there exist unordered estimators which do not depend on the order in which the units are drawn. In this unit, Horvitz- Thompson estimator along with its properties for estimating the population mean have been discussed. Two important methods Midzuno system and Narain method of sampling are explained.

10.2 Objectives

When learners go through this unit, they will able to understand about:

- The concept of unordered estimators
- Horvitz- Thompson estimators for estimating population mean
- Midzuno System of sampling and Narain Method of Sampling

10.3 Horvitz- Thompson Estimators

It has been found from the theory of the unordered estimates that such estimates have limited applicability as they lack simplicity and the expressions for the estimators and their variances becomes unmanageable when sample size is even moderately large. Horvitz and Thompson (1952) suggested an unbiased estimator for population mean, which is simpler than other estimators.

Let us consider a finite population of N units and let y_i be the value of the characteristic under study for the i^{t} unit in the population $(i = 1,2, \dots, N)$. Suppose that a sample of size 'n' is drawn without replacement, using arbitrary probabilities of selection at each draw. Thus, prior to each succeeding draw there is defined a new probability distribution at each draw may or may not depend upon the initial probabilities at the first draw.

Define a random variable $\xi_i (i = 1,2, \dots, N)$ as follows:

$$\xi_i = 1 \text{ if } y_i \text{ is included in a sample 's' of size '}n\text{'}$$

$$= 0 \quad \text{otherwise.} \qquad (3.3.1)$$

and let $$\tau_i = \frac{ny_i}{N\,E(\xi_i)} = \frac{ny_i}{N\,\pi_i}, \qquad (i = 1,2, \dots, N) \qquad (3.3.2)$$

where we assume that every unit has a positive probability of being included in the sample i.e. $E(\xi_i) > 0$ for all i.

Therefore $E(\xi_i) = 1.P(y_i \in s) + 0.P(y_i \notin s) = \pi_i$

Is the probability of including the unit 'i' in the sample and is called as **inclusion probability**.

Now, we shall show that the simple arithmetic mean of τ_i provides an unbiased estimate of the population mean $\bar{Y}_N$.

The simple arithmetic mean of τ_i is given by

$$\bar{\tau}_n = \frac{1}{n}\sum^{n} \tau_i = \frac{1}{n}\sum_{i=1}^{N} \xi_i\,\tau_i \qquad (3.3.3)$$

Hence $$E(\bar{\tau}_n) = \frac{1}{n}\sum_{i=1}^{N} \tau_i . E(\xi_i) = \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{NE(\xi_i)} . E(\xi_i) = \bar{Y}_N \qquad (3.3.4)$$

i.e. $\bar{\tau}_n$ is an unbiased estimator of population mean $\bar{Y}_N$.

To obtain sampling variance of $\bar{\tau}_n$, we have

$$V(\bar{\tau}_n) = E\left(\bar{\tau}_n{}^2\right) - \bar{Y}_N{}^2 \qquad (3.3.5)$$

Now, using equation (3.3.3), we see that

$$E\left(\bar{\tau}_n{}^2\right) = \frac{1}{n^2} E\left(\sum_{i=1}^{N} \xi_i\,\tau_i\right)^2$$

$$= \frac{1}{n^2} E\left\{\sum_{i=1}^{N} \xi_i^2 \tau_i^2 + \sum_{i\neq j=1}^{N} \xi_i \xi_j \tau_i \tau_j\right\}$$

$$= \frac{1}{n^2} \left\{\sum_{i=1}^{N} E(\xi_i)\tau_i^2 + \sum_{i\neq j=1}^{N} E(\xi_i \xi_j)\tau_i \tau_j\right\}. \quad (3.3.6)$$

Since $E(\xi_i^2) = E(\tau_i)$. Further from equation (3.3.4)

$$\bar{Y}_N{}^2 = \frac{1}{n^2} [\sum_{i=1}^{N} \tau_i E(\xi_i)]^2$$

$$= \frac{1}{n^2} \left[\sum_{i=1}^{N} \{E(\xi_i)\}^2 \tau_i^2 + \sum_{i\neq j=1}^{N} E(\xi_i) E(\xi_j) \tau_i \tau_j\right] \quad (3.3.7)$$

From equations (3.3.5), (3.3.6)and (3.3.7), we have

$$V(\bar{\tau}_n) = \frac{1}{n^2} \left[\sum_{i=1}^{N} E(\xi_i)\{1 - E(\xi_i)\}\tau_i^2 + \sum_{i\neq j=1}^{N} \{E(\xi_i \xi_j) - E(\xi_i)E(\xi_j)\tau_i \tau_j\}\right] \quad (3.3.8)$$

The expression (3.3.8)can be expressed in terms of y as

$$V(\bar{\tau}_n) = \frac{1}{N^2} \left[\sum_{i=1}^{N} \left\{\frac{1-E(\xi_i)}{E(\xi_i)}\right\} y_i^2 + \sum_{i\neq j=1}^{N} \left\{\frac{E(\xi_i \xi_j) - E(\xi_i)E(\xi_j)}{E(\xi_i)E(\xi_j)}\right\} y_i y_j\right] \quad (3.3.9)$$

An unbiased estimate of this variance can be given by

$$Est.V(\bar{\tau}_n) = \frac{1}{N^2} \left[\sum^{n} \frac{1-E(\xi_i)}{[E(\xi_i)]^2} y_i^2 + \sum_{i\neq j}^{N} \left\{\frac{E(\xi_i \xi_j) - E(\xi_i)E(\xi_j)}{E(\xi_i \xi_j)\, E(\xi_i)E(\xi_j)}\right\} y_i y_j\right] \quad (3.3.10)$$

It is important to mention here the one drawback of this estimate that it does not reduce to zero even all the τ_i are equal(a case for which the variance is necessarily zero). Consequently, this estimate may assume negative values for more samples.

A more elegant expression for the variance of the estimate is given by Yates and Grundy (1953). This can be obtained as follows.

Let there are exactly n values of ξ_i which are 1 and $(N - n)$ values which are zero,

$$\sum_{i=1}^{N} \xi_i = n \quad (3.3.11)$$

Taking expectations on both sides, we have

$$\sum_{i=1}^{N} E(\xi_i) = n \quad (3.3.12)$$

Squaring equation (3.3.11), taking expectations, we obtain by using (3.3.12) and the fact that $E(\xi_i^2) = E(\tau_i)$ as

$$\sum_{i\neq j=1}^{N} E(\xi_i \xi_j) = n(n-1) \quad (3.3.13)$$

Also, $E(\xi_i \xi_j) = P(\xi_i = 1, \xi_j = 1)$

$$= P(\xi_i = 1).P(\xi_j = 1|\xi_i = 1)$$

$$= E(\xi_i)E(\xi_j|\xi_i = 1).$$

Therefore

$$\sum_{j(\neq i)=1}^{N}\left[E\left(\xi_i\xi_j\right)-E(\xi_i)E(\xi_j)\right]=E(\xi_i)\sum_{j(\neq i)=1}^{N}\left[E\left(\xi_j|\xi_i=1\right)-E\left(\xi_j\right)\right]$$

$$=E(\xi_i)[(n-1)-\{n-E(\xi_i)\}]$$

$$=-E(\xi_i)[1-E(\xi_i)] \qquad (3.3.14)$$

Similarly

$$\sum_{i(\neq j)=1}^{N}\left[E\left(\xi_i\xi_j\right)-E(\xi_i)E(\xi_j)\right]=-E\left(\xi_j\right)\left[1-E\left(\xi_j\right)\right] \qquad (3.3.15)$$

Using equations (3.3.14)and (3.3.15), we can re-write the equation (3.3.8) for $V(\bar{\tau}_n)$ as follows:

$$V(\bar{\tau}_n)=\frac{1}{2n^2}\left[\sum_{i=1}^{N}E(\xi_i)\{1-E(\xi_i)\}\tau_i^2+\sum_{j=1}^{N}E\left(\xi_j\right)\left\{1-E\left(\xi_j\right)\right\}\tau_j^2\right.$$

$$\left.-2\sum_{i\neq j=1}^{N}[E(\xi_i)E(\xi_j)-E\left(\xi_i\xi_j\right)]\tau_i\tau_j\right]$$

$$=\frac{1}{2n^2}\sum_{i\neq j=1}^{N}\left[E(\xi_i)E(\xi_j)-E\left(\xi_i\xi_j\right)\right]\left(\tau_i^2+\tau_j^2-2\tau_i\tau_j\right)$$

$$=\frac{1}{2n^2}\sum_{i\neq j=1}^{N}\left[E(\xi_i)E(\xi_j)-E\left(\xi_i\xi_j\right)\right]\left(\tau_i-\tau_j\right)^2 \qquad (3.3.16)$$

The expression for $E(\xi_i)$ and $E\left(\xi_i\xi_j\right)$ can be written explicitly for any given sample size. For example, let $n=2$ and assume that at the second draw, the probability of selecting a unit from the units available is proportional to the probability of selecting it at the first draw. Since

$E(\xi_i)=$ Probabiltiy of selecting y_i in a sample of two

$=p_{i_1}+p_{i_2}$,

where$P_{i_r}(r=1,2)$ is the probability of selecting y_i at the r-th draw, therefore $E(\xi_i)$ can be expressed by using equations (A) and (B) of the section 1.1.3 as

$$E(\xi_i)=p_i\left[S+1-\frac{p_i}{1-p_i}\right] \qquad (3.3.17)$$

and$S=\sum_{j=1}^{N}\frac{p_j}{1-p_j}$

Again, $E\left(\xi_i\xi_j\right)=$ Probability of including both y_i and y_j in a sample of two

$$=p_{i_1}p_{j_2|i}+p_{j_1}p_{i_2|j}$$

$$=p_i\frac{p_i}{1-p_i}+p_j\frac{p_j}{1-p_j}$$

$$=p_ip_j\left[\frac{1}{1-p_i}+\frac{1}{1-p_j}\right] \qquad (3.3.18)$$

Similarly, we can obtain the expression for any sample size although they would necessarily be very complex. It may be noticed that if $E(\xi_i)$ is proportional to y_i, the Horvitz-Thompson estimate of the population mean reduces to $\bar{Y}_N$ with zero variance. Since the values of$y_i$usually not known in advance but values x_i of a character correlated with y_i may be known. It follows as a near approximation that if $E(\xi_i)$ is made proportional to x_i, we should substantial reduction in the variance of the mean.

From the expression for the variance given by equation (2.16), it follows immediately that an unbiased estimate of $V(\bar{\tau}_n)$ is given by

$$Est.[V(\bar{\tau}_n)] = \frac{1}{2n^2}\sum_{i\neq j}^{n}\frac{E(\xi_i\xi_j)-E(\xi_i)E(\xi_j)}{E(\xi_i\xi_j)}\left(\tau_i-\tau_j\right)^2, \qquad (3.3.19)$$

which is a linear function of the squares of the differences between the τ_i in the sample and vanishes when they are all equal. However, this estimate, known usually as the Yates-Grundy estimate, is not entirely satisfactory in the sense that it may also assume negative values, though under very inflexible conditions. It is easily seen that for samples of any size $2 \leq n \leq N$, a sufficient condition for the estimated variance given by equation (3.3.19) to be always non-negative is that

$$E(\xi_i)E(\xi_j) \geq E(\xi_i\xi_j) \text{ for all pairs } i,j\ (i \neq j) \qquad (3.3.20)$$

Obviously, for sample of size 2, the condition (3.3.20) is necessary.

Consider a population of 4 units given by Des Raj (1956) with $y_1 = 1$, $y_2 = 2$, $y_3 = 3$ and $y_4 = 4$ and

$$E(\xi_1\xi_2) = E(\xi_3\xi_4) = \frac{3}{8}$$

$$E(\xi_1\xi_3) = E(\xi_1\xi_4) = E(\xi_2\xi_3) = E(\xi_2\xi_4) = \frac{1}{16}$$

We then have samples of size 2,

$$E(\xi_1) = E(\xi_2) = E(\xi_3) = E(\xi_4) = \frac{1}{2}$$

Since condition (3.3.20)is not satisfied for all pairs $i,j\ (i \neq j)$, it follows that Yates Grundy estimate of the variance of Horvitz Thompson estimate as given by (3.3.19)will not always be non-negative.

It would now be very interesting to enquire whether sampling without replacement using Horvitz Thompson estimate is always more efficient than the *pps* sampling with replacement resulting in the same probabilities$E(\xi_i)$for

the i^{th} unit in the sample. Durbin (1953) has pointed out that this need not be the case.

Example 3.1: Estimate the total production for the data given in Example 1.4 of 8 orchards along with the standard error $(S.E.)$ using Horvitz-Thompson estimator.

Solution: To estimate the total production and its standard error, the table given in Example 1.4 is used.

The Horvitz-Thompson estimator for population total is given by the equations (3.3.2) and (3.3.3)as follows

$$(\hat{y}_N)_{HT} = N(\bar{\tau}_n) = \sum^n \frac{y_i}{\pi_i}$$

along with $Est.V[(\hat{y}_N)_{HT}] = \sum_i^n \sum_{j>i}^n \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}} \left[\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right]^2$

[Expression derived by Yates and Grundy. (1953)]

Also we have $S = \sum_i^8 \frac{p_i}{1-p_i} = 1.157$

For the units selected in the sample,

$$\pi_1 = p_1\left[S + 1 - \frac{p_1}{1-p_1}\right] = 0.1969$$

$$\pi_2 = p_2\left[S + 1 - \frac{p_2}{1-p_2}\right] = 0.1538$$

$$\pi_{12} = p_1p_2\left[\frac{1}{1-p_1} - \frac{1}{1-p_2}\right] = 0.0155$$

Therefore $(\hat{y}_N)_{HT} = \frac{30}{0.1969} + \frac{22}{0.1538}$

$= 295.403$ (in 10 kg units)

$$Est.V[(\hat{y}_N)_{HT}] = \left[\frac{0.1969\times0.1538-0.0155}{0.0155}\right]\left[\frac{30}{0.1969} - \frac{22}{0.1538}\right]^2$$

$$= 79.91$$

and $S.E.[(\hat{y}_N)_{HT}] = \sqrt{79.91} = 8.93$

10.4 Midzuno System of Sampling

Midzuno (1950) has suggested a system of selection probabilities in which the unit at the first draw is selected with unequal probabilities of selection while, at

all subsequent draws, the units are selected with equal probability and without replacement.

Defining ξ_i as in equation (3.3.1), it is easy to observe that under this system

$$E(\xi_i) = p_{i_1} + p_{i_2} + \cdots + p_{i_n}$$

$= p_i +$ Probability that y_i is not selected at the first draw and is selected

at any of the subsequent $(n-1)$ draws

$$= p_i + (1 - p_i)\frac{n-1}{N-1}$$

$$= \frac{N-n}{N-1}p_i + \frac{n-1}{N-1} \qquad (3.4.1)$$

$E(\xi_i\xi_j) =$ Probability that y_i and y_j are both in the sample

$=$ Probability that y_i is selected at the first draw and y_j is selected at any of the subsequent $(n-1)$ draws + Probability that y_j is selected at the first draw and y_i is selected at any of the subsequent $(n-1)$ draws+ Probability that neither y_i nor y_j is selected at the first draw but both of them are selected during the subsequent $(n-1)$ draws

$$= p_i\frac{n-1}{N-1} + p_j\frac{n-1}{N-1} + (1 - p_i - p_j)\frac{(n-1)(n-2)}{(N-1)(N-2)}$$

$$= \frac{(n-1)}{(N-1)}\left\{\frac{(N-n)}{(N-2)}(p_i + p_j + p_k) + \frac{(n-3)}{(N-3)}\right\} \qquad (3.4.2)$$

Similarly, $E(\xi_i\xi_j\xi_k) =$probability of including y_i, y_j and y_k in the sample

$$= \frac{(n-1)(n-2)}{(N-1)(N-2)}\left\{\frac{(N-n)}{(N-3)}(p_i + p_j + p_k) + \frac{(n-3)}{(N-3)}\right\} \qquad (3.4.3)$$

Similarly by the extension of the above argument, we find that if y_i, y_j, y_k,.....,y_q be the n units, the probability of including these n units in the sample is given by

$$E(\xi_i\xi_j\xi_k\ldots..\xi_q) = \frac{(n-1)(n-2)\ldots.1}{(N-1)(N-2)\ldots(N-n+1)}(p_i + p_j + p_k + \ldots + p_q) \qquad (3.4.4)$$

$$= \frac{1}{\binom{N-1}{n-1}}(p_i + p_j + p_k + \ldots + p_q) \qquad (3.4.5)$$

Thus if the p_i are considered proportional to some measure of size of the units in the population, we see that the probability of selecting a specified sample is proportional to the total measure of the size of the units included in the sample. Substituting from equations (3.4.1) and (3.4.2) in (3.4.3) and (3.2.16), we obtain the Horvitz-Thomspon estimate of the population mean and its variance.

In particular, it may be noted that the unbiased estimate of variance of Horvitz-Thompson estimate, given by equation (3.2.19), reduces to

$$\frac{N-n}{2(N-1)^2n^2}\sum_{i\neq j}^{n}\left[(N-n)p_ip_j+\frac{n-1}{N-2}(1-p_i-p_j)\right]\frac{(\tau_i-\tau_j)^2}{E(\xi_i\xi_j)} \quad (3.4.6)$$

which is always positive, except when then ξ_i are equal in which case it is zero. Thus, under Midzuno scheme of sampling, the Yates-Grundy estimate of the variance of Horvitz-Thompson estimate is never negative.

The main advantage of this method of sampling, besides of course its simplicity, is the fact that it is possible to compute a set of revised probabilities of selection p_i' such that the inclusion probabilities $E(\xi_i)$ resulting from the revised probabilities p_i' are proportional to the initial probabilities of selection $p_i(i=1,2,\ldots,N)$. It is desirable to be able to do so, since the p_i can be chosen proportional to some measure of size.

Now if $p_i'(i=1,2,\ldots,N)$ are the revised probabilities of selection, we have from equation (3.4.1),

$$E(\xi_i)=\frac{N-n}{N-1}p_i'+\frac{n-1}{N-1} \quad (3.4.7)$$

If this is proportional to the initial probability of selection p_i, we obtain, by utilizing the fact that $\sum_{i=1}^{N}E(\xi_i)=n$,

$$\frac{N-n}{N-1}p_i'+\frac{n-1}{N-1}=np_i \quad (3.4.8)$$

Hence, $$p_i'=np_i\frac{N-1}{N-n}-\frac{n-1}{N-n}(i=1,2,\ldots,N) \quad (3.4.9)$$

Since $p_i'(i=1,2,\ldots,N)$ must always be positive, the initial probabilities of selection p_i must satisfy the condition

$$p_i>\frac{n-1}{n(N-1)} \quad (3.4.10)$$

This restriction on the initial probabilities of selection naturally limits the use of the system in practice. Subject to this restriction, it has been shown by Rao (1963) for samples of size 2 that the Horvitz-Thompson estimate of the population mean under Midzuno scheme of sampling is always more efficient than the usual estimate in the case of sampling with varying probabilities and with replacement.

10.5 Narain Method of Sampling

In this section, we shall discuss a scheme of sampling suggested by Narain (1951) which does not require any restriction on the initial set of selection probabilities and at the same time leads to a more efficient estimate of the population value than in the case of sampling with replacement. The method consists in constructing revised probabilities of selection $p_i^*(i = 1,2, \ldots . , N)$ such that the inclusion probabilities $E(\xi_i)$resulting from p_i^* are proportional to the original probabilities of selection $p_i(i = 1,2, \ldots , N)$; the sampling is done without replacement and the probabilities of selection at the second and subsequent draws are proportional to the revised probabilities of selection p_i^* at the first draw. From equation (3.2.17), we have

$$E(\xi_i) = p_i^n \left[S^* + 1 - \frac{p_i^*}{1-p_i^*}\right] = np_i \ (i = 1,2, \ldots , N) \quad (3.5.1)$$

where $$S^* = \sum_{i=1}^{N} \frac{p_i^*}{1-p_i^*} \quad (3.5.2)$$

Narain (1951) and Yates and Grundy (1953) have suggested a methods of solving the set of equations (3.5.1)) to obtain the revised probabilities of selection p_i^*. The computations are tedious for n greater than 2 and the method eventually breaks down. More recently, Brewer and Undy (1962) have considered this method of sampling in detail for the case $n = 2$ and have given a faster iterative method of solution which always yields a unique solution.

Further, improvement in the sampling scheme has been introduced by Singh (1954) and shown some situations in which sampling without replacement will be less efficient as compared to that with replacement.

10.6 Summery

In this unit 3, Horvitz- Thompson estimator and its properties are explained. Different types of sampling, like- Midzuno system of sampling and Narain method of sampling are discussed. Suitable examples and questions in exercise on the discussed points are given for practice. For further study and exercise, readers are required to study the referred text books of sampling techniques.

10.7 Check Your Self (Questions)

1. If units are drawn one by one with varying probabilities and without replacement, and at any draw subsequent to the first draw, the probability of selecting a unit from the units available at that draw is proportional to the probability of selecting it at the first draw, show that for sample of size two, the Yates and Grundy estimate of variance is always positive.
2. Consider a sampling system where the first two units are selected with varying probabilities and without replacement, the probability of selecting a unit at the second draw being proportional to the probability of selecting it at the first draw, while the remaining $(n-2)$ units are selected with equal probability and without replacement from the remaining $(N-2)$ units. Obtain expressions for $E(\alpha_i)$ and $E(\alpha_i\alpha_j)$. Hence, or otherwise, show that the Yates and Grundy estimate of variance is always positive.
3. Show that under Midzuno scheme of sampling, $T_1 = \frac{\sum^n y_i}{\sum^n P_i}$(where the summation is taken over the entire sample and the P_i are the initial probabilities of selection of the units selected in the sample) is an unbiased estimate of the population total. Give also an unbiased estimate of the Variance of T_1.
4. Select a sample of 2 villages from the data given below with probability proportional to the area under crop without replacement. On the basis of these units selected in the sample, estimate total yield of 20 villages along with the standard error by using Horvitz-Thompson estimator.

S. No. of village	1	2	3	4	5	6	7	8	9	10
Area under crop (a_i)	4.8	4.1	1.3	5.2	6.9	6.0	2.0	6.3	5.2	4.2
Yield of crop(y_i)	22	19	6	25	54	43	4	40	28	29

S. No. of village	11	12	13	14	15	16	17	18	19	20
Area under crop (a_i)	4.8	5.9	5.8	5.8	5.1	4.7	5.6	5.2	4.0	4.6
Yield of crop (y_i)	22	39	39	44	30	27	34	31	18	31

10.8 References

- Basu, D. (1958) "On sampling with and without replacement", *Sankhya*, **20**, 287-294.
- Brewer, K. R. W., and Undy, G. C. (1962) "Samples of two units with unequal probabilities", *Aust. J. Stat.*, **5**, 5-13.
- Des Raj (1956) "Some estimators in sampling with varying probabilities", *J. Amer. Stat. Assoc.*, **61**, 384-390.

- Durbin, J. (1953) "Design of multi-stage surveys for the estimation of sampling errors", *Applied Statistics*, **16**, 152-164.
- Horvitz, D. G., and Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", *J. Amer. Stat. Assoc.*, **47**, 663-685.
- Midzuno, H. (1950) "An outline of the theory of sampling systems", *Ann. Inst. Stat. Math.*, Japan, **1**, 149-156.
- Murthy, M. N. (1957) "Ordered and unordered estimators in sampling without replacement", *Sankhya*, **18**, 379-390.
- Narain, R. D. (1951) "On sampling without replacement with varying probabilities", *J. Ind. Soc. Agri. Stat.*, **3**, 169-174.
- Rao, J. N. K. (1963) "On two systems of unequal probability sampling without replacement", *Ann. Inst. Stat. Math.*, Japan, **15**, 67-72.
- Singh, D. (1954) "On efficiency of sampling with varying probabilities without replacement", *J. Ind. Soc. Agr. Statist.*,**6**, 48-57.
- Yates, F. and Grundy, P. (1953) "Selection without replacement from within strata with probability proportional to size", *J. Roy. Stat. Soc., Series B*, **15**, 253-261.

10.9 Further Readings

- **Cochran, W. G. (1977)**. Sampling Techniques. New York: Wiley.
- **Mukhopadhyay, P (2009).**Theory and Methods of Survey Sampling, PHI Learning Pvt. Ltd.
- **Singh, D. and Chaudhary, F. S. (1986).** Wiley Eastern Ltd. New Age International Ltd.
- **Sukhatme, P. V. and Sukhatme, B. V. (1997).** Sampling Theory of Surveys with Applications. The Iowa State University Press, Ames, Iowa, U.S.A.; The Indian Society of Agricultural Statistics, Piyush Publications, New Delhi.