

**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 be 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 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. Organizational 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 constitute 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 regression 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

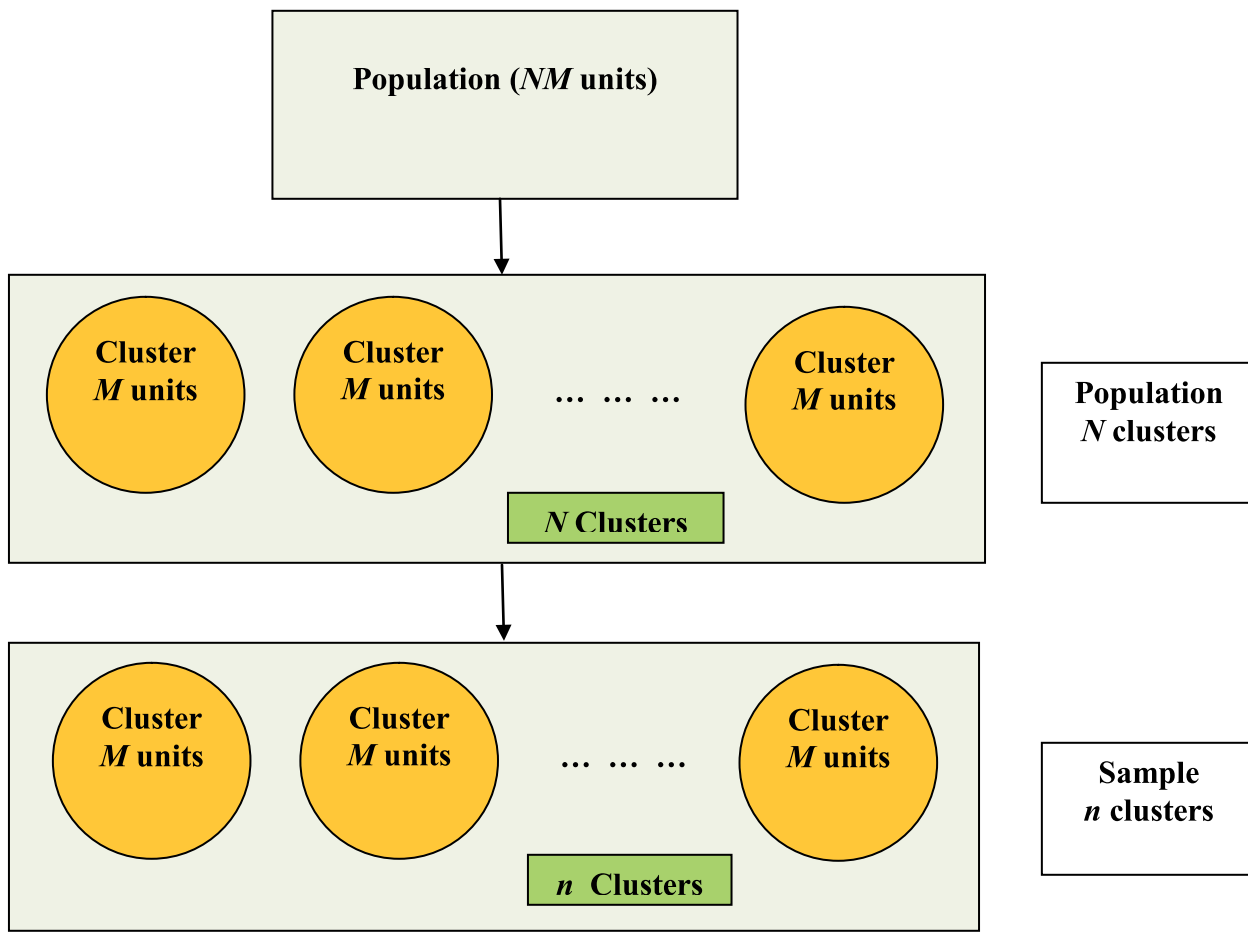
total population size =  $NM$

total sample size =  $nM$ .

Let

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

$\bar{y}_i = \frac{1}{M} \sum_{j=1}^M y_{ij}$  mean per element of  $i^{\text{th}}$  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  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n$ . 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:

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

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  $y_1, y_2, \dots, y_n$  whereas in case of  $\bar{y}_{cl}$ , the sampling units are  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_n$ .

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

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

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

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

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

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

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

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

Consider

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

where

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

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

The efficiency of cluster sampling over SRSWOR is

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

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 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

$$\begin{aligned} \rho &= \frac{E(y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{E(y_{ij} - \bar{Y})^2}; \quad -\frac{1}{M-1} \leq \rho \leq 1 \\ &= \frac{\frac{1}{MN(M-1)} \sum_{i=1}^N \sum_{j=1}^M \sum_{k(\neq j)=1}^M (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\frac{1}{MN} \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{Y})^2} \\ &= \frac{\frac{1}{MN(M-1)} \sum_{i=1}^N \sum_{j=1}^M \sum_{k(\neq j)=1}^M (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{\left(\frac{MN-1}{MN}\right) S^2} \\ &= \frac{\sum_{i=1}^N \sum_{j=1}^M \sum_{k(\neq j)=1}^M (y_{ij} - \bar{Y})(y_{ik} - \bar{Y})}{(MN-1)(M-1)S^2}. \end{aligned}$$

Consider

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

or

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

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

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

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

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

$$Var(\bar{y}_{cl}) \approx \frac{1}{nM} \frac{S^2}{M} [1+(M-1)\rho].$$

The variance of sample mean under SRSWOR for large  $N$  is

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

The relative efficiency for large  $N$  is now given by

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

- 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

$$E > 1$$

$$\text{or } (M-1)\rho < 0$$

$$\text{or } \rho < 0.$$

- 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,  $\rho$  is usually positive and  $\rho$  decreases as  $M$  increases but the rate of decrease in  $\rho$  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  $\bar{y}_i, i=1,2,\dots,N$  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  $S_i^2, i=1,2,\dots,N$ , 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

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

An unbiased estimator of  $S^2$  can be obtained as

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

So

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

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

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

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

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

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

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

and its estimate is

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

### 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  $j^{\text{th}}$  unit in the  $i^{\text{th}}$  cluster belongs to the specified category (i.e. possessing the given attribute) and  $y_{ij} = 0$  otherwise, we find that

$$\begin{aligned}
\bar{y}_i &= P_i, \\
\bar{Y} &= \frac{1}{N} \sum_{i=1}^N P_i = P, \\
S_i^2 &= \frac{MP_iQ_i}{(M-1)}, \\
S_w^2 &= \frac{M \sum_{i=1}^N P_iQ_i}{N(M-1)}, \\
S^2 &= \frac{NMPQ}{NM-1}, \\
S_b^2 &= \frac{1}{N-1} \sum_{i=1}^N (P_i - P)^2, \\
&= \frac{1}{N-1} \left[ \sum_{i=1}^N P_i^2 - NP^2 \right] \\
&= \frac{1}{(N-1)} \left[ -\sum_{i=1}^N P_i(1-P_i) + \sum_{i=1}^N P_i - NP^2 \right] \\
&= \frac{1}{(N-1)} \left[ NPQ - \sum_{i=1}^N P_iQ_i \right],
\end{aligned}$$

where  $P_i$  is the proportion of elements in the  $i^{\text{th}}$  cluster, belonging to the specified category and  $Q_i = 1 - P_i$ ,  $i = 1, 2, \dots, N$  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  $\rho$  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  $\rho$ , 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

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

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

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

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

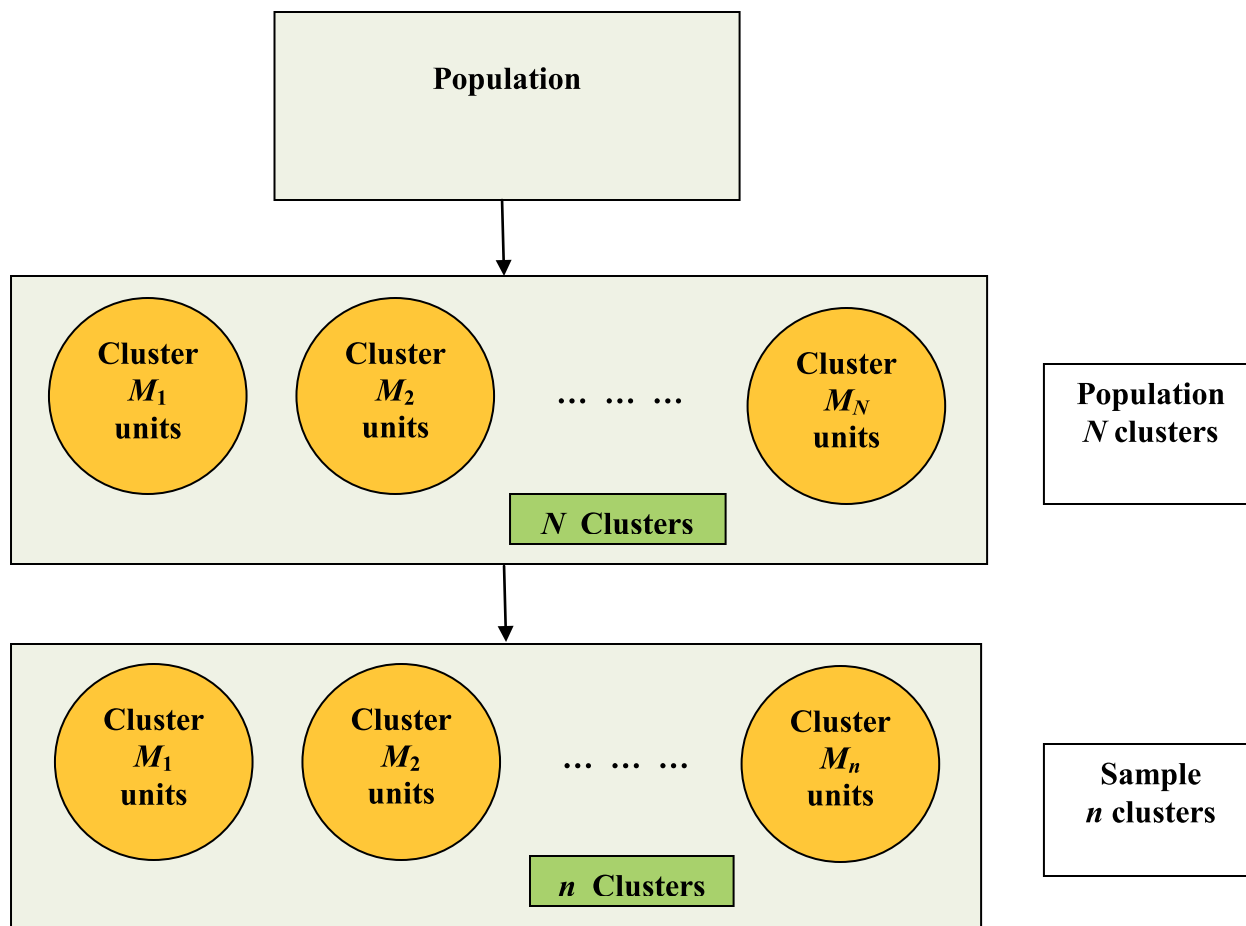
$$\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} : \text{mean of } i^{th} \text{ cluster}$$

$$\bar{Y} = \frac{1}{M_0} \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$$

$$= \sum_{i=1}^N \frac{M_i}{M_0} \bar{y}_i$$

$$= \frac{1}{N} \sum_{i=1}^N \frac{M_i}{M_0} \bar{y}_i$$

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

$$\begin{aligned}\bar{\bar{y}}_c &= \frac{1}{n} \sum_{i=1}^n \bar{y}_i \\ E(\bar{\bar{y}}_c) &= \frac{1}{N} \sum_{i=1}^N \bar{y}_i \\ &\neq \bar{Y} \quad (\text{where } \bar{Y} = \sum_{i=1}^N \frac{M_i}{M_0} \bar{y}_i).\end{aligned}$$

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

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

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

The mean squared error is

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

where

$$\begin{aligned}S_b^2 &= \frac{1}{N-1} \sum_{i=1}^N (\bar{y}_i - \bar{Y})^2 \\ S_{m\bar{y}} &= \frac{1}{N-1} \sum_{i=1}^N (M_i - \bar{M})(\bar{y}_i - \bar{Y}).\end{aligned}$$

An estimate of  $Var(\bar{y}_c)$  is

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

where  $s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_c - \bar{y}_c)^2$ .

## 2. Weighted mean of cluster means

Consider the arithmetic mean based on cluster total as

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

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

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

where

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

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

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

Since SRSWOR is followed, so

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

Since

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

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

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

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

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

$$V_i = \frac{M_i}{\bar{M}} \quad i=1,2,\dots,N$$

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

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

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

The ratio estimator based on  $U$  and  $V$  is

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

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

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

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

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

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,

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

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

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

Since SRSWOR is used, so

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

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

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

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

So

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

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  $\frac{M_i}{M} \bar{y}_i$  and auxiliary variable (earlier  $x_i$ ) by  $\frac{M_i}{M}$ . 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

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

A consistent estimate of this variance is

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

The variance of  $\bar{y}_c^{**}$  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  $\frac{M_i}{\bar{M}}$  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  $i^{\text{th}}$  cluster which is given by

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

Consider the following estimator of the population mean:

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

Then this estimator can be expressed as

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

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

$$E(\alpha_i) = nP_i, \quad \text{Var}(\alpha_i) = nP_i(1 - P_i)$$

$$\text{Cov}(\alpha_i, \alpha_j) = -nP_iP_j, \quad i \neq j.$$

Hence,

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

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



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

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

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

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

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

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

Consider

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

where  $E(\alpha_i) = nP_i$ ,  $\text{Var}(\alpha_i) = nP_i(1 - P_i)$ ,  $\text{Cov}(\alpha_i, \alpha_j) = -nP_i P_j$ ,  $i \neq j$

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

## 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  $j^{\text{th}}$  second stage units of the  $i^{\text{th}}$  first stage unit;  $i = 1, 2, \dots, N$ ;  $j = 1, 2, \dots, m$ .

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

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

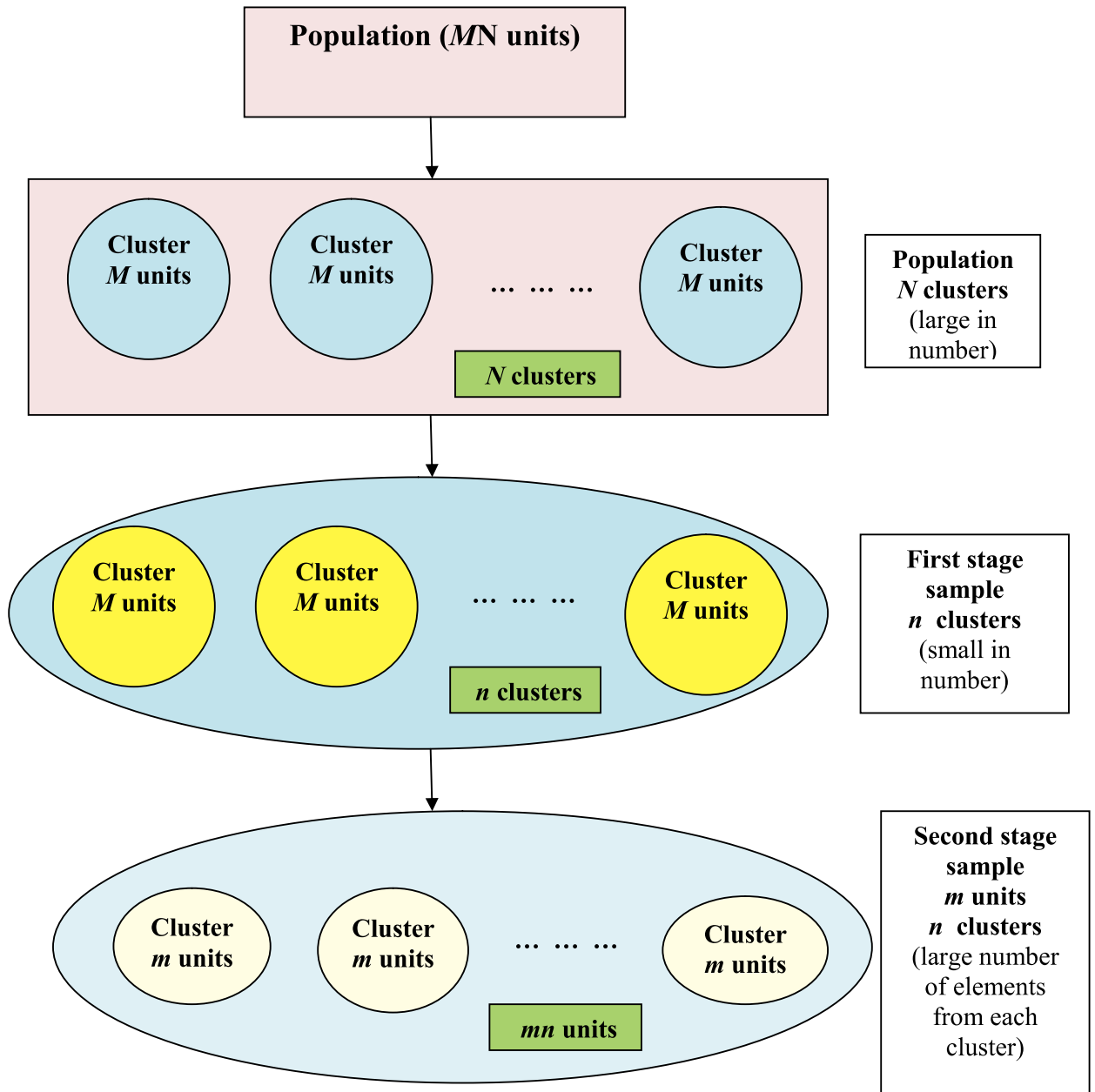
$\bar{y}_i = \frac{1}{n} \sum_{j=1}^m y_{ij}$  : mean per second stage unit in the  $i^{\text{th}}$  first stage unit in the sample.

$\bar{y} = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m y_{ij} = \frac{1}{n} \sum_{i=1}^n \bar{y}_i = \bar{y}_{mn}$  : 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,

$$E(\hat{\theta}) = E_1[E_2(\hat{\theta})]$$

↓

↓

↘

average over all samples	average over all 1 <sup>st</sup> stage samples	average over all possible 2 <sup>nd</sup> stage selections from a fixed set of units
-----------------------------------	---	---

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,

$$\begin{aligned} Var(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E_1 E_2 (\hat{\theta} - \theta)^2 \end{aligned}$$

Consider

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

Now average over first stage selection as

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

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

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

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

### Variance

$$\begin{aligned} \text{Var}(\bar{y}) &= E_1[V_2(\bar{y}|i)] + V_1[E_2(\bar{y}/i)] \\ &= E_1\left[V_2\left\{\frac{1}{n} \sum_{i=1}^n \bar{y}_i | i\right\}\right] + V_1\left[E_2\left\{\frac{1}{n} \sum_{i=1}^n \bar{y}_i / i\right\}\right] \\ &= E_1\left[\frac{1}{n^2} \sum_{i=1}^n V(\bar{y}_i | i)\right] + V_1\left[\frac{1}{n} \sum_{i=1}^n E_2(\bar{y}_i / i)\right] \\ &= E_1\left[\frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2\right] + V_1\left[\frac{1}{n} \sum_{i=1}^n \bar{Y}_i\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{m} - \frac{1}{M}\right) E_1(S_i^2) | i + V_1(\bar{y}_c) \\ &\quad (\text{where } \bar{y}_c \text{ is based on cluster means as in cluster sampling}) \\ &= \frac{1}{n^2} n \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 + \frac{N-n}{Nn} S_b^2 \\ &= \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 + \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 \end{aligned}$$

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

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

## Estimate of variance

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

Consider an estimator of

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

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

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

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

So

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

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

Consider

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

as an estimator of

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

So

$$\begin{aligned} E(s_b^2) &= \frac{1}{n-1} E \left[ \sum_{i=1}^n (\bar{y}_i - \bar{y})^2 \right] \\ (n-1)E(s_b^2) &= E \left[ \sum_{i=1}^n \bar{y}_i^2 - n\bar{y}^2 \right] \\ &= E \left[ \sum_{i=1}^n \bar{y}_i^2 \right] - nE(\bar{y}^2) \\ &= E_1 \left[ E_2 \left( \sum_{i=1}^n \bar{y}_i^2 \right) \right] - n \left[ \text{Var}(\bar{y}) + \{E(\bar{y})\}^2 \right] \\ &= E_1 \left[ \sum_{i=1}^n E_2(\bar{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} \bar{S}_w^2 + \bar{Y}^2 \right] \\ &= E_1 \left[ \sum_{i=1}^n \left\{ \text{Var}(\bar{y}_i) + (E(\bar{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} \bar{S}_w^2 + \bar{Y}^2 \right] \\ &= E_1 \left[ \sum_{i=1}^n \left\{ \left( \frac{1}{m} - \frac{1}{M} \right) S_i^2 + \bar{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} \bar{S}_w^2 + \bar{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 + \bar{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} \bar{S}_w^2 + \bar{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 \bar{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} \bar{S}_w^2 + \bar{Y}^2 \right] \\ &= n \left[ \left( \frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 + \frac{1}{N} \sum_{i=1}^N \bar{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} \bar{S}_w^2 + \bar{Y}^2 \right] \\ &= (n-1) \left( \frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 + \frac{n}{N} \sum_{i=1}^N \bar{Y}_i^2 - n\bar{Y}^2 - n \left( \frac{1}{n} - \frac{1}{N} \right) S_b^2 \\ &= (n-1) \left( \frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 + \frac{n}{N} \left[ \sum_{i=1}^N \bar{Y}_i^2 - N\bar{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) \bar{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) \bar{S}_w^2 + (n-1) S_b^2. \end{aligned}$$

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

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



Thus

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

### 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

$$C = kmn$$

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  $C = k_1n + k_2mn$**

Let cost  $C$  be fixed as  $C_0 = k_1n + k_2mn$  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  $C_0 = k_1n + k_2mn$ .

We have

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

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

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

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

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

The optimum  $n$  follows from  $C_0 = k_1n + k_2mn$  as

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

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

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

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

$$C_0 \geq k_1 + k_2M; \hat{m} = M \text{ and } \hat{n} = \frac{C_0}{k_1 + k_2M}.$$

If  $C_0 < k_1 + k_2M$ ; then  $\hat{m} = \frac{C_0 - k_1}{k_2}$  and  $\hat{n} = 1$ .

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

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

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

## Case 2: When variance is fixed

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

$$\begin{aligned} V_0 &= \frac{1}{n} \left( \frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2 + \left( \frac{1}{n} - \frac{1}{N} \right) \bar{S}_b^2 \\ \Rightarrow n &= \frac{S_b^2 + \left( \frac{1}{m} - \frac{1}{M} \right) \bar{S}_w^2}{V_0 + \frac{S_b^2}{N}} \end{aligned}$$

So

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

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

If  $\left( S_b^2 - \frac{\bar{S}_w^2}{M} \right) < 0$ ,  $C$  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

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

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

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

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

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

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  $Var(\bar{y}_{cl})$  exceeds  $Var(\bar{y}_{TS})$  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  $S_b^2 = \frac{MN-1}{M(N-1)} \frac{S^2}{M} [1 + \rho(M-1)]$

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

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  $j^{\text{th}}$  second stage unit of the  $i^{\text{th}}$  first stage unit.

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

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

$m_i$  : number of second stage units to be selected from  $i^{\text{th}}$  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.

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

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

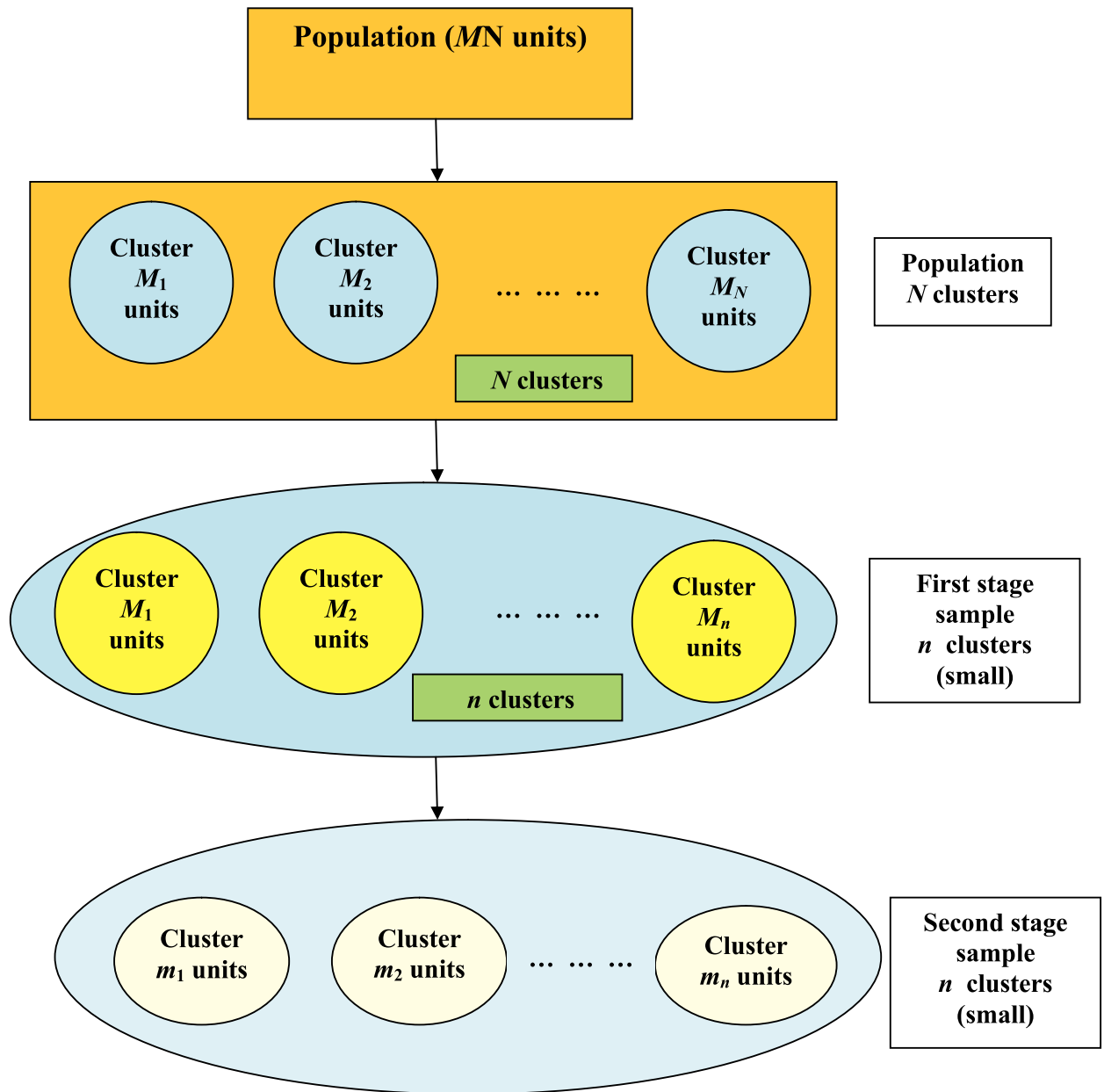
$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N \bar{y}_i = \bar{\bar{Y}}_N$$

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

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

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

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



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

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

$$\hat{Y} = \bar{y}_{S_2} = \frac{1}{n} \sum_{i=1}^n \bar{y}_{i(m_i)}$$

#### Bias

$$\begin{aligned} E(\bar{y}_{S_2}) &= E \left[ \frac{1}{n} \sum_{i=1}^n \bar{y}_{i(m_i)} \right] \\ &= E_1 \left[ \frac{1}{n} \sum_{i=1}^n E_2(\bar{y}_{i(m_i)}) \right] \\ &= E_1 \left[ \frac{1}{n} \sum_{i=1}^n \bar{Y}_i \right] \quad [\text{Since a sample of size } m_i \text{ is selected out of } M_i \text{ units by SRSWOR}] \\ &= \frac{1}{N} \sum_{i=1}^N \bar{Y}_i \\ &= \bar{\bar{Y}}_N \\ &\neq \bar{Y}. \end{aligned}$$

So  $\bar{y}_{S_2}$  is a biased estimator of  $\bar{Y}$  and its bias is given by

$$\begin{aligned} \text{Bias}(\bar{y}_{S_2}) &= E(\bar{y}_{S_2}) - \bar{Y} \\ &= \frac{1}{N} \sum_{i=1}^N \bar{Y}_i - \frac{1}{NM} \sum_{i=1}^N M_i \bar{Y}_i \\ &= -\frac{1}{NM} \left[ \sum_{i=1}^N M_i \bar{Y}_i - \frac{1}{N} \left( \sum_{i=1}^N \bar{Y}_i \right) \left( \sum_{i=1}^N M_i \right) \right] \\ &= \frac{1}{NM} \sum_{i=1}^N (M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N). \end{aligned}$$

This bias can be estimated by

$$\widehat{\text{Bias}}(\bar{y}_{S_2}) = -\frac{N-1}{NM(n-1)} \sum_{i=1}^n (M_i - \bar{m})(\bar{y}_{i(m_i)} - \bar{y}_{S_2})$$

which can be seen as follows:

$$\begin{aligned} E \left[ \widehat{\text{Bias}}(\bar{y}_{S_2}) \right] &= -\frac{N-1}{NM} E_1 \left[ \frac{1}{n-1} \sum_{i=1}^n E_2 \left\{ (M_i - \bar{m})(\bar{y}_{i(m_i)} - \bar{y}_{S_2}) / n \right\} \right] \\ &= -\frac{N-1}{NM} E \left[ \frac{1}{n-1} \sum_{i=1}^n (M_i - \bar{m})(\bar{Y}_i - \bar{y}_n) \right] \\ &= -\frac{1}{NM} \sum_{i=1}^N (M_i - \bar{M})(\bar{Y}_i - \bar{\bar{Y}}_N) \\ &= \bar{\bar{Y}}_N - \bar{Y} \end{aligned}$$

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



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

$$\bar{y}_{S_2} + \frac{N-1}{NM} \frac{1}{N-1} \sum_{i=1}^n (M_i - \bar{m})(\bar{y}_{i(mi)} - \bar{y}_{S_2}).$$

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:**

$$\begin{aligned} Var(\bar{y}_{S_2}) &= E[Var(\bar{y}_{S_2}|n)] + Var[E(\bar{y}_{S_2}|n)] \\ &= Var\left[\frac{1}{n} \sum_{i=1}^n \bar{y}_i\right] + E\left[\frac{1}{n^2} \sum_{i=1}^n Var(\bar{y}_{i(mi)}|i)\right] \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + E\left[\frac{1}{n^2} \sum_{i=1}^n \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2\right] \\ &= \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{Nn} \sum_{i=1}^N \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \end{aligned}$$

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

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

The MSE can be obtained as

$$MSE(\bar{y}_{S_2}) = Var(\bar{y}_{S_2}) + [Bias(\bar{y}_{S_2})]^2.$$

**Estimation of variance:**

Consider mean square between cluster means in the sample

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_{i(mi)} - \bar{y}_{S_2})^2.$$

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

$$\widehat{Var}(\bar{y}_{S_2}) = \left( \frac{1}{n} - \frac{1}{N} \right) \hat{S}_b^2 + \frac{1}{Nn} \sum_{i=1}^N \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \hat{S}_i^2.$$

## 2. Estimation based on first stage unit totals:

$$\begin{aligned} \hat{Y} = \bar{y}_{S_2}^* &= \frac{1}{n} \sum_{i=1}^n \frac{M_i \bar{y}_{i(mi)}}{\bar{M}} \\ &= \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_{i(mi)} \end{aligned}$$

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

### Bias

$$\begin{aligned} E(y_{S_2}^*) &= E \left[ \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_{i(mi)} \right] \\ &= E \left[ \frac{1}{n} \sum_{i=1}^n u_i E_2(\bar{y}_{i(mi)} | i) \right] \\ &= E \left[ \frac{1}{n} \sum_{i=1}^n u_i \bar{Y}_i \right] \\ &= \frac{1}{N} \sum_{i=1}^N u_i \bar{Y}_i \\ &= \bar{Y}. \end{aligned}$$

Thus  $\bar{y}_{S_2}^*$  is an unbiased estimator of  $\bar{Y}$ .

### Variance:

$$\begin{aligned} Var(\bar{y}_{S_2}^*) &= Var \left[ E(\bar{y}_{S_2}^* | n) \right] + E \left[ Var(\bar{y}_{S_2}^* | n) \right] \\ &= Var \left[ \frac{1}{n} \sum_{i=1}^n u_i \bar{Y}_i \right] + E \left[ \frac{1}{n^2} \sum_{i=1}^n u_i^2 Var(\bar{y}_{i(mi)} | i) \right] \\ &= \left( \frac{1}{n} - \frac{1}{N} \right) S_b^{*2} + \frac{1}{nN} \sum_{i=1}^N u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2 \end{aligned}$$

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

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

### 3. Estimator based on ratio estimator:

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

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

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

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

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

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

$$\begin{aligned} \bar{y}^* &= \frac{1}{n} \sum_{i=1}^n y_i^* = \bar{y}_{S2}^* \\ \bar{x}^* &= \frac{1}{n} \sum_{i=1}^n x_i^* = \bar{u}_n \\ \bar{X}^* &= \frac{1}{N} \sum_{i=1}^N X_i^* = 1. \end{aligned}$$

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

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

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

$$\begin{aligned}
Bias(\hat{y}_R) &\approx \frac{N-n}{Nn} \bar{Y} (C_x^2 - 2\rho C_x C_y) \\
&= \bar{Y} \left[ \frac{Var(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x}, \bar{y})}{\bar{X}\bar{Y}} \right] \\
MSE(\hat{Y}_R) &\approx [Var(\bar{y}) + R^2 Var(\bar{x}) - 2RCov(\bar{x}, \bar{y})]
\end{aligned}$$

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

### 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  $\bar{x}_{S2}^* = \frac{1}{n} \sum_{i=1}^n \bar{x}_{i(mi)}$ .

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

$$\begin{aligned}
Cov(\bar{x}_{S2}^*, \bar{y}_{S2}^*) &= Cov \left[ E \left( \frac{1}{n} \sum_{i=1}^n u_i \bar{x}_{i(mi)}, \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_{i(mi)} \right) \right] + E \left[ Cov \left( \frac{1}{n} \sum_{i=1}^n u_i \bar{x}_{i(mi)}, \frac{1}{n} \sum_{i=1}^n u_i \bar{y}_{i(mi)} \right) \right] \\
&= Cov \left[ \frac{1}{n} \sum_{i=1}^n u_i E(\bar{x}_{i(mi)}), \frac{1}{n} \sum_{i=1}^n u_i E(\bar{y}_{i(mi)}) \right] + E \left[ \frac{1}{n^2} \sum_{i=1}^n u_i^2 Cov(\bar{x}_{i(mi)}, \bar{y}_{i(mi)}) | i \right] \\
&= Cov \left[ \frac{1}{n} \sum_{i=1}^n u_i \bar{X}_i, \frac{1}{n} \sum_{i=1}^n u_i \bar{Y}_i \right] + E \left[ \frac{1}{n^2} \sum_{i=1}^n u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{ixy} \right] \\
&= \left( \frac{1}{n} - \frac{1}{N} \right) S_{bxy}^* + \frac{1}{nN} \sum_{i=1}^N u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{ixy}
\end{aligned}$$

where

$$\begin{aligned}
S_{bxy}^* &= \frac{1}{N-1} \sum_{i=1}^N (u_i \bar{X}_i - \bar{X})(u_i \bar{Y}_i - \bar{Y}) \\
S_{ixy} &= \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (x_{ij} - \bar{X}_i)(y_{ij} - \bar{Y}_i).
\end{aligned}$$

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

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

Substituting  $Cov(\bar{x}_{S_2}^*, \bar{y}_{S_2}^*)$  and  $Var(\bar{x}_{S_2}^*)$  in  $Bias(\bar{y}_{S_2}^{**})$ , we obtain the approximate bias as

$$Bias(\bar{y}_{S_2}^{**}) \approx \bar{Y} \left[ \left( \frac{1}{n} - \frac{1}{N} \right) \left( \frac{S_{bx}^{*2}}{\bar{X}^2} - \frac{S_{bxy}^*}{\bar{X}\bar{Y}} \right) + \frac{1}{nN} \sum_{i=1}^N \left\{ u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) \left( \frac{S_{ix}^2}{\bar{X}^2} - \frac{S_{ixy}}{\bar{X}\bar{Y}} \right) \right\} \right].$$

### Mean squared error

$$MSE(\bar{y}_{S_2}^{**}) \approx Var(\bar{y}_{S_2}^*) - 2R^* Cov(\bar{x}_{S_2}^*, \bar{y}_{S_2}^*) + R^{*2} Var(\bar{x}_{S_2}^*)$$

$$Var(\bar{y}_{S_2}^{**}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_{by}^{*2} + \frac{1}{nN} \sum_{i=1}^N u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{iy}^2$$

$$Var(\bar{x}_{S_2}^{**}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_{bx}^{*2} + \frac{1}{nN} \sum_{i=1}^N u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{ix}^2$$

$$Cov(\bar{x}_{S_2}^*, \bar{y}_{S_2}^*) = \left( \frac{1}{n} - \frac{1}{N} \right) S_{bxy}^* + \frac{1}{nN} \sum_{i=1}^N u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) S_{ixy}$$

where

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

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

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

Thus

$$MSE(\bar{y}_{S_2}^{**}) \approx \left( \frac{1}{n} - \frac{1}{N} \right) (S_{by}^{*2} - 2R^* S_{bxy}^* + R^{*2} S_{bx}^{*2}) + \frac{1}{nN} \sum_{i=1}^N \left[ u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) (S_{iy}^2 - 2R^* S_{ixy} + R^{*2} S_{ix}^2) \right].$$

Also

$$MSE(\bar{y}_{S_2}^{**}) \approx \left( \frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^N u_i^2 (\bar{Y}_i - R^* \bar{X}_i)^2 + \frac{1}{nN} \sum_{i=1}^N \left[ u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) (S_{iy}^2 - 2R^* S_{ixy} + R^{*2} S_{ix}^2) \right].$$

### Estimate of variance

Consider

$$s_{bxy}^* = \frac{1}{n-1} \sum_{i=1}^n \left[ (u_i \bar{y}_{i(mi)} - \bar{y}_{S_2}^*) (u_i \bar{x}_{i(mi)} - \bar{x}_{S_2}^*) \right]$$

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

It can be shown that

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

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

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  $\bar{y}_{S2}^{**}$  can be obtained by substituting the unbiased estimators of respective statistics in  $MSE(\bar{y}_{S2}^{**})$  as

$$\begin{aligned} \widehat{MSE}(\bar{y}_{S2}^{**}) &\approx \left( \frac{1}{n} - \frac{1}{N} \right) (s_{by}^{*2} - 2r^* s_{bxy}^* + r^{*2} s_{bx}^{*2}) \\ &\quad + \frac{1}{nN} \sum_{i=1}^n u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) (s_{iy}^2 - 2r^* s_{ixy} + r^{*2} s_{ix}^2) \\ &\approx \left( \frac{1}{n} - \frac{1}{N} \right) \frac{1}{n-1} \sum_{i=1}^n (\bar{y}_{i(mi)} - r^* \bar{x}_{i(mi)})^2 \\ &\quad + \frac{1}{nN} \sum_{i=1}^n \left[ u_i^2 \left( \frac{1}{m_i} - \frac{1}{M_i} \right) (s_{iy}^2 - 2r^* s_{ixy} + r^{*2} s_{ix}^2) \right] \end{aligned}$$

where  $r^* = \frac{\bar{y}_{S2}^*}{\bar{x}_{S2}^*}$ .

## 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.







































































































