COURSE CODE: **STA 441**

COURSE TITLE: **SAMPLING THEORY AND SURVEY METHODS II**

COURSE STATUS: **ELECTIVE**

COURSE DESCRIPTION:

Single and multi-stage cluster sampling, Equal and unequal clusters, Two-Phase sampling, Further use of auxiliary information. Multivariate ratio estimation. Unequal probability sampling with or without replacement, Ordered and unordered estimators.

READING LIST:

1. Sampling Techniques: W.G. Cochran, Wiley, Third Edition.
2. Theory and Methods of Survey Sampling: Parimal Mukhopadhyay, Prentice Hall of India.
3. Theory of Sample surveys with applications : P.V. Sukhatme, B.V Sukhatme, S. Sukhatme and C. Asok, IASRI, Delhi.
4. Sampling Methodologies and Applications: P.S.R.S. Rao, Chapman and Hall/ CRC.
5. Sampling Theory and Methods: M.N. Murthy, Statistical Publishing Society, Calcutta.
6. Elements of sampling theory and methods: Z. Govindrajalu, Prentice Hall.

Introduction

Statistics is the science of data.

Data are the numerical values containing some information.

Statistical tools can be used on a data set to draw statistical inferences. These statistical inferences are in turn used for various purposes. For example, government uses such data for policy formulation for the welfare of the people, marketing companies use the data from consumer surveys to improve the company and to provide better services to the customer, etc. Such data is obtained through sample surveys. Sample surveys are conducted throughout the world by governmental as well as non-governmental agencies. For example, "National Sample Survey Organization (NSSO)" conducts surveys in India, "Statistics Canada" conducts surveys in Canada, agencies of United Nations like "World Health Organization (WHO), "Food and Agricultural Organization (FAO)" etc. conduct surveys in different countries.

Sampling theory provides the tools and techniques for data collection keeping in mind the objectives to be fulfilled and nature of population.

There are two ways of obtaining the information

1. **Sample surveys**
2. **Complete enumeration or census**

Sample surveys collect information on a fraction of total population whereas census collect information on whole population. Some surveys e.g., economic surveys, agricultural surveys etc. are conducted regularly. Some surveys are need based and are conducted when some need arises, e.g., consumer satisfaction surveys at a newly opened shopping mall to see the satisfaction level with the amenities provided in the mall .

Sampling unit:

An element or a group of elements on which the observations can be taken is called a sampling unit. The objective of the survey helps in determining the definition of sampling unit.

For example, if the objective is to determine the total income of all the persons in the household, then the sampling unit is household. If the objective is to determine the income of any particular person in the household, then the sampling unit is the income of the particular person in the household. So the definition of sampling unit depends and varies as per the objective of the survey. Similarly, in another example, if the objective is to study the blood sugar level, then the sampling unit is the value of blood sugar level of a person. On the other hand, if the objective is to study the health conditions, then the sampling unit is the person on whom the readings on the blood sugar level, blood pressure and other factors will be obtained. These values will together classify the person as healthy or unhealthy.

Population:

Collection of all the sampling units in a given region at a particular point of time or a particular period is called the population. For example, if the medical facilities in a hospital are to be surveyed through the patients, then the total number of patients registered in the hospital during the time period of survey will the population. Similarly, if the production of wheat in a district is to be studied, then all the fields cultivating wheat in that district will be constitute the population. The total number of sampling units in the population is the population size, denoted generally by N. The population size can be finite or infinite (N is large).

Census:

The complete count of population is called census. The observations on all the sampling units in the population are collected in the census. For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

Sample:

One or more sampling units are selected from the population according to some specified procedure.
A sample consists only of a portion of the population units. Such a collection of units is called the sample.

In the context of sample surveys, a collection of units like households, people, cities, countries etc. is called a finite population.

A census is a 100% sample and it is a complete count of the population.

Representative sample:

When all the salient features of the population are present in the sample, then it is called a representative sample,

It goes without saying that every sample is considered as a representative sample.

For example, if a population has 30% males and 70% females, then we also expect the sample to have nearly 30% males and 70% females.

In another example, if we take out a handful of wheat from a 100 Kg. bag of wheat, we expect the same quality of wheat in hand as inside the bag. Similarly, it is expected that a drop of blood will give the same information as all the blood in the body.

Sampling frame:

The list of all the units of the population to be surveyed constitutes the sampling frame. All the sampling units in the sampling frame have identification particulars. For example, all the students in a particular university listed along with their roll numbers constitute the sampling frame. Similarly, the list of households with the name of head of family or house address constitutes the sampling frame. In another example, the residents of a city area may be listed in more than one frame - as per automobile registration as well as the listing in the telephone directory.

Ways to ensure representativeness:

There are two possible ways to ensure that the selected sample is representative.

1. Random sample or probability sample:

The selection of units in the sample from a population is governed by the laws of chance or probability. The probability of selection of a unit can be equal as well as unequal.

2. Non-random sample or purposive sample:
The selection of units in the sample from population is not governed by the probability laws.

For example, the units are selected on the basis of personal judgment of the surveyor. The persons volunteering to take some medical test or to drink a new type of coffee also constitute the sample on non-random laws.

Another type of sampling is Quota Sampling. The survey in this case is continued until a predetermined number of units with the characteristic under study are picked up.

For example, in order to conduct an experiment for rare type of disease, the survey is continued till the required number of patients with the disease are collected.

Advantages of sampling over complete enumeration:

1. **Reduced cost and enlarged scope.**
   Sampling involves the collection of data on smaller number of units in comparison to the complete enumeration, so the cost involved in the collection of information is reduced. Further, additional information can be obtained at little cost in comparison to conducting another separate survey. For example, when an interviewer is collecting information on health conditions, then he/she can also ask some questions on health practices. This will provide additional information on health practices and the cost involved will be much less than conducting an entirely new survey on health practices.

2. **Organizaton of work:**
   It is easier to manage the organization of collection of smaller number of units than all the units in a census. For example, in order to draw a representative sample from a state, it is easier to manage to draw small samples from every city than drawing the sample from the whole state at a time. This ultimately results in more accuracy in the statistical inferences because better organization provides better data and in turn, improved statistical inferences are obtained.

3. **Greater accuracy:**

   The persons involved in the collection of data are trained personals. They can collect the data more accurately if they have to collect smaller number of units than large number of units.

4. **Urgent information required:**

   The data from a sample can be quickly summarized.

   For example, the forecasting of the crop production can be done quickly on the basis of a sample of data than collecting first all the observation.

5. **Feasibility:**

   Conducting the experiment on smaller number of units, particularly when the units are destroyed, is more feasible. For example, in determining the life of bulbs, it is more feasible to fuse minimum number of bulbs. Similarly, in any medical experiment, it is more feasible to use less number of animals.

Type of surveys:

There are various types of surveys which are conducted on the basis of the objectives to be fulfilled.

1. Demographic surveys:

These surveys are conducted to collect the demographic data, e.g., household surveys, family size, number of males in families, etc. Such surveys are useful in the policy formulation for any city, state or country for the welfare of the people.

2. Educational surveys:

These surveys are conducted to collect the educational data, e.g., how many children go to school, how many persons are graduate, etc. Such surveys are conducted to examine the educational programs in schools and colleges. Generally, schools are selected first and then the students from each school constitue the sample.

3. Economic surveys:

These surveys are conducted to collect the economic data, e.g., data related to export and import of goods, industrial production, consumer expenditure etc. Such data is helpful in constructing the indices indicating the growth in a particular sector of economy or even the overall economic growth of the country.

4. Employment surveys:

These surveys are conducted to collect the employment related data, e.g., employment rate, labour conditions, wages, etc. in a city, state or country. Such data helps in constructing various indices to know the employment conditions among the people.

5. Health and nutrition surveys:

These surveys are conducted to collect the data related to health and nutrition issues, e.g., number of visits to doctors, food given to children, nutritional value etc. Such surveys are conducted in cities, states as well as countries by the national and international organizations like UNICEF, WHO etc.

6. Agricultural surveys:

These surveys are conducted to collect the agriculture related data to estimate, e.g., the acreage and production of crops, livestock numbers, use of fertilizers, use of pesticides and other related topics. The government bases its planning related to the food issues for the people based on such surveys.

7. Marketing surveys:

These surveys are conducted to collect the data related to marketing. They are conducted by major companies, manufacturers or those who provide services to consumer etc. Such data is used for knowing the satisfaction and opinion of consumers as well as in developing the sales, purchase and promotional activities etc.

8. Election surveys:

These surveys are conducted to study the outcome of an election or a poll. For example, such polls are conducted in democratic countries to have the opinions of people about any candidate who is contesting the election.

9. Public polls and surveys:

These surveys are conducted to collect the public opinion on any particular issue. Such surveys are generally conducted by the news media and the agencies which conduct polls and surveys on the current topics of interest to public.

10. Campus surveys:

These surveys are conducted on the students of any educational institution to study about the educational programs, living facilities, dining facilities, sports activities, etc.

Principal steps in a sample survey:

The broad steps to conduct any sample surveys are as follows:

1. Objective of the survey:

The objective of the survey has to be clearly defined and well understood by the person planning to conduct it. It is expected from the statistician to be well versed with the issues to be addressed in consultation with the person who wants to get the survey conducted. In complex surveys, sometimes the objective is forgotten and data is collected on those issues which are far away from the objectives.

2. Population to be sampled:

Based on the objectives of the survey, decide the population from which the information can be obtained. For example, population of farmers is to be sampled for an agricultural survey whereas the population of patients has to be sampled for determining the medical facilities in a hospital.

3. Data to be collected:

It is important to decide that which data is relevant for fulfilling the objectives of the survey and to note that no essential data is omitted. Sometimes, too many questions are asked and some of their outcomes are never utilized. This lowers the quality of the responses and in turn results in lower efficiency in the statistical inferences.

4. Degree of precision required:

The results of any sample survey are always subjected to some uncertainty. Such uncertainty can be reduced by taking larger samples or using superior instruments. This involves more cost and more time. So it is very important to decide about the required degree of precision in the data. This needs to be conveyed to the surveyor also.

5. Method of measurement:

The choice of measuring instrument and the method to measure the data from the population needs to be specified clearly. For example, the data has to be collected through interview, questionnaire, personal visit, combination of any of these approaches, etc. The forms in which the data is to be recorded so that the data can be transferred to mechanical equipment for easily creating the data summary etc. is also needed to be prepared accordingly.

6. The frame:

The sampling frame has to be clearly specified. The population is divided into sampling units such that the units cover the whole population and every sampling unit is tagged with identification. The list of all sampling units is called the frame. The frame must cover the whole population and the units must not overlap each other in the sense that every element in the population must belong to one and only one unit. For example, the sampling unit can be an individual member in the family or the whole family.

7. Selection of sample:

The size of the sample needs to be specified for the given sampling plan. This helps in determining and comparing the relative cost and time of different sampling plans. The method and plan adopted for drawing a representative sample should also be detailed.

8. The Pre-test:

It is advised to try the questionnaire and field methods on a small scale. This may reveal some troubles and problems beforehand which the surveyor may face in the field in large scale surveys.

9. Organization of the field work:

How to conduct the survey, how to handle business administrative issues, providing proper training to surveyors, procedures, plans for handling the non-response and missing observations etc. are some of the issues which need to be addressed for organizing the survey work in the fields. The procedure for early checking of the quality of return should be prescribed. It should be clarified how to handle the situation when the respondent is not available.

10. Summary and analysis of data:

It is to be noted that based on the objectives of the data, the suitable statistical tool is decided which can answer the relevant questions. In order to use the statistical tool, a valid data set is required and this dictates the choice of responses to be obtained for the questions in the questionnaire, e.g., the data has to be qualitative, quantitative, nominal, ordinal etc. After getting the completed questionnaire back, it needs to be edited to amend the recording errors and delete the erroneous data. The tabulating procedures, methods of estimation and tolerable amount of error in the estimation needs to be decided before the start of survey. Different methods of estimation may be available to get the answer of the same query from the same data set. So the data needs to be collected which is compatible with the chosen estimation procedure.

11. Information gained for future surveys:

The completed surveys work as guide for improved sample surveys in future. Beside this they also supply various types of prior information required to use various statistical tools, e.g., mean, variance, nature of variability, cost involved etc. Any completed sample survey acts as a potential guide for the surveys to be conducted in the future. It is generally seen that the things always do not go in the same way in any complex survey as planned earlier. Such precautions and alerts help in avoiding the mistakes in the execution of future surveys.

Variability control in sample surveys:

The variability control is an important issue in any statistical analysis. A general objective is to draw statistical inferences with minimum variability. There are various types of sampling schemes which are adopted in different conditions. These schemes help in controlling the variability at different stages. Such sampling schemes can be classified in the following way.

1. **Before selection of sampling units**
   - Stratified sampling
   - Cluster sampling
   - Two stage sampling
   - Double sampling etc.

2. **At the time of selection of sampling units**
   - Systematic sampling
   - Varying probability sampling

3. **After the selection of sampling units**
   - Ratio method of estimation
   - Regression method of estimation

Note that the ratio and regtresion methods are the methods of estimation and not the methods of drawing samples.

Methods of data collection

There are various way of data collection. Some of them are as follows:

1. **Physical observations and measurements:**

The surveyor contacts the respondent personally through the meeting. He observes the sampling unit and records the data. The surveyor can always use his prior experience to collect the data in a better way. For example, a young man telling his age as 60 years can easily be observed and corrected by the surveyor.

2. Personal interview:

The surveyor is supplied with a well prepared questionnaire. The surveyor goes to the respondents and asks the same questions mentioned in the questionnaire. The data in the questionnaire is then filled up accordingly based on the responses from the respondents.

3. Mail enquiry:

The well prepared questionnaire is sent to the respondents through postal mail, e-mail, etc. The respondents are requested to fill up the questionnaires and send it back. In case of postal mail, many times the questionnaires are accompanied by a self addressed envelope with postage stamps to avoid any non-response due to the cost of postage.

4. Web based enquiry:

The survey is conducted online through internet based web pages. There are various websites which provide such facility. The questionnaires are to be in their formats and the link is sent to the respondents through email. By clicking on the link, the respondent is brought to the concerned website and the answers are to be given online. These answers are recorded and responses as well as their statistics is sent to the surveyor. The respondents should have internet connection to support the data collection with this procedure.

5. Registration:

The respondent is required to register the data at some designated place. For example, the number of births and deaths along with the details provided by the family members are recorded at city municipal office which are provided by the family members.

6. Transcription from records:

The sample of data is collected from the already recorded information. For example, the details of the number of persons in different families or number of births/deaths in a city can be obtained from the city municipal office directly.

The methods in (1) to (5) provide primary data which means collecting the data directly from the source. The method in (6) provides the secondary data which means getting the data from the primary sources.

Cluster Sampling

It is one of the basic assumptions in any sampling procedure that the population can be divided into a finite number of distinct and identifiable units, called **sampling units.** The smallest units into which the population can be divided are called **elements** of the population. The groups of such elements are called **clusters**.

In many practical situations and many types of populations, a list of elements is not available and so the use of an element as a sampling unit is not feasible. The method of cluster sampling or area sampling can be used in such situations.

In cluster sampling

- divide the whole population into clusters according to some well defined rule.
- Treat the clusters as sampling units.
- Choose a sample of clusters according to some procedure.
- Carry out a complete enumeration of the selected clusters, i.e., collect information on all the sampling units available in selected clusters.

Area sampling

In case, the entire area containing the populations is subdivided into smaller area segments and each element in the population is associated with one and only one such area segment, the procedure is called as area sampling.

Examples:

- In a city, the list of all the individual persons staying in the houses may be difficult to obtain or even may be not available but a list of all the houses in the city may be available. So every individual person will be treated as sampling unit and every house will be a cluster.
- The list of all the agricultural farms in a village or a district may not be easily available but the list of village or districts are generally available. In this case, every farm in sampling unit and every village or district is the cluster.

Moreover, it is easier, faster, cheaper and convenient to collect information on clusters rather than on sampling units.

In both the examples, draw a sample of clusters from houses/villages and then collect the observations on all the sampling units available in the selected clusters.

Conditions under which the cluster sampling is used:

Cluster sampling is preferred when

(i) No reliable listing of elements is available and it is expensive to prepare it.

(ii) Even if the list of elements is available, the location or identification of the units may be difficult.

(iii) A necessary condition for the validity of this procedure is that every unit of the population under study must correspond to one and only one unit of the cluster so that the total number of sampling units in the frame may cover all the units of the population under study without any omission or duplication. When this condition is not satisfied, bias is introduced.

Open segment and closed segment:

It is not necessary that all the elements associated with an area segment need be located physically within its boundaries. For example, in the study of farms, the different fields of the same farm need not lie within the same area segment. Such a segment is called an open segment.

In a closed segment, the sum of the characteristic under study, i.e., area, livestock etc. for all the elements associated with the segment will account for all the area, livestock etc. within the segment.

Construction of clusters:

The clusters are constructed such that the sampling units are heterogeneous within the clusters and homogeneous among the clusters. The reason for this will become clear later. This is opposite to the construction of the strata in the stratified sampling.

There are two options to construct the clusters – equal size and unequal size. We discuss the estimation of population means and its variance in both the cases.

Case of equal clusters

- Suppose the population is divided into N clusters and each cluster is of size n.
- Select a sample of n clusters from N clusters by the method of SRS, generally WOR.

So

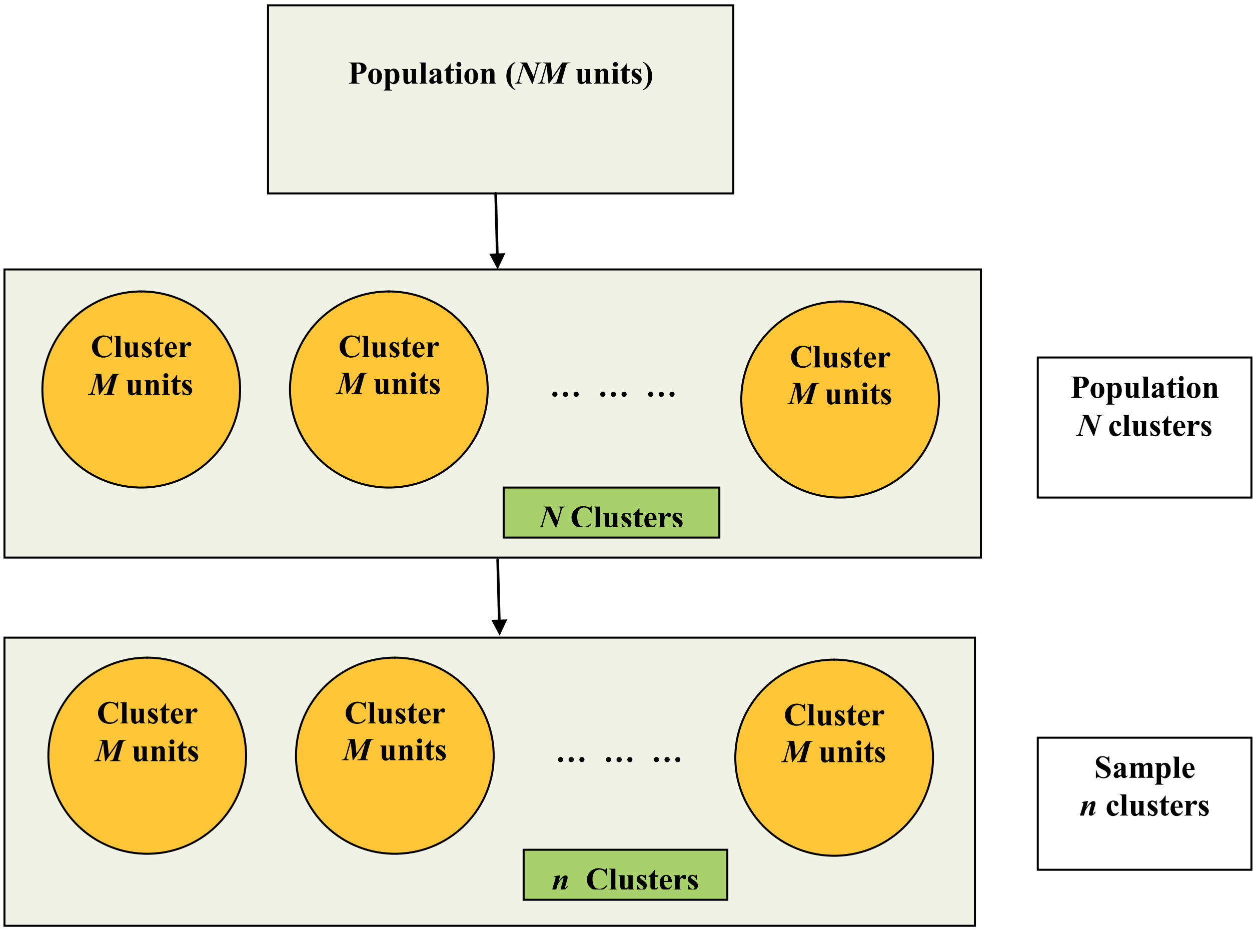
total population size = NM

total sample size = nM.

Let

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

$\bar{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ mean per element of i^{th} cluster .



Estimation of population mean:

First select n clusters from N clusters by SRSWOR.

Based on n clusters, find the mean of each cluster separately based on all the units in every cluster. So we have the cluster means as $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$. Consider the mean of all such cluster means as an estimator of population mean as

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

Bias:

$$\begin{aligned} E(\bar{y}_{cl}) &= \frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\bar{Y} \qquad \text{(since SRS is used)} \\ &= \bar{Y}. \end{aligned}$$

Thus $\bar{y}_{cl}$ is an unbiased estimator of $\bar{Y}$.

Variance:

The variance of $\bar{y}_{cl}$ can be derived on the same lines as deriving the variance of sample mean in SRSWOR. The only difference is that in SRSWOR, the sampling units are $y_1, y_2, ..., y_n$ whereas in case of $\bar{y}_{cl}$, the sampling units are $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$.

$\left[\text{Note that is case of SRSWOR, } Var(\bar{y}) = \frac{N-n}{Nn}S^2 \text{ and } \widehat{Var}(\bar{y}) = \frac{N-n}{Nn}s^2\right]$,

$$\begin{aligned} Var(\bar{y}_{cl}) &= E(\bar{y}_{cl} - \bar{Y})^2 \\ &= \frac{N-n}{Nn}S_b^2 \end{aligned}$$

where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \bar{Y})^2$ which is the mean sum of square between the cluster means in the population.

Estimate of variance:

Using again the philosophy of estimate of variance in case of SRSWOR, we can find

$$\widehat{Var}(\bar{y}_{cl}) = \frac{N-n}{Nn}s_b^2$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{cl})^2$ is the mean sum of squares between cluster means in the sample .

Comparison with SRS :

If an equivalent sample of nM units were to be selected from the population of NM units by SRSWOR, the variance of the mean per element would be

$$\begin{aligned} Var(\bar{y}_{nM}) &= \frac{NM - nM}{NM} . \frac{S^2}{nM} \\ &= \frac{f}{n} . \frac{S^2}{M} \end{aligned}$$

where $f = \dfrac{N\text{ - }n}{N}$ and $S^2 = \dfrac{1}{NM-1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2.$

Also $$\begin{aligned} Var(\bar{y}_{cl}) &= \frac{N-n}{Nn} S_b^2 \\ &= \frac{f}{n} S_b^2. \end{aligned}$$

Consider

$$\begin{aligned} (NM-1)S^2 &= \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} \left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{Y}) \right]^2 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{N} \sum_{j=1}^{M} (\bar{y}_i - \bar{Y})^2 \\ &= N(M-1)\bar{S}_w^2 + M(N-1)S_b^2 \end{aligned}$$

where

$\bar{S}_w^2 = \dfrac{1}{N} \sum_{i=1}^{N} S_i^2$ is the mean sum of squares within clusters in the population

$S_i^2 = \dfrac{1}{M-1} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2$ is the mean sum of squares for the i^{th} cluster.

The efficiency of cluster sampling over SRSWOR is

$$\begin{aligned} E &= \frac{Var(\bar{y}_{nM})}{Var(\bar{y}_{cl})} \\ &= \frac{S^2}{MS_b^2} \\ &= \frac{1}{(NM-1)} \left[\frac{N(M-1)}{M} \frac{\bar{S}_w^2}{S_b^2} + (N-1) \right]. \end{aligned}$$

Thus the relative efficiency increases when $\bar{S}_w^2$ is large and S_b^2 is small. So cluster sampling will be efficient if clusters are so formed that the variation the between cluster means is as small as possible while variation within the clusters is as large as possible.

Efficiency in terms of intra class correlation

The intra class correlation between the elements within a cluster is given by

$$\rho = \frac{E(y_{ij} - Y)(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})}; -\frac{1}{M-1} \le \rho \le 1$$

$$= \frac{\dfrac{1}{MN(M-1)} \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\dfrac{1}{MN} \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{Y})^2}$$

$$= \frac{\dfrac{1}{MN(M-1)} \displaystyle\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\left(\dfrac{MN-1}{MN}\right) S^2}$$

$$= \frac{\displaystyle\sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k(\neq j)=1}^{M} (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(MN-1)(M-1)S^2}.$$

Consider

$$\sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2 = \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{j=1}^{M} (y_{ij} - \bar{Y}) \right]^2$$

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

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

or

$$\rho(MN-1)(M-1)S^2 = M^2(N-1)S_b^2 - (NM-1)S^2$$

or $$S_b^2 = \frac{(MN-1)}{M^2(N-1)} \left[1 + \rho(M-1)S^2 \right].$$

The variance of $\bar{y}_{cl}$ now becomes

$$\begin{aligned} Var(\bar{y}_{cl}) &= \frac{N-n}{N} S_b^2 \\ &= \frac{N-n}{Nn} \frac{MN-1}{N-1} \frac{S^2}{M^2} \left[1+(M-1)\rho\right]. \end{aligned}$$

For large $N, \frac{MN-1}{MN} \simeq 1, \frac{N-n}{N} \simeq 1$ and so

$$Var(\bar{y}_{cl}) \simeq \frac{1}{n} \frac{S^2}{M} \left[1+(M-1)\rho\right].$$

The variance of sample mean under SRSWOR for large N is

$$Var(\bar{y}_{nM}) \simeq \frac{S^2}{nM}.$$

The relative efficiency for large N is now given by

$$\begin{aligned} E &= \frac{Var(\bar{y}_{nM})}{Var(\bar{y}_{cl})} \\ &= \frac{\frac{S^2}{nM}}{\frac{S^2}{nM}\left[1+(M-1)\rho\right]} \\ &= \frac{1}{1+(M-1)\rho}; \quad 1 \le \rho \le -\frac{1}{(M-1)}. \end{aligned}$$

- If $M=1$ then $E=1,$ i.e., SRS and cluster sampling are equally efficient. Each cluster will consist of one unit, i.e., SRS.
- If $M>1,$ then cluster sampling is more efficient when

  $$\begin{aligned} & E > 1 \\ \text{or} \quad & (M-1)\rho < 0 \\ \text{or} \quad & \rho < 0. \end{aligned}$$

- If $\rho = 0,$ then $E=1$, i.e., there is no error which means that the units in each cluster are arranged randomly. So sample is heterogeneous.
- In practice, ρ is usually positive and ρ decreases as M increases but the rate of decrease in ρ is much lower in comparison to the rate of increase in $M.$ The situation that $\rho > 0$ is possible when the nearby units are grouped together to form cluster and which are completely enumerated.
- There are situations when $\rho < 0.$

Estimation of relative efficiency:

The relative efficiency of cluster sampling relative to an equivalent SRSWOR is obtained as

$$E = \frac{S^2}{MS_b^2}.$$

An estimator of E can be obtained by substituting the estimates of S^2 and S_b^2.

Since $\bar{y}_{cl} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ is the mean of n means $\bar{y}_i$ from a population of N means $\bar{y}_i, i = 1,2,...,N$ which are drawn by SRSWOR, so from the theory of SRSWOR,

$$\begin{aligned} E(s_b^2) &= E\left[\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_c)^2\right] \\ &= \frac{1}{N-1}\sum_{i=1}^{N}(\bar{y}_i - \bar{Y})^2 \\ &= S_b^2 . \end{aligned}$$

Thus s_b^2 is an unbiased estimator of S_b^2.

Since $s_w^2 = \frac{1}{n}\sum_{i=1}^{n}S_i^2$ is the mean of n mean sum of squares S_i^2 drawn from the population of N mean sums of squares $S_i^2, i = 1,2,...,N,$ so it follows from the theory of SRSWOR that

$$\begin{aligned} E(s_w^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}S_i^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}S_i^2 \\ &= \bar{S}_w^2 . \end{aligned}$$

Thus $\bar{s}_w^2$ is an unbiased estimator of $\bar{S}_w^2$.

Consider

$$S^2 = \frac{1}{MN-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y})^2$$

$$\begin{aligned} \text{or } (MN-1)S^2 &= \sum_{i=1}^{N}\sum_{j=1}^{M}\left[(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{Y})\right]^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}\left[(y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{Y})^2\right] \\ &= \sum_{i=1}^{N}(M-1)S_i^2 + M(N-1)S_b^2 \\ &= N(M-1)\bar{S}_w^2 + M(N-1)S_b^2. \end{aligned}$$

An unbiased estimator of S^2 can be obtained as

$$\hat{S}^2 = \frac{1}{MN-1}\left[N(M-1)\overline{s}_w^2 + M(N-1)s_b^2\right].$$

So

$$\widehat{Var}(\overline{y}_{cl}) = \frac{N-n}{Nn}s_b^2$$

$$\widehat{Var}(\overline{y}_{nM}) = \frac{N-n}{Nn}\frac{\hat{S}^2}{M}$$

where $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{y}_i - \overline{y}_{cl})^2.$

An estimate of efficiency $E = \frac{S^2}{MS_b^2}$ is

$$\hat{E} = \frac{N(M-1)\overline{s}_w^2 + M(N-1)s_b^2}{M(NM-1)s_b^2}.$$

If N is large so that $M(N-1) \simeq MN$ and $MN-1 \simeq MN,$ then

$$E = \frac{1}{M} + \left(\frac{M-1}{M}\right)\frac{\overline{S}_w^2}{MS_b^2}$$

and its estimate is

$$\hat{E} = \frac{1}{M} + \left(\frac{M-1}{M}\right)\frac{\overline{s}_w^2}{Ms_b^2}.$$

Estimation of a proportion in case of equal cluster

Now, we consider the problem of estimation of the proportion of units in the population having a specified attribute on the basis of a sample of clusters. Let this proportion be P.

Suppose that a sample of n clusters is drawn from N clusters by SRSWOR. Defining $y_{ij} = 1$ if the j^{th} unit in the i^{th} cluster belongs to the specified category (i.e. possessing the given attribute) and $y_{ij} = 0$ otherwise, we find that

$$\bar{y}_i = P_i,$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} P_i = P,$$

$$S_i^2 = \frac{MP_iQ_i}{(M-1)},$$

$$S_w^2 = \frac{M\sum_{i=1}^{N} P_iQ_i}{N(M-1)},$$

$$S^2 = \frac{NMPQ}{NM-1)},$$

$$\begin{aligned}
S_b^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(P_i - P)^2, \\
&= \frac{1}{N-1}\left[\sum_{i=1}^{N} P_i^2 - NP^2\right] \\
&= \frac{1}{(N-1)}\left[-\sum_{i=1}^{N} P_i(1-P_i) + \sum_{i=1}^{N} P_i \; - NP^2\right] \\
&= \frac{1}{(N-1)}\left[NPQ - \sum_{i=1}^{N} P_iQ_i\right],
\end{aligned}$$

where P_i is the proportion of elements in the i^{th} cluster, belonging to the specified category and $Q_i = 1 - P_i,\ i = 1,2,...,N$ and $Q = 1 - P.$ Then, using the result that $\bar{y}_{cl}$ is an unbiased estimator of $\bar{Y}$, we find that

$$\hat{P}_{cl} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

is an unbiased estimator of P and

$$Var(\hat{P}_{cl}) = \frac{(N-n)}{Nn}\frac{\left[NPQ - \sum_{i=1}^{N} P_iQ_i\right]}{(N-1)}.$$

This variance of $\hat{P}_{cl}$ can be expressed as

$$Var(\hat{P}_{cl}) = \frac{N-n}{N-1}\frac{PQ}{nM}[1 + (M-1)\rho],$$

where the value of ρ can be obtained from where

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

by substituting $S_b^2, \bar{S}_w^2$ and S^2 in ρ, we obtain

$$\rho = 1 - \frac{M}{(M-1)} \frac{1}{N} \frac{\sum_{i=1}^{N} P_i Q_i}{PQ} \ .$$

The variance of $\hat{P}_{cl}$ can be estimated unbiasedly by

$$\begin{aligned}
\widehat{Var}(\hat{P}_{cl}) &= \frac{N-n}{nN} s_b^2 \\
&= \frac{N-n}{nN} \frac{1}{(n-1)} \sum_{i=1}^{n} (P_i - \hat{P}_{cl})^2 \\
&= \frac{N-n}{Nn(n-1)} \left[n\hat{P}_{cl}\hat{Q}_{cl} - \sum_{i=1}^{n} P_i Q_i \right]
\end{aligned}$$

where $\hat{Q}_{cl} = I - \hat{P}_{cl}$. The efficiency of cluster sampling relative to SRSWOR is given by

$$\begin{aligned}
E &= \frac{M(N-1)}{(MN-1)} \frac{1}{\left[1+(M-1)\rho\right]} \\
&= \frac{(N-1)}{NM-1} \frac{NPQ}{\left(NPQ - \sum_{i=1}^{N} P_i Q_i \right)}.
\end{aligned}$$

If N is large, then $E \cong \frac{1}{M}$.

An estimator of the total number of elements belonging to a specified category is obtained by multiplying $\hat{P}_{cl}$ by NM, *i.e.* by $NM\hat{P}_{cl}$. The expressions of variance and its estimator are obtained by multiplying the corresponding expressions for $\hat{P}_{cl}$ by N^2M^2.

Case of unequal clusters:

In practice, the equal size of clusters are available only when planned. For example, in a screw manufacturing company, the packets of screws can be prepared such that every packet contains same number of screws. In real applications, it is hard to get clusters of equal size. For example, the villages with equal areas are difficult to find, the districts with same number of persons are difficult to find, the number of members in a household may not be same in each household in a given area.

Let there be N clusters and M_i be the size of i^{th} cluster, let

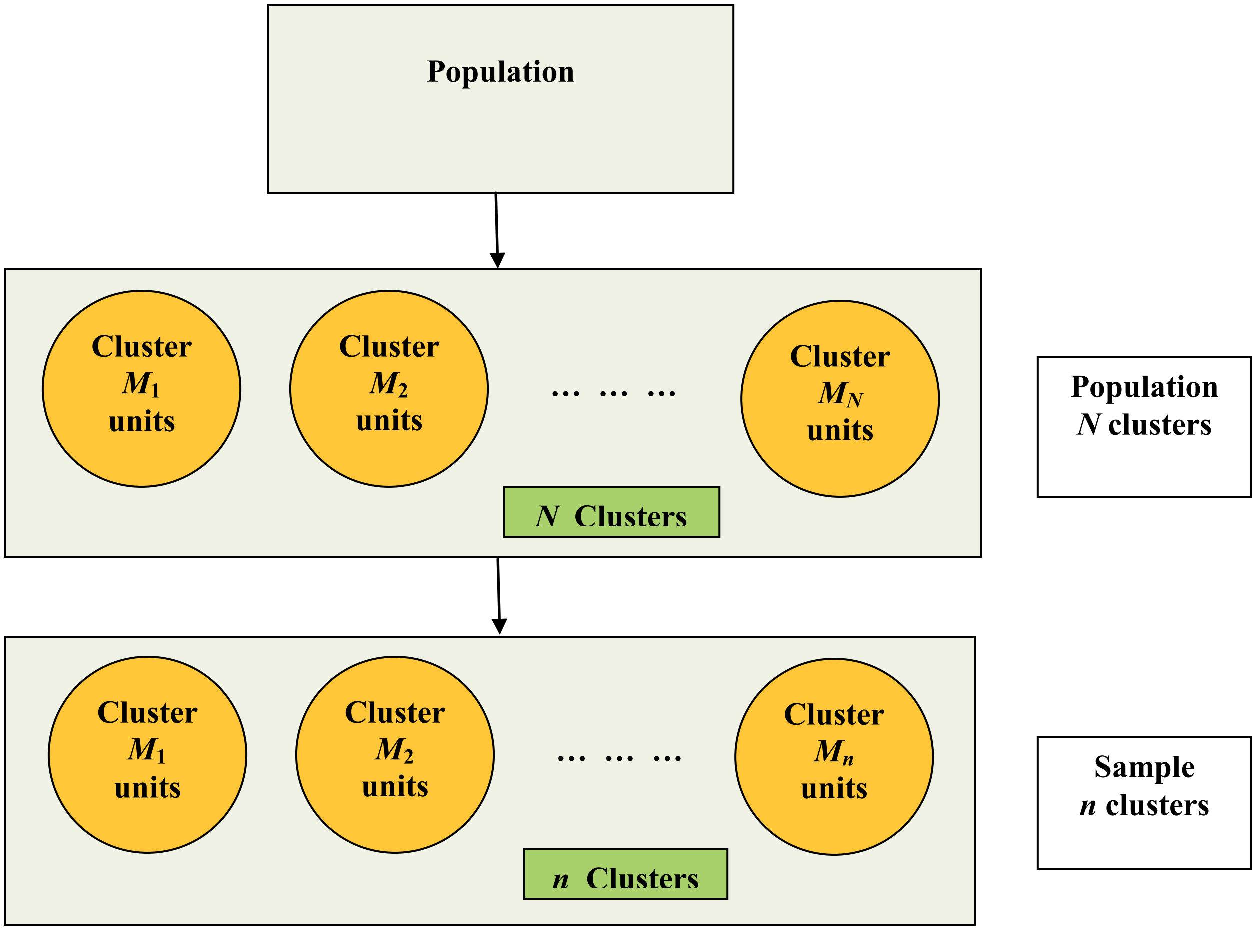
$$M_0 = \sum_{i=1}^{N} M_i$$

$$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i$$

$$\bar{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} \text{ : mean of } i^{th} \text{ cluster}$$

$$\begin{aligned}\bar{Y} &= \frac{1}{M_0}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \\ &= \sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_i\end{aligned}$$

Suppose that n clusters are selected with SRSWOR and all the elements in these selected clusters are surveyed. Assume that M_i's $(i = 1, 2, ..., N)$ are known.



Based on this scheme, several estimators can be obtained to estimate the population mean. We consider four type of such estimators.

1. Mean of cluster means:

Consider the simple arithmetic mean of the cluster means as

$$\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$E\left(\bar{\bar{y}}_c\right) = \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i$$

$$\neq \bar{Y} \ \text{ (where } \bar{Y} = \sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_i\text{)}.$$

The bias of $\bar{\bar{y}}_c$ is

$$\begin{aligned}
Bias\left(\bar{\bar{y}}_c\right) &= E\left(\bar{\bar{y}}_c\right) - \bar{Y} \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i - \sum_{i=1}^{N}\left(\frac{M_i}{M_0}\right)\bar{y}_i \\
&= -\frac{1}{M_0}\left[\sum_{i=1}^{N}M_i\bar{y}_i - \frac{M_0}{N}\sum_{i=1}^{N}\bar{y}_i\right] \\
&= -\frac{1}{M_0}\left[\sum_{i=1}^{N}M_i\bar{y}_i - \frac{\left(\sum_{i=1}^{N}M_i\right)\left(\sum_{i=1}^{N}\bar{y}_i\right)}{N}\right] \\
&= -\frac{1}{M_0}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y}) \\
&= -\left(\frac{N-1}{M_0}\right)S_{m\bar{y}}
\end{aligned}$$

$Bias\left(\bar{\bar{y}}_c\right) = 0$ if M_i and $\bar{y}_i$ are uncorrelated .

The mean squared error is

$$\begin{aligned}
MSE\left(\bar{\bar{y}}_c\right) &= Var\left(\bar{\bar{y}}_c\right) + \left[Bias\left(\bar{\bar{y}}_c\right)\right]^2 \\
&= \frac{N-n}{Nn}S_b^2 + \left(\frac{N-1}{M_0}\right)^2 S_{m\bar{y}}^2
\end{aligned}$$

where

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

$$S_{m\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y}).$$

An estimate of $Var\left(\overline{\overline{y}}_c\right)$ is

$$\widehat{Var}\left(\overline{\overline{y}}_c\right) = \frac{N-n}{Nn} s_b^2$$

where $s_b^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(\overline{y}_c - \overline{\overline{y}}_c\right)^2$.

2. Weighted mean of cluster means

Consider the arithmetic mean based on cluster total as

$$\begin{aligned}
\overline{y}_c^* &= \frac{1}{n\overline{M}}\sum_{i=1}^{n} M_i \overline{y}_i \\
E(\overline{y}_c^*) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\overline{M}}E(\overline{y}_i M_i) \\
&= \frac{n}{n}\frac{1}{M_0}\sum_{i=1}^{N} M_i \overline{y}_i \\
&= \frac{1}{M_0}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \\
&= \overline{Y}.
\end{aligned}$$

Thus $\overline{y}_c^*$ is an unbiased estimator of $\overline{Y}$. The variance of $\overline{y}_c^*$ and its estimate are given by

$$\begin{aligned}
Var(\overline{y}_c^*) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\overline{M}}\overline{y}_i\right) \\
&= \frac{N-n}{Nn} S_b^{*2} \\
\widehat{Var}(\overline{y}_c^*) &= \frac{N-n}{Nn} s_b^{*2}
\end{aligned}$$

where

$$\begin{aligned}
S_b^{*2} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\overline{M}}\overline{y}_i - \overline{Y}\right)^2 \\
s_b^{*2} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\overline{M}}\overline{y}_i - \overline{y}_c^*\right)^2 \\
E(s_b^{*2}) &= S_b^{*2}.
\end{aligned}$$

Note that the expressions of variance of $\overline{y}_c^*$ and its estimate can be derived using directly the theory of SRSWOR as follows:

Let $z_i = \frac{M_i}{\overline{M}} \overline{y}_i$, then $\overline{y}_c^* = \frac{1}{n} \sum_{i=1}^{n} z_i = \overline{z}$.

Since SRSWOR is followed, so

$$\begin{aligned}
Var(\overline{y}_c^*) = Var(\overline{z}) &= \frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^{n} (z_i - \overline{Y})^2 \\
&= \frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\overline{M}} \overline{y}_i - \overline{Y} \right)^2 \\
&= \frac{N-n}{Nn} S_b^{*2}.
\end{aligned}$$

Since

$$\begin{aligned}
E(s_b^{*2}) &= E\left[\frac{1}{n-1} \sum_{i=1}^{n} (z_i - \overline{z})^2 \right] \\
&= E\left[\frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{M_i}{\overline{M}} \overline{y}_i - \overline{y}_c^* \right)^2 \right] \\
&= \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{M_i}{\overline{M}} \overline{y}_i - \overline{Y} \right)^2 \\
&= S_b^{*2}
\end{aligned}$$

So an unbiased estimator of variance can be easily derived.

3. Estimator based on ratio method of estimation

Consider the weighted mean of the cluster means as

$$\overline{y}_c^{**} = \frac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i}$$

It is easy to see that this estimator is a biased estimator of population mean. Before deriving its bias and mean squared error, we note that this estimator can be derived using the philosophy of ratio method of estimation. To see this, consider the study variable U_i and auxiliary variable V_i as

$$U_i = \frac{M_i \bar{y}_i}{\bar{M}}$$

$$V_i = \frac{M_i}{\bar{M}} \; i = 1, 2, ..., N$$

$$\bar{V} = \frac{1}{N}\sum_{i=1}^{N} V_i = \frac{1}{N}\frac{\sum_{i=1}^{N} M_i}{\bar{M}} = 1$$

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i.$$

The ratio estimator based on U and V is

$$\begin{aligned}\hat{\bar{Y}}_R &= \frac{\bar{u}}{\bar{v}}\bar{V} \\ &= \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} v_i} \\ &= \frac{\sum_{i=1}^{n} \frac{M_i \bar{y}_i}{\bar{M}}}{\sum_{i=1}^{n} \frac{M_i}{\bar{M}}} \\ &= \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}.\end{aligned}$$

Since the ratio estimator is biased, so $\bar{y}_c^{**}$ is also a biased estimator. The approximate bias and mean squared errors of $\bar{y}_c^{**}$ can be derived directly by using the bias and *MSE* of ratio estimator. So using the results from the ratio method of estimation, the bias up to second order of approximation is given as follows

$$\begin{aligned}Bias(\bar{y}_c^{**}) &= \frac{N-n}{Nn}\left(\frac{S_v^2}{\bar{V}^2} - \frac{S_{uv}}{\bar{U}\bar{V}}\right)\bar{U} \\ &= \frac{N-n}{Nn}\left(S_v^2 - \frac{S_{uv}}{\bar{U}}\right)\bar{U}\end{aligned}$$

where $\bar{U} = \frac{1}{N}\sum_{i=1}^{N} U_i = \frac{1}{N\bar{M}}\sum_{i=1}^{N} M_i \bar{y}_i$

$$
\begin{aligned}
S_v^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(V_i-\bar{V})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}-1\right)^2 \\
S_{uv} &= \frac{1}{N-1}\sum_{i=1}^{N}(U_i-\bar{U})(V_i-\bar{V}) \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\bar{y}_i}{\bar{M}}-\frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)\left(\frac{M_i}{\bar{M}}-1\right) \\
R_{uv} &= \frac{\bar{U}}{\bar{V}} = \bar{U} = \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i.
\end{aligned}
$$

The *MSE* of $\bar{y}_c^{**}$ up to second order of approximation can be obtained as follows:

$$MSE(\bar{y}_c^{**}) = \frac{N-n}{Nn}\left(S_u^2 + R^2S_v^2 - 2RS_{uv}\right)$$

where $S_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i\bar{y}_i}{\bar{M}}-\frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)^2$

Alternatively,

$$
\begin{aligned}
MSE(\bar{y}_c^{**}) &= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(U_i - R_{uv}V_i\right)^2 \\
&= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{M_i\bar{y}_i}{\bar{M}}-\left(\frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i\right)\frac{M_i}{\bar{M}}\right]^2 \\
&= \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\right)^2\left[\bar{y}_i - \frac{\sum_{i=1}^{N}M_i\bar{y}_i}{N\bar{M}}\right]^2.
\end{aligned}
$$

An estimator of *MSE* can be obtained as

$$\widehat{MSE}(\bar{y}_c^{**}) = \frac{N-n}{Nn}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{M_i}{\bar{M}}\right)^2(\bar{y}_i - \bar{y}_c^{**})^2.$$

The estimator $\bar{y}_c^{**}$ is biased but consistent.

4. Estimator based on unbiased ratio type estimation

Since $\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$ (where $\bar{y}_i = \frac{1}{M_i}\sum_{i=1}^{M_i} y_{ij}$) is a biased estimator of population mean and

$$\begin{aligned} Bias(\bar{\bar{y}}_c) &= -\left(\frac{N-1}{M_0}\right)S_{m\bar{y}} \\ &= -\left(\frac{N-1}{N\bar{M}}\right)S_{m\bar{y}} \end{aligned}$$

Since SRSWOR is used, so

$$S_{m\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{n}(M_i - \bar{m})(\bar{y}_i - \bar{\bar{y}}_c),\ \bar{m} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

is an unbiased estimator of

$$S_{m\bar{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M})(\bar{y}_i - \bar{Y}),$$

i.e., $E(s_{m\bar{y}}) = S_{m\bar{y}}$.

So it follow that

$$E(\bar{\bar{y}}_c) - \bar{Y} = -\left(\frac{N-1}{N\bar{M}}\right)E(s_{m\bar{y}})$$

or $\quad E\left[\bar{\bar{y}}_c + \left(\frac{N-1}{N\bar{M}}\right)s_{m\bar{y}}\right] = \bar{Y}$.

So

$$\bar{\bar{y}}_c^{**} = \bar{\bar{y}}_c + \left(\frac{N-1}{N\bar{M}}\right)s_{m\bar{y}}$$

is an unbiased estimator of the population mean $\bar{Y}$.

This estimator is based on unbiased ratio type estimator. This can be obtained by replacing the study variable (earlier y_i) by $\frac{M_i}{\bar{M}}\bar{y}_i$ and auxiliary variable (earlier x_i) by $\frac{M_i}{\bar{M}}$. The exact variance of this estimate is complicated and does not reduces to a simple form. The approximate variance upto first order of approximation is

$$Var\left(\bar{\bar{y}}_c^{**}\right) = \frac{1}{n(N-1)}\sum_{i=1}^{N}\left[\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{Y}\right) - \left(\frac{1}{N\bar{M}}\sum_{i=1}^{N}\bar{y}_i\right)(M_i - \bar{M})\right]^2.$$

A consistent estimate of this variance is

$$\widehat{Var}\left(\bar{\bar{y}}_{\boldsymbol{c}}^{**}\right)=\frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\left(\frac{M_i}{\bar{M}}\bar{y}_i-\bar{y}_c)\right)-\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n}\bar{y}_i\right)\left(M_i-\frac{\sum_{i=1}^{n}M_i}{n}\right)\right]^2 .$$

The variance of $\bar{\bar{y}}_{\boldsymbol{c}}^{**}$ will be smaller than that of $\bar{y}_c^{**}$ (based on the ratio method of estimation) provided the regression coefficient of $\frac{M_i\bar{y}_i}{\bar{M}}$ on $\frac{\mathrm{M_i}}{\bar{M}}$ is nearer to $\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i$ than to $\frac{1}{M_0}\sum_{i=1}^{N}M_i\bar{y}_i$.

Comparison between SRS and cluster sampling:

In case of unequal clusters, $\sum_{i=1}^{n}M_i$ is a random variable such that

$$E\left(\sum_{i=1}^{n}M_i\right)=n\bar{M}.$$

Now if a sample of size $n\bar{M}$ is drawn from a population of size $N\bar{M}$, then the variance of corresponding sample mean based on SRSWOR is

$$\begin{aligned}Var(\bar{y}_{SRS})&=\frac{N\bar{M}-n\bar{M}}{N\bar{M}}\frac{S^2}{n\bar{M}}\\&=\frac{N-n}{Nn}\frac{S^2}{\bar{M}}.\end{aligned}$$

This variance can be compared with any of the four proposed estimators.

For example, in case of

$$\begin{aligned}\bar{y}_c^* &=\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\bar{y}_i\\ Var(\bar{y}_c^*)&=\frac{N-n}{Nn}S_b^{*2}\\&=\frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\bar{y}_i-\bar{Y}\right)^2.\end{aligned}$$

The relative efficiency of $\bar{y}_c^{**}$ relative to SRS based sample mean

$$\begin{aligned}E&=\frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_c^*)}\\&=\frac{S^2}{\bar{M}S_b^{*2}}.\end{aligned}$$

For $Var(\bar{y}_c^*)<Var(\bar{y}_{SRS})$, the variance between the clusters (S_b^{*2}) should be less. So the clusters should be formed in such a way that the variation between them is as small as possible.

Sampling with replacement and unequal probabilities (PPSWR)

In many practical situations, the cluster total for the study variable is likely to be positively correlated with the number of units in the cluster. In this situation, it is advantageous to select the clusters with probability proportional to the number of units in the cluster instead of with equal probability, or to stratify the clusters according to their sizes and then to draw a SRSWOR of clusters from each of the stratum. We consider here the case where clusters are selected with probability proportional to the number of units in the cluster and with replacement.

Suppose that n clusters are selected with ppswr, the size being the number of units in the cluster. Here P_i is the probability of selection assigned to the i^{th} cluster which is given by

$$P_i = \frac{M_i}{M_0} = \frac{M_i}{N\bar{M}}, \quad i = 1, 2, ..., N.$$

Consider the following estimator of the population mean:

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i.$$

Then this estimator can be expressed as

$$\hat{\bar{Y}}_c = \frac{1}{n} \sum_{i=1}^{N} \alpha_i \bar{y}_i$$

where α_i denotes the number of times the i^{th} cluster occurs in the sample. The random variables $\alpha_1, \alpha_2, ..., \alpha_N$ follow a multinomial probability distribution with

$$E(\alpha_i) = nP_i, \ Var(\alpha_i) = nP_i(1 - P_i)$$
$$Cov(\alpha_i, \alpha_j) = -nP_iP_j, \quad i \neq j.$$

Hence,

$$\begin{aligned} E(\hat{\bar{Y}}_c) &= \frac{1}{n} \sum_{i=1}^{N} E(\alpha_i) \bar{y}_i \\ &= \frac{1}{n} \sum_{i=1}^{N} nP_i \bar{y}_i \\ &= \sum_{i=1}^{N} \frac{M_i}{N\bar{M}} \bar{y}_i \\ &= \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{N\bar{M}} = \bar{Y}. \end{aligned}$$

Thus $\hat{\bar{Y}}_c$ is an unbiased estimator of $\bar{Y}$.

We now derive the variance of $\hat{\bar{Y}}_c$.

From $\hat{\bar{Y}}_c = \frac{1}{n}\sum_{i=1}^{N} \alpha_i \bar{y}_i,$

$$\begin{aligned}
Var(\hat{\bar{Y}}_c) &= \frac{1}{n^2}\left[\sum_{i=1}^{N} Var(\alpha_i)\bar{y}_i^2 + \sum_{i\neq j}^{N} Cov(\alpha_i,\alpha_j)\bar{y}_i\bar{y}_j\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} P_i(1-P_i)\bar{y}_i^2 - \sum_{i\neq j}^{N} P_iP_j\bar{y}_i\bar{y}_j\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} P_i\bar{y}_i^2 - \left(\sum_{i\neq j}^{N} P_i\bar{y}_i\right)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N} P_i\left(\bar{y}_i - \bar{Y}\right)^2 \\
&= \frac{1}{nN\bar{M}}\sum_{i=1}^{N} M_i(\bar{y}_i - \bar{Y})^2.
\end{aligned}$$

An unbiased estimator of the variance of $\hat{\bar{Y}}_c$ is

$$\widehat{Var}(\hat{\bar{Y}}_c) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2$$

which can be seen to satisfy the unbiasedness property as follows:

Consider

$$\begin{aligned}
&E\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2\right] \\
&\quad = E\left[\frac{1}{n(n-1)}\left(\sum_{i=1}^{n}(\bar{y}_i^2 - n\hat{\bar{Y}}_c)^2\right)\right] \\
&\quad = \frac{1}{n(n-1)}\left[E\left(\sum_{i=1}^{n}\alpha_i\bar{y}_i^2\right) - nVar(\hat{\bar{Y}}_c) - n\bar{Y}^2\right]
\end{aligned}$$

where $E(\alpha_i) = nP_i, Var(\alpha_i) = nP_i(1-P_i), Cov(\alpha_i,\alpha_j) = -nP_iP_j, i \neq j$

$$\begin{aligned}
E\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\bar{y}_i - \hat{\bar{Y}}_c)^2\right] &= \frac{1}{n(n-1)}\left[\sum_{i=1}^{N} n_iP_i\bar{y}_i^2 - n\frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 - n\bar{Y}^2\right] \\
&= \frac{1}{(n-1)}\left[\sum_{i=1}^{N} P_i(\bar{y}_i^2 - \bar{Y}^2) - \frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2\right] \\
&= \frac{1}{(n-1)}\left[\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2\right] \\
&= \frac{1}{(n-1)}\sum_{i=1}^{N} P_i(\bar{y}_i - \bar{Y})^2 \\
&= Var(\hat{\bar{Y}}_c).
\end{aligned}$$

Two Stage Sampling (Subsampling)

In cluster sampling, all the elements in the selected clusters are surveyed. Moreover, the efficiency in cluster sampling depends on size of the cluster. As the size increases, the efficiency decreases. It suggests that higher precision can be attained by distributing a given number of elements over a large number of clusters and then by taking a small number of clusters and enumerating all elements within them. This is achieved in subsampling.

In subsampling

- divide the population into clusters.
- Select a sample of clusters [first stage}
- From each of the selected cluster, select a sample of specified number of elements [second stage]

The clusters which form the units of sampling at the first stage are called the **first stage units** and the units or group of units within clusters which form the unit of clusters are called the **second stage units** or **subunits.**

The procedure is generalized to three or more stages and is then termed as **multistage sampling**.

For example, in a crop survey

- villages are the first stage units,
- fields within the villages are the second stage units and
- plots within the fields are the third stage units.

In another example, to obtain a sample of fishes from a commercial fishery

- first take a sample of boats and
- then take a sample of fishes from each selected boat.

Two stage sampling with equal first stage units:

Assume that

- population consists of NM elements.
- NM elements are grouped into N first stage units of M second stage units each, (i.e., N clusters, each cluster is of size M)
- Sample of n first stage units is selected (i.e., choose n clusters)

- Sample of m second stage units is selected from each selected first stage unit (i.e., choose m units from each cluster).
- Units at each stage are selected with SRSWOR.

Cluster sampling is a special case of two stage sampling in the sense that from a population of N clusters of equal size $m = M$, a sample of n clusters are chosen.

If further $M = m = 1,$ we get SRSWOR.

If $n = N,$ we have the case of stratified sampling.

y_{ij} : Value of the characteristic under study for the j^{th} second stage units of the i^{th} first stage unit; $i = 1, 2, ..., N;\ j = 1, 2, .., m.$

$\overline{Y}_i = \frac{1}{M}\sum_{j=1}^{m} y_{ij}$: mean per 2^{nd} stage unit of i^{th} 1^{st} stage units in the population.

$\overline{Y} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i = \overline{Y}_{MN}$: mean per second stage unit in the population

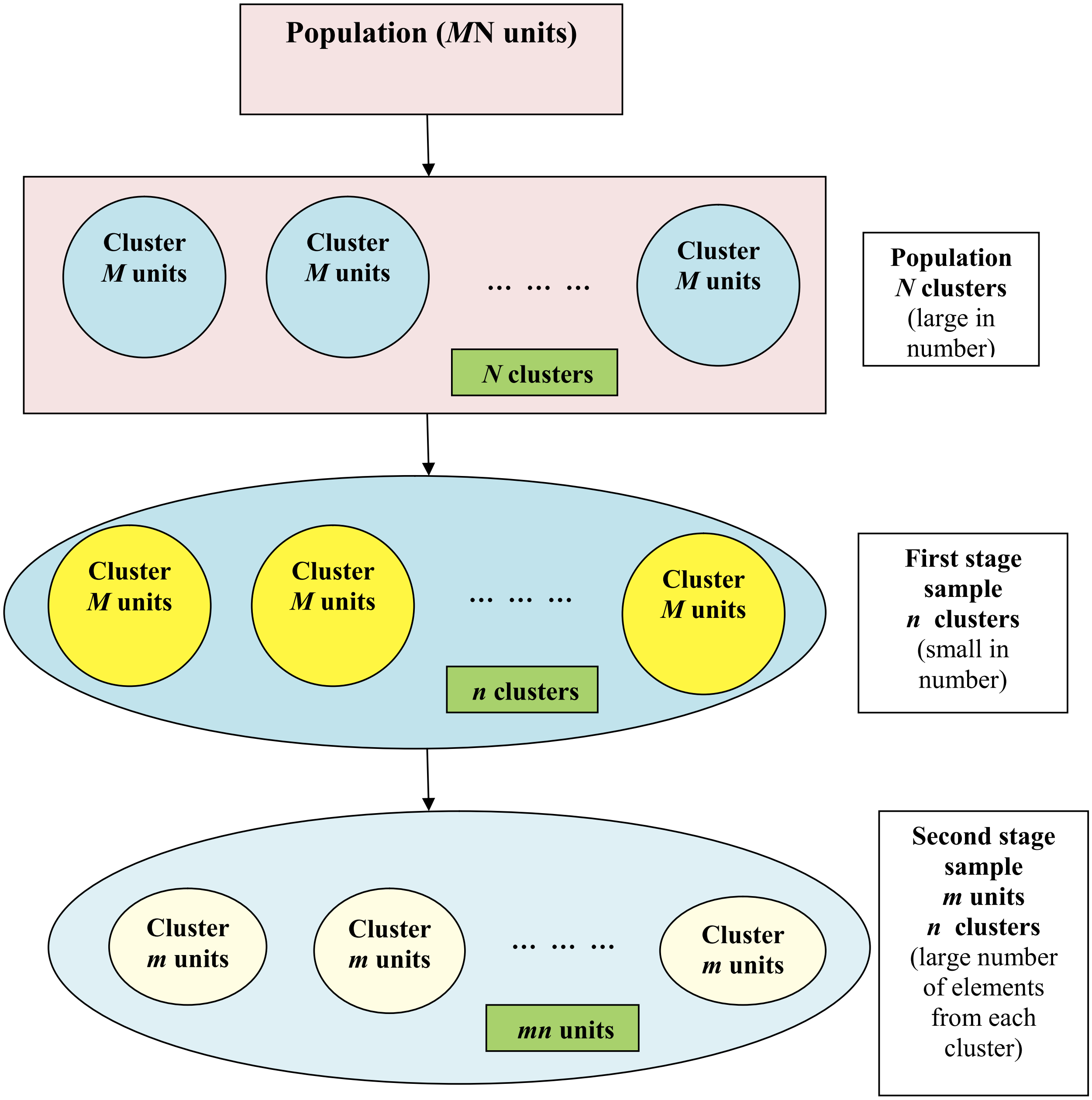
$\overline{y}_i = \frac{1}{n}\sum_{j=1}^{m} y_{ij}$: mean per second stage unit in the i^{th} first stage unit in the sample.

$\overline{y} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_i = \overline{y}_{mn}$: mean per second stage in the sample.

Advantages:

The principle advantage of two stage sampling is that it is more flexible than the one stage sampling. It reduces to one stage sampling when $m = M$ but unless this is the best choice of m, we have the opportunity of taking some smaller value that appears more efficient. As usual, this choice reduces to a balance between statistical precision and cost. When units of the first stage agree very closely, then consideration of precision suggests a small value of m. On the other hand, it is sometimes as cheap to measure the whole of a unit as to a sample. For example, when the unit is a household and a single respondent can give as accurate data as all the members of the household.

A pictorial scheme of two stage sampling scheme is as follows:



Note: The expectations under two stage sampling scheme depend on the stages. For example, the expectation at second stage unit will be dependent on first stage unit in the sense that second stage unit will be in the sample provided it was selected in the first stage.

To calculate the average

- First average the estimator over all the second stage selections that can be drawn from a fixed set of n units that the plan selects.
- Then average over all the possible selections of n units by the plan.

In case of two stage sampling,

$$\begin{array}{ccc} E(\hat{\theta}) & = & E_1[E_2(\hat{\theta})] \\ \downarrow & & \downarrow \qquad \searrow \end{array}$$

$$\boxed{\begin{array}{l}\text{average}\\ \text{over}\\ \text{all}\\ \text{samples}\end{array}} \quad \boxed{\begin{array}{l}\text{average}\\ \text{over all}\\ 1^{st}\text{ stage}\\ \text{samples}\end{array}} \quad \boxed{\begin{array}{l}\text{average over}\\ \text{all possible } 2^{nd}\text{stage}\\ \text{selections from}\\ \text{a fixed set of units}\end{array}}$$

In case of three stage sampling,

$$E(\hat{\theta}) = E_1\left[E_2\left\{E_3(\hat{\theta})\right\}\right].$$

To calculate the variance, we proceed as follows:

In case of two stage sampling,

$$\begin{aligned} Var(\hat{\theta}) &= E(\hat{\theta}-\theta)^2 \\ &= E_1E_2(\hat{\theta}-\theta)^2 \end{aligned}$$

Consider

$$\begin{aligned} E_2(\hat{\theta}-\theta)^2 &= E_2(\hat{\theta}^2) - 2\theta E_2(\hat{\theta}) + \theta^2 \\ &= \left[\left\{E_2(\hat{\theta}\right\}^2 + V_2(\hat{\theta})\right] - 2\theta E_2(\hat{\theta}) + \theta^2 \end{aligned}.$$

Now average over first stage selection as

$$\begin{aligned} E_1E_2(\hat{\theta}-\theta)^2 &= E_1\left[E_2(\hat{\theta})\right]^2 + E_1\left[V_2(\hat{\theta})\right] - 2\theta E_1E_2(\hat{\theta}) + E_1(\theta^2) \\ &= E_1\left[E_1\left\{E_2(\hat{\theta})\right\}^2 - \theta^2\right] + E_1\left[V_2(\hat{\theta})\right] \\ Var(\hat{\theta}) &= V_1\left[E_2(\hat{\theta})\right] + E_1\left[V_2(\hat{\theta})\right]. \end{aligned}$$

In case of three stage sampling,

$$Var(\hat{\theta}) = V_1\left[E_{\,2}\left\{E_3(\hat{\theta})\right\}\right] + E_1\left[V_2\left\{E_3(\hat{\theta})\right\}\right] + E_1\left[E_2\left\{V_3(\hat{\theta})\right\}\right].$$

Estimation of population mean:

Consider $\bar{y} = \bar{y}_{mn}$ as an estimator of the population mean $\bar{Y}$.

Bias:

Consider

$$
\begin{aligned}
E(\bar{y}) &= E_1\left[E_2(\bar{y}_{mn})\right] \\
&= E_1\left[E_2(\bar{y}_{im}|i)\right] \quad \text{(as } 2^{nd} \text{ stage is dependent on } 1^{st} \text{ stage)} \\
&= E_1\left[E_2(\bar{y}_{im}|i)\right] \quad \text{(as } y_i \text{ is unbiased for } \bar{Y}_i \text{ due to SRSWOR)} \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i \\
&= \bar{Y}.
\end{aligned}
$$

Thus $\bar{y}_{mn}$ is an unbiased estimator of the population mean.

Variance

$$
\begin{aligned}
Var(\bar{y}) &= E_1\left[V_2(\bar{y}|i)\right] + V_1\left[E_2(\bar{y}/i)\right] \\
&= E_1\left[V_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i|i\right\}\right] + V_1\left[E_2\left\{\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i/i\right\}\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}V(\bar{y}_i|i)\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\bar{y}_i/i)\right] \\
&= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2\right] + V_1\left[\frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)E_1(S_i^2)|i + V_1(\bar{y}_c)
\end{aligned}
$$

(where $\bar{y}_c$ is based on cluster means as in cluster sampling)

$$
\begin{aligned}
&= \frac{1}{n^2}n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \frac{N-n}{Nn}S_b^2 \\
&= \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2
\end{aligned}
$$

where $\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N}S_i^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij}-\bar{Y}_i\right)^2$

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

Estimate of variance

An unbiased estimator of variance of $\overline{y}$ can be obtained by replacing S_b^2 and $\overline{S}_w^2$ by their unbiased estimators in the expression of variance of $\overline{y}$.

Consider an estimator of

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

where $S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \overline{Y}_i\right)^2$

as $\overline{s}_w^2 = \frac{1}{n}\sum_{i=1}^{n} s_i^2$

where $s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(y_{ij} - \overline{y}_i)^2$.

So

$$\begin{aligned}
E(\overline{s}_w^2) &= E_1 E_2\left(\overline{s}_w^2 \middle| i\right) \\
&= E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n} s_i^2 \middle| i\right] \\
&= E_1 \frac{1}{n}\sum_{i=1}^{n}\left[E_2(s_i^2 | i)\right] \\
&= E_1 \frac{1}{n}\sum_{i=1}^{n} S_i^2 \qquad \text{(as SRSWOR is used)} \\
&= \frac{1}{n}\sum_{i=1}^{n} E_1(S_i^2) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N}\sum_{i=1}^{N} S_i^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} S_i^2 \\
&= \overline{S}_w^2
\end{aligned}$$

so $\overline{s}_w^2$ is an unbiased estimator of $\overline{S}_w^2$.

Consider

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2$$

as an estimator of

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\overline{Y}_i - \overline{Y})^2 \ .$$

So

$$E(s_b^2) = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2\right]$$

$$(n-1)E(s_b^2) = E\left[\sum_{i=1}^{n}\overline{y}_i^2 - n\overline{y}^2\right]$$

$$= E\left[\sum_{i=1}^{n}\overline{y}_i^2\right] - nE(\overline{y}^2)$$

$$= E_1\left[E_2\left(\sum_{i=1}^{n}\overline{y}_i^2\right)\right] - n\left[Var(\overline{y}) + \left\{E(\overline{y})\right\}^2\right]$$

$$= E_1\left[\sum_{i=1}^{n}E_2(\overline{y}_i^2)|i)\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= E_1\left[\sum_{i=1}^{n}\left\{Var(\overline{y}_i) + \left(E(\overline{y}_i)\right)^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= E_1\left[\sum_{i=1}^{n}\left\{\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \overline{Y}_i^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= nE_1\left[\frac{1}{n}\left\{\sum_{i=1}^{n}\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \overline{Y}_i^2\right\}\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= n\left[\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{N}\sum_{i=1}^{N}S_i^2 + \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_i^2\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= n\left[\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_i^2\right] - n\left[\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{n}\overline{S}_w^2 + \overline{Y}^2\right]$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + \frac{n}{N}\sum_{i=1}^{N}\overline{Y}_i^2 - n\overline{Y}^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + \frac{n}{N}\left[\sum_{i=1}^{N}\overline{Y}_i^2 - N\overline{Y}^2\right] - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + \frac{n}{N}(N-1)S_b^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2$$

$$= (n-1)\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + (n-1)S_b^2.$$

$$\Rightarrow E(s_b^2) = \left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + S_b^2$$

or $E\left[s_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\overline{s}_w^2\right] = S_b^2.$

Thus

$$\begin{aligned}\widehat{Var}(\bar{y}) &= \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\hat{\bar{S}}_{\omega}^{2} + \left(\frac{1}{n}-\frac{1}{N}\right)\hat{S}_b^2 \\ &= \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)\left[s_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2\right] \\ &= \frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)s_b^2.\end{aligned}$$

Allocation of sample to the two stages: Equal first stage units:

The variance of sample mean in the case of two stage sampling is

$$\widehat{Var}(\bar{y}) = \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2.$$

It depends on $S_b^2, \bar{S}_w^2,$ *n* and *m*. So the cost of survey of units in the two stage sample depends on *n* and *m*.

Case 1. When cost is fixed

We find the values of n and m so that the variance is minimum for given cost.

(I) **When cost function is *C* = *kmn***

Let the cost of survey be proportional to sample size as

$$C = knm$$

where C is the total cost and k is constant.

When cost is fixed as $C = C_0$. Substituting $m = \frac{C_0}{kn}$ in $Var(\bar{y}),$ we get

$$\begin{aligned}Var(\bar{y}) &= \frac{1}{n}\left[S_b^2 - \frac{\bar{S}_w^2}{M}\right] - \frac{S_b^2}{N} + \frac{1}{n}\frac{kn}{C_0}\bar{S}_w^2 \\ &= \frac{1}{n}\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) - \left(\frac{S_b^2}{N} - \frac{k\bar{S}_w^2}{C_0}\right).\end{aligned}$$

This variance is monotonic decreasing function of n if $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0.$ The variance is minimum when n assumes maximum value, i.e.,

$$\hat{n} = \frac{C_0}{k} \text{ corresponding to } m = 1.$$

If $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) < 0$ (i.e., intraclass correlation is negative for large N), then the variance is a monotonic increasing function of n, It reaches minimum when n assumes the minimum value, *i.e.,* $\hat{n} = \frac{C_0}{kM}$ (i.e., no subsampling).

(II) **When cost function is** $C = k_1 n + k_2 mn$

Let cost C be fixed as $C_0 = k_1 n + k_2 mn$ where k_1 and k_2 are positive constants. The terms k_1 and k_2 denote the costs of per unit observations in first and second stages respectively. Minimize the variance of sample mean under the two stage with respect to m subject to the restriction $C_0 = k_1 n + k_2 mn$.

We have

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + k_2\bar{S}_w^2 + mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + \frac{k_1\bar{S}_w^2}{m}.$$

When $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0$, then

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = \left[\sqrt{k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)} + \sqrt{k_2\bar{S}_w^2}\right]^2 + \left[\sqrt{mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)} - \sqrt{\frac{k_1\bar{S}_w^2}{m}}\right]^2$$

which is minimum when the second term of right hand side is zero. So we obtain

$$\hat{m} = \sqrt{\frac{k_1}{k_2}\frac{\bar{S}_w^2}{\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right)}}.$$

The optimum n follows from $C_0 = k_1 n + k_2 mn$ as

$$\hat{n} = \frac{C_0}{k_1 + k_2\hat{m}}.$$

When $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) \le 0$ then

$$C_0\left[Var(\bar{y}) + \frac{S_b^2}{N}\right] = k_1\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + k_2\bar{S}_w^2 + mk_2\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) + \frac{k_1\bar{S}_w^2}{m}$$

is minimum if m is the greatest attainable integer. Hence in this case, when

$$C_0 \ge k_1 + k_2 M;\ \hat{m} = M \text{ and } \hat{n} = \frac{C_0}{k_1 + k_2 M}.$$

If $C_0 \ge k_1 + k_2 M$; then $\hat{m} = \frac{C_0 - k_1}{k_2}$ and $\hat{n} = 1$.

If N is large, then $\overline{S}_w^2 \approx S^2(1-\rho)$

$$\overline{S}_w^2 - \frac{\overline{S}_w^2}{M} \approx \rho S^2$$

$$\hat{m} \approx \sqrt{\frac{k_1}{k_2}\left(\frac{1}{\rho}-1\right)}.$$

Case 2: When variance is fixed

Now we find the sample sizes when variance is fixed, say as V_0.

$$V_0 = \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2 + \left(\frac{1}{n}-\frac{1}{N}\right)\overline{S}_b^2$$

$$\Rightarrow n = \frac{S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\overline{S}_w^2}{V_0 + \frac{S_b^2}{N}}$$

So

$$C = kmn = km\left(\frac{S_b^2 - \frac{\overline{S}_w^2}{M}}{V_0 + \frac{S_b^2}{N}}\right) + \frac{k\overline{S}_w^2}{V_0 + \frac{S_b^2}{N}} \ .$$

If $\left(S_b^2 - \frac{\overline{S}_w^2}{M}\right) > 0$, C attains minimum when m assumes the smallest integral value, *i.e.*, 1.

If $\left(S_b^2 - \frac{\overline{S}_w^2}{M}\right) < 0$, C attains minimum when $\hat{m} = M$.

Comparison of two stage sampling with one stage sampling

One stage sampling procedures are comparable with two stage sampling procedures when either

(i) sampling mn elements in one single stage or

(ii) sampling $\frac{mn}{M}$ first stage units as cluster without sub-sampling.

We consider both the cases.

Case 1: Sampling mn elements in one single stage

The variance of sample mean based on

- mn elements selected by SRSWOR (one stage) is given by

$$V(\bar{y}_{SRS}) = \left(\frac{1}{mn} - \frac{1}{MN}\right) S^2$$

- two stage sampling is given by

$$V(\bar{y}_{TS}) = \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 .$$

The intraclass correlation coefficient is

$$\rho = \frac{M(N-1)S_b^2 - N\bar{S}_w^2}{(MN-1)S^2}; \; -\frac{1}{M-1} \le \rho \le 1$$

and using the identity

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y})^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_i - \bar{Y})^2$$

where $\bar{Y} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}, \bar{Y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}.$

We have

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

and

$$(MN-1)S^2 = -N\bar{S}_w^2 + M(N-1)S_b^2$$

$$\Rightarrow S_b^2 = \frac{(MN-1)S^2}{M^2(N-1)}\left[1 + (M-1)\rho\right] \qquad \text{(Eliminating } \bar{S}_w^2\text{)}$$

$$\bar{S}_w^2 = \left(\frac{MN-1}{MN}\right) S^2 (1-\rho).$$

Substituting S_b^2 and $\bar{S}_w^2$ in $Var(\bar{y}_{TS})$

$$V(\bar{y}_{TS}) = \left(\frac{MN-1}{MN}\right)\frac{S^2}{mn}\left[1 - \frac{m(n-1)}{M(N-1)} + \rho\left\{\frac{N-n}{N-1}\frac{m}{M}(M-1) - \frac{M-m}{M}\right\}\right].$$

When subsampling rate $\frac{m}{M}$ is small, $MN - 1 \approx MN$ and $M - 1 \approx M$, then

$$V(\bar{y}_{SRS}) = \frac{S^2}{mn}$$

$$V(\bar{y}_{TS}) = \frac{S^2}{mn}\left[1 + \rho\left(\frac{N-n}{N-1}m - 1\right)\right].$$

The relative efficiency of the two stage in relation to one stage sampling of SRSWOR is

$$RE = \frac{Var(\bar{y}_{TS})}{Var(\bar{y}_{SRS})} = 1 + \rho\left(\frac{N-n}{N-1}m - 1\right).$$

If $N-1 \approx N$ and finite population correction is ignorable, then $\frac{N-n}{N-1} \approx \frac{N-n}{N} \approx 1,$ then

$$RE = 1 + \rho(m-1).$$

Case 2: Comparison with cluster sampling

Suppose a random sample of $\frac{mn}{M}$ clusters, without further subsampling is selected.

The variance of the sample mean of equivalent mn/M clusters is

$$Var(\bar{y}_{cl}) = \left(\frac{M}{mn} - \frac{1}{N}\right)S_b^2 .$$

The variance of sample mean under the two stage sampling is

$$Var(\bar{y}_{TS}) = \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2 + \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 .$$

So $Var(\bar{y}_{cl})$ exceedes $Var(\bar{y}_{TS})$ by

$$\frac{1}{n}\left(\frac{M}{m} - 1\right)\left(S_b^2 - \frac{1}{M}\bar{S}_w^2\right)$$

which is approximately

$$\frac{1}{n}\left(\frac{M}{m} - 1\right)\rho S^2 \text{ for large } N \text{ and } \left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) > 0.$$

where $S_b^2 = \frac{MN-1}{M(N-1)}\frac{S^2}{M}\left[1 + \rho(M-1)\right]$

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

So smaller the m/M, larger the reduction in the variance of two stage sample over a cluster sample.

When $\left(S_b^2 - \frac{\bar{S}_w^2}{M}\right) < 0$ then the subsampling will lead to loss in precision.

Two stage sampling with unequal first stage units:

Consider two stage sampling when the first stage units are of unequal size and SRSWOR is employed at each stage.

Let

y_{ij} : value of j^{th} second stage unit of the i^{th} first stage unit.

M_i : number of second stage units in i^{th} first stage units $(i = 1, 2, ..., N)$.

$M_0 = \sum_{i=1}^{N} M_i$: total number of second stage units in the population.

m_i : number of second stage units to be selected from i^{th} first stage unit, if it is in the sample.

$m_0 = \sum_{i=1}^{n} m_i$: total number of second stage units in the sample.

$$\bar{y}_{i(m_i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$$

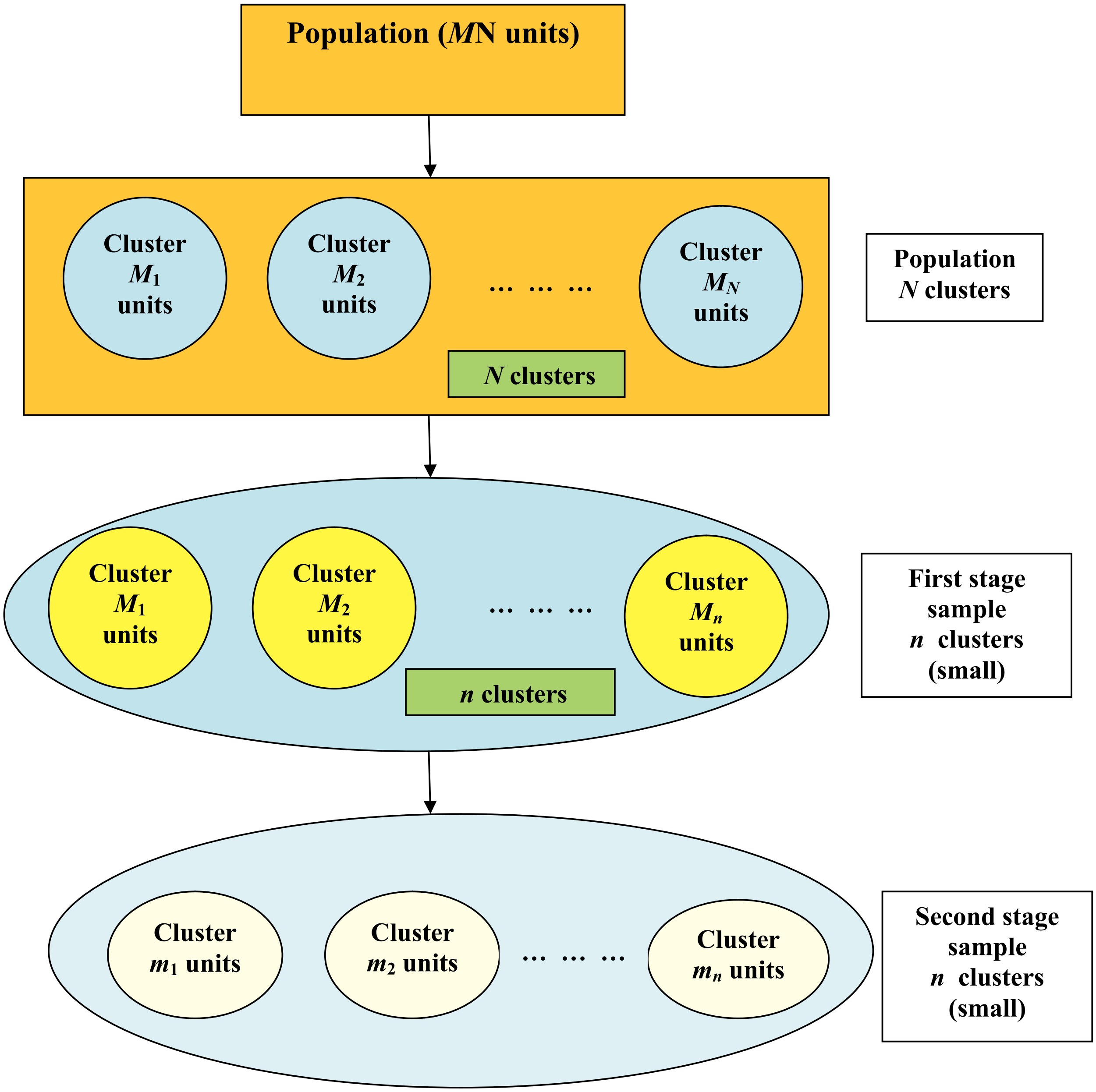
$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i = \bar{\bar{Y}}_N$$

$$\bar{Y} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i \bar{Y}_i}{\bar{M} N} = \frac{1}{N} \sum_{i=1}^{N} u_i \bar{Y}_i$$

$$u_i = \frac{M_i}{\bar{M}}$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

The pictorial scheme of two stage sampling with unequal first stage units case is as follows:


Population (MN units)
Cluster M_1 units
Cluster M_2 units
… … …
Cluster M_N units
N clusters
Population N clusters
Cluster M_1 units
Cluster M_2 units
… … …
Cluster M_n units
n clusters
First stage sample n clusters (small)
Cluster m_1 units
Cluster m_2 units
… … …
Cluster m_n units
Second stage sample n clusters (small)

Now we consider different estimators for the estimation of population mean.

1. Estimator based on the first stage unit means in the sample:

$$\hat{\overline{Y}} = \overline{y}_{S2} = \frac{1}{n}\sum_{i=1}^{n}\overline{y}_{i(m_i)}$$

Bias

$$\begin{aligned}
E(\overline{y}_{S2}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}\overline{y}_{i(m_i)}\right] \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}E_2(\overline{y}_{i(m_i)})\right] \\
&= E_1\left[\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i\right] \quad \text{[Since a sample of size } m_i \text{ is selected out of } M_i \text{ units by SRSWOR]} \\
&= \frac{1}{N}\sum_{i=1}^{N}\overline{Y}_i \\
&= \overline{\overline{Y}}_N \\
&\neq \overline{Y}.
\end{aligned}$$

So $\overline{y}_{S2}$ is a biased estimator of $\overline{Y}$ and its bias is given by

$$\begin{aligned}
Bias\ (\overline{y}_{S2}) &= E(\overline{y}_{S2}) - \overline{Y} \\
&= \frac{1}{N}\sum_{i=1}^{N}Y_i - \frac{1}{N\overline{M}}\sum_{i=1}^{N}M_i\overline{Y}_i \\
&= -\frac{1}{N\overline{M}}\left[\sum_{i=1}^{N}M_i\overline{Y}_i - \frac{1}{N}\left(\sum_{i=1}^{N}\overline{Y}_i\right)\left(\sum_{i=1}^{N}M_i\right)\right] \\
&= \frac{1}{N\overline{M}}\sum_{i=1}^{N}(M_i - \overline{M})(\overline{Y}_i - \overline{\overline{Y}}_N).
\end{aligned}$$

This bias can be estimated by

$$\widehat{Bias}(\overline{y}_{S2}) = -\frac{N-1}{N\overline{M}(n-1)}\sum_{i=1}^{n}(M_i - \overline{m})(\overline{y}_{i(mi)} - \overline{y}_{S2})$$

which can be seen as follows:

$$\begin{aligned}
E\left[\widehat{Bias}(\overline{y}_{S2})\right] &= -\frac{N-1}{N\overline{M}}E_1\left[\frac{1}{n-1}\sum_{i=1}^{n}E_2\left\{(M_i - \overline{m})(\overline{y}_{i(mi)} - \overline{y}_{S2})/n\right\}\right] \\
&= -\frac{N-1}{N\overline{M}}E\left[\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{m})(\overline{Y}_i - \overline{\overline{y}}_n)\right] \\
&= -\frac{1}{N\overline{M}}\sum_{i=1}^{N}(M_i - \overline{M})(\overline{Y}_i - \overline{\overline{Y}}_N) \\
&= \overline{\overline{Y}}_N - \overline{Y}
\end{aligned}$$

where $\overline{\overline{y}}_n = \frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i.$

An unbiased estimator of the population mean $\bar{Y}$ is thus obtained as

$$\bar{y}_{S2}+\frac{N-1}{N\bar{M}}\frac{1}{N-1}\sum_{i=1}^{n}(M_i-\bar{m})(\bar{y}_{i(mi)}-\bar{y}_{S2}).$$

Note that the bias arises due to the inequality of sizes of the first stage units and probability of selection of second stage units varies from one first stage to another.

Variance:

$$\begin{aligned}
Var(\bar{y}_{S2}) &= E\left[Var(\bar{y}_{S2}|n)\right]+Var\left[E(\bar{y}_{S2}|n)\right]\\
&= Var\left[\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right]+E\left[\frac{1}{n^2}\sum_{i=1}^{n}Var(\bar{y}_{i(mi)}|i)\right]\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2+E\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2\right]\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2+\frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2
\end{aligned}$$

where $S_b^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i-\bar{\bar{Y}}_N\right)^2$

$$S_i^2=\frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(y_{ij}-\bar{Y}_i\right)^2.$$

The *MSE* can be obtained as

$$MSE(\bar{y}_{S2})=Var(\bar{y}_{S2})+\left[Bias(\bar{y}_{S2})\right]^2.$$

Estimation of variance:

Consider mean square between cluster means in the sample

$$s_b^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_{i(mi)}-\bar{y}_{S2}\right)^2.$$

It can be shown that

$$E(s_b^2)=S_b^2+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2.$$

Also $s_i^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}(y_{ij}-\bar{y}_{i(mi)})^2$

$$E(s_i^2)=S_i^2=\frac{1}{M_i-1}\sum_{j=1}^{M_i}(y_{ij}-\bar{Y}_i)^2$$

So $E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)s_i^2\right]=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2.$

Thus

$$E(s_b^2)=S_b^2+E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i}-\frac{1}{M_i}\right)s_i^2\right]$$

and an unbiased estimator of S_b^2 is

$$\hat{S}_b^2 = s_b^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2 .$$

So an estimator of the variance can be obtained by replacing S_b^2 and S_i^2 by their unbiased estimators as

$$\widehat{Var}(\bar{y}_{S2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{S}_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)\hat{S}_i^2 .$$

2. Estimation based on first stage unit totals:

$$\begin{aligned}\hat{\bar{Y}} = \bar{y}^*_{S2} &= \frac{1}{n}\sum_{i=1}^{n}\frac{M_i\bar{y}_{i(mi)}}{\bar{M}} \\ &= \frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\end{aligned}$$

where $u_i = \dfrac{M_i}{\bar{M}}$.

Bias

$$\begin{aligned}E(y^*_{S2}) &= E\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}u_iE_2(\bar{y}_{i(mi)}|i)\right] \\ &= E\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{Y}_i\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}u_i\bar{Y}_i \\ &= \bar{Y}.\end{aligned}$$

Thus $\bar{y}^*_{S2}$ is an unbiased estimator of $\bar{Y}$.

Variance:

$$\begin{aligned}Var(\bar{y}^*_{S2}) &= Var\left[E(\bar{y}^*_{S2}|n)\right] + E\left[Var(\bar{y}^*_{S2}|n)\right] \\ &= Var\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{Y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2Var(\bar{y}_{i(mi)}|i)\right] \\ &= \left(\frac{1}{n} - \frac{1}{N}\right)S_b^{*2} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_i^2\end{aligned}$$

where $S_i^2 = \dfrac{1}{M_i - 1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$

$$S_b^{*2} = \frac{1}{N-1}\sum_{j=1}^{N}(u_i\bar{Y}_i - \bar{Y})^2.$$

3. Estimator based on ratio estimator:

$$\begin{aligned}\hat{\bar{Y}} = \bar{y}_{S2}^{**} &= \frac{\sum_{i=1}^{n} M_i \bar{y}_{i(mi)}}{\sum_{i=1}^{n} M_i} \\ &= \frac{\sum_{i=1}^{n} u_i \bar{y}_{i(mi)}}{\sum_{i=1}^{n} u_i} \\ &= \frac{\bar{y}_{S2}^{*}}{\bar{u}_n}\end{aligned}$$

where $u_i = \dfrac{M_i}{\bar{M}}$, $\bar{u}_n = \dfrac{1}{n}\sum_{i=1}^{n} u_i.$

This estimator can be seen as if arising by the ratio method of estimation as follows:

Let $y_i^* = u_i \bar{y}_{i(mi)}$

$$x_i^* = \frac{M_i}{\bar{M}},\ i = 1, 2, ..., N$$

be the values of study variable and auxiliary variable in reference to the ratio method of estimation. Then

$$\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i^* = \bar{y}_{S2}^*$$

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^* = \bar{u}_n$$

$$\bar{X}* = \frac{1}{N}\sum_{i=1}^{N} X_i^* = 1.$$

The corresponding ratio estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_R = \frac{\bar{y}*}{\bar{x}*}\bar{X}* = \frac{\bar{y}_{S2}^*}{\bar{u}_n}1 = \bar{y}_{S2}^{**}.$$

So the bias and mean squared error of $\bar{y}_{S2}^{**}$ can be obtained directly from the results of ratio estimator.

Recall that in ratio method of estimation, the bias and MSE of the ratio estimator upto second order of approximation

is

$$Bias\,(\hat{\bar{y}}_R) \approx \frac{N-n}{Nn}\bar{Y}(C_x^2 - 2\rho C_x C_y)$$

$$= \bar{Y}\left[\frac{Var(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}}\right]$$

$$MSE\,(\hat{\bar{Y}}_R) \approx \left[Var(\bar{y}) + R^2 Var(\bar{x}) - 2RCov(\bar{x},\bar{y})\right]$$

where $R = \dfrac{\bar{Y}}{\bar{X}}.$

Bias:

The bias of $\bar{y}_{S2}^{**}$ up to second order of approximation is

$$Bias(\bar{y}_{S2}^{**}) = \bar{Y}\left[\frac{Var(\bar{x}_{S2}^{*})}{\bar{X}^2} - \frac{Cov(\bar{x}_{S2}^{*},\bar{y}_{S2}^{*})}{\bar{X}\bar{Y}}\right]$$

where $\bar{x}_{S2}^{*}$ is the mean of auxiliary variable similar to $\bar{y}_{S2}^{*}$ as $\bar{x}_{S2}^{*} = \dfrac{1}{n}\sum_{i=1}^{n}\bar{x}_{i(mi)}.$

Now we find $Cov(\bar{x}_{S2}^{*},\bar{y}_{S2}^{*}).$

$$Cov(\bar{x}_{S2}^{*},\bar{y}_{S2}^{*}) = Cov\left[E\left(\frac{1}{n}\sum_{i=1}^{n}u_i\bar{x}_{i(mi)},\frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right)\right] + E\left[Cov\left(\frac{1}{n}\sum_{i=1}^{n}u_i\bar{x}_{i(mi)},\frac{1}{n}\sum_{i=1}^{n}u_i\bar{y}_{i(mi)}\right)\right]$$

$$= Cov\left[\frac{1}{n}\sum_{i=1}^{n}u_i E(\bar{x}_{i(mi)}),\frac{1}{n}\sum_{i=1}^{n}u_i E(\bar{y}_{i(mi)})\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2 Cov(\bar{x}_{i(mi)},\bar{y}_{i(mi)})\,\middle|\, i\right]$$

$$= Cov\left[\frac{1}{n}\sum_{i=1}^{n}u_i\bar{X}_i,\frac{1}{n}\sum_{i=1}^{n}u_i\bar{Y}_i\right] + E\left[\frac{1}{n^2}\sum_{i=1}^{n}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{ixy}\right]$$

$$= \left(\frac{1}{n}-\frac{1}{N}\right)S_{bxy}^{*} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{ixy}$$

where

$$S_{bxy}^{*} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{X}_i - \bar{X})(u_i\bar{Y}_i - \bar{Y})$$

$$S_{ixy} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}(x_{ij} - \bar{X}_i)(y_{ij} - \bar{Y}_i).$$

Similarly, $Var(\bar{x}_{S2}^{*})$ can be obtained by replacing x in place of y in $Cov(\bar{x}_{S2}^{*},\bar{y}_{S2}^{*})$ as

$$Var(\bar{x}_{S2}^{*}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{bx}^{*2} + \frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{ix}^2$$

where $S_{bx}^{*2} = \dfrac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{X}_i - \bar{X})^2$

$$S_{ix}^{*2} = \frac{1}{M_i - 1}\sum_{i=1}^{M_i}(x_{ij} - \bar{X}_i)^2.$$

Substituting $Cov(\overline{x}_{S2}^{*},\overline{y}_{S2}^{*})$ and $Var(\overline{x}_{S2}^{*})$ in $Bias(\overline{y}_{S2}^{**})$, we obtain the approximate bias as

$$Bias(\overline{y}_{S2}^{**}) \approx \overline{Y}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{S_{bx}^{*2}}{\overline{\overline{X}}^{2}}-\frac{S_{bxy}^{*}}{\overline{\overline{X}}\,\overline{\overline{Y}}}\right)+\frac{1}{nN}\sum_{i=1}^{N}\left\{u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\left(\frac{S_{ix}^{2}}{\overline{\overline{X}}^{2}}-\frac{S_{ixy}}{\overline{\overline{X}}\,\overline{\overline{Y}}}\right)\right\}\right].$$

Mean squared error

$$MSE(\overline{y}_{S2}^{**}) \approx Var(\overline{y}_{S2}^{*})-2R^{*}Cov(\overline{x}_{S2}^{*},\overline{y}_{S2}^{*})+R^{*2}Var(\overline{x}_{S2}^{*})$$

$$Var(\overline{y}_{S2}^{**}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{by}^{*2}+\frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{iy}^{2}$$

$$Var(\overline{x}_{S2}^{**}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{bx}^{*2}+\frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{ix}^{2}$$

$$Cov(\overline{x}_{S2}^{*},\overline{y}_{S2}^{**}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{bxy}^{*}+\frac{1}{nN}\sum_{i=1}^{N}u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_{ixy}$$

where

$$S_{by}^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\overline{Y}_i-\overline{Y})^2$$

$$S_{iy}^{*2} = \frac{1}{M_i-1}\sum_{j=1}^{M_i}(y_{ij}-\overline{Y}_i)^2$$

$$R^{*} = \frac{\overline{Y}}{\overline{\overline{X}}} = \overline{Y}.$$

Thus

$$MSE(\overline{y}_{S2}^{**}) \approx \left(\frac{1}{n}-\frac{1}{N}\right)\left(S_{by}^{*2}-2R^{*}S_{bxy}^{*}+R^{*2}S_{bx}^{*2}\right)+\frac{1}{nN}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\left(S_{iy}^{2}-2R^{*}S_{ixy}+R^{*2}S_{ix}^{2}\right)\right].$$

Also

$$MSE(\overline{y}_{S2}^{**}) \approx \left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}u_i^2\left(\overline{Y}_i-R^{*}\overline{X}_i\right)^2+\frac{1}{nN}\sum_{i=1}^{N}\left[u_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)\left(S_{iy}^{2}-2R^{*}S_{ixy}+R^{*2}S_{ix}^{2}\right)\right].$$

Estimate of variance

Consider

$$s_{bxy}^{*} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(u_i\overline{y}_{i(mi)}-\overline{y}_{S2}^{*}\right)\left(u_i\overline{x}_{i(mi)}-\overline{x}_{S2}^{*}\right)\right]$$

$$s_{ixy} = \frac{1}{m_i-1}\sum_{j=1}^{n}\left[\left(x_{ij}-\overline{x}_{i(mi)}\right)\left(y_{ij}-\overline{y}_{i(mi)}\right)\right].$$

It can be shown that

$$E\left(s_{bxy}^{*}\right) = S_{bxy}^{*} + \frac{1}{N}\sum_{i=1}^{N} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{ixy}$$

$$E(s_{ixy}) = S_{ixy}.$$

So

$$E\left[\frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ixy}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{ixy}\right].$$

Thus

$$\hat{S}_{bxy}^{*} = s_{bxy}^{*} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ixy}$$

$$\hat{S}_{bx}^{*2} = s_{bx}^{*2} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ix}^2$$

$$\hat{S}_{by}^{*2} = s_{by}^{*2} - \frac{1}{n}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iy}^2.$$

Also

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{ix}^2\right\}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{ix}^2\right]$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left\{u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_{iy}^2\right\}\right] = \frac{1}{N}\sum_{i=1}^{N}\left[u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iy}^2\right].$$

A consistent estimator of *MSE* of $\overline{y}_{S2}^{**}$ can be obtained by substituting the unbiased estimators of respective

statistics in $MSE(\overline{y}_{S2}^{**})$ as

$$\widehat{MSE}(\overline{y}_{S2}^{**}) \approx \left(\frac{1}{n} - \frac{1}{N}\right)\left(s_{by}^{*2} - 2r^{*}s_{bxy}^{*} + r^{*2}s_{bx}^{*2}\right)$$

$$+ \frac{1}{nN}\sum_{i=1}^{n} u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(s_{iy}^2 - 2r^{*}s_{ixy} + r^{*2}s_{ix}^2\right)$$

$$\approx \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{y}_{i(mi)} - r^{*}\overline{x}_{i(mi)}\right)^2$$

$$+ \frac{1}{nN}\sum_{i=1}^{n}\left[u_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right)\left(s_{iy}^2 - 2r^{*}s_{ixy} + r^{*2}s_{ix}^2\right)\right]$$

where $r^{*} = \dfrac{\overline{y}_{S2}^{*}}{\overline{x}_{S2}^{*}}$.

Double Sampling (Two Phase Sampling)

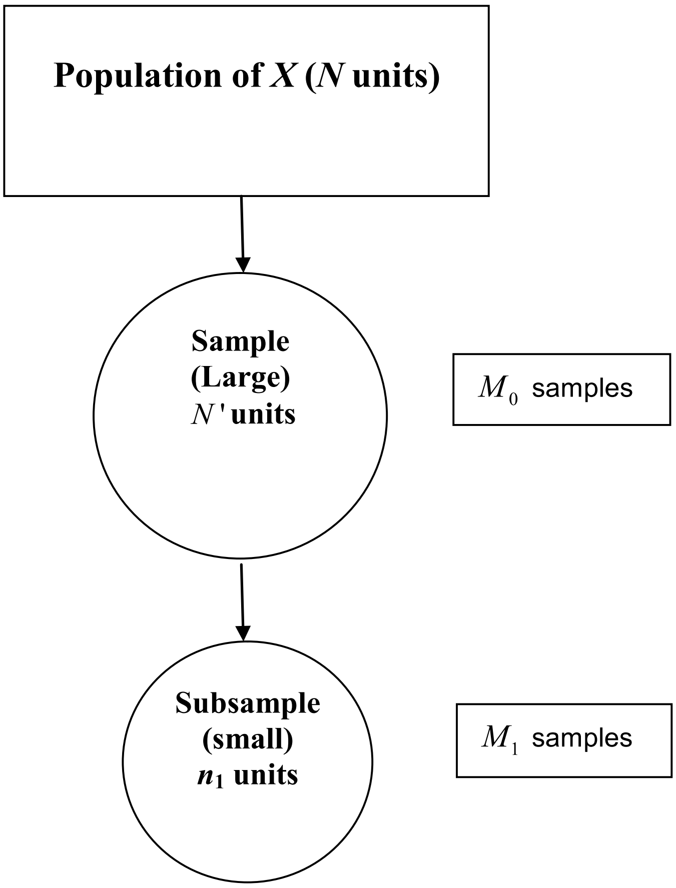
The ratio and regression methods of estimation require the knowledge of population mean of auxiliary variable $(\bar{X})$ to estimate the population mean of study variable $(\bar{Y})$. If information on the auxiliary variable is not available, then there are two options – one option is to collect a sample only on study variable and use sample mean as an estimator of population mean.

An alternative solution is to use a part of the budget for collecting information on auxiliary variable to collect a large preliminary sample in which x_i alone is measured. The purpose of this sampling is to furnish a good estimate of $\bar{X}$. This method is appropriate when the information about x_i is on file cards that have not been tabulated. After collecting a large preliminary sample of size n' units from the population, select a smaller sample of size n from it and collect the information on y. These two estimates are then used to obtain an estimator of population mean $\bar{Y}$. This procedure of selecting a large sample for collecting information on auxiliary variable x and then selecting a sub-sample from it for collecting the information on the study variable y is called double sampling or two phase sampling. It is useful when it is considerably cheaper and quicker to collect data on x than y and there is high correlation between x and y.

In this sampling, the randomization is done twice. First a random sample of size n' is drawn from a population of size N and then again a random sample of size n is drawn from the first sample of size n'

So the sample mean in this sampling is a function of the two phases of sampling. If SRSWOR is utilized to draw the samples at both the phases, then

- number of possible samples at the first phase when a sample of size n is drawn from a population of size N is $\binom{N}{n'} = M_0$, say.
- number of possible samples at the second phase where a sample of size n is drawn from the first phase sample of size n' is $\binom{n'}{n} = M_1$, say.



Population of X (N units)
Sample (Large) N' units
M_0 samples
Subsample (small) n_1 units
M_1 samples

Then the sample mean is a function of two variables. If τ is the statistic calculated at the second phase such that $\tau_{ij}, i = 1, 2, ..., M_0,\ j = 1, 2, ..., M_1$ with P_{ij} being the probability that i^{th} sample is chosen at first phase and j^{th} sample is chosen at second phase, then

$$E(\tau) = E_1\left[E_2(\tau)\right]$$

where $E_2(\tau)$ denotes the expectation over second phase and E_1 denotes the expectation over the first phase. Thus

$$\begin{aligned} E(\tau) &= \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_{ij}\tau_{ij} \\ &= \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_i P_{j/i}\,\tau_{ij} \quad (\text{using } P(A \cap B) = P(A)P(B/A)) \\ &= \underbrace{\sum_{i=1}^{M_0} P_i}_{1^{st}\text{ stage}}\ \underbrace{\sum_{j=1}^{M_1} P_{j/i}\,\tau_{ij}}_{2^{nd}\text{ stage}} \end{aligned}$$

2

Variance of τ

$$
\begin{aligned}
Var(\tau) &= E\left[\tau - E(\tau)\right]^2 \\
&= E\left[(\tau - E_2(\tau)) + (E_2(\tau) - E(\tau))\right]^2 \\
&= E\left[\tau - E_2(\tau)\right]^2 + \left[E_2(\tau) - E(\tau)\right]^2 + 0 \\
&= E_1 E_2\left[\tau - E_2(\tau)\right]^2 + [E_2(\tau) - E(\tau)\left.\right]^2 \\
&= E_1 E_2\left[\tau - E_2(\tau)\right]^2 + \underbrace{E_1 E_2\left[E_2(\tau) - E(\tau)\right]^2}_{\substack{\downarrow \\ \text{constant for } E_2}} \\
&= E_1\left[V_2(\tau)\right] + E_1\left[E_2(\tau) - E_1(E_2(\tau))\right]^2 \\
&= E_1\left[V_2(\tau)\right] + V_1\left[E_2(\tau)\right]
\end{aligned}
$$

Note: The two phase sampling can be extended to more than two phases depending upon the need and objective of the experiment. Various expectations can also be extended on the similar lines .

Double sampling in ratio method of estimation

If the population mean $\bar{X}$ is not known then double sampling technique is applied. Take a large initial sample of size n' by SRSWOR to estimate the population mean $\bar{X}$ as

$$\hat{\bar{X}} = \bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i \,.$$

Then a second sample is a subsample of size n selected from the initial sample by SRSWOR. Let $\bar{y}$ and $\bar{x}$ be the means of y and x based on the subsample. Then $E(\bar{x}') = \bar{X},\ E(\bar{x}) = \bar{X},\ E(\bar{y}) = \bar{Y}.$

The ratio estimator under double sampling now becomes

$$\hat{\bar{Y}}_{Rd} = \frac{\bar{y}}{\bar{x}}\bar{x}'.$$

The exact expressions for the bias and mean squared error of $\hat{\bar{Y}}_{Rd}$ are difficult to derive. So we find their approximate expressions using the same approach mentioned while describing the ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \quad \varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}, \quad \varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2$$

$$\begin{aligned}
E(\varepsilon_1 \varepsilon_2) &= \frac{1}{\bar{X}^2} E(\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \\
&= \frac{1}{\bar{X}^2} E_1\left[E_2(\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \mid n' \right] \\
&= \frac{1}{\bar{X}^2} E_1\left[(\bar{x}' - \bar{X})^2 \right] \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2} \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) C_x^2 \\
&= E(\varepsilon_2^2).
\end{aligned}$$

$$\begin{aligned}
E(\varepsilon_0 \varepsilon_2) &= Cov(\bar{y}, \bar{x}') \\
&= Cov\left[E(\bar{y} \mid n'), E(\bar{x}' \mid n') \right] + E\left[Cov(\bar{y}, \bar{x}') \mid n' \right] \\
&= Cov\left[\bar{Y}, \bar{X} \right] + E\left[Cov(\bar{y}', \bar{x}') \right] \\
&= Cov\left[(\bar{y}', \bar{x}' \right] \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}} \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}} \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho C_x C_y
\end{aligned}$$

where $\bar{y}'$ is the sample mean of $y's$ based on the sample size n'.

$$
\begin{aligned}
E(\varepsilon_0\varepsilon_1) &= \frac{1}{\overline{x}\,\overline{y}}Cov(\overline{y},\overline{x})\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xy}}{\overline{X}\,\overline{Y}}\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)\rho\frac{S_x}{\overline{X}}\frac{S_y}{\overline{Y}}\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y
\end{aligned}
$$

$$
\begin{aligned}
E(\varepsilon_2^2) &= \frac{1}{\overline{Y}^2}Var(\overline{y})\\
&= \frac{1}{\overline{Y}^2}\left[V_1\left\{E_2(\overline{y}\,|\,n')\right\}+E_1\left\{V_2(\overline{y}_n\,|\,n')\right\}\right]\\
&= \frac{1}{\overline{Y}^2}\left[V_1(\overline{y}_n^{'})+E_1\left\{\left(\frac{1}{n}-\frac{1}{n'}\right)s_y^{'2}\right\}\right]\\
&= \frac{1}{\overline{Y}^2}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2+\left(\frac{1}{n}-\frac{1}{n'}\right)S_y^2\right]\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_y^2}{\overline{Y}^2}\\
&= \left(\frac{1}{n}-\frac{1}{N}\right)C_y^2
\end{aligned}
$$

where $s_y^{'2}$ is the mean sum of squares of y based on initial sample of size n'.

$$
\begin{aligned}
E(\varepsilon_1\varepsilon_2) &= \frac{1}{\overline{X}^2}Cov(\overline{x},\overline{x}')\\
&= \frac{1}{\overline{X}^2}\left[Cov\left\{E(\overline{x}\,|\,n'),E(\overline{x}'\,|\,n')\right\}+0\right]\\
&= \frac{1}{\overline{X}^2}Var(\overline{X}')
\end{aligned}
$$

where $Var(\overline{X}')$ is the variance of mean of x based on initial sample of size n'.

Estimation error of $\hat{\bar{Y}}_{Rd}$

Write $\hat{\bar{Y}}_{Rd}$ *as*

$$\begin{aligned}\hat{\bar{Y}}_{Rd} &= \frac{(1+\varepsilon_0)}{(1+\varepsilon_1)}(1+\varepsilon_2)\frac{\bar{Y}}{\bar{X}}\bar{X}\\ &= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1+\varepsilon_1)^{-1}\\ &= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1-\varepsilon_1+\varepsilon_1^2-...)\\ &\simeq \bar{Y}(1+\varepsilon_0+\varepsilon_2+\varepsilon_0\varepsilon_2-\varepsilon_1-\varepsilon_o\varepsilon_1-\varepsilon_1\varepsilon_2+\varepsilon_1^2)\end{aligned}$$

upto the terms of order two. Other terms of degree greater than two are assumed to be negligible.

Bias of $\bar{Y}_{Rd}$

$$\begin{aligned}E(\hat{\bar{Y}}_{Rd}) &= \bar{Y}\left[1+0+0+E(\varepsilon_0\varepsilon_2)-0-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]\\ Bias(\hat{\bar{Y}}_{Rd}) &= E(\hat{\bar{Y}}_{Rd})-\bar{Y}\\ &= \bar{Y}\left[E(\varepsilon_0\varepsilon_2)-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]\\ &= \bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_xC_y-\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_xC_y-\left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2+\left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right]\\ &= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_x^2-\rho C_xC_y\right)\\ &= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)C_x(C_x-\rho C_y).\end{aligned}$$

The bias is negligible if n is large and relative bias vanishes if $C_x^2=C_{xy}$, i.e., the regression line passes through origin.

MSE of $\hat{\bar{Y}}_{Rd}$:

$$\begin{aligned}MSE(\hat{\bar{Y}}_{Rd}) &= E(\hat{\bar{Y}}_{Rd}-\bar{Y})^{2`}\\ &\simeq \bar{Y}^2E(\varepsilon_0+\varepsilon_2-\varepsilon_1)^2 \quad \text{(retaining the terms upto order two)}\\ &= \bar{Y}^2E\left[\varepsilon_0^2+\varepsilon_1^2+\varepsilon_2^2+2\varepsilon_0\varepsilon_2-2\varepsilon_0\varepsilon_1-2\varepsilon_1\varepsilon_2\right]\\ &= \bar{Y}^2E\left[\varepsilon_0^2+\varepsilon_1^2+\varepsilon_2^2+2\varepsilon_0\varepsilon_2-2\varepsilon_0\varepsilon_1-2\varepsilon_2^2\right]\\ &= \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2+\left(\frac{1}{n}-\frac{1}{N}\right)C_x^2-\left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2+2\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_xC_y-2\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_xC_y\right]\\ &= \bar{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(C_x^2+C_y^2-2\rho C_xC_y\right)+\bar{Y}^2\left(\frac{1}{n'}-\frac{1}{N}\right)C_x(2\rho C_y-C_x)\\ &= MSE\text{(ratio estimator)} \; +\bar{Y}^2\left(\frac{1}{n'}-\frac{1}{n}\right)\left(2\rho C_xC_y-C_x^2\right).\end{aligned}$$

The second term is the contribution of second phase of sampling. This method is preferred over ratio method if

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

$$\text{or} \quad \rho > \frac{1}{2}\frac{C_x}{C_y}$$

Choice of n and n'

Write

$$MSE(\hat{\hat{Y}}_{Rd}) = \frac{V}{n} + \frac{V'}{n'}$$

where V and V' contain all the terms containing n and n' respectively.

The cost function is $C_0 = nC + n'C'$ where C and C' are the costs per unit for selecting the samples n and n' respectively.

Now we find the optimum sample sizes n and n' for fixed cost C_0. The Lagrangian function is

$$\varphi = \frac{V}{n} + \frac{V'}{n'} + \lambda(nC + n'C' - C_0)$$

$$\frac{\partial\varphi}{\partial n} = 0 \Rightarrow \lambda C = \frac{V}{n^2}$$

$$\frac{\partial\varphi}{\partial n'} = 0 \Rightarrow \lambda C' = \frac{V'}{n'^2}.$$

Thus $\lambda C n^2 = V$

or $n = \sqrt{\frac{V}{\lambda C}}$

or $\sqrt{\lambda}\; nC = \sqrt{VC}.$

Similarly $\sqrt{\lambda}\, n'C' = \sqrt{V'C'}.$

Thus

$$\sqrt{\lambda} = \frac{\sqrt{VC} + \sqrt{V'C'}}{C_0}$$

and so

$$\text{Optimum } n = \frac{C_0}{\sqrt{VC}+\sqrt{V'C'}}\sqrt{\frac{V}{C}} = n_{opt}, \text{ say}$$

$$\text{Optimum } n' = \frac{C_0}{\sqrt{VC}+\sqrt{V'C'}}\sqrt{\frac{V'}{C'}} = n_{opt}^{'}, \text{ say}$$

$$Var_{opt}(\hat{\bar{Y}}_{Rd}) = \frac{V}{n_{opt}} + \frac{V'}{n_{opt}^{'}}$$

$$= \frac{(\sqrt{VC}+\sqrt{V'C'})^2}{C_0}$$

Comparison with SRS

If X is ignored and all resources are used to estimate $\bar{Y}$ by $\bar{y}$, then required sample size $= \dfrac{C_0}{C}$.

$$Var(\bar{y}) = \frac{S_y^2}{C_0/C} = \frac{CS_y^2}{C_0}$$

$$\text{Relative effiiency } = \frac{Var(\bar{y})}{Var_{opt}(\hat{\bar{Y}}_{Rd})} = \frac{CS_y^2}{(\sqrt{VC}+\sqrt{V'C'})^2}$$

Double sampling in regression method of estimation

When the population mean of auxiliary variable $\bar{X}$ is not known, then double sampling is used as follows:

- A large sample of size n' is taken from of the population by SRSWOR from which the population mean $\bar{X}$ is estimated as $\bar{x}'$, i.e. $\hat{\bar{X}} = \bar{x}'$.
- Then a subsample of size n is chosen from the larger sample and both the variables x and y are measured from it by taking $\bar{x}'$ in place of $\bar{X}$ and treat it as if it is known.

Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$. The regression estimate of $\bar{Y}$ in this case is given by

$$\hat{\bar{Y}}_{regd} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$$

where $\hat{\beta} = \dfrac{s_{xy}}{s_x^2} = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ is an estimator of $\beta = \dfrac{S_{xy}}{S_x^2}$ based on the sample of size n.

It is difficult to find the exact properties like bias and mean squared error of $\hat{\hat{Y}}_{regd}$, so we derive the approximate expressions.

Let

$$\begin{aligned}
\varepsilon_1 &= \frac{\overline{x}-\overline{X}}{\overline{X}} \Rightarrow \overline{x} = (1+\varepsilon_1)\overline{X} \\
\varepsilon_2 &= \frac{\overline{x}'-\overline{X}}{\overline{X}} \Rightarrow \overline{x}' = (1+\varepsilon_2)\overline{X} \\
\varepsilon_3 &= \frac{s_{xy}-S_{xy}}{S_{xy}} \Rightarrow s_{xy} = (1+\varepsilon_3)S_{xy} \\
\varepsilon_4 &= \frac{s_x^2-S_x^2}{S_x^2} \Rightarrow s_x^2 = (1+\varepsilon_4)S_x^2 \\
& E(\varepsilon_1)=0, E(\varepsilon_2)=0, E(\varepsilon_3)=0, E(\varepsilon_4)=0
\end{aligned}$$

Define

$$\mu_{21} = E\left[(\overline{x}-\overline{X})^2(y-\overline{Y})\right]$$

$$\mu_{30} = E\left[\overline{x}-\overline{X}\right]^3$$

Estimation error:

Then

$$\begin{aligned}
\hat{\hat{Y}}_{regd} &= \overline{y}+\hat{\beta}(\overline{x}'-\overline{x}) \\
&= \overline{y}+\frac{S_{xy}(1+\varepsilon_3)}{S_x^2(1+\varepsilon_4)}(\varepsilon_2-\varepsilon_1)\overline{X} \\
&= \overline{y}+\overline{X}\frac{S_{xy}}{S_x^2}(1+\varepsilon_3)(\varepsilon_2-\varepsilon_1)(1+\varepsilon_4)^{-1} \\
&= \overline{y}+\overline{X}\beta(1+\varepsilon_3)(\varepsilon_2-\varepsilon_1)(1-\varepsilon_4+\varepsilon_4^2-...)
\end{aligned}$$

Retaining the powers of $\varepsilon's$ upto order two assuming $|\varepsilon_3|$<1, (using the same concept as detailed in the case of ratio method of estimation)

$$\hat{\hat{Y}}_{regd} \simeq \overline{y}+\overline{X}\beta(\varepsilon_2+\varepsilon_2\varepsilon_3-\varepsilon_2\varepsilon_4-\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4).$$

Bias:

The bias of $\hat{\bar{Y}}_{regd}$ upto the second order of approximation is

$$E(\hat{\bar{Y}}_{regd}) = \bar{Y} + \bar{X}\beta\left[E(\varepsilon_2\varepsilon_3) - E(\varepsilon_2\varepsilon_4) - E(\varepsilon_1\varepsilon_3) + E(\varepsilon_1\varepsilon_4)\right]$$

$$Bias(\hat{\bar{Y}}_{regd}) = E(\hat{\bar{Y}}_{regd}) - \bar{Y}$$

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)\right]$$

$$-\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_x^2 - S_X^2)}{\bar{X}S_X^2}\right)$$

$$-\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)$$

$$+\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_x^2 - S_x^2)}{\bar{X}S_x^2}\right)$$

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} - \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2} - \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} + \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2}\right]$$

$$= -\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right).$$

Mean squared error:

$$MSE(\hat{\bar{Y}}_{regd}) = E(\bar{Y}_{regd} - \bar{Y})^2$$

$$= \left[\bar{y} + \hat{\beta}(\bar{x}' - \bar{x}) - \bar{Y}\right]^2$$

$$= E\left[(\bar{y} - \bar{Y}) + \bar{X}\beta(1 + \varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - ...)\right]^2$$

Retaining the powers of $\varepsilon's$ upto order two, the mean squared error upto the second order of approximation is

$$\begin{aligned}
MSE(\hat{\hat{Y}}_{regd}) &\simeq E\left[(\bar{y}-\bar{Y})+\bar{X}\beta(\varepsilon_2+\varepsilon_2\varepsilon_3-\varepsilon_2\varepsilon_4-\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4)\right]^2 \\
&\simeq E(\bar{y}-\bar{Y})^2+\bar{X}^2\beta^2E(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2)+2\bar{X}\beta E[(\bar{y}-\bar{Y})(\varepsilon_1-\varepsilon_2)] \\
&= Var(\bar{y})+\bar{X}^2\beta^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}-2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}\right] \\
&\quad +2\beta\bar{X}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}}-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}}\right] \\
&= Var(\bar{y})+\beta^2\left(\frac{1}{n}-\frac{1}{n'}\right)S_x^2-2\beta\left(\frac{1}{n}-\frac{1}{n'}\right)S_{xy} \\
&= Var(\bar{y})+\beta^2\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\beta^2S_x^2-2\beta S_{xy}\right) \\
&= Var(\bar{y})+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}^2}{S_x^4}S_x^2-2\frac{S_{xy}}{S_x^2}S_{xy}\right) \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}}{S_x}\right)^2 \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho^2S_y^2 \quad \text{(using } S_{xy}=\rho S_xS_y\text{)} \\
&\approx \frac{(1-\rho^2)S_y^2}{n}+\frac{\rho^2S_y^2}{n'}. \quad \text{(Ignoring the finite population correction)}
\end{aligned}$$

Clearly, $\hat{\hat{Y}}_{regd}$ is more efficient than sample mean SRS, i.e. when no auxiliary variable is used.

Now we address the issue that whether the reduction in variability is worth the extra expenditure required to observe the auxiliary variable.

Let the total cost of survey is

$$C_0 = C_1n + C_2n'$$

where C_1 and C_2 are the costs per unit observing the study variable y and auxiliary variable x respectively.

Now minimize the $MSE\ (\hat{\hat{Y}}_{regd})$ for fixed cost C_0 using Lagrangian function with Lagranagian multiplier λ as

$$\varphi = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} + \lambda(C_1 n + C_2 n' - C_0)$$

$$\frac{\partial \varphi}{\partial n} = 0 \Rightarrow -\frac{1}{n^2} S_y^2(1-\rho^2) + \lambda C_1 = 0$$

$$\frac{\partial \varphi}{\partial n'} = 0 \Rightarrow -\frac{1}{n'^2} S_y^2 \rho^2 + \lambda C_2 = 0$$

Thus $\quad n = \sqrt{\dfrac{S_y^2(1-\rho^2)}{\lambda C_1}}$

and $\quad n' = \dfrac{\rho S_y}{\sqrt{\lambda C_2}}.$

Substituting these values in the cost function, we have

$$\begin{aligned} C_0 &= C_1 n + C_2 n' \\ &= C_1 \sqrt{\frac{S_y^2(1-\rho^2)}{C_1 \lambda}} + C_2 \sqrt{\frac{\rho^2 S_y^2}{\lambda C_2}} \end{aligned}$$

$$\text{or } C_0\sqrt{\lambda} = \sqrt{C_1 S_y^2(1-\rho^2)} + \sqrt{C_2 \rho^2 S_y^2}$$

$$\text{or } \lambda = \frac{1}{C_0^2}\left[S_y\sqrt{C_1(1-\rho^2)} \; + \; \rho S_y \sqrt{C_2}\right]^2.$$

Thus the optimum values of n and n' are

$$n^{'}_{opt} = \frac{\rho S_y C_0}{\sqrt{C_2}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]}$$

$$n_{opt} = \frac{C_0 S_y \sqrt{1-\rho^2}}{\sqrt{C_1}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]}.$$

The optimum mean squared error of $\hat{\bar{Y}}_{regd}$ is obtained by substituting $n = n_{opt}$ and $n' = n^{'}_{opt}$ as

$$\begin{aligned} MSE(\hat{\bar{Y}}_{regd})_{opt} &= \frac{S_y^2(1-\rho^2)\left[\sqrt{C_1}\left(\sqrt{C_1 S_y^2(1-\rho^2)} + \rho S_y\sqrt{C_2}\right)\right]}{C_0\sqrt{S_y^2(1-\rho^2)}} \\ &\quad + \frac{S_y^2\rho^2\sqrt{C_2}\left[S_y\left(\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right)\right]}{\rho S_y C_0} \\ &= \frac{1}{C_0}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]^2 \\ &= \frac{S_y^2}{C_0}\left[\sqrt{C_1(1-\rho^2)} + \rho\sqrt{C_2}\right]^2 \end{aligned}$$

The optimum variance of $\bar{y}$ under SRS for SRS where no auxiliary information is used is

$$Var(\bar{y}_{SRS})_{opt} = \frac{C_1 S_y^2}{C_0}$$

which is obtained by substituting $\rho = 0, \; _2 = 0$ in $MSE(\hat{\bar{Y}}_{SRS})_{opt}$. The relative efficiency is

$$RE = \frac{Var(\bar{y}_{SRS})_{opt}}{MSE(\hat{\bar{Y}}_{regd})_{opt}} = \frac{C_1 S_y^2}{S_y^2\left[\sqrt{C_1(1-\rho^2)}\rho+ \;\; \sqrt{C_2}\right]^2}$$

$$= \frac{1}{\left[\sqrt{1-\rho^2}\rho+ \;\; \sqrt{\frac{C_2}{C_1}}\right]^2}$$

$$\leq 1.$$

Thus the double sampling in regression estimator will lead to gain in precision if

$$\frac{C_1}{C_2} > \frac{\rho^2}{\left[1-\sqrt{1-\rho^2}\right]^2}.$$

x

x

x

n

x y x y

n

Y

$$\hat{Y} = {}' \sum_{i=1}^{n} {}_i ,$$

x' x

$\hat{}$ $\frac{f}{n}$

$(\ )$ $\frac{f}{n}$ ^{2} $\hat{Y}_p$ y

$\frac{1}{2}\frac{C_x}{C_y}$ R

$\frac{1}{2}\frac{C_x}{C_y}$ R

Multivariate Ratio Estimator

Let y be the study variable and $X_1, X_2, ..., X_p$ be p auxiliary variables assumed to be corrected with y. Further it is assumed that $X_1, X_2, ..., X_p$ are independent. Let $\bar{Y}, \bar{X}_1, \bar{X}_2, ..., \bar{X}_p$ be the population means of the variables y, $X_1, X_2, ..., X_p$. We assume that a SRSWOR of size n is selected from the population of N units. The following notations will be used.

S_i^2 : the population mean sum of squares for the variate X_i,

s_i^2 : the sample mean sum of squares for the variate X_i,

S_0^2 : the population mean sum of squares for the study variable y,

s_0^2 : the sample mean sum of squares for the study variable y,

$C_i = \frac{S_i}{X_i}$: coefficient of variation of the variate X_i,

$C_0 = \frac{S_0}{Y}$: coefficient of variation of the variate y,

$\rho_i = \frac{S_{iy}}{S_i S_0}$: coefficient of correlation between y and X_i,

$\hat{\bar{Y}}_{Ri} = \frac{\bar{y}}{\bar{X}_i}$: ratio estimator of $\bar{Y}$, based on X_i

where $i = 1, 2, ..., p$. Then the multivariate ratio estimator of $\bar{Y}$ is given as follows.

$$\hat{\bar{Y}}_{MR} = \sum_{i=1}^{p} w_i \hat{\bar{Y}}_{Ri}, \quad \sum_{i=1}^{p} w_i = 1$$

$$= \bar{y} \sum_{i=1}^{p} w_i \frac{\bar{X}_i}{\bar{x}_i}.$$

(i) Bias of the multivariate ratio estimator:

The approximate bias of $\hat{\bar{Y}}_{Ri}$ upto the second order of approximation is

$$Bias(\hat{\bar{Y}}_{Ri}) = \frac{f}{n}\bar{Y}(C_i^2 - \rho_i C_i C_0).$$

The bias of $\hat{\bar{Y}}_{MR}$ is obtained as

$$\begin{aligned} Bias(\hat{\bar{Y}}_{MR}) &= \sum_{i=1}^{p} w_i \frac{\bar{Y}f}{n}(C_i^2 - \rho_i C_i C_0) \\ &= \frac{\bar{Y}f}{n}\sum_{i=1}^{p} w_i C_i\,(C_i - \rho_i C_0). \end{aligned}$$

(ii) Variance of the multivariate ratio estimator:

The variance of $\hat{\bar{Y}}_{Ri}$ upto the second order of approximation is given by

$$Var(\hat{\bar{Y}}_{Ri}) = \frac{f}{n}\bar{Y}^2(C_0^2 + C_i^2 - 2\rho_i C_0 C_i).$$

The variance of $\hat{\bar{Y}}_{MR}$ upto the second order of approximation is obtained as

$$Var(\hat{\bar{Y}}_{MR}) = \frac{f}{n}\bar{Y}^2\sum_{i=1}^{p} w_i^2(C_0^2 + C_i^2 - 2\rho_i C_0 C_i).$$

Varying Probability Sampling

The simple random sampling scheme provides a random sample where every unit in the population has equal probability of selection. Under certain circumstances, more efficient estimators are obtained by assigning unequal probabilities of selection to the units in the population. This type of sampling is known as varying probability sampling scheme.

If Y is the variable under study and X is an auxiliary variable related to Y, then in the most commonly used varying probability scheme, the units are selected with probability proportional to the value of X, called as size. This is termed as probability proportional to a given measure of size (pps) sampling. If the sampling units vary considerably in size, then SRS does not takes into account the possible importance of the larger units in the population. A large unit, i.e., a unit with large value of Y contributes more to the population total than the units with smaller values, so it is natural to expect that a selection scheme which assigns more probability of inclusion in a sample to the larger units than to the smaller units would provide more efficient estimators than the estimators which provide equal probability to all the units. This is accomplished through pps sampling.

Note that the "size" considered is the value of auxiliary variable X and not the value of study variable Y. For example in an agriculture survey, the yield depends on the area under cultivation. So bigger areas are likely to have larger population and they will contribute more towards the population total, so the value of the area can be considered as the size of auxiliary variable. Also, the cultivated area for a previous period can also be taken as the size while estimating the yield of crop. Similarly, in an industrial survey, the number of workers in a factory can be considered as the measure of size when studying the industrial output from the respective factory.

Difference between the methods of SRS and varying probability scheme:

In SRS, the probability of drawing a specified unit at any given draw is the same. In varying probability scheme, the probability of drawing a specified unit differs from draw to draw.

It appears in pps sampling that such procedure would give biased estimators as the larger units are over-represented and the smaller units are under-represented in the sample. This will happen in case of sample mean as an estimator of population mean where all the units are given equal weight. Instead of giving equal weights to all the units, if the sample observations are suitably weighted at the estimation stage by taking the probabilities of selection into account, then it is possible to obtain unbiased estimators.

In pps sampling, there are two possibilities to draw the sample, i.e., with replacement and without replacement.

Selection of units with replacement:

The probability of selection of a unit will not change and the probability of selecting a specified unit is same at any stage. There is no redistribution of the probabilities after a draw.

Selection of units without replacement:

The probability of selection of a unit will change at any stage and the probabilities are redistributed after each draw.

PPS without replacement (WOR) is more complex than PPS with replacement (WR) . We consider both the cases separately.

PPS sampling with replacement (WR):

First we discuss the two methods to draw a sample with PPS and WR.

1. Cumulative total method:

The procedure of selection a simple random sample of size n consists of

- associating the natural numbers from 1 to N units in the population and
- then selecting those n units whose serial numbers correspond to a set of n numbers where each number is less than or equal to N which is drawn from a random number table.

In selection of a sample with varying probabilities, the procedure is to associate with each unit a set of consecutive natural numbers, the size of the set being proportional to the desired probability.

If $X_1, X_2, ..., X_N$ are the positive integers proportional to the probabilities assigned to the N units in the population, then a possible way to associate the cumulative totals of the units. Then the units are selected based on the values of cumulative totals. This is illustrated in the following table:

Units	Size	Cumulative		
1	X_1	$T_1 = X_1$	Select a random number R between 1 and T_N by using random number table.	• If $T_{i-1} \le R \le T_i$, then i^{th} unit is selected with probability $\frac{X_i}{T_N}$, $i = 1,2,\ldots,N$. • Repeat the procedure n times to get a sample of size n.
2	X_2	$T_2 = X_1 + X_2$		
$\vdots$	$\vdots$	$\vdots$		
$i-1$	X_{i-1}	$T_{i-1} = \sum_{j=1}^{i-1} X_j$		
i	X_i	$T_i = \sum_{j=1}^{i} X_j$		
$\vdots$	$\vdots$	$\vdots$		
N	$X_N = \sum_{j=1}^{N} X_j$	$T_N = \sum_{j=1}^{N} X_j$		

In this case, the probability of selection of i^{th} unit is

$$P_i = \frac{T_i - T_{i-1}}{T_N} = \frac{X_i}{T_N}$$

$$\Rightarrow P_i \propto X_i.$$

Note that T_N is the population total which remains constant.

Drawback : This procedure involves writing down the successive cumulative totals. This is time consuming and tedious if the number of units in the population is large.

This problem is overcome in the Lahiri's method.

Lahiri's method:

Let $M = \underset{i=1,2,\ldots,N}{Max} X_i,$ i.e., maximum of the sizes of N units in the population or some convenient number greater than M .

The sampling procedure has following steps:

1. Select a pair of random number (*i, j*) such that $1 \le i \le N,\ 1 \le j \le M$.
2. If $j \le X_i$, then i^{th} unit is selected otherwise rejected and another pair of random number is chosen.
3. To get a sample of size n, this procedure is repeated till n units are selected.

Now we see how this method ensures that the probabilities of selection of units are varying and are proportional to size.

Probability of selection of i^{th} unit at a trial depends on two possible outcomes

- either it is selected at the first draw
- or it is selected in the subsequent draws preceded by ineffective draws. Such probability is given by

$$P(1 \le i \le N)P(1 \le j \le M \mid i)$$
$$= \frac{1}{N}.\frac{X_i}{M} = P_i^*, \text{ say.}$$

Probability that no unit is selected at a trial
$$= \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{X_i}{M}\right)$$
$$= \frac{1}{N}\left(N - \frac{N\overline{X}}{M}\right)$$
$$= 1 - \frac{\overline{X}}{M} = Q, \text{ say.}$$

Probability that unit i is selected at a given draw (all other previous draws result in the non selection of unit i)

$$= P_i^* + QP_i^* + Q^2P_i^* + ...$$
$$= \frac{P_i^*}{1-Q}$$
$$= \frac{X_i / NM}{\overline{X} / M} = \frac{X_i}{N\overline{X}} = \frac{X_i}{X_{total}} \propto X_i.$$

Thus the probability of selection of unit i is proportional to the size X_i. So this method generates a pps sample.

Advantage:

1. It does not require writing down all cumulative totals for each unit.
2. Sizes of all the units need not be known before hand. We need only some number greater than the maximum size and the sizes of those units which are selected by the choice of the first set of random numbers 1 to N for drawing sample under this scheme.

Disadvantage: It results in the wastage of time and efforts if units get rejected.

The probability of rejection $= 1 - \frac{\overline{X}}{M}$.

The expected numbers of draws required to draw one unit $= \frac{M}{\overline{X}}$.

This number is large if M is much larger than $\overline{X}$.

Example: Consider the following data set of 10 number of workers in the factory and its output. We illustrate the selection of units using the cumulative total method.

Factory no.	Number of workers (X) (in thousands)	Industrial production (in metric tonns) (Y)	Cumulative total of sizes
1	2	30	$T_1 = 2$
2	5	60	$T_2 = 2+5=7$
3	10	12	$T_3 = 2+5+10=17$
4	4	6	$T_4 = 17+4=21$
5	7	8	$T_5 = 21+7=28$
6	12	13	$T_6 = 28+12=30$
7	3	4	$T_7 = 30+3=33$
8	14	17	$T_8 = 33+14=47$
9	11	13	$T_9 = 47+11=58$
10	6	8	$T_{10} = 58+6=64$

Selection of sample using cumulative total method:

1.First draw: - Draw a random number between 1 and 64.

- Suppose it is 23
- $T_4 < 23 < T_5$
- Unit Y is selected and $Y_5 = 8$ enters in the sample.

2. Second draw:

- Draw a random number between 1 and 64
- Suppose it is 38
- $T_7 < 38 < T_8$
- Unit 8 is selected and $Y_8 = 17$ enters in the sample
- and so on.
- This procedure is repeated till the sample of required size is obtained.

Selection of sample using Lahiri's Method

In this case

$$M = \underset{i=1,2,...,10}{Max} \; X_i = 14$$

So we need to select a pair of random number (i, j) such that $1 \le i \le 10, 1 \le j \le 14$.

Following table shows the sample obtained by Lahiri's scheme:

Random no $1 \le i \le 10$	Random no $1 \le j \le 14$	Observation	Selection of unit
3	7	$j = 7 < X_3 = 10$	trial accepted (y_3)
6	13	$j = 13 > X_6 = 12$	trial rejected
4	7	$j = 7 > X_4 = 4$	trial rejected
2	9	$j = 9 > X_2 = 5$	trial rejected
9	2	$j = 2 < X_9 = 11$	trial accepted (y_9)

and so on. Here (y_3, y_9) are selected into the sample.

Varying probability scheme with replacement: Estimation of population mean

Let

Y_i: value of study variable for the i^{th} unit of the population, i = 1, 2,…,N.
X_i: known value of auxiliary variable (size) for the i^{th} unit of the population.

P_i: probability of selection of i^{th} unit in the population at any given draw and is proportional to size X_i.

Consider the varying probability scheme and with replacement for a sample of size *n*. Let y_r be the value of r^{th} observation on study variable in the sample and p_r be its initial probability of selection. Define

$$z_r = \frac{y_r}{Np_r}, \; r = 1, 2, ..., n,$$

then

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

is an unbiased estimator of population mean $\overline{Y}$, variance of $\overline{z}$ is $\frac{\sigma_z^2}{n}$ where $\sigma_z^2 = \sum_{i=1}^{N} P_i \left(\frac{Y_i}{NP_i} - \overline{Y} \right)^2$ and an unbiased estimate of variance of $\overline{z}$ is $\frac{s_z^2}{n} = \frac{1}{n-1} \sum_{r=1}^{n} (z_r - \overline{z})^2$.

Proof:

Note that z_r can take any one of the *N* values out of $Z_1, Z_2, ..., Z_N$ with corresponding initial probabilities $P_1, P_2, ..., P_N$, respectively. So

$$
\begin{aligned}
E(z_r) &= \sum_{i=1}^{N} Z_i P_i \\
&= \sum_{i=1}^{N} \frac{Y_i}{NP_i} P_i \\
&= \overline{Y}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
E(\overline{z}) &= \frac{1}{n} \sum_{i=1}^{n} E(z_r) \\
&= \frac{1}{n} \sum_{i=1}^{n} \overline{Y} \\
&= \overline{Y}.
\end{aligned}
$$

So $\overline{z}$ is an unbiased estimator of population mean $\overline{Y}$.

The variance of $\overline{z}$ is

$$
\begin{aligned}
Var(\overline{z}) &= \frac{1}{n^2} Var\left(\sum_{r=1}^{n} z_r \right) \\
&= \frac{1}{n^2} \sum_{r=1}^{n} Var(z_r) \qquad (z_r^{'} s \text{ are independent in WR case}).
\end{aligned}
$$

Now

$$
\begin{aligned}
Var(z_r) &= E\left[z_r - E(z_r) \right]^2 \\
&= E\left[z_r - \overline{Y} \right]^2 \\
&= \sum_{i=1}^{N} \left(Z_i - \overline{Y} \right)^2 P_i \\
&= \sum_{i=1}^{N} \left(\frac{Y_i}{NP_i} - \overline{Y} \right)^2 P_i \\
&= \sigma_z^2 \quad \text{(say)} .
\end{aligned}
$$

Thus

$$Var(\overline{z}) = \frac{1}{n^2}\sum_{r=1}^{n}\sigma_z^2$$

$$= \frac{\sigma_z^2}{n}.$$

To show that $\frac{s_z^2}{n}$ is an unbiased estimator of variance of $\overline{z}$, consider

$$(n-1)E(s_z^2) = E\left[\sum_{r=1}^{n}(z_r - \overline{z})^2\right]$$

$$= E\left[\sum_{r=1}^{n} z_r^2 - n\overline{z}^2\right]$$

$$= \left[\sum_{r=1}^{n} E(z_r^2) - nE(\overline{z})^2\right]$$

$$= \sum_{r=1}^{n}\left[Var(z_r) + \left\{E(z_r)\right\}^2\right] - n\left[Var(\overline{z}) + \left\{E(\overline{z})\right\}^2\right]$$

$$= \sum_{r=1}^{n}\left(\sigma_z^2 + \overline{Y}^2\right) - n\left(\tfrac{\sigma_z^2}{n} + \overline{Y}^2\right) \quad \left(\text{using } Var(z_r) = \sum_{i=1}^{N}\left(\frac{Y_i}{NP_i} - \overline{Y}\right)^2 P_i = \sigma_z^2\right)$$

$$= (n-1)\sigma_z^2$$

$$E(s_z^2) = \sigma_z^2$$

or $E\left(\frac{s_z^2}{n}\right) = \frac{\sigma_z^2}{n} = Var(\overline{z})$

$$\Rightarrow \widehat{Var}(\overline{z}) = \frac{s_z^2}{n} = \frac{1}{n(n-1)}\left[\sum_{r=1}^{n}\left(\frac{y_r}{Np_r}\right)^2 - n\overline{z}^2\right].$$

Note: If $P_i = \frac{1}{N}$, then $\overline{z} = \overline{y}$,

$$Var(\overline{z}) = \frac{1}{n}\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{N.\frac{1}{N}} - \overline{Y}\right)^2 = \frac{\sigma_y^2}{n}$$

which is the same as in the case of SRSWR.

Estimation of population total:

An estimate of population total is

$$\hat{Y}_{tot} = \frac{1}{n}\sum_{r=1}^{n}\left(\frac{y_r}{p_r}\right) = N\,\bar{z}..$$

Taking expectation, we get

$$\begin{aligned}
E(\hat{Y}_{tot}) &= \frac{1}{n}\sum_{r=1}^{n}\left[\frac{Y_1}{P_1}P_1 + \frac{Y_2}{P_2}P_2 + \ldots + \frac{Y_N}{P_N}P_N\right] \\
&= \frac{1}{n}\sum_{r=1}^{n}\left[\sum_{i=1}^{N} Y_i\right] \\
&= \frac{1}{n}\sum_{r=1}^{n} Y_{tot} \\
&= Y_{tot}.
\end{aligned}$$

Thus $\hat{Y}_{tot}$ is an unbiased estimator of population total. Its variance is

$$\begin{aligned}
Var(\hat{Y}_{tot}) &= N^2 Var(\bar{z}) \\
&= N^2\frac{1}{n}\sum_{i=1}^{N}\frac{1}{N^2}\left(\frac{Y_i}{P_i} - N\bar{Y}\right)^2 P_i \\
&= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 P_i \\
&= \frac{1}{n}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - Y_{tot}^2\right].
\end{aligned}$$

An estimate of the variance

$$\widehat{Var}(\hat{Y}_{tot}) = N^2\frac{s_z^2}{n}.$$

Varying probability scheme without replacement

In varying probability scheme without replacement, when the initial probabilities of selection are unequal, then the probability of drawing a specified unit of the population at a given draw changes with the draw. Generally, the sampling WOR provides a more efficient estimator than sampling WR. The estimators for population mean and variance are more complicated. So this scheme is not commonly used in practice, especially in large scale sample surveys with small sampling fractions.

Let U_i : i^{th} unit,

P_i :Probability of selection of U_i at the first draw, $i = 1, 2, ..., N$

$$\sum_{i=1}^{N} P_i = 1$$

$P_{i(r)}$: Probability of selecting U_i at the r^{th} draw

$P_{i(1)} = P_i$.

Consider

$P_{i(2)} =$ Probability of selection of U_i at 2nd draw.

Such an event can occur in the following possible ways:

U_i is selected at 2nd draw when

- U_1 is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_2 is selected at 1^{st} draw and U_i is selected at 2^{nd} draw

$\vdots$

- $U_{i\text{-}1}$ is selected at 1^{st} draw and U_i is selected at 2^{nd} draw
- U_{i+1} is selected at 1^{st} draw and U_i is selected at 2^{nd} draw

$\vdots$

- U_N is selected at 1^{st} draw and U_i is selected at 2^{nd} draw

So $P_{i(2)}$ can be expressed as

$$\begin{aligned}
P_{i(2)} &= P_1 \frac{P_i}{1-P_1} + P_2 \frac{P_i}{1-P_2} + ... + P_{i-1} \frac{P_i}{1-P_{i-1}} + P_{i+1} \frac{P_i}{1+P_{i+1}} + ... + P_N \frac{P_i}{1-P_N} \\
&= \sum_{j(\neq i)=1}^{N} P_j \frac{P_i}{1-P_j} \\
&= \sum_{j(\neq i)=1}^{N} P_j \frac{P_i}{1-P_j} + P_i \frac{P_i}{1-P_i} - P_i \frac{P_i}{1-P_i} \\
&= \sum_{j=1}^{N} P_j \frac{P_i}{1-P_j} - P_i \frac{P_i}{1-P_i} \\
&= P_i \left[\sum_{j=1}^{N} \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i} \right]
\end{aligned}$$

$P_{i(2)} \neq P_{i(1)}$ for all i unless $P_i = \dfrac{1}{N}$.

$P_{i(2)}$ will, in general, be different for each $i = 1,2,\ldots,N$. So $E\left(\frac{y_i}{p_i}\right)$ will change with successive draws. This makes the varying probability scheme WOR more complex. Only $\frac{y_1}{Np_1}$ will provide an unbiased estimator of $\overline{Y}$. In general, $\frac{y_i}{Np_i}(i \neq 1)$ will not provide an unbiased estimator of $\overline{Y}$.

Ordered estimates

To overcome the difficulty of changing expectation with each draw, associate a new variate with each draw such that its expectation is equal to the population value of the variate under study. Such estimators take into account the order of the draw. They are called the ordered estimates. The order of the value obtained at previous draw will affect the unbiasedness of population mean.

We consider the ordered estimators proposed by Des Raj, first for the case of two draws and then generalize the result.

Des Raj ordered estimator

Case 1: Case of two draws:

Let y_1 and y_2 denote the values of units $U_{i(1)}$ and $U_{i(2)}$ drawn at the first and second draws respectively. Note that any one out of the N units can be the first unit or second unit, so we use the notations $U_{i(1)}$ and $U_{i(2)}$ instead of U_1 and U_2. Also note that y_1 and y_2 are not the values of the first two units in the population. Further, let p_1 and p_2 denote the initial probabilities of selection of $U_{i(1)}$ and $U_{i(2)}$, respectively.

Consider the estimators

$$z_1 = \frac{y_1}{Np_1}$$

$$z_2 = \frac{1}{N}\left[y_1 + \frac{y_2}{p_2/(1-p_1)}\right]$$

$$= \frac{1}{N}\left[y_1 + y_2\frac{(1-p_1)}{p_2}\right]$$

$$\overline{z} = \frac{z_1 + z_2}{2}.$$

Note that $\frac{p_2}{1-p_1}$ is the probability $P(U_{i(2)} \mid U_{i(1)})$.

Estimation of Population Mean:

First we show that $\bar{z}$ is an unbiased estimator of $\bar{Y}$.

$$E(\bar{z}) = \bar{Y}.$$

Note that $\sum_{i=1}^{N} P_i = 1.$

Consider

$$E(z_1) = \frac{1}{N} E\left(\frac{y_1}{p_1}\right) \quad \left(\text{Note that } \frac{y_1}{p_1} \text{ can take any one of out of the } N \text{ values } \frac{Y_1}{P_1}, \frac{Y_2}{P_2}, ..., \frac{Y_N}{P_N}\right)$$

$$= \frac{1}{N}\left[\frac{Y_1}{P_1}P_1 + \frac{Y_2}{P_2}P_2 + ... + \frac{Y_N}{P_N}P_N\right]$$

$$= \bar{Y}$$

$$E(z_2) = \frac{1}{N} E\left[y_1 + y_2 \frac{(1-p_1)}{p_2}\right]$$

$$= \frac{1}{N}\left[E(y_1) + E_1\left\{E_2\left(y_2 \frac{(1-P_1)}{p_2}\bigg| U_{i(1)}\right)\right\}\right] \quad (\text{Using } E(Y) = E_X[E_Y(Y \mid X)].$$

where E_2 is the conditional expectation after fixing the unit $U_{i(1)}$ selected in the first draw.

Since $\frac{y_2}{p_2}$ can take any one of the (N – 1) values (except the value selected in the first draw) $\frac{Y_j}{P_j}$ with probability $\frac{P_j}{1-P_1}$, so

$$E_2\left[y_2 \frac{(1-P_1)}{p_2}\bigg| U_{i(1)}\right] = (1-P_1)E_2\left[\frac{y_2}{p_2}\bigg| U_{i(1)}\right] = (1-P_1)\sum_j^{*}\left[\frac{Y_j}{P_j}.\frac{P_j}{1-P_1}\right].$$

where the summation is taken over all the values of Y except the value y_1 which is selected at the first draw. So

$$E_2\left[y_2 \frac{(1-P_1)}{p_2}\bigg| U_{i(1)}\right] = \sum_j^{*} Y_j = Y_{tot} - y_1.$$

Substituting it is $E(z_2)$, we have

$$E(z_2) = \frac{1}{N}\left[E(y_1) + E_1(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N}\left[E(y_1) + E(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N}E(Y_{tot}) = \frac{Y_{tot}}{N} = \bar{Y}.$$

Thus $E(\bar{z}) = \dfrac{E(z_1) + E(z_2)}{2}$

$$= \frac{\bar{Y} + \bar{Y}}{2}$$

$$= \bar{Y}.$$

Variance:

The variance of $\bar{z}$ for the case of two draws is given as

$$Var(\bar{z}) = \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2\right)\left[\frac{1}{2N^2}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2\right] - \frac{1}{4N^2}\sum_{i=1}^{N} P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

Proof: Before starting the proof, we note the following property

$$\sum_{i \neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i\left[\sum_{j=1}^{N} b_j - b_i\right]$$

which is used in the proof.

The variance of $\bar{z}$ is

$$Var(\bar{z}) = E(\bar{z}^2) - \left[E(\bar{z})\right]^2$$

$$= E\left[\frac{1}{2N}\left\{\frac{y_1}{p_1} + y_1 + \frac{y_2(1-p_1)}{p_2}\right\}\right]^2 - \bar{Y}^2$$

$$= \frac{1}{4N^2}E\left[\underbrace{\frac{y_1(1+p_1)}{p_1}}_{\downarrow\ \boxed{\text{nature of variable depends only on } 1^{st} \text{ draw}}} + \underbrace{\frac{y_2(1-p_1)}{p_2}}_{\downarrow\ \boxed{\text{nature of variable depends upon } 1^{st} \text{ and } 2^{nd} \text{ draw}}}\right]^2 - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i \neq j=1}^{N}\left\{\frac{Y_i(1+P_i)}{P_i} + \frac{Y_j(1-P_i)}{P_j}\right\}^2 \frac{P_iP_j}{1-P_i}\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i \neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i^2}\frac{P_iP_j}{1-P_i} + \frac{Y_j^2(1-P_i)^2}{P_j^2}\frac{P_iP_j}{1-P_i} + 2Y_iY_j\frac{(1-P_i^2)}{P_iP_j}\frac{P_iP_j}{1-P_i}\right\}\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i \neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i}\frac{P_j}{1-P_i} + \frac{Y_j^2(1-P_i)^2}{P_j}\frac{P_i}{1-P_i} + 2Y_iY_j(1+P_i)\right\}\right. - \bar{Y}^2.$$

Using the property

$$\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i \left[\sum_{j=1}^{N} b_j - b_i \right], \text{ we can write}$$

$$Var(\overline{z}) = \frac{1}{4N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2(1+P_i)^2}{P_i(1-P_i)} \left\{ \sum_{j=1}^{N} P_j - P_i \right\} + \sum_{i=1}^{N} P_i(1-P_i) \left\{ \sum_{j=1}^{N} \frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i} \right\} + 2\sum_{i=1}^{N} Y_i(1+P_i)(\sum_{j=1}^{N} Y_j - Y_i)\right] - \overline{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2}{P_i}(1+P_i^2+2P_i) + \sum_{i=1}^{N} P_i(1-P_i) \left\{ \sum_{j=1}^{N} \frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i} \right\} + 2\sum_{i=1}^{N} Y_i(1+P_i)(\sum_{j=1}^{N} Y_j - Y_i) \right] - \overline{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N} \frac{Y_i^2}{P_i} + \sum_{i=1}^{N} Y_i^2 P + 2\sum_{i=1}^{N} Y_i^2 + \sum_{i=1}^{N} P_i \sum_{j=1}^{N} \frac{Y_j^2}{P_j} - \sum_{i=1}^{N} Y_i^2 - \sum_{i=1}^{N} P_i^2 \sum_{j=1}^{N} \frac{Y_j^2}{P_j} \right.$$

$$\left. + \sum_{i} P_i Y_i^2 + 2\sum_{i=1}^{N} Y_i \sum_{j=1}^{N} Y_j - 2\sum_{i=1}^{N} Y_i^2 P_i + 2\sum_{i=1}^{N} Y_i P_i \sum_{j=1}^{N} Y_j - 2\sum_{i=1}^{N} Y_i^2 \right] - \overline{Y}^2$$

$$= \frac{1}{4N^2}\left[2\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - \sum_{i=1}^{N} P_i^2 \sum_{j=1}^{N} \frac{Y_j^2}{P_j} - \sum_{i=1}^{N} Y_i^2 + 2Y_{tot}^2 + 2Y_{tot} \sum_{i=1}^{N} Y_i P_i \right] - \overline{Y}^2$$

$$= 2\left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \frac{1}{4N^2} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y_{tot}^2 + Y_{tot}^2 \right) - \frac{1}{4N^2}\left[\sum_{i=1}^{N} Y_i^2 - 2Y_{tot}^2 - 2Y_{tot} \sum_{i=1}^{N} Y_i P_i + 4N^2 \overline{Y}^2 \right]$$

$$= \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \frac{1}{2N^2} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2 - \frac{1}{4N^2} (\sum_{i=1}^{N} Y_i^2 - 2Y_{tot} \sum_{i=1}^{N} Y_i P_i - 2Y_{tot}^2 + 4Y_{tot}^2)$$

$$+ \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \frac{1}{2N^2} Y_{tot}^2$$

$$= \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \frac{1}{2N^2} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2 - \frac{1}{4N^2} (\sum_{i=1}^{N} Y_i^2 - 2Y_{tot} \sum_{i=1}^{N} Y_i P_i + 2Y_{tot}^2 - 2Y_{tot}^2 + \sum_{i} P_i^2 Y_{tot}^2)$$

$$= \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \frac{1}{2N^2} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2 - \frac{1}{4N^2} \sum_{i=1}^{N} \left(Y_i^2 - 2Y_{tot} Y_i P_i + P_i^2 Y_{tot}^2 \right)$$

$$= \frac{1}{2N^2} \left(1 - \frac{1}{2}\sum_{i=1}^{N} P_i^2 \right) \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y_{tot} \right) - \frac{1}{4N^2} \sum_{i=1}^{N} P_i^2 \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2$$

$$= \frac{1}{2} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{NP_i} - \overline{Y} \right)^2 - \frac{1}{4N^2} \sum_{i=1}^{N} P_i^2 \sum_{i=1}^{N} \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2 - \frac{1}{4N^2} \sum_{i=1}^{N} P_i^2 \left(\frac{Y_i}{P_i} - Y_{tot} \right)^2$$

↓ ↓

variance of WR
case for $n = 2$

reduction of variance
in WR with varying
probability

Estimation of $Var(\overline{z})$

$$\begin{aligned} Var(\overline{z}) &= E(\overline{z}^2) - (E(\overline{z}))^2 \\ &= E(\overline{z}^2) - \overline{Y}^2 \end{aligned}$$

Since

$$\begin{aligned} E(z_1 z_2) &= E\left[z_1 E(z_2 \mid u_1)\right] \\ &= E\left[z_1 \overline{Y}\right] \\ &= \overline{Y}E(z_1) \\ &= \overline{Y}^2. \end{aligned}$$

Consider

$$\begin{aligned} E\left[\overline{z}^2 - z_1 z_2\right] &= E(\overline{z}^2) - E(z_1 z_2) \\ &= E(\overline{z}^2) - \overline{Y}^2 \\ &= Var(\overline{z}) \end{aligned}$$

$\Rightarrow \widehat{Var}(\overline{z}) = \overline{z}^2 - z_1 z_2$ is an unbiased estimator of $Var(\overline{z})$

Alternative form

$$\begin{aligned} \widehat{Var}(\overline{z}) &= \overline{z}^2 - z_1 z_2 \\ &= \left(\frac{z_1 + z_2}{2}\right)^2 - z_1 z_2 \\ &= \frac{(z_1 - z_2)^2}{4} \\ &= \frac{1}{4}\left[\frac{y_1}{Np_1} - \frac{y_1}{N} - \frac{y_2}{N}\frac{1-p_1}{p_2}\right]^2 \\ &= \frac{1}{4N^2}\left[(1-p_1)\frac{y_1}{p_1} - \frac{y_2(1-p_1)}{p_2}\right]^2 \\ &= \frac{(1-p_1)^2}{4N^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2. \end{aligned}$$

Case 2: General Case

Let $(U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)})$ be the units selected in the order in which they are drawn in n draws where $U_{i(r)}$ denotes that the i^{th} unit is drawn at the r^{th} draw. Let $(y_1, y_2, .., y_r, ..., y_n)$ and $(p_1, p_2, ..., p_r, ..., p_n)$ be the values of study variable and corresponding initial probabilities of selection, respectively. Further, let $P_{i(1)}, P_{i(2)}, ..., P_{i(r)}, ..., P_{i(n)}$ be the initial probabilities of $U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)}$, respectively.

Further, let

$$z_1 = \frac{y_1}{Np_1}$$

$$z_r = \frac{1}{N}\left[y_1 + y_2 + ... + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right] \text{ for } r = 2,3,...,n.$$

Consider $\bar{z} = \frac{1}{n}\sum_{r=1}^{n} z_r$ as an estimator of population mean $\bar{Y}$.

We already have shown in case 1 that $E(z_1) = \bar{Y}$.

Now we consider $E(z_r), r = 2,3,...,n.$ We can write

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[z_r \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right]$$

where E_2 is the conditional expectation after fixing the units $U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)}$ drawn in the first (r - 1) draws.

Consider

$$E\left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1})\right] = E_1 E_2 \left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right]$$

$$= E_1 \left[(1 - P_{i(1)} - P_{i(2)} ... - P_{i(r-1)}) E_2 \left(\frac{y_r}{p_r} \left| U_{i(1)}, U_{i(2)}, ..., U_{i(r-1)} \right. \right) \right].$$

Since conditionally $\frac{y_r}{p_r}$ can take any one of the (N - r - 1) values $\frac{Y_j}{P_j}, j = 1,2,...,N$ with probabilities

$\frac{P_j}{1 - P_{i(1)} - P_{i(2)} ... - P_{i(r-1)}}$, so

$$E\left[\frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1})\right] = E_1 \left[(1 - P_{i(1)} - P_{i(2)} ... - P_{i(r-1)}) \sum_{j=1}^{N}{}^{*} \frac{Y_j}{P_j} . \frac{P_j}{(1 - P_{i(1)} - P_{i(2)} ... - P_{i(r-1)})} \right]$$

$$= E_1 \left[\sum_{j=1}^{N}{}^{*} Y_j \right]$$

where $\sum_{j=1}^{N}{}^{*}$ denotes that the summation is taken over all the values of y except the y values selected in the first (r -1) draws

like as $\sum_{j=1(\neq i(1), i(2), ..., i(r-1))}^{N}$, i.e., except the values $y_1, y_2, ..., y_{r-1}$ which are selected in the first (r-1) draws.

Thus now we can express

$$
\begin{aligned}
E(z_r) &= \frac{1}{N} E_1 E_2 \left[y_1 + y_2 + \ldots + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \right] \\
&= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1}^{N}{}^{*} Y_j \right] \\
&= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1(\neq i(1), i(2), \ldots, i(r-1))}^{N} Y_j \right] \\
&= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \left\{ Y_{tot} - \left(Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} \right) \right\} \right] \\
&= \frac{1}{N} E_1 \left[Y_{tot} \right] \\
&= \frac{Y_{tot}}{N} \\
&= \overline{Y} \quad \text{for all} \quad r = 1, 2, \ldots, n.
\end{aligned}
$$

Then

$$
\begin{aligned}
E(\overline{z}) &= \frac{1}{n} \sum_{r=1}^{n} E(z_r) \\
&= \frac{1}{n} \sum_{r=1}^{n} \overline{Y} \\
&= \overline{Y}.
\end{aligned}
$$

Thus $\overline{z}$ is an unbiased estimator of population mean $\overline{Y}$.

The expression for variance of $\overline{z}$ in general case is complex but its estimate is simple.

Estimate of variance:

$Var(\overline{z}) = E(\overline{z}^2) - \overline{Y}^2.$

Consider for $r < s,$

$$
\begin{aligned}
E(z_r z_s) &= E\left[z_r E(z_s \mid U_1, U_2, \ldots, U_{s-1}) \right] \\
&= E\left[z_r \overline{Y} \right] \\
&= \overline{Y} E(z_r) \\
&= \overline{Y}^2
\end{aligned}
$$

because for $r < s,$ z_r will not contribute

and similarly for $s < r, z_s$ will not contribute in the expectation.

Further, for $s < r$,

$$\begin{aligned}E(z_r z_s) &= E\left[z_s E(z_r \mid U_1, U_2, ..., U_{r-1})\right]\\ &= E\left[z_s \overline{Y}\right]\\ &= \overline{Y} E(z_s)\\ &= \overline{Y}^2.\end{aligned}$$

Consider

$$\begin{aligned}E\left[\frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s\right] &= \frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} E(z_r z_s)\\ &= \frac{1}{n(n-1)} n(n-1)\overline{Y}^2\\ &= \overline{Y}^2.\end{aligned}$$

Substituting $\overline{Y}^2$ in $Var(\overline{z})$, we get

$$\begin{aligned}Var(\overline{z}) &= E(\overline{z}^2) - \overline{Y}^2\\ &= E(\overline{z}^2) - E\left[\frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} E(z_r z_s)\right]\\ \Rightarrow \widehat{Var}(\overline{z}) &= \overline{z}^2 - \frac{1}{n(n-1)}\sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s.\end{aligned}$$

Using $\left(\sum_{r=1}^{n} z_r\right)^2 = \sum_{r=1}^{n} z_r^2 + \sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s$

$$\Rightarrow \sum_{r(\neq s)=1}^{n}\sum_{s=1}^{n} z_r z_s = n^2\overline{z}^2 - \sum_{r=1}^{n} z_r^2,$$

The expression of $\widehat{Var}(\overline{z})$ can be further simplified as

$$\begin{aligned}\widehat{Var}(\overline{z}) &= \overline{z}^2 - \frac{1}{n(n-1)}\left[n^2\overline{z}^2 - \sum_{r=1}^{n} z_r^2\right]\\ &= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} z_r^2 - n\overline{z}^2\right]\\ &= \frac{1}{n(n-1)}\sum_{r=1}^{n}(z_r - \overline{z})^2.\end{aligned}$$

Unordered estimator:

In ordered estimator, the order in which the units are drawn is considered. Corresponding to any ordered estimator, there exist an unordered estimator which does not depend on the order in which the units are drawn and has smaller variance than the ordered estimator.

In case of sampling WOR from a population of size N, there are $\binom{N}{n}$ unordered sample(s) of size n.

Corresponding to any unordered sample(s) of size n units, there are $n!$ ordered samples.

For example, for $n = 2$ if the units are u_1 and u_2, then

- there are 2! ordered samples - (u_1, u_2) and (u_2, u_1)
- there is one unordered sample (u_1, u_2).

Moreover,

$$\begin{pmatrix}\text{Probability of unordered} \\ \text{sample } (u_1, u_2)\end{pmatrix} = \begin{pmatrix}\text{Probability of ordered} \\ \text{sample } (u_1, u_2)\end{pmatrix} + \begin{pmatrix}\text{Probability of ordered} \\ \text{sample } (u_2, u_1)\end{pmatrix}$$

For $n = 3,$ there are three units u_1, u_2, u_3 and

-there are following 3! = 6 ordered samples:

$$(u_1,u_2,u_3),(u_1,u_3,u_2),(u_2,u_1,u_3),(u_2,u_3,u_1),(u_3,u_1,u_2),(u_3,u_2,u_1)$$

- there is one unordered sample (u_1, u_2, u_3).

Moreover,

Probability of unordered sample
= Sum of probability of ordered sample, i.e.

$$P(u_1,u_2,u_3)+P(u_1,u_3,u_2)+P(u_2,u_1,u_3)+P(u_2,u_3,u_1)+P(u_3,u_1,u_2)+P(u_3,u_2,u_1),$$

Let z_{si}, $s = 1, 2, .., \binom{N}{n}$, $i = 1, 2, ..., n!(= M)$ be an estimator of population parameter θ based on ordered sample s_i. Consider a scheme of selection in which the probability of selecting the ordered sample (s_i) is p_{si}. The probability of getting the unordered sample(s) is the sum of the probabilities, i.e.,

$$p_s = \sum_{i=1}^{M} p_{si}.$$

For a population of size N with units denoted as 1, 2,…, N, the samples of size n are $n-$tuples. In the n^{th} draw, the sample space will consist of $N(N-1)...(N-n+1)$ unordered sample points.

$$p_{sio} = P\left[\text{selection of any ordered sample}\right] = \frac{1}{N(N-1)...(N-n+1)}$$

$$p_{siu} = P\left[\text{selection of any unordered sample}\right] = \frac{n!}{N(N-1)...(N-n+1)} = n!P\begin{bmatrix}\text{selection of any}\\ \text{ordered sample}\end{bmatrix}$$

then $\quad p_s = \sum_{i=1}^{M(=n!)} p_{sio} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$

Theorem: $\hat{\theta}_0 = z_{si},\ s = 1,2,...,\binom{N}{n};\ i = 1,2,...,M(=n!)$

and $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p_{si}^{'}$

are the ordered and unordered estimators of θ, then

(i) $E(\hat{\theta}_u) = E(\hat{\theta}_0)$

(ii) $Var(\hat{\theta}_u) \leq Var(\hat{\theta}_0)$

where z_{s_i} is a function of s_i^{th} ordered sample (hence a random variable) and p_{s_i} is the probability of selection of s_i^{th} ordered sample and $p_{s_i}^{'} = \frac{p_{si}}{p_s}$.

Proof: Total number of ordered sample = $n!\binom{N}{n}$

$$(i)\ E(\hat{\theta}_0) = \sum_{s=1}^{\binom{N}{n}}\sum_{i=1}^{M} z_{si}p_{si}$$

$$\begin{aligned}E(\hat{\theta}_u) &= \sum_{s=1}^{\binom{N}{n}}\left(\sum_{i=1}^{M} z_{si}p_{si}^{'}\right)p_s\\ &= \sum_s\left(\sum_i z_{si}\frac{p_{si}}{p_s}\right)p_s\\ &= \sum_s\sum_i z_{si}p_{si}\\ &= E(\hat{\theta}_0)\end{aligned}$$

(ii) Since $\hat{\theta}_0 = z_{si}$, so $\hat{\theta}_0^2 = z_{si}^2$ with probability $p_{si},\ i = 1,2,...,M,\ s = 1,2,...,\binom{N}{n}$.

Similarly, $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p_{si}^{'}$, so $\hat{\theta}_u^2 = \left(\sum_{i=1}^{M} z_{si}p_{si}^{'}\right)^2$ with probability p_s

Consider

$$Var(\hat{\theta}_0) = E(\hat{\theta}_0^2) - \left[E(\hat{\theta}_0)\right]^2$$

$$= \sum_s \sum_i z_{si}^2 p_{si} - \left[E(\hat{\theta}_0)\right]^2$$

$$Var(\hat{\theta}_u) = E(\hat{\theta}_u^2) - \left[E(\hat{\theta}_u)\right]^2$$

$$= \sum_s \left(\sum_i z_{si} p_{si}^{'} \right)^2 p_s - \left[E(\hat{\theta}_0)\right]^2$$

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i z_{si}^2 \, p_{si} - \sum_s \left(\sum_i z_{si} p_{si}^{'} \right)^2 p_s$$

$$= \sum_s \sum_i z_{si}^2 \, p_{si} + \sum_s \left(\sum_i z_{si} p'_{si} \right)^2 p_s$$

$$-2 \sum_s \left(\sum_i z_{si} \, p_{si}^{'} \right) \left(\sum_i z_{si} \, p'_{si} \right) p_s$$

$$= \sum_s \left[\sum_i z_{si}^2 \, p_{si} + \left(\sum_i z_{si} \, p_{si}^{'} \right)^2 \left(\sum_i p_{si} \right) - 2 \left(\sum_i z_{si} \, p_{si}^{'} \right) \left(\sum_i z_{si} \, p_{si} \right) p_s \right]$$

$$= \sum_s \left[\sum_i \left\{ z_{si}^2 \, p_{si} + \left(\sum_i z_{si} \, p_{si}^{'} \right)^2 p_{si} - 2 \left(\sum_i z_{si} \, p_{si}^{'} \right) z_{si} p_{si} \right\} \right]$$

$$= \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p_{si}^{'})^2 \, p_{si} \right] \geq 0$$

$$\Rightarrow Var(\hat{\theta}_0) - Var(\hat{\theta}_u) \geq 0$$

or $Var(\hat{\theta}_u) \leq Var(\hat{\theta}_0)$

Estimate of $Var(\hat{\theta}_u)$

Since

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p_{si}^{'})^2 \, p_{si} \right]$$

$$\widehat{Var}(\hat{\theta}_u) = \widehat{Var}(\hat{\theta}_0) - \sum_s \sum_i \left[\overline{(z_{si} - \sum_i z_{si} p_{si}^{'})^2 p_{si}} \right]$$

$$= \sum_i p_{si}^{'} \, \widehat{Var}(\hat{\theta}_0) - \sum_i p_{si}^{'} \overline{(z_{si} - \sum_i z_{si} p_{si}^{'})^2}.$$

Based on this result, now we use the ordered estimators to construct an unordered estimator. It follows from this theorem that the unordered estimator will be more efficient than the corresponding ordered estimators.

Murthy's unordered estimator corresponding to Des Raj's ordered estimator for the sample size 2

Suppose y_i and y_j are the values of units U_i and U_j selected in the first and second draws respectively with varying probability and WOR in a sample of size 2 and let p_i and p_j be the corresponding initial probabilities of selection. So now we have two ordered estimates corresponding to the ordered samples s_1^* and s_2^* as follows

$s_1^* = (y_i, y_j)$ with (U_i, U_j)

$s_2^* = (y_j, y_i)$ with (U_j, U_i)

which are given as

$$\bar{z}(s_1^*) = \frac{1}{2N}\left[(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_i + \frac{y_i}{p_i} + \frac{y_j(1-p_i)}{p_j}\right]$$

and

$$\bar{z}(s_2^*) = \frac{1}{2N}\left[(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_j + \frac{y_j}{p_j} + \frac{y_i(1-p_j)}{p_i}\right].$$

The probabilities corresponding to $\bar{z}(s_1^*)$ and $\bar{z}(s_2^*)$ are

$$p(s_1^*) = \frac{p_i p_j}{1-p_i}$$

$$p(s_2^*) = \frac{p_j p_i}{1-p_j}$$

$$\begin{aligned} p(s) &= p(s_1^*) + p(s_2^*) \\ &= \frac{p_i p_j (2 - p_i - p_j)}{(1-p_i)(1-p_j)} \end{aligned}$$

$$p'(s_1^*) = \frac{1-p_j}{2-p_i-p_j}$$

$$p'(s_2^*) = \frac{1-p_i}{2-p_i-p_j}.$$

Murthy's unordered estimate $\bar{z}(u)$ corresponding to the Des Raj's ordered estimate is given as

$$\begin{aligned}
\bar{z}(u) &= \bar{z}(s_1^*)p'(s_1) + \bar{z}(s_2^*)p'(s_2) \\
&= \frac{\bar{z}(s_1^*)p(s_1^*) + \bar{z}(s_2^*)p(s_2^*)}{p(s_1^*)+p(s_2^*)} \\
&= \frac{\left[\frac{1}{2N}\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}\left(\frac{p_ip_j}{1-p_i}\right)\right] + \left[\frac{1}{2N}\left\{(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right\}\left(\frac{p_jp_i}{1-p_j}\right)\right]}{\frac{p_ip_j}{1-p_i} + \frac{p_jp_i}{1-p_j}} \\
&= \frac{\frac{1}{2N}\left[\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}(1-p_j) + \left\{(1+p_j)\frac{y_j}{p_i} + (1-p_j)\frac{y_i}{p_i}\right\}(1-p_i)\right]}{(1-p_j)+(1-p_i)} \\
&= \frac{\frac{1}{2N}\left[(1-p_j)\frac{y_i}{p_i}\{(1+p_i)+(1-p_i)\} + (1-p_i)\frac{y_j}{p_j}\{(1-p_j)+(1+p_j\}\right]}{2-p_i-p_j} \\
&= \frac{(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}}{N(2-p_i-p_j)}.
\end{aligned}$$

Unbiasedness:

Note that y_i and p_i can take any one of the values out of $Y_1, Y_2, ..., Y_N$ and $P_1, P_2, ..., P_N$, respectively. Then y_j and p_j can take any one of the remaining values out of $Y_1, Y_2, ..., Y_N$ and $P_1, P_2, ..., P_N$, respectively, i.e., all the values except the values taken at the first draw. Now

$$E\left[\bar{z}(u)\right]=\frac{1}{N}\sum_{i<j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_iP_j}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}2\sum_{i<j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_jP_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}\sum_{i\neq j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_iP_j}{1-P_i}+\frac{P_jP_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$=\frac{1}{2N}\sum_{i\neq j}\left[\left\{(1-P_j)\frac{Y_i}{P_i}+(1-P_i)\frac{Y_j}{P_j}\right\}\left\{\frac{P_iP_j}{(1-P_i)(1-P_j)}\right\}\right]$$

$$=\frac{1}{2N}\sum_{i\neq j}\left[\frac{Y_iP_j}{1-P_i}+\frac{Y_jP_i}{1-P_j}\right]$$

Using result $\sum_{i\neq j=1}^{N}a_ib_j=\sum_{i=1}^{N}a_i\left\{\sum_{j=1}^{N}b_j-b_i\right\}$, we have

$$E\left[\bar{z}(u)\right]=\frac{1}{2N}\left[\left\{\sum_{i=1}^{N}\frac{Y_i}{1-P_i}(\sum_{j=1}^{N}P_j-P_i)\right\}+\left\{\sum_{j=1}^{N}\frac{Y_j}{1-P_j}(\sum_{i=1}^{N}P_i-P_j)\right\}\right]$$

$$=\frac{1}{2N}\left[\left\{\sum_{i=1}^{N}\frac{Y_i}{1-P_i}(1-P_i)\right\}+\sum_{j=1}^{N}\frac{Y_j}{1-P_j}(1-P_j)\right]$$

$$=\frac{1}{2N}\left\{\sum_{i=1}^{N}Y_i+\sum_{j=1}^{N}Y_j\right\}$$

$$=\frac{\bar{Y}+\bar{Y}}{2}$$

$$=\bar{Y}.$$

Variance: The variance of $\bar{z}(u)$ can be found as

$$Var\left[\bar{z}(u)\right]=\frac{1}{2}\sum_{i\neq j=1}^{N}\frac{(1-P_i-P_j)(1-P_i)(1-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2\frac{P_iP_j(2-P_i-P_j)}{(1-P_i)(1-P_j)}$$

$$=\frac{1}{2}\sum_{i\neq j=1}^{N}\frac{P_iP_j(1-P_i-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2$$

Using the theorem that $Var(\hat{\theta}_u)\leq Var(\hat{\theta}_0)$ we get

$$Var\left[\bar{z}(u)\right]\leq Var\left[\bar{z}(s_1^*)\right]$$

and $Var\left[\bar{z}(u)\right]\leq Var\left[\bar{z}(s_2^*)\right]$

Unbiased estimator of $V\left[\bar{z}(u)\right]$

An unbiased estimator of $Var\left(\bar{z}\,|\,u\right)$ is

$$\widehat{Var}\left[\bar{z}(u)\right]=\frac{(1-p_i-p_j)(1-p_i)(1-p_j)}{N^2(2-p_i-p_j)^2}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2.$$

Horvitz Thompson (HT) estimate

The unordered estimates have limited applicability as they lack simplicity and the expressions for the estimators and their variance becomes unmanageable when sample size is even moderately large. The HT estimate is simpler than other estimators. Let N be the population size and $y_i,\ (i=1,2,...,N)$ be the value of characteristic under study and a sample of size n is drawn by WOR using arbitrary probability of selection at each draw.

Thus prior to each succeeding draw, there is defined a new probability distribution for the units available at that draw. The probability distribution at each draw may or may not depend upon the initial probability at the first draw.

Define a random variable $\alpha_i (i=1,2,..,N)$ as

$$\alpha_i=\begin{cases}1 \text{ if } Y_i \text{ is included in a sample } 's' \text{ of size } n\\ 0 \text{ otherwise.}\end{cases}$$

Let $z_i = \dfrac{ny_i}{NE(\alpha_i)}$, $i = 1...N$ assuming $E(\alpha_i) > 0$ for all i

where

$$\begin{aligned} E(\alpha_i) &= 1.P(Y_i \in s) + 0.P(Y_i \notin s) \\ &= \pi_i \end{aligned}$$

is the probability of including the unit i in the sample and is called as **inclusion probability.**

The HT estimator of $\overline{Y}$ based on $y_1, y_2, ..., y_n$ is

$$\begin{aligned} \overline{z}_n = \hat{\overline{Y}}_{HT} &= \frac{1}{n}\sum_{i=1}^{n} z_i \\ &= \frac{1}{n}\sum_{i=1}^{N} \alpha_i z_i. \end{aligned}$$

Unbiasedness

$$\begin{aligned} E(\hat{\overline{Y}}_{HT}) &= \frac{1}{n}\sum_{i=1}^{N} E(z_i \alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} z_i E(\alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{NE(\alpha_i)} E(\alpha_i) \\ &= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{N} = \overline{Y} \end{aligned}$$

which shows that HT estimator is an unbiased estimator of population mean.

Variance

$$\begin{aligned} V(\hat{\overline{Y}}_{HT}) &= V(\overline{z}_n) \\ &= E(\overline{z}_n^2) - \left[E(\overline{z}_n)\right]^2 \\ &= E(\overline{z}_n^2) - \overline{Y}^2. \end{aligned}$$

Consider

$$\begin{aligned} E(\overline{z}_n^2) &= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i z_i\right]^2 \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j z_i z_j\right] \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_i z_j E(\alpha_i \alpha_j)\right]. \end{aligned}$$

If $S=\{s\}$ is the set of all possible samples and π_i is probability of selection of i^{th} unit in the sample s then

$$
\begin{aligned}
E(\alpha_i) &= 1\ P(y_i \in s) + 0.P(y_i \notin s)\\
&= 1.\pi_i\ + 0.(1-\pi_i) = \pi_i\\
E(\alpha_i^2) &= 1^2.P(y_i \in s) + 0^2.P(y_i \notin s)\\
&= \pi_i.
\end{aligned}
$$

So

$$
\begin{aligned}
E(\alpha_i) &= E(\alpha_i^2)\\
E(\overline{z}_n^2) &= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\pi_i + \sum_{i(\# j)}^{N}\sum_{i=1}^{N}\pi_{ij}z_iz_j\right]
\end{aligned}
$$

where π_{ij} is the probability of inclusion of i^{th} and j^{th} unit in the sample. This is called as **second order inclusion probability**.

Now

$$
\begin{aligned}
\overline{Y}^2 &= \left[E(\overline{z}_n)\right]^2\\
&= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{N}\alpha_i z_i\right)\right]^2\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\left[E(\alpha_i)\right]^2\right] + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_iz_jE(\alpha_i)E(\alpha_j)\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\pi_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_i\pi_jz_iz_j\right].
\end{aligned}
$$

Thus

$$
\begin{aligned}
Var(\hat{\overline{Y}}_{HT}) &= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_{ij}z_iz_j\right]\\
&\quad -\frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_i\pi_jz_iz_j\right]\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i\ (1-\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_iz_j\right]\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i\ (1-\pi_i)\frac{n^2y_i^2}{N^2\pi_i^2} + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)\frac{n^2y_iy_j}{N^2\pi_i\pi_j}\right]\\
&= \frac{1}{N^2}\left[\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_i}{\pi_i\pi_j}\right)y_iy_j\right]
\end{aligned}
$$

Estimate of variance

$$\hat{V}_1 = \widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2}\left[\sum_{i=1}^{n}\frac{y_i^2(1-\pi_i)}{\pi_i^2} + \sum_{i(\neq j)=1}^{n}\sum_{j=1}^{n}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\frac{y_i y_j}{\pi_i\pi_j}\right].$$

This is an unbiased estimator of variance .

Drawback: It does not reduces to zero when all $\frac{y_i}{\pi_i}$ are same, i.e., when $y_i \propto \pi_i$.

Consequently, this may assume negative values for some samples.

A more elegant expression for the variance of $\hat{\bar{y}}_{HT}$ has been obtained by Yates and Grundy.

Yates and Grundy form of variance

Since there are exactly n values of α_i which are 1 and $(N-n)$ values which are zero, so

$$\sum_{i=1}^{N}\alpha_i = n.$$

Taking expectation on both sides

$$\sum_{i=1}^{N}E(\alpha_i) = n.$$

Also

$$E\left(\sum_{i=1}^{N}\alpha_i\right)^2 = \sum_{i=1}^{N}E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_j)$$

$$E(n)^2 = \sum_{i=1}^{N}E(\alpha_i) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_J) \text{ (using } E(\alpha_i) = E(\alpha_i^2))$$

$$n^2 = n + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_J)$$

$$\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_J) = n(n-1)$$

Thus
$$\begin{aligned} E(\alpha_i\alpha_j) &= P(\alpha_i = 1, \alpha_j = 1) \\ &= P(\alpha_i = 1)P(\alpha_j = 1 | \alpha_i = 1) \\ &= E(\alpha_i)E(\alpha_j | \alpha_i = 1) \end{aligned}$$

Therefore

$$\begin{aligned}
&\sum_{j(\neq i)=1}^{N}\left[E(\alpha_i\,\alpha_j)-E(\alpha_i)E(\alpha_j)\right]\\
&\quad=\sum_{j(\neq i)=1}^{N}\left[E(\alpha_i)E(\alpha_j\,|\,\alpha_i=1)-E(\alpha_i)E(\alpha_j)\right]\\
&\quad=E(\alpha_i)\sum_{j(\neq i)=1}^{N}\left[E(\alpha_j\,|\,\alpha_i=1)-E(\alpha_j)\right]\\
&\quad=E(\alpha_i)\left[(n-1)-(n-E(\alpha_i)\right]\\
&\quad=-E(\alpha_i)\left[1-E(\alpha_i)\right]\\
&\quad=-\pi_i(1-\pi_i) \qquad\qquad\qquad\qquad\qquad (1)
\end{aligned}$$

Similarly

$$\sum_{i(\neq j)=1}^{N}\left[E(\alpha_i\,\alpha_j)-E(\alpha_i)E(\alpha_j)\right]=-\pi_j(1-\pi_j). \qquad (2)$$

We had earlier derived the variance of HT estimator as

$$Var(\hat{\bar{Y}}_{HT})=\frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2+\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)z_iz_j\right]$$

Using (1) and (2) in this expression, we get

$$\begin{aligned}
Var(\hat{\bar{Y}}_{HT})&=\frac{1}{2n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2+\sum_{j=1}^{N}\pi_j(1-\pi_j)z_j^2-2\sum_{i\neq j=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})z_iz_j\right]\\
&=\frac{1}{2n^2}\Bigg[-\sum_{i=1}^{N}\left\{\sum_{j(\neq i)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_i^2\\
&\qquad-\sum_{j=1}^{N}\left\{\sum_{i(\neq j)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_j^2-2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{n}\left\{E(\alpha_i)E(\alpha_j)-E(\alpha_i\alpha_j)\right\}z_iz_j\Bigg]
\end{aligned}$$

$$\begin{aligned}
&=\frac{1}{2n^2}\Bigg[\Bigg[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_i^2+\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_j^2+2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_iz_j\Bigg]\\
&=\frac{1}{2n^2}\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})(z_i^2+z_j^2-2z_iz_j)\right].
\end{aligned}$$

The expression for π_i and π_{ij} can be written for any given sample size.

For example, for $n=2$, 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

$E(\alpha_i) =$ Probability of selecting Y_i in a sample of two

$$= P_{i1} + P_{i2}$$

where P_{ir} is the probability of selecting Y_i at r^{th} draw $(r = 1,2)$. If P_i is the probability of selecting the i^{th} unit at first draw $(i = 1,2,...,N)$ then we had earlier derived that

$$P_{i1} = P_i$$

$$P_{i2} = P\begin{bmatrix} y_i \text{ is not selected} \\ \text{at } 1^{st} \text{ draw} \end{bmatrix} P\begin{bmatrix} y_{\text{i}} \text{ is selected at } 2^{nd} \text{ draw}| \\ y_{\text{i}} \text{ is not selected at } 1^{st} \text{ draw} \end{bmatrix}$$

$$= \sum_{j(\neq i)=1}^{N} \frac{P_j P_i}{1-P_j}$$

$$= \left[\sum_{j=1}^{N} \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i}\right] P_i.$$

So

$$E(\alpha_i) = P_i \left[\sum_{j=1}^{N} \frac{P_j}{1-P_j} - \frac{P_i}{1-P_i}\right]$$

Again

$E(\alpha_i \alpha_j) =$ Probability of including both y_i and y_j in a sample of size two

$$= P_{i1} \, P_{j2|i} + P_{j1} P_{i2|j}$$

$$= P_i \frac{P_j}{1-P_i} + P_j \frac{P_i}{1-P_j}$$

$$= P_i P_j \left[\frac{1}{1-P_i} + \frac{1}{1-P_j}\right].$$

Estimate of Variance

The estimate of variance is given by

$$\widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2} \sum_{i(\neq j)}^{n} \sum_{j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} (z_i - z_j)^2.$$

Midzuno system of sampling:

Under this system of selection of probabilities, the unit in the first draw is selected with unequal probabilities of selection (i.e., pps) and remaining all the units are selected with SRSWOR at all subsequent draws.

Under this system

$$E(\alpha_i) = \pi_i = P \text{ (unit } i \ (U_i) \text{ is included in the sample)}$$

$$= P(U_i \text{ is included in } 1^{st} \text{ draw}) + P(U_i \text{ is included in any other draw})$$

$$= P_i + \begin{pmatrix} \text{Probability that } U_i \text{ is not selected at the first draw and} \\ \text{is selected at any of subsequent } (n\text{-}1) \text{ draws} \end{pmatrix}$$

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

$$= \frac{N-n}{N-1}P_i + \frac{n-1}{N-1}.$$

Similarly,

$$E(\alpha_i\alpha_j) = \text{Probability that both the units } U_i \text{ and } U_j \text{ are in the sample}$$

$$= \begin{pmatrix} \text{Probability that } U_i \text{ is selected at the first draw and} \\ U_j \text{ is selected at any of the subsequent draws } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that } U_j \text{ is selected at the first draw and} \\ U_i \text{ is selected at any of the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that neither } U_i \text{ nor } U_j \text{ is selected at the first draw but} \\ \text{both of them are selected during the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$= P_i\frac{n-1}{N-1} + P_j\frac{n-1}{N-1} + (1 - P_i - P_j)\frac{(n-1)(n-2)}{(N-1)(N-2)}$$

$$= \frac{(n-1)}{(N-1)}\left[\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right]$$

$$\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right].$$

Similarly,

$$E(\alpha_i\alpha_j\alpha_k) = \pi_{ijk} = \text{Probability of including } U_i, U_j \text{ and } U_k \text{ in the sample}$$

$$= \frac{(n-1)(n-2)}{(N-1)(N-2)}\left[\frac{N-n}{N-3}(P_i + P_j + P_k) + \frac{n-3}{N-3}\right].$$

By an extension of this argument, if $U_i, U_j, ..., U_r$ are the r units in the sample of size $n(r < n)$, the probability of including these r units in the sample is

$$E(\alpha_i \alpha_j ... \alpha_r) = \pi_{ij...r} = \frac{(n-1)(n-2)...(n-r+1)}{(N-1)(N-2)...(N-r+1)}\left[\frac{N-n}{N-r}(P_i + P_j + ... + P_r) + \frac{n-r}{N-r}\right]$$

Similarly, if $U_1, U_2, ..., U_q$ be the n units, the probability of including these units in the sample is

$$E(\alpha_i \alpha_j ... \alpha_q) = \pi_{ij...q} = \frac{(n-1)(n-2)...1}{(N-1)(N-2)...(N-n+1)}(P_i + P_j + ... + P_q)$$

$$= \frac{1}{\binom{N-1}{n-1}}(P_i + P_j + ... + P_q)$$

which is obtained by substituting $r = n$.

Thus if P_i's are proportional to some measure of size of units in the population then the probability of selecting a specified sample is proportional to the total measure of the size of units included in the sample.

Substituting these $\pi_i, \pi_{ij}, \pi_{ijk}$ etc. in the HT estimator, we can obtain the estimator of population's mean and variance. In particular, an unbiased estimate of variance of HT estimator given by

$$\widehat{Var}(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2}\sum_{i \neq j=1}^{n}\sum_{j=1}^{n}\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}(z_i - z_j)^2$$

where

$$\pi_i \pi_j - \pi_{ij} = \frac{N-n}{(N-1)^2}\left[(N-n)P_i P_j + \frac{n-1}{N-2}(1 - P_i - P_j)\right].$$

The main advantage of this method of sampling is that it is possible to compute a set of revised probabilities of selection such that the inclusion probabilities resulting from the revised probabilities are proportional to the initial probabilities of selection. It is desirable to do so since the initial probabilities can be chosen proportional to some measure of size.