Notes On Sample Survey

Chapter 1
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.

Chapter -2

Simple Random Sampling

Simple random sampling (SRS) is a method of selection of a sample comprising of n number of sampling units out of the population having N number of sampling units such that every sampling unit has an equal chance of being chosen.

The samples can be drawn in two possible ways.

- The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population .
- The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

1. Simple random sampling without replacement (SRSWOR):

SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, anyone of the remaining units have same chance of being selected, i.e. $1/N$.

2. Simple random sampling with replacement (SRSWOR):

SRSWR is a method of selection of n units out of the N units one by one such that at each stage of selection each unit has equal chance of being selected, i.e., $1/N$..

Procedure of selection of a random sample:

The procedure of selection of a random sample follows the following steps:

1. Identify the N units in the population with the numbers 1 to N.
2. Choose any random number arbitrarily in the random number table and start reading numbers.
3. Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.
4. In case of SRSWR, all the random numbers are accepted ever if repeated more than once.

In case of SRSWOR, if any random number is repeated, then it is ignored and more numbers are drawn.

Such process can be implemented through programming and using the discrete uniform distribution. Any number between 1 and N can be generated from this distribution and corresponding unit can be seleced into the sample by associating an index with each sampling unit. Many statistical softwares like R, SAS, etc. have inbuilt functions for drawing a sample using SRSWOR or SRSWR.

Notations:

The following notations will be used in further notes:

N : Number of sampling units in the population (Population size).

n : Number of sampling units in the sample (sample size)

Y : The characteristic under consideration

Y_i : Value of the characteristic for the i^{th} unit of the population

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: sample mean

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$: population mean

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

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

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

Probability of drawing a sample :

1.SRSWOR:

If n units are selected by SRSWOR, the total number of possible samples are $\binom{N}{n}$.

So the probability of selecting any one of these samples is $\frac{1}{\binom{N}{n}}$.

Note that a unit can be selected at any one of the n draws. Let u_i be the i^{th} unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, …, or n^{th} draw.

Let $P_j(i)$ denotes the probability of selection of u_i at the j^{th} draw, $j = 1,2,...,n$. Then

$$P_j(i) = P_1(i) + P_2(i) + ... + P_n(i)$$
$$= \frac{1}{N} + \frac{1}{N} + ... + \frac{1}{N} \ (n \ times)$$
$$= \frac{n}{N}$$

Now if $u_1, u_2, ..., u_n$ are the n units selected in the sample, then the probability of their selection is

$$P(u_1, u_2, ..., u_n) = P(u_1).P(u_2), ..., P(u_n)$$

Note that when the second unit is to be selected, then there are $(n-1)$ units left to be selected in the sample from the population of $(N-1)$ units. Similarly, when the third unit is to be selected, then there are $(n-2)$ units left to be selected in the sample from the population of $(N-2)$ units and so on.

If $P(u_1) = \frac{n}{N}$, then

$$P(u_2) = \frac{n-1}{N-1}, ..., P(u_n) = \frac{1}{N-n+1}.$$

Thus

$$P(u_1, u_2, .., u_n) = \frac{n}{N}.\frac{n-1}{N-1}.\frac{n-2}{N-2}...\frac{1}{N-n+1} = \frac{1}{\binom{N}{n}}.$$

Alternative approach:

The probability of drawing a sample in SRSWOR can alternatively be found as follows:

Let $u_{i(k)}$ denotes the i^{th} unit drawn at the k^{th} draw. Note that the i^{th} unit can be any unit out of the N units. Then $s_o = (u_{i(1)}, u_{i(2)}, ..., u_{i(n)})$ is an ordered sample in which the order of the units in which they are drawn, i.e., $u_{i(1)}$ drawn at the first draw, $u_{i(2)}$ drawn at the second draw and so on, is also considered. The probability of selection of such an ordered sample is

$$P(s_o) = P(u_{i(1)})P(u_{i(2)} \mid u_{i(1)})P(u_{i(3)} \mid u_{i(1)}u_{i(2)})...P(u_{i(n)} \mid u_{i(1)}u_{i(2)}...u_{i(n-1)}).$$

Here $P(u_{i(k)} \mid u_{i(1)}u_{i(2)}...u_{i(k-1)})$ is the probability of drawing $u_{i(k)}$ at the k^{th} draw given that $u_{i(1)}, u_{i(2)}, ..., u_{i(k-1)}$ have already been drawn in the first $(k-1)$ draws.

Such probability is obtained as

$$P(u_{i(k)} \mid u_{i(1)}u_{i(2)}...u_{i(k-1)}) = \frac{1}{N-k+1}.$$

So

$$P(s_o) = \prod_{k=1}^{n} \frac{1}{N-k+1} = \frac{(N-n)!}{N!}.$$

The number of ways in which a sample of size n can be drawn $= n!$

Probability of drawing a sample in a given order $= \frac{(N-n)!}{N!}$

So the probability of drawing a sample in which the order of units in which they are drawn is

irrelevant $= n!\frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$

2. SRSWR

When n units are selected with SRSWR, the total number of possible samples are N^n. The Probability of drawing a sample is $\frac{1}{N^n}$.

Alternatively, let u_i be the i^{th} unit selected in the sample. This unit can be selected in the sample either at first draw, second draw, …, or n^{th} draw. At any stage, there are always N units in the population in case of SRSWR, so the probability of selection of u_i at any stage is $1/N$ for all $i = 1,2,\ldots,n$. Then the probability of selection of n units $u_1, u_2, ..., u_n$ in the sample is

$$\begin{aligned} P(u_1, u_2, .., u_n) &= P(u_1).P(u_2)...P(u_n) \\ &= \frac{1}{N}.\frac{1}{N}...\frac{1}{N} \\ &= \frac{1}{N^n} \end{aligned}$$

Probability of drawing an unit

1. SRSWOR

Let A_e denotes an event that a particular unit u_j is not selected at the ℓ^{th} draw. The probability of selecting, say, j^{th} unit at k^{th} draw is

$$P \text{ (selection of } u_j \text{ at } k^{th} \text{ draw)} = P(A_1 \cap A_2 \cap .... \cap A_{k-1} \cap \bar{A}_k)$$

$$= P(A_1)P(A_2|A_1)\,P(A_3|A_1A_2).....P(A_{k-1}|A_1, A_2.....A_{k-2})\,P(\bar{A}_k|A_1, A_2.....A_{k-1})$$

$$= \left(1-\frac{1}{N}\right)\left(1-\frac{1}{N-1}\right)\left(1-\frac{1}{N-2}\right)...\left(1-\frac{1}{N-k+2}\right)\frac{1}{N-k+1}$$

$$= \frac{N-1}{N}.\frac{N-2}{N-1}...\frac{N-k+1}{N-+2}.\frac{1}{N-k+1}$$

$$= \frac{1}{N}$$

2. SRSWR

P[selection of u_j at kth draw] = $\frac{1}{N}$.

Estimation of population mean and population variance

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatterdness of the data around the central value. Among various indicators of central tendency and dispersion, the popular choices are arithmetic mean and variance. So the population mean and population variability are generally measured by the arithmetic mean (or weighted arithmetic mean) and variance, respectively. There are various popular estimators for estimating the population mean and population variance. Among them, sample arithmetic mean and sample variance are more popular than other estimators. One of the reason to use these estimators is that they possess nice statistical properties. Moreover, they are also obtained through well established statistical estimation procedures like maximum likelihood estimation, least squares estimation, method of moments etc. under several standard statistical distributions. One may also consider other indicators like median, mode, geometric mean, harmonic mean for measuring the central tendency and mean deviation, absolute deviation, Pitman nearness etc. for measuring the dispersion. The properties of such estimators can be studied by numerical procedures like bootstraping.

1. Estimation of population mean

Let us consider the sample arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as an estimator of population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and verify $\bar{y}$ is an unbiased estimator of $\bar{Y}$ under the two cases.

SRSWOR

Let $t_i = \sum_{i=1}^{n} y_i$. Then

$$
\begin{aligned}
E(\bar{y}) &= \frac{1}{n} E\left(\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n} E\left(t_i\right) \\
&= \frac{1}{n}\left(\frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} t_i \right) \\
&= \frac{1}{n} \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \left(\sum_{i=1}^{n} y_i \right).
\end{aligned}
$$

When n units are sampled from N units by without replacement , then each unit of the population can occur with other units selected out of the remaining $(N-1)$ units is the population and each unit occurs in $\binom{N-1}{n-1}$ of the $\binom{N}{n}$ possible samples. So

So $\sum_{i=1}^{\binom{N}{n}} \left(\sum_{i=1}^{n} y_i \right) = \binom{N-1}{n-1} \sum_{i=1}^{N} y_i$.

Now

$$
\begin{aligned}
E(\bar{y}) &= \frac{(N-1)!}{(n-1)!(N-n)!} \frac{n!(N-n)!}{n\,N!} \sum_{i=1}^{N} y_i \\
&= \frac{1}{N} \sum_{i=1}^{N} y_i \\
&= \bar{Y}.
\end{aligned}
$$

Thus $\bar{y}$ is an unbiased estimator of $\bar{Y}$. Alternatively, the following approach can also be adopted to show the unbiasedness property.

$$
\begin{aligned}
E(\bar{y}) &= \frac{1}{n}\sum_{j=1}^{n} E(y_j) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=1}^{N} Y_i P_j(i)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\left[\sum_{i=1}^{N} Y_i \cdot \frac{1}{N}\right] \\
&= \frac{1}{n}\sum_{j=1}^{n} \bar{Y} \\
&= \bar{Y}
\end{aligned}
$$

where $P_j(i)$ denotes the probability of selection of i^{th} unit at j^{th} stage.

SRSWR

$$
\begin{aligned}
E(\bar{y}) &= \frac{1}{n}E\left(\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} (Y_1 P_1 + .. + Y_N P) \\
&= \frac{1}{n}\sum^{n} \bar{Y} \\
&= \bar{Y}.
\end{aligned}
$$

where $P_i = \frac{1}{N}$ for all $i = 1, 2, ..., N$ is the probability of selection of a unit. Thus $\bar{y}$ is an unbiased estimator of population mean under SRSWR also.

Variance of the estimate

Assume that each observation has some variance σ^2. Then

$$
\begin{aligned}
V(\bar{y}) &= E(\bar{y}-\bar{Y})^2 \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{Y})\right]^2 \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{n}(y_i-\bar{Y})^2+\frac{1}{n^2}\sum_{i}^{n}\sum_{\neq j}^{n}(y_i-\bar{Y})(y_j-\bar{Y})\right] \\
&= \frac{1}{n^2}\sum^{n}E(y_i-\bar{Y})^2+\frac{1}{n^2}\sum_{i}^{n}\sum_{\neq j}^{n}E(y_i-\bar{Y})(y_j-\bar{Y}) \\
&= \frac{1}{n^2}\sum^{n}\sigma^2+\frac{K}{n^2} \\
&= \frac{N-1}{Nn}S^2+\frac{K}{n^2}
\end{aligned}
$$

where $K=\sum_{i}^{n}\sum_{\neq j}^{n}E(y_i-\bar{Y})(y_i-\bar{Y})$ assuming that each observation has variance σ^2. Now we find K under the setups of SRSWR and SRSWOR.

SRSWOR

$K=\sum_{i}^{n}\sum_{\neq j}^{n}E(y_i-\bar{Y})(y_i-\bar{Y})$.

Consider

$$E(y_i-\bar{Y})(y_j-\bar{Y})=\frac{1}{N(N-1)}\sum_{k}^{N}\sum_{\neq \ell}^{N}(y_k-\bar{Y})(y_e-\bar{Y})$$

Since

$$
\begin{aligned}
\left[\sum_{k=1}^{N}(y_k-\bar{Y})\right]^2 &= \sum_{i=1}^{N}(y_k-\bar{Y})^2+\sum_{k}^{N}\sum_{\neq \ell}^{N}(y_k-\bar{Y})(y_\ell-\bar{Y})) \\
0 &= (N-1)S^2+\sum_{k}^{N}\sum_{\neq \ell}^{N}(y_k-\bar{Y})(y_\ell-\bar{Y}) \\
\sum_{k}^{N}\sum_{\neq \ell}^{N}(y_k-\bar{Y})(y_\ell-\bar{Y}) &= \frac{1}{N(N-1)}[-(N-1)S^2] \\
&= -\frac{S^2}{N}.
\end{aligned}
$$

Thus $K = -n(n-1)\frac{S^2}{N}$ and so substituting the value of K, the variance of $\bar{y}$ under SRSWOR is

$$V(\bar{y}_{WOR}) = \frac{N-1}{Nn}S^2 - \frac{1}{n^2}n(n-1)\frac{S^2}{N}$$
$$= \frac{N-n}{Nn}S^2.$$

SRSWR

$$K = \sum_{i}^{N}\sum_{\neq j}^{N} E(y_i - \bar{Y})(y_i - \bar{Y})$$
$$= \sum_{i}^{N}\sum_{\neq j}^{N} E(y_i - \bar{Y})E(y_{je} - \bar{Y})$$
$$= 0$$

because the ith and jth draws $(i \neq j)$ are independent.

Thus the variance of $\bar{y}$ under SRSWR is

$$V(\bar{y}_{WR}) = \frac{N-1}{Nn}S^2.$$

It is to be noted that if N is infinite (large enough), then

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

is both the cases of SRSWOR and SRSWR. So the factor $\frac{N-n}{N}$ is responsible for changing the variance of $\bar{y}$ when the sample is drawn from a finite population in comparison to an infinite population. This is why $\frac{N-n}{N}$ is called a finite population correction (fpc) . It may be noted that $\frac{N-n}{N} = 1 - \frac{n}{N}$, so $\frac{N-n}{N}$ is close to 1 if the ratio of sample size to population $\frac{n}{N}$, is very small or negligible. The term $\frac{n}{N}$ is called sampling fraction. In practice, fpc can be ignored whenever $\frac{n}{N} < 5\%$ and for many purposes even if it is as high as 10%. Ignoring fpc will result in the overestimation of variance of $\bar{y}$.

Efficiency of $\bar{y}$ under SRSWOR over SRSWR

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

$$V(\bar{y}_{WR}) = \frac{N-1}{Nn}S^2$$

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

$$= V(\bar{y}_{WOR}) + a \ \textit{positive quantity}$$

Thus

$$V(\bar{y}_{WR}) > V(\bar{y}_{WOR})$$

and so, SRSWOR is more efficient than SRSWR.

Estimation of variance from a sample

Since the expressions of variances of sample mean involve S^2 which is based on population values, so these expressions can not be used in real life applications. In order to estimate the variance of $\bar{y}$ on the basis of a sample, an estimator of S^2 (or equivalently σ^2) is needed. Consider S^2 as an estimator of s^2 (or σ^2) and we investigate its biasedness for S^2 in the cases of SRSWOR and SRSWR,

Consider

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

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

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

$$E(s^2) = \frac{1}{n-1}\left[\sum_{i=1}^{n}E(y_i - \bar{Y})^2 - nE(\bar{y} - \bar{Y})^2\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}Var(y_i) - nVar(\bar{y})\right] = \frac{1}{n-1}\left[n\sigma^2 - nVar(\bar{y})\right]$$

In case of SRSWOR

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

and so

$$\begin{aligned} E(s^2) &= \frac{n}{n-1}\left[\sigma^2 - \frac{N-n}{Nn} S^2\right] \\ &= \frac{n}{n-1}\left[\frac{N-1}{N} S^2 - \frac{N-n}{Nn} S^2\right] \\ &= S^2 \end{aligned}$$

In case of SRSWR

$$V(\bar{y}_{WR}) = \frac{N-1}{Nn} S^2$$

and so

$$\begin{aligned} E(s^2) &= \frac{n}{n-1}\left[\sigma^2 - \frac{N-n}{Nn} S^2\right] \\ &= \frac{n}{n-1}\left[\frac{N-1}{N} S^2 - \frac{N-n}{Nn} S^2\right] \\ &= \frac{N-1}{N} S^2 \\ &= \sigma^2 \end{aligned}$$

Hence

$$E(s^2) = \begin{cases} S^2 & is\ SRSWOR \\ \sigma^2 & is\ SRSWR \end{cases}$$

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

$\hat{V}(\bar{y}_{WOR}) = \frac{N-n}{Nn} s^2$ in case of SRSWOR and

$$\begin{aligned} \hat{V}(\bar{y}_{WR}) &= \frac{N-1}{Nn} \cdot \frac{N}{N-1} s^2 \\ &= \frac{s^2}{n} \text{ in case of SRSWR.} \end{aligned}$$

Standard errors

The standard error of $\bar{y}$ is defined as $\sqrt{Var(\bar{y})}$.

In order to estimate the standard error, one simple option is to consider the square root of estimate of variance of sample mean.

- under SRSWOR, a possible estimator is $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-n}{Nn}}s$.
- under SRSWR, a possible estimator is $\hat{\sigma}(\bar{y}) = \sqrt{\frac{N-1}{Nn}}s$.

It is to be noted that this estimator does not possess the same properties as of $\widehat{Var}(\bar{y})$.

Reason being if $\hat{\theta}$ is an estimator of θ, then $\widehat{\sqrt{\theta}}$ is not necessarily an estimator of $\sqrt{\theta}$.

In fact, the $\hat{\sigma}(\bar{y})$ is a negatively biased estimator under SRSWOR.

The approximate expressions for large *N* case are as follows:

(Reference: Sampling Theory of Surveys with Applications, P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme, C. Asok, Iowa State University Press and Indian Society of Agricultural Statistics, 1984, India)

Consider *s* as an estimator of *S*.

Let

$$s^2 = S^2 + \varepsilon \text{ with } E(\varepsilon) = 0, E(\varepsilon^2) = S^2.$$

Write

$$s = (S^2 + \varepsilon)^{1/2}$$

$$= S\left(1 + \frac{\varepsilon}{S^2}\right)^{1/2}$$

$$= S\left(1 + \frac{\varepsilon}{2S^2} - \frac{\varepsilon^2}{8S^4} + \ldots\right)$$

assuming ε will be small as compared to S^2 and as n becomes large, the probability of such an event approaches one. Neglecting the powers of ε higher than two and taking expectation, we have

$$E(s) = \left[1 - \frac{Var(s^2)}{8S^4}\right] S$$

where

$$Var\left(s^2\right) = \frac{2S^4}{(n-1)}\left[1 + \left(\frac{n-1}{2n}\right)\left(\beta_2 - 3\right)\right] \text{ for large } N.$$

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^j$$

$\beta_2 = \frac{\mu_4}{S^4}$: coefficient of kurtosis.

Thus

$$E\left(s\right) = S\left[1 - \frac{1}{4(n-1)} - \frac{\beta_2 - 3}{8n}\right]$$

$$Var(s) = S^2 - S^2\left[1 - \frac{1}{8}\frac{Var(s^2)}{S^4}\right]^2$$

$$= \frac{Var(s^2)}{4S^2}$$

$$= \frac{S^2}{2\left(n-1\right)}\left[1 + \left(\frac{n-1}{2n}\right)\left(\beta_2 - 3\right)\right].$$

Note that for a normal distribution, $\beta_2 = 3$ and we obtain

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

Both $Var(s)$ and $Var(s^2)$ are inflated due to nonnormality to the same extent, by the inflation factor

$$\left[1 + \left(\frac{n-1}{2n}\right)\left(\beta_2 - 3\right)\right]$$

and this does not depends on coefficient of skewness.

This is an important result to be kept in mind while determining the sample size in which it is assumed that S^2 is known. If inflation factor is ignored and population is non-normal, then the reliability on s^2 may be misleading.

Alternative approach:

The results for the unbiasedness property and the variance of sample mean can also be proved in an alternative way as follows:

(i) SRSWOR

With the i^{th} unit of the population, we associate a random variable a_i defined as follows:

$$a_i = \begin{cases} 1, \text{if the } i^{th} \text{ unit occurs in the sample} \\ 0, \text{ if the } i^{th} \text{ unit does not occurs in the sample } (i = 1, 2, ..., N) \end{cases}$$

Then,

$E(a_i) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample}$

$$= \frac{n}{N},\ i = 1, 2, ..., N.$$

$E(a_i^2) = 1 \times \text{Probability that the } i^{th} \text{ unit is included in the sample}$

$$= \frac{n}{N},\ i = 1, 2, ..., N$$

$E(a_i a_j) = 1 \times \text{Probability that the } i^{th} \text{ and } j^{th} \text{ units are included in the sample}$

$$= \frac{n(n-1)}{N(N-1)},\ i \neq j = 1, 2, ..., N.$$

From these results, we can obtain

$$Var(a_i) = E(a_i^2) - \left(E(a_i)\right)^2 = \frac{n(N-n)}{N^2},\ i = 1, 2, ..., N$$

$$Cov(a_i, a_j) = E(a_i a_j) - E(a_i)E(a_j) = \frac{n(N-n)}{N^2(N-1)},\ i \neq j = 1, 2, ..., N.$$

We can rewrite the sample mean as

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

Then

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

and

$$Var(\bar{y}) = \frac{1}{n^2} Var\left(\sum_{i=1}^{N} a_i y_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{N} Var(a_i) y_i^2 + \sum_{i \neq j}^{N} Cov(a_i, a_j) y_i y_j\right].$$

Substituting the values of $Var(a_i)$ and $Cov(a_i, a_j)$ in the expression of $Var(\bar{y})$ and simplifying, we get

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

To show that $E(s^2) = S^2$, consider

$$s^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] = \frac{1}{(n-1)}\left[\sum_{i=1}^{N} a_i y_i^2 - n\bar{y}^2\right].$$

Hence, taking, expectation, we get

$$E(s^2) = \frac{1}{(n-1)}\left[\sum_{i=1}^{N} E(a_i) y_i^2 - n\left\{Var(\bar{y}) + \bar{Y}^2\right\}\right]$$

Substituting the values of $E(a_i)$ and $Var(\bar{y})$ in this expression and simplifying, we get $E(s^2) = S^2$.

(ii) SRSWR

Let a random variable a_i associated with the i^{th} unit of the population denotes the number of times the i^{th} unit occurs in the sample $i = 1, 2, ..., N$. So a_i assumes values 0, 1, 2,…,*n*. The joint distribution of $a_1, a_2, ..., a_N$ is the multinomial distribution given by

$$P(a_1, a_2, ..., a_N) = \frac{n!}{\prod_{i=1}^{N} a_i!} . \frac{1}{N^n}$$

where $\sum_{i=1}^{N} a_i = n.$ For this multinomial distribution, we have

$E(a_i) = \frac{n}{N}$,

$Var(a_i) = \frac{n(N-1)}{N^2}$, $i = 1, 2, ..., N.$

$Cov(a_i, a_j) = -\frac{n}{N^2}$, $i \neq j = 1, 2, ..., N.$

We rewrite the sample mean as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i y_i.$$

Hence, taking expectation of $\bar{y}$ and substituting the value of $E(a_i) = n/N$ we obtain that

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

Further,

$$Var(\bar{y}) = \frac{1}{n^2}\left[\sum_{i=1}^{N} Var(a_i)y_i^2 + \sum_{i=1}^{N} Cov(a_i, a_j)y_i y_j\right]$$

Substituting, the values of $Var(a_i) = n(N-1)/N^2$ and $Cov(a_i, a_j) = -n/N^2$ and simplifying, we get

$$Var(\bar{y}) = \frac{N-1}{Nn}S^2.$$

To prove that $E(s^2) = \frac{N-1}{N}S^2 = \sigma^2$ in SRSWR, consider

$$(n-1)s^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{N} a_i y_i^2 - n\bar{y}^2,$$

$$(n-1)E(s^2) = \sum_{i=1}^{N} E(a_i)y_i^2 - n\left\{Var(\bar{y}) + \bar{Y}^2\right\}$$

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

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

$$E(s^2) = \frac{N-1}{N}S^2 = \sigma^2$$

Estimator of population total:

Sometimes, it is also of interest to estimate the population total, e.g. total household income, total expenditures etc. Let denotes the population total

$$Y_T = \sum_{i=1}^{N} Y_i = N\bar{Y}$$

which can be estimated by

$$\hat{Y}_T = N\hat{\bar{Y}}$$

$$= N\bar{y}.$$

Obviously

$$E\left(\hat{Y}_T\right) = NE\left(\bar{y}\right)$$
$$= N\bar{Y}$$
$$Var\left(\hat{Y}_T\right) = N^2\left(\bar{y}\right)$$
$$= \begin{cases} N^2\left(\frac{N-n}{Nn}\right)S^2 = \frac{N(N-n)}{n}S^2 & for\ SRSWOR \\ N^2\left(\frac{N-1}{Nn}\right)S^2 = \frac{N(N-1)}{n}S^2 & for\ SRSWOR \end{cases}$$

and the estimates of variance of $\hat{Y}_T$ are

$$\widehat{Var}(\hat{Y}_T) = \begin{cases} \frac{N(N-n)}{n}s^2 & for\ SRSWOR \\ \frac{N}{n}s^2 & for\ SRSWOR \end{cases}$$

Confidence limits for the population mean

Now we construct the $100(1-\alpha)\%$ confidence interval for the population mean. Assume that the population is normally distributed $N(\mu,\sigma^2)$ with mean μ and variance σ^2. then $\frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}}$ follows $N(0,1)$ when σ^2 is known. If σ^2 is unknown and is estimated from the sample then $\frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}}$ follows a t-distribution with $(n-1)$ degrees of freedom. When σ^2 is known, then the $100(1-\alpha)\%$ confidence interval is given by

$$P\left[-Z_{\frac{\alpha}{2}} \le \frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}} \le Z_{\frac{\alpha}{2}}\right] = 1-\alpha$$

$$or\quad P\left[\bar{y} - Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})} \le \bar{y} \le \bar{y} + Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})}\right] = 1-\alpha$$

and the confidence limits are

$$\left(\bar{y} - Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})},\ \bar{y} + Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y}}\right)$$

when $Z_{\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}\%$ points on $N(0,1)$ distribution. Similarly, when σ^2 is unknown, then the 100(1-1$-\alpha$)% confidence interval is

$$P\left[-t_{\frac{\alpha}{2}} \le \frac{\bar{y}-\bar{Y}}{\sqrt{\hat{Var}(\bar{y})}} \le t_{\frac{\alpha}{2}}\right] = 1-\alpha$$

or $$P\left[\bar{y} - t_{\frac{\alpha}{2}} \le \sqrt{\hat{Var}(\bar{y})} \le \bar{y} \le \bar{y} + t_{\frac{\alpha}{2}}\sqrt{\hat{Var}(\bar{y})}\right] = 1-\alpha$$

and the confidence limits are

$$\left[\bar{y} - t_{\frac{\alpha}{2}} \le \sqrt{\hat{Var}(\bar{y})} \le \bar{y} + t_{\frac{\alpha}{2}}\sqrt{\hat{Var}(\bar{y})}\right]$$

where $t_{\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}\%$ points on t-distribution with $(n-1)$ degrees of freedom.

Determination of sample size

The size of the sample is needed before the survey starts and goes into operation. One point to be kept is mind is that when the sample size increases, the variance of estimators decreases but the cost of survey increases and vice versa. So there has to be a balance between the two aspects. The sample size can be determined on the basis of prescribed values of standard error of sample mean, error of estimation, width of the confidence interval, coefficient of variation of sample mean, relative error of sample mean or total cost among several others.

An important constraint or need to determine the sample size is that the information regarding the population standard derivation S should be known for these criterion. The reason and need for this will be clear when we derive the sample size in the next section. A question arises about how to have information about S before hand? The possible solutions to this issue are to conduct a pilot survey and collect a preliminary sample of small size, estimate S and use it as known value of S it. Alternatively, such information can also be collected from past data, past experience, long association of experimenter with the experiment, prior information etc.

Now we find the sample size under different criteria assuming that the samples have been drawn using SRSWOR. The case for SRSWR can be derived similarly.

1. Prespecified variance

The sample size is to be determined such that the variance of $\bar{y}$ should not exceed a given value, say V. In this case, find n such that

$Var(\bar{y}) \leq V$

or $\frac{N-n}{Nn}(\bar{y}) \leq V$

or $\frac{N-n}{Nn}S^2 \leq V$

or $\frac{1}{n}-\frac{1}{N} \leq \frac{V}{S^2}$

or $\frac{1}{n}-\frac{1}{N} \leq \frac{1}{n_e}$

$n \geq \frac{n_e}{1+\frac{n_e}{N}}$

where $n_e = \frac{S^2}{v}$.

It may be noted here that n_e can be known only when S^2 is known. This reason compels to assume that S should be known. The same reason will also be seen in other cases.

The smallest sample size needed in this case is

$n_{smallest} = \frac{n_e}{1+\frac{n_e}{N}}$.

It N is large, then the required *n* is

$n \geq n_e$ and $n_{smallest} = n_e$.

2. Pre-specified estimation error

It may be possible to have some prior knowledge of population mean $\bar{Y}$ and it may be required that the sample mean $\bar{y}$ should not differ from it by more than a specified amount of absolute estimation error, i.e., which is a small quantity. Such requirement can be satisfied by associating a probability $(1-\alpha)$ with it and can be expressed as

$P\left[\left|\bar{y}-\bar{Y}\right| \leq e\right] = (1-\alpha).$

Since $\bar{y}$ follows $N(\bar{Y}, \frac{N-n}{Nn}S^2)$ assuming the normal distribution for the population, we can write

$$P\left[\frac{|\bar{y}-\bar{Y}|}{\sqrt{Var(\bar{y})}} \le \frac{e}{\sqrt{Var(\bar{y})}}\right] = 1-\alpha$$

which implies that

$$\frac{e}{\sqrt{Var(\bar{y})}} = Z_{\frac{\alpha}{2}}$$

or $Z^2_{\frac{\alpha}{2}} Var(\bar{y}) = e^2$

or $Z^2_{\frac{\alpha}{2}} \frac{N-n}{Nn} S^2 = e^2$

or $$n = \left(\frac{\left(\frac{Z_{\frac{\alpha}{2}} S}{e}\right)^2}{1+\frac{1}{N}\left(\frac{Z_{\frac{\alpha}{2}} S}{e}\right)^2} \right)$$

which is the required sample size. If N is large then

$$n = \left(\frac{Z_{\frac{\alpha}{2}} S}{e}\right)^2 .$$

3. Prespecified width of confidence interval

If the requirement is that the width of the confidence interval of $\bar{y}$ with confidence coefficient $(1-\alpha)$ should not exceed a prespecified amount W, then the sample size n is determined such that

$$2Z_{\frac{\alpha}{2}}\sqrt{Var(\bar{y})} \le W$$

assuming σ^2 is known and population is normally distributed. This can be expressed as

$$2Z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S \le W$$

or $4Z^2_{\frac{\alpha}{2}}\left(\frac{1}{n}-\frac{1}{N}\right)S^2 \le W^2$

or $\frac{1}{n} \leq \frac{1}{N} + \frac{W^2}{4Z^2_{\frac{\alpha}{2}} S^2}$

or $n \geq \frac{\frac{4Z^2_{\frac{\alpha}{2}} S^2}{W^2}}{1+\frac{4Z^2_{\frac{\alpha}{2}} S^2}{NW^2}}.$

The minimum sample size required is

$$n_{smallest} = \frac{\frac{4Z^2_{\frac{\alpha}{2}} S^2}{W^2}}{1+\frac{4Z^2_{\frac{\alpha}{2}} S^2}{NW^2}}$$

If N is large then

$$n \geq \frac{4Z^2_{\frac{\alpha}{2}} S^2}{W^2}$$

and the minimum sample size needed is

$$n_{smallest} = \frac{4Z^2_{\frac{\alpha}{2}} S^2}{W^2}.$$

4. Prespecified coefficient of variation

The coefficient of variation (CV) is defined as the ratio of standard error (or standard deviation) and mean. The know ledge of coefficient of variation has played an important role in the sampling theory as this information has helped in deriving efficient estimators.

If it is desired that the the coefficient of variation of $\bar{y}$ should not exceed a given or prespecified value of coefficient of variation, say C_0 , then the required sample size n is to be determined such that

$CV(\bar{y}) \leq C_0$

or $\frac{\sqrt{Var(\bar{y})}}{\bar{Y}} \leq C_0$

$$\text{or} \quad \frac{\frac{N-n}{Nn}S^2}{\bar{Y}^2} \le C_0^2$$

$$\text{or} \quad \frac{1}{n} - \frac{1}{N} \le \frac{C_0^2}{C^2}$$

$$\text{or} \quad n \ge \frac{\frac{C^2}{C_o^2}}{1+\frac{C^2}{NC_0^2}}$$

is the required sample size where $C = \frac{S}{\bar{Y}}$ is the population coefficient of variation.

The smallest sample size needed in this case is

$$n_{smallest} = \frac{\frac{C^2}{C_0^2}}{1+\frac{C^2}{NC_0^2}}.$$

If N is large, then

$$n \ge \frac{C^2}{C_0^2}$$

$$and \quad n_{smalest} = \frac{C^2}{C_0^2}$$

5. Prespecified relative error

When $\bar{y}$ is used for estimating the population mean $\bar{Y}$, then the relative estimation error is defined as $\frac{\bar{y}-\bar{Y}}{\bar{Y}}$. If it is required that such relative estimation error should not exceed a prespecified value R with probability $(1-\alpha)$, then such requirement can be satisfied by expressing it like such requirement can be satisfied by expressing it like

$$P\left[\frac{\left|\bar{y}-\bar{Y}\right|}{\sqrt{Var(\bar{y})}} \le \frac{R\bar{Y}}{\sqrt{Var(\bar{y})}}\right] = 1-\alpha.$$

Assuming the population to be normally distributed, $\bar{y}$ follows $N\left(\bar{Y}, \frac{N-n}{Nn}S^2\right)$.

So it can be written that

$$\frac{R\bar{Y}}{\sqrt{Var(\bar{y})}} = Z_{\frac{\alpha}{2}}.$$

or $Z^2_{\frac{\alpha}{2}}\left(\frac{N-n}{Nn}\right)S^2 = R^2\bar{Y}^2$

or $\left(\frac{1}{n}-\frac{1}{N}\right) = \frac{R^2}{C^2 Z^2_{\frac{\alpha}{2}}}$

or $n = \frac{\left(\frac{Z_{\frac{\alpha}{2}}C}{R}\right)^2}{1+\frac{1}{N}\left(\frac{Z_{\frac{\alpha}{2}}C}{R}\right)^2}$

where $C = \frac{S}{\bar{Y}}$ is the population coefficient of variation and should be known.

If N is large, then

$$n = \left(\frac{z_{\frac{\alpha}{2}}C}{R}\right)^2.$$

6. Prespecified cost

Let an amount of money C is being designated for sample survey to called n observations, C_0 be the overhead cost and C_1 be the cost of collection of one unit in the sample. Then the total cost C can be expressed as

$$C = C_0 + nC_1$$

Or $n = \frac{C - C_0}{C_1}$

is the required sample size.

Chapter 3

Sampling For Proportions and Percentages

In many situations, the characteristic under study on which the observations are collected are qualitative in nature. For example, the responses of customers in many marketing surveys are based on replies like 'yes' or 'no' , 'agree' or 'disagree' etc. Sometimes the respondents are asked to arrange several options in the order like first choice, second choice etc. Sometimes the objective of the survey is to estimate the proportion or the percentage of brown eyed persons, unemployed persons, graduate persons or persons favoring a proposal, etc. In such situations, the first question arises how to do the sampling and secondly how to estimate the population parameters like population mean, population variance, etc.

Sampling procedure:

The same sampling procedures that are used for drawing a sample in case of quantitative characteristics can also be used for drawing a sample for qualitative characteristic. So, the sampling procedures remain same irrespective of the nature of characteristic under study - either qualitative or quantitative. For example, the SRSWOR and SRSWR procedures for drawing the samples remain the same for qualitative and quantitative characteristics. Similarly, other sampling schemes like stratified sampling, two stage sampling etc. also remain same.

Estimation of population proportion:

The population proportion in case of qualitative characteristic can be estimated in a similar way as the estimation of population mean in case of quantitative characteristic.

Consider a qualitative characteristic based on which the population can be divided into two mutually exclusive classes, say C and C^*. For example, if C is the part of population of persons saying 'yes' or 'agreeing' with the proposal then C^* is the part of population of persons saying 'no' or 'disagreeing' with the proposal. Let A be the number of units in C and $(N - A)$ units in C^* be in a population of size N. Then the proportion of units in C is

$$P = \frac{A}{N}$$

and the proportion of units in C^* is

$$Q = \frac{N - A}{N} = 1 - P.$$

An indicator variable Y can be associated with the characteristic under study and then for i = 1,2,..,N

$$Y_i = \begin{cases} 1 & i^{th} \text{ unit belongs to } C \\ 0 & i^{th} \text{ unit belongs to } C^*. \end{cases}$$

Now the population total is

$$Y_{TOTAL} = \sum_{i=1}^{N} Y_i = A$$

and population mean is

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{A}{N} = P.$$

Suppose a sample of size n is drawn from a population of size N by simple random sampling .

Let a be the number of units in the sample which fall into class C and $(n-a)$ units fall in class C^*, then the sample proportion of units in C is

$$p = \frac{a}{n}.$$

which can be written as

$$p = \frac{a}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}.$$

Since $\sum_{i=1}^{N} Y_i^2 = A = NP,$ so we can write S^2 and s^2 in terms of P and Q as follows:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \\ &= \frac{1}{N-1}(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2) \\ &= \frac{1}{N-1}(NP - NP^2) \\ &= \frac{N}{N-1}PQ. \end{aligned}$$

Similarly, $\sum_{i=1}^{n} y_i^2 = a = np$ and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$= \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)$$
$$= \frac{1}{n-1}(np - np^2)$$
$$= \frac{n}{n-1}pq.$$

Note that the quantities $\bar{y}, \bar{Y}, s^2$ and S^2 have been expressed as functions of sample and population proportions. Since the sample has been drawn by simple random sampling and sample proportion is same as the sample mean, so the properties of sample proportion in SRSWOR and SRSWR can be derived using the properties of sample mean directly.

1. SRSWOR

Since sample mean $\bar{y}$ an unbiased estimator of population mean $\bar{Y}$, *i.e.* $E(\bar{y}) = \bar{Y}$ in case of SRSWOR, so

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

and p is an unbiased estimator of P.

Using the expression of $Var(\bar{y})$, the variance of p can be derived as

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

Similarly, using the estimate of $Var(\bar{y})$, the estimate of variance can be derived as

$$\widehat{Var}(p) = \widehat{Var}(\bar{y}) = \frac{N-n}{Nn}s^2$$
$$= \frac{N-n}{Nn}\frac{n}{n-1}pq$$
$$= \frac{N-n}{N(n-1)}pq.$$

(ii) SRSWR

Since the sample mean $\bar{y}$ is an unbiased estimator of population mean $\bar{Y}$ in case of SRSWR, so the sample proportion,

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

i.e., p is an unbiased estimator of P.

Using the expression of variance of $\bar{y}$ and its estimate in case of SRSWR, the variance of p and its estimate can be derived as follows:

$$\begin{aligned} Var(p) = Var(\bar{y}) &= \frac{N-1}{Nn} S^2 \\ &= \frac{N-1}{Nn} \frac{N}{N-1} PQ \\ &= \frac{PQ}{n} \end{aligned}$$

$$\begin{aligned} \widehat{Var}(p) &= \frac{n}{n-1} . \frac{pq}{n} \\ &= \frac{pq}{n-1}. \end{aligned}$$

Estimation of population total or total number of count

It is easy to see that an estimate of population total A (or total number of count) is

$$\hat{A} = Np = \frac{Na}{n},$$

its variance is

$$Var(\hat{A}) = N^2 Var(p)$$

and the estimate of variance is

$$\widehat{Var}(\hat{A}) = N^2 \widehat{Var}(p).$$

Confidence interval estimation of *P*

If N and n are large then $\frac{p-P}{\sqrt{Var(p)}}$ approximately follows $N(0,1)$. With this approximation, we can write

$$P\left[-Z_{\frac{\alpha}{2}} \le \frac{p-P}{\sqrt{Var(p)}} \le Z_{\frac{\alpha}{2}} \right] = 1-\alpha$$

and the $100(1-\alpha)\%$ confidence interval of P is

$$\left(p - Z_{\frac{\alpha}{2}} \sqrt{Var(p)}, p + Z_{\frac{\alpha}{2}} \sqrt{Var(p)} \right).$$

It may be noted that in this case, a discrete random variable is being approximated by a continuous random variable, so a continuity correction $n/2$ can be introduced in the confidence limits and the limits become

$$\left(p - Z_{\frac{\alpha}{2}} \sqrt{Var(p)} + \frac{n}{2}, p + Z_{\frac{\alpha}{2}} \sqrt{Var(p)} - \frac{n}{2} \right)$$

Use of Hypergeometric distribution :

When SRS is applied for the sampling of a qualitative characteristic, the methodology is to draw the units one-by-one and so the probability of selection of every unit remains the same at every step. If n sampling units are selected together from N units, then the probability of selection of units does not remains the same as in the case of SRS.

Consider a situation in which the sampling units in a population are divided into two mutually exclusive classes. Let P and Q be the proportions of sampling units in the population belonging to classes '1' and '2' respectively. Then NP and NQ are the total number of sampling units in the population belonging to class '1' and '2', respectively and so $NP + NQ = N$. The probability that in a sample of n selected units out of N units by SRS such that n_1 selected units belongs to class '1' and n_2 selected units belongs to class '2' is governed by the hypergeometric distribution and

$$P(n_1) = \frac{\binom{NP}{n_1}\binom{NQ}{n_2}}{\binom{N}{n}}.$$

As N grows large, the hypergeometric distribution tends to Binomial distribution and $P(n_1)$ is approximated by

$$P(n_1) = \binom{n}{n_1} p^{n_1} (1-p)^{n_2}$$

Inverse sampling

In general, it is understood in the SRS methodology for qualitative characteristic that the attribute under study is not a rare attribute. If the attribute is rare, then the procedure of estimating the population proportion P by sample proportion n/N is not suitable. Some such situations are, e.g., estimation of frequency of rare type of genes, proportion of some rare type

of cancer cells in a biopsy, proportion of rare type of blood cells affecting the red blood cells etc. In such cases, the methodology of inverse sampling can be used.

In the methodology of inverse sampling, the sampling is continued until a predetermined number of units possessing the attribute under study occur in the sampling which is useful for estimating the population proportion. The sampling units are drawn one-by-one with equal probability and without replacement. The sampling is discontinued as soon as the number of units in the sample possessing the characteristic or attribute equals a predetermined number.

Let m denotes the predetermined number indicating the number of units possessing the characteristic. The sampling is continued **till *m* number** of units are obtained. Therefore, the sample size n required to attain m becomes a random variable.

Probability distribution function of *n*

In order to find the probability distribution function of n, consider the stage of drawing of samples t such that at $t = n$, the sample size n completes the m units with attribute. Thus the first $(t - 1)$ draws would contain $(m - 1)$ units in the sample possessing the characteristic out of NP units. Equivalently, there are $(t - m)$ units which do not possess the characteristic out of NQ such units in the population. Note that the last draw must ensure that the units selected possess the characteristic.

So the probability distribution function of n can be expressed as

$$P(n) = P\begin{pmatrix}\text{In a sample of } (n-1) \text{ units}\\ \text{drawn from } N,\ (m-1) \text{ units}\\ \text{will possess the attribute}\end{pmatrix} \times P\begin{pmatrix}\text{The unit drawn at}\\ \text{the } n^{th} \text{ draw will}\\ \text{possess the attribute}\end{pmatrix}$$

$$= \left[\frac{\binom{NP}{m-1}\binom{NQ}{n-m}}{\binom{N}{n-1}}\right]\left(\frac{NP-m+1}{N-n+1}\right), \quad n = m, m+1, ..., m+NQ.$$

Note that the first term (in square brackets) is derived using hypergeometric distribution as the probability for deriving a sample of size $(n - 1)$ in which $(m - 1)$ units are from NP units and $(n - m)$ units are from NQ units. The second term $\frac{NP-m+1}{N-n+1}$ is the probability associated with the last draw where it is assumed that we get the unit possessing the characteristic.

Note that $\sum_{n=m}^{m+NQ} P(n) = 1.$

Estimate of population proportion

Consider the expectation of $\frac{m-1}{n-1}$.

$$E\left(\frac{m-1}{n-1}\right)=\sum_{n=m}^{m+NQ}\left(\frac{m-1}{n-1}\right)P(n)$$

$$=\sum_{n=m}^{m+NQ}\left(\frac{m-1}{n-1}\right)\frac{\binom{NP}{m-1}\binom{NQ}{n-m}}{\binom{N}{n-1}}\cdot\frac{Np-m+1}{N-n+1}$$

$$=\sum_{n=m}^{m+NQ-1}\left(\frac{NP-m+1}{N-n+1}\right)\frac{\binom{NP-1}{m-2}\binom{NQ}{n-m}}{\binom{N-1}{n-2}}$$

which is obtained by replacing NP by $NP-1$, m by $(m-1)$ and n by $(n-1)$ in the earlier step. Thus

$$E\left(\frac{m-1}{n-1}\right)=P.$$

So $\hat{P}=\frac{m-1}{n-1}$ is an unbiased estimator of P.

Estimate of variance of $\hat{P}$

Now we derive an estimate of variance of $\hat{P}$. By definition

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

$$=E(\hat{P}^2)-P^2.$$

Thus

$$\widehat{Var}(\hat{P})=\hat{P}^2-\text{Estimate of } P^2.$$

In order to obtain an estimate of P^2, consider the expectation of $\frac{(m-1)(m-2)}{(n-1)(n-2)}$, *i.e.*,

$$E\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right]=\sum_{n\geq m}\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right]P(n)$$

$$=\frac{P(NP-1)}{N-1}\sum_{n\geq m}\left(\frac{NP-m+1}{N-n+1}\right)\left[\frac{\binom{NP-2}{m-3}\binom{NQ}{n-m}}{\binom{N-2}{n-3}}\right]$$

where the last term inside the square bracket is obtained by replacing NP by $(NP-2)$, N by $(n-2)$ and m by $(m-2)$ in the probability distribution function of hypergeometric distribution. This solves further to

$$E\left[\frac{(m-1)(m-2)}{(n-1)(n-2)}\right]=\frac{NP^2}{N-1}-\frac{P}{N-1}.$$

Thus an unbiased estimate of P^2 is

$$\begin{aligned}\text{Estimate of } P^2 &= \left(\frac{N-1}{N}\right)\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{\hat{P}}{N}\\ &= \left(\frac{N-1}{N}\right)\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{1}{N}.\frac{m-1}{n-1}.\end{aligned}$$

Finally, an estimate of variance of $\hat{P}$ is

$$\begin{aligned}\widehat{Var}(\hat{P}) &= \hat{P}^2-\text{Estimate of } P^2\\ &= \left(\frac{m-1}{n-1}\right)^2-\left[\frac{N-1}{N}.\frac{(m-1)(m-2)}{(n-1)(n-2)}+\frac{1}{N}\left(\frac{m-1}{n-1}\right)\right]\\ &= \left(\frac{m-1}{n-1}\right)\left[\left(\frac{m-1}{n-1}\right)+\frac{1}{N}\left(1-\frac{(N-1)(m-2)}{n-2}\right)\right].\end{aligned}$$

For large N, the hypergeometric distribution tends to negative Binomial distribution with probability density function $\binom{n-1}{m-1}P^mQ^{n-m}$. So

$$\hat{P}=\frac{m-1}{n-1}$$

and

$$\widehat{Var}(\hat{P})=\frac{(m-1)(n-m)}{(n-1)^2(n-2)}=\frac{\hat{P}(1-\hat{P})}{n-2}.$$

Estimation of proportion for more than two classes

We have assumed up to now that there are only two classes in which the population can be divided based on a qualitative characteristic. There can be situations when the population is to be divided into more than two classes. For example, the taste of a coffee can be divided into four categories very strong, strong, mild and very mild. Similarly in another example the damage to crop due to storm can be classified into categories like heavily damaged, damaged, minor damage and no damage etc.

These type of situations can be represented by dividing the population of size N into, say k, mutually exclusive classes $C_1, C_2, ..., C_k$. Corresponding to these classes, let $P_1 = \frac{C_1}{N}, P_2 = \frac{C_2}{N}, ..., P_k = \frac{C_k}{n},$ be the proportions of units in the classes $C_1, C_2, ..., C_k$ respectively.

Let a sample of size n is observed such that $c_1, c_2, ..., c_k$ number of units have been drawn from $C_1, C_2, ..., C_k$. respectively. Then the probability of observing $c_1, c_2, ..., c_k$ is

$$P(c_1, c_2, ..., c_k) = \frac{\binom{C_1}{c_1}\binom{C_2}{c_2}...\binom{C_k}{c_k}}{\binom{N}{n}}.$$

The population proportions P_i can be estimated by $p_i = \frac{c_i}{n}, i = 1, 2, ..., k.$

It can be easily shown that

$E(p_i) = P_i, \quad i = 1, 2, ..., k,$

$Var(p_i) = \frac{N-n}{N-1}\frac{P_i Q_i}{n}$

and

$\widehat{Var}(p_i) = \frac{N-n}{N}\frac{p_i q_i}{n-1}$

For estimating the number of units in the i^{th} class,

$\hat{C}_i = Np_i$

$Var(\hat{C}_i) = N^2 Var(p_i)$

and

$\widehat{Var}(\hat{C}_i) = N^2 \widehat{Var}(p_i).$

The confidence intervals can be obtained based on single p_i as in the case of two classes.

If N is large, then the probability of observing $c_1, c_2, ..., c_k$ can be approximated by multinomial distribution given by

$$P(c_1, c_2, ..., c_k) = \frac{n!}{c_1! c_2! ... c_k!} P_1^{c_1} P_2^{c_2} ... P_k^{c_k} .$$

For this distribution

$$E(p_i) = P_i, \quad i = 1, 2, .., k,$$

$$Var(p_i) = \frac{P_i(1-P_i)}{n}$$

and

$$\widehat{Var}(\hat{p}_i) = \frac{p_i(1-p_i)}{n}.$$

Chapter 4

Stratified Sampling

An important objective in any estimation problems is to obtain an estimator of a population parameter which can take care of all salient features of the population. If the population is homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample and sample mean will serve as a good estimator of population mean. Thus if the population is homogeneous with respect to the characteristic under study, then the sample drawn through simple random sampling is expected to provide a representative sample. Moreover, the variance of sample mean not only depends on the sample size and sampling fraction but also on the population variance. In order to increase the precision of an estimator is to use a sampling scheme which reduces the heterogeneity in the population. If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is stratified sampling.

The basic idea behind the stratified sampling is to

- divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and
- heterogeneous with respect to the characteristic under study between/among the subpopulation. Such subpopulations are termed as **strata.**
- treat each subpopulation as separate population and draw a sample by SRS from each stratum.

[Note: Stratum is singular and strata is plural].

Example: In order to find the average height of students in a school of class 1 to class 12, the height varies a lot as the students in class 1 are of age around 6 years and students in class 10 one of age around 16 years. So one can divide all the students into different subpopulations or strata such as

Students of class 1, 2 and 3: Stratum 1
Students of class 4, 5 and 6: Stratum 2
Students of class 7, 8 and 9: Stratum 3
Students of class 10, 11 and 12: Stratum 4

Notations:

We use the following symbols and notations:

N : Population size

k : Number of strata

N_1 : Number of sampling units in i^{th} strata

$N = \sum_{i=1}^{k} N_i$

n_i : Numbers of sampling units to be drawn from i^{th} stratum.

$n = \sum_{i=1}^{k} n_i$: Total sample size

SAURABH, TYPE FLOW CHART PAGE 5

Procedure of stratified sampling

Divide the population of N units into k strata. Let the i^{th} stratum has $N_1, i = 1,2,...,k$ number of units.

- Stratam are constructed such that they are nonoverlapping and homogeneous with respect to the characteristic under study such that $\sum_{i=1}^{k} N_i = N.$
- Draw a sample of size n_i from i^{th} $(i = 1,2,...,k)$ stratum using SRS (preferably WOR) independently from each stratum.
- All the sampling units drawn from each stratum will constitute a stratified sample of size $n = \sum_{i=1}^{k} n_i.$

Difference between stratified and cluster sampling schemes

In stratified sampling, the strata are constructed such that they are

- within homogeneous and
- among heterogeneous

In cluster sampling, the clusters are constructed such that they are

- within heterogeneous and
- among homogeneous.

[Note: We consider cluster sampling later]

Issue in estimation in stratified sampling

Note that there are *k* independent samples drawn through SRS of sizes $n_1, n_2, ..., n_k$. So one can have *k* estimators of a parameter based sizes $n_1, n_2, ..., n_k$. The ultimate goal is not to have *k* different estimators of the parameters but a single estimator. In this case, the issue is how to combine the different sample information together into one estimator which is good enough to provide the information about the parameter.

Estimation of population mean and its variance

Let

Y : characteristic under study

y_{ij} : value of j^{th} unit in i^{th} stratum $j = 1,2,\ldots,n_i$, $i = 1,2,...,k$.

$\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$: population mean of i^{th} stratum

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: sample mean from i^{th} stratum or stratum mean.

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{Y}_i = \sum_{i=1}^{k} w_i\bar{Y}_i$: population mean

Note that the population mean is defined as the weighted arithmetic mean of stratum means in case of stratified sampling where the weights are provided in terms of strata sizes.

Based on the expression $\bar{Y} = \frac{1}{N}\sum_{i=1}^{k} N_i\bar{Y}_i$, one may choose the sample mean

$\bar{y} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{y}_i$,

as a possible estimator of $\bar{Y}$.

Since the sample in each stratum is drawn by SRS, so

$E(\bar{y}_i) = \bar{Y}_i$,

thus

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{k} n_i E(\bar{y}_i)$$
$$= \frac{1}{n}\sum_{i=1}^{k} n_i \bar{Y}_i$$
$$\neq \frac{1}{N}\sum_{i=1}^{k} n_i \bar{Y}_i$$
$$\neq \bar{Y}$$

and $\bar{y}$ turns out to be a biased estimator of $\bar{Y}$. Based on this, one can modify $\bar{y}$ so as to obtain an unbiased estimator of $\bar{Y}$. Consider the stratum mean which is defined as the weight arithmetic mean of strata sample means with strata sizes as weights.

$$\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{k} N_i \bar{y}_i.$$

Now

$$E(\bar{y}_{st}) = \frac{1}{N}\sum_{i=1}^{k} N_i E(\bar{y}_i)$$
$$= \frac{1}{N}\sum_{i=1}^{k} N_i \bar{Y}_i$$
$$= \bar{Y}$$

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

Variance of $\bar{y}_{st}$

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \sum_{i\neq j}^{k}\sum^{k} w_i w_j Cov(\bar{y}_i, \bar{y}_j)$$

Since all the samples have been drawn independently from each strata by SRSWOR so

$$Cov(\bar{y}_i, \bar{y}_j) = 0$$

$$Var(\bar{y}_i) = \frac{N_i - n_i}{N_i n_i} S_i^2$$

where

$$S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2$$

Thus

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2$$

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

Observe that $Var(\bar{y}_{st})$ is small when S_i^2 is small. This observation suggest how to construct the strata . If S_i^2 is small for all I = 1,2,...,k, the $Var(\bar{y}_{st})$ will also be small . That is why it was mentioned earlier that the strata are to be constructed such that they are within homogeneous, i.e., S_i^2 is small and among heterogeneous.

For example, the units in geographical proximity will tend to be more close. The consumption pattern in households will be similar within a lower income group housing society and within a higher income group housing society whereas they will differ a lot between the two housing societies based on income.

Estimate of Variance

Since the samples have been drawn by SRSWOR, so

$$E(s_i^2) = S_i^2$$

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

and $Var(\bar{y}_i) = \frac{N_i - n_i}{w_i n_i} s_i^2$

so $Var(\bar{y}_{st}) = \sum_{j=1}^{k} w_i^2 \; Var(\bar{y}_i)$

$$= \sum_{i=1} w_i^2 \left(\frac{N_i - n_i}{N_i n_i}\right) s_i^2$$

Note: If SRSWR is used instead of SRSWOR for drawing the samples from stratum, then appropriate changes can be made at required steps.

In this case

$$\bar{y}_{st} = \sum_{i=1}^{k} w_i \bar{y}_i$$

$$E(\bar{y}_{st}) = \bar{Y}$$

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \left(\frac{N_i - 1}{N_i n_i} \right) S_i^2 = \sum_{=1}^{k} w_i^2 \frac{\sigma_i^2}{n_i}$$

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i}$$

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

Advantages of stratified sampling

1. Data of known precision may be required for certain parts of the population.
   This can be accomplished with a more careful investigation to few strata.
   Example: In order to know the direct impact of hike is petrol prices, the population can be divided into strata like lower income group, middle income group and higher income group. Obviously, the higher income group is more affected than the lower income group. So more careful investigation can be made only in the higher income group strata.
2. Sampling problems may differ in different parts of the population.
   Example: To study the consumption pattern of households, the people living is houses, hotels, hospitals, prison etc. are to be treated differently.
3. Administrative convenience can be exercised in stratified sampling.
   Example: In taking a sample of villages from a big state, it a more administratively convenient to consider the districts as strata so that the administrative setup at district level may be used for this purpose.
4. Full cross-section of population can be obtained through stratified sampling. It may be possible is SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different segments of the population to any desired extent. The desired degree of representation of some specified parts of population is also possible.
5. Substantial gain in efficiency is achieved of strata are formed intelligently.
6. In case of skewed population, use of stratification is of importance since larger weight may have to be given for the few extremely large units for reducing the sampling variability.

7. If population is large, then it is convenient to sample separately from the strata rather then the entire population.
8. The population mean or population total can be estimated with higher precision by suitably providing the weights to the estimates obtained from each stratum.

Allocation problem and choice of sample sizes is different strata

Question: How to choose the sample sizes $n_1, n_2, ..., n_k$ so that the available resources are used in an effective way?

There are two aspects of choosing the sample sizes:

(i) Minimize the cost of survey for a specified precision.

(ii) Maximize the precision for a given cost.

Note: The sample size can not be determined by minimizing both the cost and variability simultaneously. The cost function is directly proportional to the sample size whereas variability is inversely proportional to the sample size.

1. Equal allocation

Choose the sample size n_i to be same for all strata.

Draw sample of equal size from each strata.

Let *n* be the sample size and *k* be the number of strata.

$n_i = \frac{n}{k}$ for all $i = 1, 2, ..., k$.

2. Proportional allocation

For fixed *k*, select n_i such that it is proportional to stratum size N_i, i.e.,

$n_i \ \alpha \ N_i$

or $n_i = CN_i$

Where *C* is constant of proportionality.

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} CN_i$$

or $n = CN$

$$\Rightarrow C = \frac{n}{N}$$

Thus $n_i = \left(\frac{n}{N}\right) N_i$

Such allocation arises from the considerations like operational convenience..

3. Neyman or optimum allocation

This allocation considers the size of strata and variability both

$n_i \propto N_i S_i$

$n_i = C^* N_i S_i$

where C^* is the constant of proportionality.

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} C^* N_i S_i$$

or $n = C^* \sum_{i=1}^{k} N_i S_i$

or $C^* = \frac{n}{\sum_{i=1}^{k} N_i S_i}$

Thus $n_i = \frac{n N_i S_i}{\sum_{i=1}^{k} N_i S_i}$

This allocation arises when the $Var\left(\bar{y}^{st}\right)$ is minimized subject to the constraint $\sum_{i=1}^{k} n_i$ (prespecified).

The knowledge of $S_i (i = 1, 2, ..., k)$ is needed to know n_i.

Choice of sample size based on cost of survey and variability

The cost of survey depends upon the nature of survey. A simple choice of cost function is

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

where

C: total cost

C_0 : overhead cost, e.g., setting up of office, training people etc

C_i : cost per unit in the i^{th} stratum

$\sum_{i=1}^{k} C_i n_i$: total cost within sample.

To find n_i under this cost function, consider the Lagrangian function with Lagrangian multiplier λ as

$$\begin{aligned}\phi &= Var(\bar{y}_{st}) + \lambda^2 (C - C_0) \\ &= \sum_{i=1}^{k} w_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i} \right) S_i^2 + \lambda^2 \sum_{i=1}^{k} C_i n_i \\ &= \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{n_i} + \lambda^2 \sum_{i=1}^{k} C_i n_i - \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i} \\ &= \sum_{i=1}^{k} \left[\frac{w_i S_i}{\sqrt{n_i}} - \lambda \sqrt{C_i n_i} \right]^2 + \text{terms independent of } n_i.\end{aligned}$$

Thus ϕ is minimum when

$$\frac{w_i S_i}{\sqrt{n_i}} = \lambda \sqrt{C_i n_i} \text{ for all } i.$$

$$\text{or } n_i = \frac{1}{\lambda} \frac{w_i S_i}{\sqrt{C_i}}. \quad (*)$$

How to determine λ ?

There are two ways to determine λ .

(i) Minimize variability for fixed cost and

(ii) Minimize cost for given variability.

We consider both the cases

(i) Minimize variability for fixed cost

Let $C = C_0^*$ be fixed.

$$\text{so } \sum_{i=1}^{k} C_i n_i = C_0^*$$

$$\text{So} \quad \text{or } \sum_{i=1}^{k} C_i \frac{w_i S_i}{\lambda \sqrt{C_i}} = C_0^*$$

$$\text{or } \lambda = \frac{\sum_{i=1}^{k} \sqrt{C_i}\, w_i S_i}{C_0^*}$$

Substituting λ in the expression for n_i, the optimum n_i is obtained as

$$n_i^* = \frac{w_i S_i}{\sqrt{C_i}} \left(\frac{C_0^*}{\sum_{i=1}^{k} \sqrt{C_i}\, w_i S_i} \right).$$

The required sample size to estimate $\bar{Y}$ such that the variance is minimum for given cost $C = C_0^*$ is

$$n = \sum_{i=1}^{k} n_i^*$$

(ii) Minimize cost for given variability

Let $V = V_0$ be prespecified variance. Now determine n_i such that

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

$$\text{or } \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{n_i} = V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}$$

$$\text{or } \sum_{i=1}^{k} \frac{\lambda \sqrt{C_i}}{W_i S_i} w_i^2 S_i^2 = V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}$$

(Substituting $n_i = \frac{1}{\lambda} \frac{w_i S_i}{\sqrt{C_i}}$ from equation (*)).

$$\text{or } \lambda = \frac{V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}}{\sum_{i=1}^{k} w_i S_i \sqrt{C_i}}.$$

Thus the optimum n_i is

$$\tilde{n}_i = \frac{w_i S_i}{\sqrt{C_i}} \left(\frac{\sum_{i=1}^{k} w_i^2 S_i^2 \sqrt{C_i}}{V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}} \right) ..$$

So the required sample size to estimate $\bar{Y}$ such that cost C is minimum for a prespecified variance V_0 is $n = \sum_{i=1}^{k} \tilde{n}_i$.

Sample size under proportional allocation

(i) If cost $C = C_0$ is fixed then $C_0 = \sum_{i=1}^{k} C_i n_i$.

Under proportional allocation, $n_i = \frac{n}{N} N_i = n w_i$

So $C_0 = n \sum_{i=1}^{k} w_i C_i$

or $n = \frac{C_0}{\sum_{i=1}^{k} w_i C_i}$.

Thus $n_i = \frac{C_o w_i}{\sum w_i C_i}$.

(ii) If variance = V_0 is fixed, then

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

or $$\sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} = V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

or $$\sum_{i=1}^{k}\frac{w_i^2S_i^2}{nW_i} = V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i} \quad (\text{using } n_i = nw_i)$$

or $$n = \frac{\sum_{i=1}^{k} w_i^2S_i^2}{V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}}$$

or $$n_i = w_i\frac{\sum_{i=1}^{k} w_iS_i^2}{V_0 + \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}}$$

This is known **Bowley's allocation**.

Variances under different allocation

Now we derive the variance of $\bar{y}_{st}$ under proportional and optimum allocations.

(i) Proportional allocation

Under proportional allocation

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

$$Var(\bar{y})_{st} = \sum_{i=1}^{k}\left(\frac{N_i-n_i}{N_in_i}\right)w_i^2S_i^2$$

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

$$= \frac{N-n}{N_n}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}$$

$$= \frac{N-n}{Nn}\sum_{i=1}^{k}w_iS_i^2$$

(ii) Optimum allocation

$$n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$$

$$V_{opt}(\bar{y}_{st}) = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)w_i^2S_i^2$$

$$= \sum_{i=1}^{k}\frac{w_i^2S_i^2}{n_i} - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[w_i^2S_i^2\left(\frac{\sum_{i=1}^{k} N_iS_i}{nN_iS_i}\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \sum_{i=1}^{k}\left[\frac{1}{n}.\frac{N_iS_i}{N^2}\left(\sum_{i=1}^{k} N_iS_i\right)\right] - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{N_iS_i}{N}\right)^2 - \sum_{i=1}^{k}\frac{w_i^2S_i^2}{N_i}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k} w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i^2S_i^2.$$

Comparison of variance of sample mean under SRS with stratified mean under proportional and optimal allocation:

(a.) Proportional allocation:

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

$$V_{\text{Prop}}(\bar{y}_{st}) = \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_iS_i^2}{N}.$$

In order to compare $V_{SRS}(\bar{y})$ and $V_{prop}(\bar{y}_{st})$, first we attempt to express S^2 as a function of S_i^2.

Consider

$$(N-1)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij}-\bar{Y})+(Y_i-\bar{Y})\right]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_i-\bar{Y})^2$$

$$= \sum_{i=1}^{k}(N_i-1)S_i^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2$$

$$\frac{N-1}{N}S^2 = \sum_{i=1}^{k}\frac{N_i-1}{N}S_i^2 + \sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2.$$

For simplification, we assume that N_i is large enough to permit the approximation

$$\frac{N_i-1}{N_i}\approx 1 \text{ and } \frac{N-1}{N}\approx 1.$$

Thus

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

or $$\frac{N-n}{Nn}S^2 = \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}S_i^2 + \frac{N-n}{Nn}\sum_{i=1}^{k}\frac{N_i}{N}(\bar{Y}_i-\bar{Y})^2$$

$$Var_{SRS}(\bar{Y}) = V_{prop}(\bar{y}_{st}) + \frac{N-n}{Nn}\sum_{i=1}^{k}w_i(\bar{Y}_i-\bar{Y})^2$$

Since $\sum_{i=1}^{k}w_i(\bar{Y}_i-\bar{Y})^2 \geq 0,$

$\Rightarrow Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y}).$

Larger gain in the difference is achieved when $\bar{Y}_i$ differs from $\bar{Y}$ more.

(b.) Optimum allocation

$$V_{opt}(\bar{y}_{st}) = \frac{1}{n}\left(\sum_{i=1}^{k}w_iS_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k}w_iS_i^2.$$

Consider

$$V_{prop}(\bar{y}_{st}) - V_{opt}(\bar{y}_{st}) = \left[\left(\frac{N-n}{Nn}\right)\sum_{i=1}^{k} w_i S_i^2\right] - \left[\frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i\right)^2 - \frac{1}{N}\sum_{i=1}^{k} w_i S_i^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{k} w_i S_i^2 - \left(\sum_{i=1}^{k} w_i S_i\right)^2\right]$$

$$= \frac{1}{n}\left(\sum_{i=1}^{k} w_i S_i^2 - \frac{1}{n}\bar{S}^2\right)$$

$$= \frac{1}{n}\sum w_i (S_i - \bar{S})^2$$

where

$$\bar{S} = \sum_{i=1}^{k} w_i S_i$$

$$\Rightarrow Var_{prop}(\bar{y}_{st}) - Var_{opt}(\bar{y}_{st}) \geq 0.$$

Larger gain in the difference is achieved when S_i differ from $\bar{S}$ more,

Combining (a) and (b), we have

$$Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y})$$

Estimate of variance and confidence intervals

Under SRSWOR, an unbiased estimate of S_i^2 for the i^{th} stratum (i = 1,2,...,k) is

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

In stratified sampling,

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} S_i^2.$$

So an unbiased estimate of $Var(\bar{y}_{st})$ is

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} w_i^2 \frac{N_i - n_i}{N_i n_i} s_i^2$$

$$= \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i} - \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{N_i}$$

$$= \sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i} - \frac{1}{N}\sum_{i=1}^{k} w_i s_i^2$$

The second term represents the reduction due to finite population correction.

The confidence limits of $\bar{Y}$ can be obtained as

$$\bar{y}_{st} \pm t\sqrt{Var(\bar{y}_{st})}$$

Assuming $\bar{y}_{st}$ is normally distributed and $\sqrt{Var(\bar{y}_{st})}$ is well determined so that t can be read from normal distribution tables. If only few degrees of freedom are provided by each stratum, then *t* values are obtained from the table of student's *t*-distribution.

The distribution of $\sqrt{Var(\bar{y}_{st})}$ is generally complex. An approximate method of assigning an effective number of degrees of freedom (n_e) to $\sqrt{Var(\bar{y}_{st})}$ is

$$n_e = \frac{\left(\sum_{i=1}^{k} g_i s_i^2\right)^2}{\sum_{i=1}^{k} \frac{g_i^2 s_i^4}{n_i - 1}}$$

where $g_i = \frac{N_i(N_i - n_i)}{n_i}$

and $Min(n_i - 1) \le n_e \le \sum_{i=1}^{k}(n_i - 1)$

assuming y_{ij} are normal.

Modification of optimal allocation

Sometimes is optimal allocation, the size of subsample exceeds the stratum size. In such a case,

replace n_i by N_i

and recompute the rest of $n_i's$ by the revised allocation.

For example, if $n_i > N_1,$ then take the revised $n_i's$ as

$$\tilde{n}_1 = N_1$$

and $\tilde{n}_i = \frac{(n - N_1)w_i S_i}{\sum_{i=2}^{k} w_i S_i}$; $i = 1, 2, 3, ..., k.$

provided $\tilde{n}_i \le N_i$ for all $i = 2,3,\ldots,k.$

Suppose in revised allocation, $\tilde{n}_2 > N_2$ then the re-revised allocation would be

$$\tilde{n}_1 = N_1$$
$$\tilde{n}_2 = N_2$$
$$\tilde{n}_i = \frac{(n - N_1 - N_2) w_i S_i}{\sum_{i=3}^{k} w_i S_i}; i = 3, 4 ... k.$$

provided $\tilde{n}_i < N_i$ for all $i = 3, 4, ..., k.$

We continue this process until every $\tilde{n}_i < N_i$.

In such cases, the formula for minimum variance of $\bar{y}_{st}$ need to be modified as

$$Min\ Var(\bar{y}_{st}) = \frac{(\sum^{*} w_i S_i)^2}{n^*} - \frac{\sum^{*} w_i S_i^2}{N}$$

where $\sum^{*}$ denotes the summation over the strata in which $\tilde{n}_i \leq N_i$ and n^* is the revised total sample size in the strata.

Stratified sampling for proportions

If the characteristic under study is qualitative is nature, then its values will fall into one of the two mutually exclusive complementary class C and C'. Ideally, only two strata are needed in which all the units can be divided depending on whether they belong to C or its complement C'. It is difficult to achieve in practice. So the strata are constructed such that the proportion in C varies as much as possible among strata.

Let

$P_i = \frac{A_i}{N_i}$: Proportion of units in C in i^{th} stratum

$p_i = \frac{a_i}{n_i}$: Proportion of units in C in i^{th} sample.

As estimate of population proportion based on stratified sampling is

$$p_{st} = \sum_{i=1}^{k} \frac{N_i p_i}{N}.$$

which is based on

$$Y_{ij} = \begin{cases} 1 & \text{when } j^{th} \text{ unit belongs to } i^{th} \text{ stratum is in } C \\ 0 & \text{otherwise} \end{cases}$$

and $\bar{y}_{st} = p_{st}$.

Here $S_i^2 = \frac{N_i}{N_i - 1} P_i Q_i$

where $Q_i = 1 - P_i$.

Also $Var(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} w_i^2 S_i^2$.

So $Var(p_{st}) = \frac{1}{N^2} \sum_{i=1}^{k} \frac{N_i^2 (N_i - n_i)}{N_i - 1} \frac{P_i Q_i}{n_i}$.

If the finite population correction can be ignored, then

$$Var(p_{st}) = \sum_{i=1}^{k} w_i^2 \frac{P_i Q_i}{n_i}.$$

If proportional allocation is used for n_i , then variance of p_{st} is

$$\begin{aligned} Var(p_{st})_{prop} &= \frac{N-n}{N} \frac{1}{Nn} \sum_{i=1}^{k} \frac{N_i^2 P_i Q_i}{N_i - 1} \\ &= \frac{N-n}{N} \sum_{i=1}^{k} w_i P_i Q_i \end{aligned}$$

and its estimate is

$$Var(p_{st})_{prop} = \frac{N-n}{N} \sum_{i=1}^{k} w_i \frac{p_i q_i}{n_i - 1}.$$

The best choice of n_i such that it minimizes the variance for fixed total sample size is

$$\begin{aligned} n_i &\propto N_i \sqrt{\frac{N_i P_i Q_i}{N_i - 1}} \\ &= N_i \sqrt{P_i Q_i} \end{aligned}$$

Thus $n_i = n \frac{N_i \sqrt{P_i Q_i}}{\sum_{i=1}^{k} N_i \sqrt{P_i Q_i}}$.

Similarly, the best choice of n_i such that the variance is minimum for fixed cost $C = C_0 + \sum_{i=1}^{k} C_i n_i$ is

$$n_i = \frac{nN_i\sqrt{\frac{P_iQ_i}{C_i}}}{\sum_{i=1}^{k} N_i\sqrt{\frac{P_iQ_i}{C_i}}}.$$

Estimation of the gain in precision due to stratification

What is the advantage of stratifying a population in the sense that instead of using SRS, the population is divided into various strata is the question of interest. This is answered by estimating the variance of estimators of population mean under SRS (without stratification) and stratified sampling by evaluating

$$\frac{Var_{SRS}(\bar{y}) - Var(\bar{y}_{st})}{Var(\bar{y}_{st})}.$$

Since $Var_{SRS}(\bar{y}) = \frac{N-n}{Nn}S^2.$

How to estimate S^2 based on a stratified sample?

$$\begin{aligned}
(N-1)S^2 &= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}\left[(Y_{ij}-\bar{Y}) + (Y_i-\bar{Y})\right]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N_i}(Y_{ij}-\bar{Y})^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2 + \sum_{i=1}^{k}N_i(\bar{Y}_i-\bar{Y})^2 \\
&= \sum_{i=1}^{k}(N_i-1)S_i^2 + N\left[\sum_{i=1}^{k}W_i\bar{Y}_i^2 - \bar{Y}^2\right].
\end{aligned}$$

In order to estimate S^2, we need to estimates of S_i^2, $\bar{Y}_i^2$ and $\bar{Y}^2$.

For estimate of S_i^2, we have

$$E(s_i^2) = S_i^2$$

So $\hat{S}_i^2 = s_i^2.$

For estimate of $\bar{Y}_i^2$, we know

$$Var(\bar{y}_i) = E(\bar{y}_i^2) = -[E(\bar{y}_i)]^2$$
$$= E(\bar{y}_i^2) = -\bar{Y}_i^2$$
$$\Rightarrow \bar{Y}_i^2 = E(\bar{y}_i^2) - Var\,\bar{y}_i$$

An unbiased estimate of $\bar{Y}_i^2$ is

$$\hat{\bar{Y}}_i^2 = \bar{y}_i^2 - Var(\bar{y}_i)$$
$$= \bar{y}_i^2 - \left(\frac{N_i - n_i}{N_i n_i}\right) s_i^2$$

For example of $\bar{Y}^2$, we know

$$Var(\bar{y}_{st}) = E(\bar{y}_{st}^2) = -[E(\bar{y}_{st})]^2$$
$$= E(\bar{y}_{st}^2) = -\bar{Y}^2$$
$$\Rightarrow \bar{Y}^2 = E(\bar{y}_{st}^2) - Var(\bar{y}_{st})$$

So an estimate of $\bar{Y}^2$ is

$$\hat{\bar{Y}}^2 = \bar{y}_{st}^2 - Var(\bar{y}_{st})$$
$$= \bar{y}_{st}^2 - \sum_{i=1}^{k}\left(\frac{N_i - n_i}{N_i n_i}\right) w_i^2 s_i^2$$

Substituting these estimates, the estimate of S^2 is obtained from

$$(N-1)S^2 = \sum_{i=1}^{k}(N_i - 1)S_i^2 + N\left[\sum_{i=1}^{k} w_i \bar{Y}_i^2 - \bar{Y}^2\right]$$

as
$$\hat{S}^2 = \frac{1}{N-1}\sum_{i=1}^{k}(N_i - 1)\hat{S}_i^2 + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i \hat{\bar{Y}}_i^2 - \hat{\bar{Y}}^2\right]$$
$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2 + \frac{N}{N-1}\left(\sum_{i=1}^{k} w_i\left(\frac{N_i - n_i}{N_i n_i} s_i^2\right)\right) - \left(\sum_{i=1}^{k}\frac{N_i - n_i}{N_i n_i} w_i^2 s_i^2\right)\right]$$
$$= \frac{1}{N-1}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2\right] + \frac{N}{N-1}\left[\sum_{i=1}^{k} w_i^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i - n_i}{N_i n_i} s_i^2\right].$$

Thus

$$Var_{SRS}(\bar{y}) = \frac{N-n}{N_n}\hat{S}^2$$
$$= \frac{N-n}{N(N-1)n}\left[\sum_{i=1}^{k}(N_i - 1)s_i^2\right] + \frac{N(N-n)}{nN(N-1)}\left[\sum_{i=1}^{k} w_i^2 - \sum_{i=1}^{k} w_i(1-w_i)\frac{N_i - n_i}{N_i n_i} s_i^2\right]$$

and

$$Var(\bar{y}_{st}) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} w_i^2 s_i^2.$$

Substituting these expressions in

$$\frac{Var_{SRS}(\bar{y}) - Var(\bar{y}_{st})}{Var(\bar{y}_{st})},$$

the gain is efficiency due to stratification can be obtained.

If any other particular allocation is used, then substituting appropriate n_i, such gain can be estimated.

Interpenetrating subsamping

Suppose a sample consists of two or more subsamples which are drawn according to the same sampling scheme. The samples are such that each subsample yields an estimate of parameter. Such subsamples are called interpenetrating subsamples.

The subsamples need not necessary be independent. The assumption of independent subsamples helps in obtaining an unbiased estimate of the variance of the composite estimator. This is even helpful if the sample design is complicated and the expression for variance of the composite estimator is complex.

Let there be g independent interpenetrating subsamples and $t_1, t_2, ..., t_g$ be g unbiased estimators of parameter θ where $t_j (j = 1, 2, ..., g)$ is based on j^{th} interpenetrating subsample.

Then an unbiased estimator of θ is given by

$$\hat{\theta} = \frac{1}{g} \sum_{j=1}^{g} t_j = \bar{t}, \text{ say.}$$

Then

$$E(\hat{\theta}) = E(\bar{t}) = \theta$$

and

$$\hat{Var}(\hat{\theta}) = \hat{Var}(\bar{t}) = \frac{1}{g(g-1)} \sum_{j=1}^{g} (t_j - \bar{t})^2 .$$

Note that

$$E\left[Var(\bar{t})\right] = \frac{1}{g(g-1)} E\left[\sum_{j=1}^{g}(t_j-\theta)^2 - g(\bar{t}-\theta)^2\right]$$
$$= \frac{1}{g(g-1)} E\left[\sum_{j=1}^{g} Var(t_j) - g\,Var(\bar{t})\right]$$
$$= \frac{1}{g(g-1)}(g^2-g)Var(\bar{t})$$
$$= Var(\bar{t})$$

If distribution of each estimator t_j is symmetric as about θ, then the confidence interval of θ can be obtained by

$$P\left[Min(t_1,t_2,...,t_g) < \theta < Max(t_1,t_2,...,t_g)\right] = 1-\left(\frac{1}{2}\right)^{g-1}.$$

Implementation of interpenetrating subsamples in stratified samping

Consider the set up of stratified sampling. Suppose each stratum provides an independent interpenetrating subsample. So based on each stratum, there are L independent interpenetrating subsamples drawn according to same sampling scheme.

Let $\hat{Y}_{ij(tot)}$ be the unbiased estimator of total of j^{th} stratum based on the i^{th} subsample ,

i = 1,2,...,L; j = 1,2,...,k.

An unbiased estimator of the j^{th} stratum total is given by

$$\hat{Y}_{j(tot)} = \frac{1}{L}\sum_{i=1}^{J}\hat{Y}_{ij(tot)}$$

and an unbiased estimator of the variance of $\hat{Y}_{ij(tot)}$ is given by

$$Var(\hat{Y}_{j(tot)}) = \frac{1}{L(L-1)}\sum_{i=1}^{L}(\hat{Y}_{ij(tot)} - \hat{Y}_{j(tot)})^2.$$

Thus an unbiased estimator of population total Y_{tot} is

$$\hat{Y}_{tot} = \sum_{j=1}^{k}\hat{Y}_{j(tot)} = \frac{1}{k}\sum_{i=1}^{L}\sum_{j=1}^{k}\hat{Y}_{ij(tot)}$$

and unbiased estimator of its variance is

$$Var(\hat{Y}_{tot}) = \sum_{j=1}^{k} Var(\hat{Y}_{j(tot)})$$

$$= \frac{1}{L(L-1)} \sum_{i=1}^{L} \sum_{j=1}^{k} (\hat{Y}_{ij(tot)} - \hat{Y}_{j(tot)})^2 .$$

Post Stratifications

Sometimes the stratum to which a unit belongs to may be known after the field survey only. For example, the age of persons, their educational qualifications etc. can not be known in advance. In such cases, we adopt the post stratification procedure to increase the precision of the estimates.

In post stratification

- draw a sample by simple random sampling from the population and carry out the survey.
- after the completion of survey, stratify the sampling units to increase the precision of the estimates.

Assume the stratum size N_i is fairly accurately known. Let

m_i : number of sampling units from i^{th} stratum, $i = 1,2,...,k$.

$$\sum_{i=1}^{k} m_i = n.$$

Note that m_i is a random variable (and that is why we are not using the symbol n_i as earlier).

Assume n is large enough or the stratification is such that the probability that some $m_i = 0$ is negligibly small. In case, $m_i = 0$ for some strata, two or more strata can be combined to make the sample size to be non-zero before evaluating the final estimates.

A post stratified estimator of population mean $\bar{Y}$ is

$$\bar{y}_{post} = \frac{1}{N} \sum_{i=1}^{k} N_i \bar{y}_i .$$

Now

$$E(\bar{y}_{post}) = \frac{1}{N} E\left[\sum_{i=1}^{k} N_i E(\bar{y}_i / m_1, m_2, ..., m_k)\right]$$

$$= \frac{1}{N} E\left[\sum_{i=1}^{k} N_i \bar{y}_i\right]$$

$$= \bar{Y}$$

$$Var(\bar{y}_{post}) = E\left[Var(\bar{y}_{post} / m_1, m_2, ..., m_k)\right] + Var\left[E(\bar{y}_{post} / m_1, m_2, ..., m_k)\right]$$

$$= E\left[\sum_{i=1}^{k} w_i^2 \left(\frac{1}{m_i} - \frac{1}{N_i}\right) S_i^2\right] + Var(\bar{Y})$$

$$= \sum_{i=1}^{k} w_i^2 \left[E\left(\frac{1}{m_i}\right) - \left(\frac{1}{N_i}\right)\right] S_i^2 \quad (Var(\bar{Y}) = 0).$$

To find $E\left(\frac{1}{m_i}\right) - \frac{1}{N_i}$, proceed as follows:

Consider the estimate of ratio based on ratio method of estimation as

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{j=1}^{n} y_i}{\sum_{j=1}^{n} x_j}, \qquad R = \frac{\bar{Y}}{\bar{X}} = \frac{\sum_{j=1}^{N} Y_i}{\sum_{j=1}^{N} X_i}.$$

We know that

$$E(\hat{R}) - R = \frac{N-n}{Nn} . \frac{RS_X^2 - S_{XY}}{\bar{X}^2}.$$

Let $x_j = \begin{cases} 1 \text{ of } j^{th} \text{ unit belongs to } i^{th} \text{ stratum} \\ 0 \text{ otherwise} \end{cases}$

$y_j = 1$ for all $j = 1,2,...,N.$

Then $R, \hat{R}$ and S_x^2 reduces to

$$\hat{R} = \frac{n}{n_i}$$

$$R = \frac{N}{N_i}$$

$$S_x^2 = \frac{1}{N-1}\left(N_i - \frac{N_i^2}{N}\right)$$

Using these values in $E(\hat{R}) - R,$ we have

$$E(\hat{R}) - R = E\left(\frac{n}{n_i}\right) - \frac{N}{N_i} = \frac{N(N-n)(N-N_i)}{nN_i^2(N-1)}.$$

Thus

$$E\left(\frac{1}{n_i}\right) - \frac{1}{N_i} = \frac{N}{nN_i} + \frac{N(N-n)(N-N_i)}{n^2N_i^2(N-1)} - \frac{1}{N_i}$$

$$= \frac{(N-n)N}{n(N-1)N_i}\left(1 + \frac{N}{N_i n} - \frac{1}{n}\right).$$

Replacing m_i is place of n_i, we obtain

$$E\left(\frac{1}{m_i}\right) - \frac{1}{N_i} = \frac{(N-n)N}{n(N-1)N_i}\left(1 + \frac{N}{N_i n} - \frac{1}{n}\right)$$

Now substrate this is the expression of $Var(\bar{y}_{post})$ as

$$Var(\bar{y}_{post}) = \sum_{i=1}^{k} w_i^2 \left[E\left(\frac{1}{m_i}\right) - \frac{1}{N_i}\right] S_i^2$$

$$= \sum_{i=1}^{k} w_i^2 S_i^2 \left[\frac{N-n}{(N-1)n} \cdot \frac{N}{N_i}\left(1 + \frac{N}{nN_i} - \frac{1}{n}\right)\right]$$

$$= \frac{N-n}{n(N-1)} \sum_{i=1}^{k} w_i^2 S_i^2 \left[\frac{1}{w_i}\left(1 + \frac{1}{nw_i} - \frac{1}{n}\right)\right]$$

$$= \frac{N-n}{n^2(N-1)} \sum_{i=1}^{k} w_i^2 S_i^2 \left[n - 1 + \frac{1}{w_i}\right]$$

$$= \frac{N-n}{n^2(N-1)} \sum_{i=1}^{k} (nw_i + 1 - w_i) S_i^2$$

$$= \frac{N-n}{n^2(N-1)} \sum_{i=1}^{k} w_i S_i^2 + \frac{N-n}{n^2(N-1)} \sum_{i=1}^{k} (1-w_i) S_i^2$$

Assuming $N - 1 \approx N.$

$$V(\bar{y}_{post}) = \frac{N-n}{Nn} \sum_{i=1}^{n} w_i^2 S_i^2 + \frac{N-n}{N^2 n} \sum_{i=1}^{n} (1-w_i) S_i^2$$

$$= V_{prop}(\bar{y}_{st}) + \frac{N-n}{Nn^2} \sum_{i=1}^{n} (1-w_i) S_i^2$$

The second term is the contribution in the variance of $\bar{y}_{post}$ due to $m_i's$ not being proportionately distributed.

If $S_i^2 \approx S_w^2,$ say for all *i*, then the last term is

$$\frac{N-n}{Nn^2}\sum_{i=1}^{k}(1-w_i)S_i^2 = \frac{N-n}{Nn^2}S_i^2(k-1) \quad (Since \ \sum_{i=1}^{k} w_i = 1)$$

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

$$= \frac{k-1}{n}Var(\bar{y}_{st}).$$

The increase in variance over $Var_{prop}(\bar{y}_{st})$ is small if the average sample size $\bar{n} = \frac{n}{2}$ per stratum.

Thus a post stratification with a large sample produces an estimator which is almost precise as an estimator in stratified sampling with proportional allocation.

Chapter 5

Ratio and Product Methods of Estimation

An important objective is any statistical estimation procedure is to obtain the estimators of parameters of interest with more precision. It is also well understood that incorporation of more information in the estimation procedure yields better estimators, provided the information is valid and proper. Use of such auxiliary information through made through the ratio method of estimation to obtain an improved estimator of population mean. In ratio method of estimation, auxiliary information on a variable is available which is linearly related to the variable under study and is utilized to estimate the population mean.

Let Y be the variable under study and X be any auxiliary variable which is correlated with Y. The observation x_i on X and y_i on Y are obtained for each sampling unit. The population mean $\bar{X}$ of X (or equivalently the population total X_{tot}) must be known. For example, $x_i's$ may be the values of $y_i's$ from .

- some earlier completed census,
- some earlier surveys,
- some characteristic on which it is easy to obtain information etc.

For example, if y_i is the quantity of fruits produced in the i^{th} plot, then x_i can be the area of i^{th} plot or the production of fruit in the same plot is previous error.

Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be the random sample of size n on paired variable (X, Y) drawn, preferably by SRSWOR, from a population of size N. The ratio estimate of population mean $\bar{Y}$ is

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R}\bar{X}$$

assuming the population mean $\bar{X}$ is known. The ratio estimator of population total $Y_{tot} = \sum_{i=1}^{N} Y_i$ is

$$\hat{Y}_{R(tot)} = \frac{y_{tot}}{x_{tot}} X_{tot}$$

where $X_{tot} = \sum_{i=1}^{N} X_i$ is the population total of X which is assumed to be known, $y_{tot} = \sum_{i=1}^{n} y_i$ and $x_{tot} = \sum_{i=1}^{n} x_i$ are the sample totals of Y and X respectively. The $\hat{Y}_{R(tot)}$ can be equivalently expressed as

$$\begin{aligned}\hat{Y}_{R(tot)} &= \frac{\bar{y}}{\bar{x}} X_{tot} \\ &= \hat{R} X_{tot}.\end{aligned}$$

Looking at the structure of ratio estimators, note that the ratio method estimates the relative change $\frac{Y_{tot}}{X_{tot}}$ that occurred after (x_i, y_i) were observed. It is clear that if the variation among the values of $\frac{y_i}{x_i}$ and is nearly same for all i = 1,2,...,n then values of $\frac{y_{tot}}{x_{tot}}$ (or equivalently $\frac{\bar{y}}{\bar{x}}$) varies little from sample to sample and ratio estimate will be of high precision.

Bias and mean squared error of ratio estimator:

Assume that the random sample $(x_i, y_i), i = 1, 2, ..., n$ is drawn by SRSWOR and population mean $\bar{X}$ is known. Then

$$\begin{aligned}E(\hat{\bar{Y}}_R) &= \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \frac{\bar{y}_i}{\bar{x}_i} \bar{X} \\ &\neq \bar{Y} \text{ (in general).}\end{aligned}$$

Moreover it is difficult to find the exact expression for $E\left(\frac{\bar{y}}{\bar{x}}\right)$ and $E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)$. So we approximate them and proceed as follows:

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = (1 + \varepsilon_o)\bar{Y}$$

$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} \Rightarrow \bar{x} = (1 + \varepsilon_1)\bar{X}.$$

Since SRSWOR is being followed , so

$$E(\varepsilon_0) = 0$$

$$E(\varepsilon_1) = 0$$

$$E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2} E(\bar{y} - \bar{Y})^2$$

$$= \frac{1}{\bar{Y}^2} \frac{N-n}{Nn} S_Y^2$$

$$= \frac{f}{n} \frac{S_Y^2}{\bar{Y}^2}$$

$$= \frac{f}{n} C_Y^2$$

where $f = \frac{N-n}{N}$, $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ and $C_Y = \frac{S_Y}{\bar{Y}}$ is the coefficient of variation related to *Y*.

Similarly,

$$E(\varepsilon_1^2) = \frac{f}{n} C_X^2$$

$$E(\varepsilon_0 \varepsilon_1) = \frac{1}{\bar{X}\bar{Y}} E[(\bar{x} - \bar{X})(\bar{y} - \bar{Y})]$$

$$= \frac{1}{\bar{X}\bar{Y}} \frac{N-n}{Nn} \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \frac{1}{\bar{X}\bar{Y}} . \frac{f}{n} S_{XY}$$

$$= \frac{1}{\bar{X}\bar{Y}} \frac{f}{n} \rho S_X S_Y$$

$$= \frac{f}{n} \rho \frac{S_X}{\bar{X}} \frac{S_Y}{\bar{Y}}$$

$$= \frac{f}{n} \rho C_X C_Y$$

where $C_X = \frac{S_X}{\bar{X}}$ is the coefficient of variation related to *X* and ρ is the correlation coefficient between *X* and *Y*.

Writing $\hat{\bar{Y}}_R$ in terms of $\varepsilon's$, we get

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$$

$$= \frac{(1+\varepsilon_0)\bar{Y}}{(1+\varepsilon_1)\bar{X}} \bar{X}$$

$$= (1+\varepsilon_0)(1+\varepsilon_1)^{-1}\bar{Y}$$

Assuming $|\varepsilon_1| < 1$, the term $(1+\varepsilon_1)^{-1}$ may be expanded as an infinite series and it would be convergent. Such assumption means that $\left|\frac{\bar{x}-\bar{X}}{\bar{X}}\right| < 1$, i.e., all possible estimate $\bar{x}$ of population mean $\bar{X}$ lies between 0 and $2\bar{X}$, This is likely to hold true if the variation in $\bar{x}$ is not large. In order to ensures that variation in $\bar{x}$ is small, assume that the sample size n it is fairly large. With this assumption,

$$\hat{\bar{Y}}_R = \bar{Y}(1+\varepsilon_0)(1-\varepsilon_1+\varepsilon_1^2-...)$$
$$= \bar{Y}(1+\varepsilon_0-\varepsilon_1+\varepsilon_1^2-\varepsilon_1\varepsilon_0+...)$$

So the estimation error of $\hat{\bar{Y}}_R$ is

$$\hat{\bar{Y}}_R - \bar{Y} = \bar{Y}(\varepsilon_0-\varepsilon_1+\varepsilon_1^2-\varepsilon_1\varepsilon_0+...).$$

In case, when sample size is large, then ε_0 and ε_1 are likely to be small quantities and so the terms involving second and higher powers of ε_0 and ε_1 would be negligibly small. In such a case

$$\hat{\bar{Y}}_R - \bar{Y} \simeq \bar{Y}(\varepsilon_0-\varepsilon_1)$$

and

$$E(\hat{\bar{Y}}_R - \bar{Y}) = 0.$$

So the ratio estimator is an unbiased estimator of population mean to the first order of approximation.

If we assume that only terms of ε_0 and ε_1 involving powers more than two are negligibly small (which is more realistic them assuming powers more than one are negligibly small), then

$$\hat{\bar{Y}}_R - \bar{Y} \simeq \bar{Y}(\varepsilon_0-\varepsilon_1+\varepsilon_1^2-\varepsilon_1\varepsilon_0)$$

and

$$E(\hat{\bar{Y}}_R - \bar{Y}) = \bar{Y}\left(0-0+\frac{f}{n}C_X^2-\frac{f}{n}\rho C_X C_y\right)$$

$$Bias(\hat{\bar{Y}}) = E(\hat{\bar{Y}}_R - \bar{Y}) = \frac{f}{n}\bar{Y}C_X(C_X-\rho C_Y).$$

Upto second order of approximation, the bias generally decreases as the sample size grows large.

The bias of $\hat{\bar{Y}}_R$ is zero, i.e.,

$$Bias(\hat{\bar{Y}}_R) = 0$$

if $E(\varepsilon_1^2 - \varepsilon_0\varepsilon_1) = 0$

or if $\frac{Var(\bar{x})}{\bar{X}^2} - \frac{Cov(\bar{x},\bar{y})}{\bar{X}\bar{Y}} = 0$

or if $\frac{1}{\bar{X}^2}\left[Var(\bar{x}) - \frac{\bar{Y}}{\bar{X}}Cov(\bar{x},\bar{y})\right] = 0$

or if $Var(\bar{x}) - RCov(\bar{x},\bar{y}) = 0$ (assuming $\bar{X} \neq 0$)

or if $R = \frac{\bar{Y}}{\bar{X}} = \frac{Cov(\bar{x},\bar{y})}{Var(\bar{x})}$

which is satisfied when the regression line of Y on X passes through origin.

Now, to find the mean squared error, consider

$$\begin{aligned} MSE(\hat{\bar{Y}}_R) &= E(\hat{\bar{Y}}_R - \bar{Y})^2 \\ &= E\left[\bar{Y}^2(\varepsilon_0 - \varepsilon_1 + \varepsilon_1^2 - \varepsilon_1\varepsilon_0 + ...)^2\right] \\ &\simeq E\left[\bar{Y}^2(\varepsilon_0^2 + \varepsilon_1^2 - 2\varepsilon_0\varepsilon_1)\right] \end{aligned}$$

Under the assumption $|\varepsilon_1| < 1$ and the terms of ε_0 and ε_1 involving powers more than two are negligible small,

$$\begin{aligned} MSE(\hat{\bar{Y}}_R) &= \bar{Y}^2\left[\frac{f}{n}C_X^2 + \frac{f}{n}C_Y^2 - \frac{2f}{n}\rho C_X C_Y\right] \\ &= \frac{\bar{Y}^2 f}{n}\left[C_X^2 + C_Y^2 - 2\rho C_X C_y\right] \end{aligned}$$

Up to the second order of approximation.

Efficiency of ratio estimator in comparison to SRSWOR

Ratio estimator is better estimate of $\bar{Y}$ than sample mean based on SRSWOR if

$$MSE(\hat{\bar{Y}}_R) < Var_{SRS}(\bar{y})$$

or if $\bar{Y}^2\frac{f}{n}(C_X^2 + C_Y^2 - 2\rho C_X C_Y) < \bar{Y}^2\frac{f}{n}C_Y^2$

or if $C_X^2 - 2\rho C_X C_Y < 0$

or if $\rho > \frac{1}{2}\frac{C_X}{C_Y}$.

Thus ratio estimator is more efficient than sample mean based on SRSWOR if

$$\rho > \frac{1}{2}\frac{C_X}{C_Y} \quad \text{if} \quad R > 0$$

and $\rho < -\frac{1}{2}\frac{C_X}{C_Y}$ if $R > 0$.

It is clear from this expression that the success of ratio estimator depends on how close is the auxiliary information to the variable under study.

Upper limit of ratio estimator:

Consider

$$\begin{aligned} Cov(\hat{R}, \bar{x}) &= E(\hat{R}\bar{x}) - E(\hat{R})E(\bar{x}) \\ &= E\left(\frac{\bar{y}}{\bar{x}}\bar{x}\right) - E(\hat{R})E(\bar{x}) \\ &= \bar{Y} - E(\hat{R})\bar{X}. \end{aligned}$$

Thus

$$\begin{aligned} E(\hat{R}) &= \frac{\bar{Y}}{\bar{X}} - \frac{Cov(\hat{R}, \bar{x})}{\bar{X}} \\ &= R - \frac{Cov(\hat{R}, \bar{x})}{\bar{X}} \\ Bias(\hat{R}) &= E(\hat{R}) - R \\ &= -\frac{Cov(\hat{R}, \bar{x})}{\bar{X}} \\ &= -\frac{\rho_{\hat{R},\bar{x}}\sigma_{\hat{R}}\sigma_{\bar{x}}}{\bar{X}} \end{aligned}$$

where $\rho_{\hat{R},\bar{x}}$ is the correlation between $\hat{R}$ and $\bar{x}$; $\sigma_{\hat{R}}$ and $\sigma_{\bar{x}}$ are the standard errors of $\hat{R}$ and $\bar{x}$ respectively.

Thus

$$\begin{aligned} \left|Bias(\hat{R})\right| &= \frac{\left|-\rho_{\hat{R},\bar{x}}\right|\sigma_{\hat{R}}\sigma_{\bar{x}}}{\bar{X}} \\ &\le \frac{\sigma_{\hat{R}}\sigma_{\bar{x}}}{\bar{X}} \quad \left(\left|\rho_{\hat{R},\bar{x}}\right| \le 1\right). \end{aligned}$$

assuming $\bar{X} > 0$. Thus

$$\left|\frac{Bias(\hat{R})}{\sigma_{\hat{R}}}\right| \leq \frac{\sigma_{\bar{x}}}{\bar{X}}$$

or $$\left|\frac{Bias(\hat{R})}{\sigma_{\hat{R}}}\right| \leq C_X$$

when C_X is the coefficient of variation of X. If $C_X < 0.1$, then the bias in $\hat{R}$ may be safely regarded as negligible in relation to standard error of $\hat{R}$.

Alternative form of *MSE* $(\hat{Y}_R)$

Consider

$$\sum_{i=1}^{N}(Y_i - RX_i)^2 = \sum_{i=1}^{N}\left[(Y_i - \bar{Y}) + (\bar{Y} - RX_i)\right]^2$$

$$= \sum_{i=1}^{N}\left[(Y_i - \bar{Y}) + R(X_i - \bar{X})\right]^2 \text{ (Using } \bar{Y} = R\bar{X})$$

$$= \sum_{i=1}^{N}(Y_i - \bar{Y})^2 + R^2\sum_{i=1}^{N}(X_i - \bar{X})^2 - 2R\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - RX_i)^2 = S_Y^2 + R^2S_X^2 - 2RS_{XY}.$$

The *MSE* of $\hat{\bar{Y}}_R$ has already been derived which is now expressed again as follows:

$$MSE(\hat{\bar{Y}}_R) = \frac{f\bar{Y}^2}{n}(C_Y^2 + C_X^2 - 2\rho C_X C_Y)$$

$$= \frac{f}{n}\bar{Y}^2\left(\frac{S_Y^2}{\bar{Y}^2} + \frac{S_X^2}{\bar{X}^2} - 2\frac{S_{XY}}{\bar{X}\bar{Y}}\right)$$

$$= \frac{f}{n}\frac{\bar{Y}^2}{\bar{Y}^2}\left(S_Y^2 + \frac{\bar{Y}^2}{\bar{X}^2}S_X^2 - 2\frac{\bar{Y}}{\bar{X}}S_{XY}\right)$$

$$= \frac{f}{n}\left(S_Y^2 + R^2S_X^2 - 2RS_{XY}\right)$$

$$= \frac{f}{n(N-1)}\sum_{i=1}^{N}(Y_i - RX_i)^2$$

$$= \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(Y_i - RX_i)^2.$$

Estimate of $MSE(\hat{Y}_R)$

Let $U_i = Y_i - RX_i,\ i = 1,2,..,N$ then MSE of $\hat{\bar{Y}}_R$ can be expressed as

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n}\frac{1}{N-1}\sum_{i=1}^{N}(U_i - \bar{U})^2$$

$$= \frac{f}{n}S_U^2$$

where $S_U^2 = \frac{1}{N-1}\sum_{i=1}^{N}(U_i - \bar{U})^2.$

Based on this, a natural estimator of $MSE(\hat{\bar{Y}}_R)$ is

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n}s_u^2$$

where $s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[(u_i - \bar{u}) - \hat{R}(x_i - \bar{x})\right]$$

$$= s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy},$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}}.$$

Based on the expression

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n(N-1)}\sum_{i=1}^{N}(Y_i - RX_i)^2,$$

an estimate of $MSE(\hat{\bar{Y}}_R)$ is

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n(n-1)}\sum_{i=1}^{n}(y_i - \hat{R}x_i)^2$$

$$= \frac{f}{n}(s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}).$$

.

Confidence interval of ratio estimator

If the sample is large so that the normal approximation is applicable, then the 100(1-α)% confidence interval of $\bar{Y}$ and R are

$$\left(\hat{\bar{Y}}_R - Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\bar{Y}}_R)},\ \ \hat{\bar{Y}}_R + Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{\bar{Y}}_R)}\right)$$

and

$$\left(\hat{R} - Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{R})},\quad \hat{R} + Z_{\frac{\alpha}{2}}\sqrt{Var(\hat{R})}\ \right)$$

respectively where $Z_{\frac{\alpha}{2}}$ is the normal derivate to be chosen for given value of confidence coefficient $(1-\alpha)$.

If $(\bar{x}, \bar{y})$ follows a bivariate normal distributions, then $(\bar{Y} - R\bar{x})$ is normally distributed. It SRS is followed for drawing the sample, then assuming R is known

$$\frac{\bar{y} - R\bar{x}}{\sqrt{\frac{N-n}{Nn}(s_y^2 + R^2 s_x^2 - 2R\, s_{xy})}}$$

is approximately $N(0,1)$.

This can also be used for finding confidence limits, see Cochran (1997, Chapter 6, page 156) for more details.

Conditions under which the ratio estimate is optimum

The ratio estimate $\hat{\bar{Y}}_R$ is best linear unbiased estimator of $\bar{Y}$ when

(i) the relationship between y_i and x_i is linear passing through origin., i.e.

$$y_i = \beta x_i + e_i,$$

where e_i's are independent with $E(e_i / x_i) = 0$ and β is the slope parameter

(ii) this line is proportional to x_i, i.e.

$$Var(y_i / x_i) = E(e_i^2) = Cx_i$$

where C is constant.

Proof. Consider the linear estimate of β as $\hat{\beta} = \sum_{i=1}^{n} \ell_i y_i$ where $y_i = \beta x_i + e_i$.

Then $\hat{\beta}$ is unbiased of $\bar{Y} = \beta \bar{X}$ as $E(y) = \beta \bar{X} + E(e_i / x_i)$.

If *n* sample values of x_i are kept fixed and then in repeated sampling

$$E(\hat{\beta}) = \sum_{i=1}^{n} \ell_i x_i \beta$$

$$\text{and } Var(\hat{\beta}) = \sum_{i=1}^{n} \ell_i^2 Var(y_i / x_i) = C \sum_{i=1}^{n} \ell_i^2 x_i$$

So $E(\hat{\beta}) = \beta$ when $\sum_{i=1}^{n} \ell_i x_i = 1.$

Consider the minimization of it $Var(y_i / x_i)$ subject to condition for unbiased estimator $\sum_{i=1}^{n} \ell_i x_i = 1$ using Lagrangian function. Thus

$$\varphi = Var(y_i / x_i) - 2\lambda(\sum_{i=1}^{n} \ell_i x_i - 1.)$$

$$= C(\sum_{i=1}^{n} \ell_1^2 x_i - 2\lambda(\sum_{i=1}^{n} \ell_i x_i - 1).$$

Now

$$\frac{\partial \varphi}{\partial \ell_i} = 0 \Rightarrow \ell_i x_i = \lambda x_i,\ i = 1, 2, .., n$$

$$\frac{\partial \varphi}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{n} \ell_i x_i = 1$$

Using $\sum_{i=1}^{n} \ell_i x_i = 1$

or $\sum_{i=1}^{n} \lambda x_i = 1$

or $\lambda = \frac{1}{n\bar{x}}.$

$$\ell_i = \frac{1}{n\bar{x}}$$

and so $\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{n\bar{x}} = \frac{\bar{y}}{\bar{x}}.$

Thus $\hat{\beta}$ is not only superior to $\bar{y}$ but also best in the class of linear and unbiased estimators.

Alternative approach:

This result can alternatively be derived as follows:

The ratio estimator $\hat{R} = \frac{\bar{y}}{\bar{x}}$ is the best linear unbiased estimator of $R = \frac{\bar{Y}}{\bar{x}}$ if the following two conditions a hold:

(i) For fixed x, $E(y) = \beta x$, i.e., the line of regression of y on x is a straight line passing through the origin.

(ii) For fixed x, $Var(x) \propto x$, i.e., $Var(x) = \lambda x$ where λ is constant of proportionality.

Proof: Let $\underline{y} = (y_{1)}, y_2, ..., y_n)'$ and $\underline{x} = (x_1, x_2, ..., x_n)'$ be two vectors of observations on $y's$ and $x's$. Hence for any fixed $\underline{x}$,

$$E(\underline{y}) = \beta \underline{x}$$
$$Var(\underline{y}) = \Omega = \lambda \, \text{diag}(x_1, x_2, ..., x_n)$$

where $\text{diag}(x_1, x_2, ..., x_n)$ is the diagonal matrix with $x_1, x_2, ..., x_n$ as the diagonal elements.

The best linear unbiased estimator of β is obtained by minimizing

$$S^2 = (\underline{y} - \beta \underline{x})' \Omega^{-1} (\underline{y} - \beta \underline{x})$$
$$= \sum_{i=1}^{n} \frac{(y_i - \beta x_i)^2}{\lambda x_i}.$$

Solving

$$\frac{\partial S^2}{\partial \beta} = 0$$
$$\Rightarrow \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) = 0$$

or $\hat{\beta} = \frac{\bar{y}}{\bar{x}} = \hat{R}$.

Thus $\hat{R}$ is the best linear unbiased estimator of R Consequently, $\hat{R}\bar{X} = \hat{\bar{Y}}_R$ is the best linear unbiased estimator of $\bar{Y}$.

Ratio estimator in stratified sampling

Suppose a population of size N is divided into k strata. The objective is to estimate the population mean $\bar{Y}$ using ratio method of estimation.

In such situation, a random sample of size n_i is being drawn from i^{th} strata of size N_i on variable under study Y and auxiliary variable X using SRSWOR.

Let

y_{ij} : j^{th} observation on Y from i^{th} strata

x_{ij} : j^{th} observation on X from i^{th} strata $\quad i = 1, 2, \ldots, k; \quad j = 1, 2, ..., n_i.$

An estimator of $\bar{Y}$ based on the philosophy of stratified sampling can be devised in following two possible ways:

1. Separate ratio estimator

- Employ first the ratio method of estimation separately in each strata and obtain ratio estimator $\hat{\bar{Y}}_{R_i} \; i = 1, 2, .., k$ assuming the stratum mean $\bar{X}_i$ to be known.
- Then combine all the estimates using weighted arithmetic mean.

This gives the separate ratio estimator as

$$\hat{\bar{Y}}_{Rs} = \sum_{i=1}^{k} \frac{N_i \hat{\bar{Y}}_{R_i}}{N}$$

$$= \sum_{i=1}^{k} w_i \hat{\bar{Y}}_{R_i}$$

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

where $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$: sample mean of Y from i^{th} strata

$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$: sample mean of X from i^{th} strata

$\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$: mean of all the units in i^{th} stratum

No assumption is made that the true ratio remains constant from stratum to stratum. It depends on information on each $\bar{X}_i$.

2. Combined ratio estimator:

- Find first the stratum mean of $Y's$ and $X's$ as

$$\bar{y}_{st} = \sum_{i=1}^{k} w_i \bar{y}_i$$

$$\bar{x}_{st} = \sum_{i=1}^{k} w_i \bar{x}_i.$$

- Then define the combine ratio estimator as

$$\hat{\bar{Y}}_{Rc} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X}$$

where $\bar{X}$ is the population mean of X based on all the $N = \sum_{i=1}^{N} N_i$ units. It does not depend on individual stratum units. It does not depend on information on each $\bar{X}_i$ but only on $\bar{X}$.

Properties of separate ratio estimator:

Note that there is an analogy between $\bar{Y} = \sum_{i=1}^{k} w_i \bar{Y}_i$ and $\bar{Y}_{Rs} = \sum_{i=1}^{k} w_i \bar{Y}_{Ri}$.

We already have derived the bias of $\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$ as

$$E(\hat{\bar{Y}}_R) = \bar{Y} + \frac{\bar{Y} f}{n} (C_x^2 - \rho C_X C_Y).$$

So for $\hat{\bar{Y}}_{Ri}$, we can write

$$E(\hat{\bar{Y}}_{Ri}) = \bar{Y}_i + \bar{Y}_i \frac{f_i}{n_i} (C_{ix}^2 - \rho_i C_{iX} C_{iY})$$

where $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$, $\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}$

$$f_i = \frac{N_i - n_i}{N_i}, C_{iy}^2 = \frac{S_{iy}^2}{\bar{Y}_i^2}, \; C_{ix}^2 = \frac{S_{ix}^2}{\bar{X}_i^2},$$

$$S_{iy}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2, \; S_{ix}^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2,$$

ρ_i : correlation coefficient between the observation on X and Y in i^{th} stratum

C_{ix} : coefficient of variation of X values in i^{th} sample.

Thus

$$E(\bar{Y}_{Rs}) = \sum_{i=1}^{k} w_i E(\hat{\bar{Y}}_{Ri})$$

$$= \sum_{i=1}^{k} w_i \left[\bar{Y}_i + \bar{Y}_i \frac{f_i}{n_i} (C_{ix}^2 - \rho_{ix} C_{ix} C_{iy} \right]$$

$$= \bar{Y} + \sum_{i=1}^{k} \frac{w_i \bar{Y}_i f_i}{n_i} (C_{ix}^2 - \rho_i C_{ix} C_{iy})$$

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

$$= \sum_{i=1}^{k} \frac{w_i \bar{Y}_i f_i}{n_i} C_{ix} (C_{ix} - \rho_i C_{iy}).$$

Assuming finite population correction to be approximately 1, $n_i = n/k$ and C_{ix}, C_{iy} and ρ_i are all same for the i^{th} stratum as C_x, C_y and ρ respectively.

$$Bias(\hat{\bar{Y}}_{Rs}) = \frac{k}{n} (C_x^2 - \rho C_x C_y) .$$

Thus the bias is negligible when the sample size within each stratum should be sufficiently large and is unbiased when $C_{ix} = \rho C_{iy}$.

Now we derive the *MSE* of $\hat{\bar{Y}}_{Rs}$. We already have derived the *MSE* of $\hat{\bar{Y}}_R$ earlier as

$$MSE(\hat{\bar{Y}}_R) = \frac{\bar{Y}^2 f}{n} (C_X^2 - C_Y^2 - 2\rho C_x C_y)$$

$$= \frac{f}{n(N-1)} \sum_{i=1}^{N} (Y_i - R X_i)^2$$

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

Thus for i^{th} stratum

$$MSE(\hat{\bar{Y}}_{Ri}) = \frac{f_i}{n_i (N_i - 1)} (C_{iX}^2 - C_{iY}^2 - 2\rho C_{iX} C_{iY})$$

$$= \frac{f_i}{n_i (N_i - 1)} \sum_{i=1}^{N_i} (Y_{ij} - R_i X_{ij})^2$$

and so

$$MSE(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} w_i^2 \, MSE(\hat{\bar{Y}}_{Ri})$$

$$= \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} \bar{Y}_i^2 (C_{iX}^2 + C_{iY}^2 - 2\rho C_{iX} C_{iY}) \right]$$

$$= \sum_{i=1}^{k} \left[w_i^2 \frac{f_i}{n_i (N_i - 1)} \sum_{j=1}^{N_i} (Y_{ij} - R_i X_{ij})^2 \right]$$

An estimate of $MSE(\hat{\bar{Y}}_{Rs})$ can be found by substituting the unbiased estimators of S_{iX}^2, S_{iY}^2 and S_{iY}^2 as s_{ix}^2, s_{iy}^2 and s_{ixy} respectively for i^{th} stratum and $R_i = \bar{Y}_i / \bar{X}_i$ can be estimated by $r_i = \bar{y}_i / \bar{x}_i$.

$$MSE(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 + r_i^2 s_{ix}^2 - 2 r_i s_{ixy}) \right].$$

Also .

$$MSE(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i (n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - r_i x_{ij})^2 \right]$$

Properties of combined ratio estimator:

Here

$$\hat{\bar{Y}}_{RC} = \frac{\sum_{i=1}^{k} w_i \bar{y}_i}{\sum_{i=1}^{k} w_i \bar{x}_i} \bar{X} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} = \hat{R}_c \bar{X}.$$

It is difficult to find the exact expression of bias and mean squared error of $\hat{\bar{Y}}_{Rc}$, so we find their approximate expressions.

Define

$$\varepsilon_1 = \frac{\bar{y}_{st} - \bar{Y}}{\bar{Y}}$$

$$\varepsilon_2 = \frac{\bar{x}_{st} - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_1) = 0$$

$$E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} \frac{w_i^2 S_{iY}^2}{\bar{Y}^2} = \sum_{i=1}^{k} \frac{f_i}{n_i} \frac{w_i^2 S_{iY}^2}{\bar{Y}^2}$$

$$E(\varepsilon_2^2) = \sum_{i=1}^{k} \frac{f_i}{n_i} \frac{w_i^2 S_{iX}^2}{\bar{Y}^2}$$

$$E(\varepsilon_1 \varepsilon_2) = \sum_{i=1}^{k} \frac{f_i}{n_i} \frac{S_{iXY}}{\bar{X}\bar{Y}}.$$

Thus assuming $|\varepsilon_2| < 1$,

$$\hat{\bar{Y}}_{RC} = \frac{(1+\varepsilon_1)\bar{Y}}{(1+\varepsilon_2)\bar{X}} \bar{X}$$
$$= \bar{Y}(1+\varepsilon_1)(1-\varepsilon_2+\varepsilon_2^2-...)$$
$$= \bar{Y}(1+\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2-...)$$

Retaining the terms upto order two due to same reason as in the case of $\hat{\bar{Y}}_R$,

$$\hat{\bar{Y}}_{RC} \simeq \bar{Y}(1+\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2)$$
$$\hat{\bar{Y}}_{RC} - \bar{Y} \simeq (\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2)$$

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

$$Bias(\hat{\bar{Y}}_{Rc}) = E(\hat{\bar{Y}}_{Rc} - \bar{Y})$$
$$\simeq \bar{Y}E(\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2)$$
$$= \bar{Y}E(0-0-\varepsilon_1\varepsilon_2+\varepsilon_2^2)$$
$$= \bar{Y}\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} - \frac{S_{iXY}}{\bar{X}\bar{Y}}\right)\right]$$
$$= \bar{Y}\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} - \frac{\rho_i S_{iX} S_{iY}}{\bar{X}\bar{Y}}\right)\right]$$
$$= \frac{\bar{Y}}{\bar{X}}\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 S_{iX} \left(\frac{S_{iX}}{\bar{X}} - \frac{\rho_i S_{iY}}{\bar{Y}}\right)\right]$$
$$= R\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 S_{iX} \left(C_{iX} - \rho_i C_{iY}\right)\right]$$

where $R = \frac{\bar{Y}}{\bar{X}}$, ρ_i is the correlation coefficient between the observations on Y and X in the i^{th} stratum, C_{ix} and C_{iy} are the coefficients of variation of X and Y respectively is ith stratum.

The mean squared error upto second order of approximation is

$$
\begin{aligned}
MSE(\hat{\bar{Y}}_{Rc}) &= E(\hat{\bar{Y}}_{Rc} - \bar{Y})^2 \\
&\simeq \bar{Y}^2 E(\varepsilon_1 - \varepsilon_2 - \varepsilon_1\varepsilon_2 + \varepsilon_2)^2 \\
&\simeq \bar{Y}^2 E(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2) \\
&= \bar{Y}\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} + \frac{S_{iY}^2}{\bar{Y}^2} - \frac{2S_{iXY}}{\bar{X}\bar{Y}}\right)\right] \\
&= \bar{Y}^2\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} + \frac{S_{iY}^2}{\bar{Y}^2} - \frac{2\rho_i S_{iX}}{\bar{X}}\frac{S_{iY}}{\bar{Y}}\right)\right] \\
&= \frac{\bar{Y}^2}{\bar{Y}^2}\sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 \left(\frac{\bar{Y}^2}{\bar{X}^2} S_{iX}^2 + S_{iY}^2 - 2\rho_i \frac{\bar{Y}}{\bar{X}} S_{iX} S_{iY}\right)\right] \\
&= \sum_{i=1}^{k}\left[\frac{f_i}{n_i} w_i^2 (R^2 S_{iX}^2 + S_{iY}^2 - 2\rho_i R S_{iX} S_{iY})\right].
\end{aligned}
$$

An estimate of $MSE(\bar{Y}_{Rc})$ can be obtained by replacing S_{iX}^2, S_{iY}^2 and S_{iXY} by their unbiased estimators s_{ix}^2, s_{iy}^2 and s_{ixy} respectively whereas $R = \frac{\bar{Y}}{\bar{X}}$ is replaced by $r = \frac{\bar{y}}{\bar{x}}$ as follows: Thus the following estimate is obtained:

$$
\begin{aligned}
MSE(\bar{Y}_{Rc}) &= \bar{Y}^2\sum_{i=1}^{k}\left[\frac{w_1^2 f_i}{n_i}\left(\frac{s_{ix}^2}{\bar{X}^2} + \frac{s_{iy}^2}{\bar{Y}^2} - 2\frac{s_{ixy}}{\bar{X}\bar{Y}}\right)\right] \\
&= \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}\left(r^2 s_{ix}^2 + s_{iy}^2 - 2rs_{ixy}\right)\right]
\end{aligned}
$$

where $\bar{X}$ is known.

Comparison of combined and separate ratio estimators

An obvious question arises that which of the estimates $\hat{Y}_{Rs}$ or $\hat{\bar{Y}}_{Rc}$ is better. So we compare their *MSE*s. Note that the only difference in the term of these *MSE*s is due to the form of ratio estimate. It is

- $R_i = \frac{\bar{y}_i}{\bar{x}_i}$ in $MSE(\hat{\bar{Y}}_{Rs})$
- $\bar{R} = \frac{\bar{Y}}{\bar{X}}$ in $MSE(\hat{\bar{Y}}_{Rc})$.

Thus

$$\Delta = MSE(\hat{\bar{Y}}_{Rc}) - MSE(\hat{\bar{Y}}_{Rs})$$
$$= \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}\left[(R^2 - R_i^2)S_{iX}^2 + 2(R_i - R)\rho_i S_{iX} S_{iY}\right]\right]$$
$$= \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}\left[(R - R_i^2)^2 S_{iX}^2 + 2(R - R_i)(R_i S_{iX}^2 - \rho_i S_{iX} S_{iY})\right]\right].$$

The difference Δ depends on

(i) The magnitude of the difference between the strata ratios (R_i) and population ratio as whole (R).

(ii) The value of $(R_i S_{ix}^2 - \rho_i S_{ix} S_{iy})$ is usually small and vanishes when the regression line of y on x is linear and passes through origin within each stratum. In such a case

$$MSE(\hat{\bar{Y}}_{Rc}) > MSE(\hat{\bar{Y}}_{Rs})$$

but $$Bias(\hat{\bar{Y}}_{Rc}) < Bias(\hat{\bar{Y}}_{Rs}).$$

So unless R_i varies considerably, the use of $\hat{\bar{Y}}_{Rc}$ would provide an estimate of $\bar{Y}$ with negligible bias and precision as good as $\hat{\bar{Y}}_{Rs}$.

- If $R_i \neq R$, $\hat{\bar{Y}}_{Rs}$ can be more precise but bias may be large.
- If $R_i \simeq R$, $\hat{\bar{Y}}_{Rc}$ can be as precise as $\hat{\bar{Y}}_{Rs}$ but its bias will be small. It also does not require knowledge of $\bar{X}_1, \bar{X}_2, ..., \bar{X}_k$.

Ratio estimators with reduced bias:

The ratio type estimators that are unbiased or have smaller bias them $\hat{R}, \hat{\bar{Y}}_R$ or $\hat{Y}_{Rc(tot)}$ are useful in sample surveys . There are several approaches to derive such estimators. We consider here two such approaches:

1. Unbiased ratio – type estimators:

Under SRS, the ratio estimator has form $\frac{\bar{Y}}{\bar{x}}\bar{X}$ to estimate the population mean $\bar{Y}$. As an alternative to this, we consider following as an estimator of population mean

$$\hat{\bar{Y}}_{Ro} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i}{X_i}\right)\bar{X} .$$

Let $R = \frac{Y_i}{X_i},\ i = 1, 2, .., N,$

then

$$\hat{\bar{Y}}_{R0} = \frac{1}{n}\sum_{i=1}^{n} R_i \bar{X}$$
$$= \bar{r}\bar{X}$$

where

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$Bias\ (\hat{\bar{Y}}_{R0}) = E(\hat{\bar{Y}}_{R0}) - \bar{Y}$$
$$= E(\bar{r}\bar{X}) - \bar{Y}$$
$$= E(\bar{r})\bar{X} - \bar{Y}.$$

Since

$$E(\bar{r}) = \frac{1}{n}\sum_{i=1}^{n}(\frac{1}{N}\sum_{i=1}^{N} R_i)$$
$$= \frac{1}{N}\sum_{i=1}^{n} R_i$$
$$= \bar{R}.$$

So $Bias(\hat{\bar{Y}}_{R0}) = \bar{R}\bar{X} - \bar{Y}.$

Using the result that under SRSWOR, $Cov(\bar{x}, \bar{y}) = \frac{N-n}{Nn} S_{XY}$, it also follows that

$$Cov(\bar{r}, \bar{x}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(R_i - \bar{R})(X_i - \bar{X})$$
$$= \frac{N-n}{Nn}\frac{1}{N-1}(\sum_{i=1}^{N} R_i X_i - N\bar{R}\bar{X})$$
$$= \frac{N-n}{Nn}\frac{1}{N-1}(\sum_{i=1}^{N} \frac{Y_i}{X_i} X_i - N\bar{R}\bar{X})$$
$$= \frac{N-n}{Nn}\frac{1}{N-1}(N\bar{Y} - N\bar{R}\bar{X})$$
$$= \frac{N-n}{Nn}\frac{1}{N-1}[-Bias(\hat{\bar{Y}}_{R0})].$$

Thus using the result that in SRSWOR, $Cov(\bar{x}, \bar{y}) = \frac{N-n}{Nn} S_{XY}$, we have

$$Bias(\hat{\bar{Y}}_{Ro}) = -\frac{Nn(N-1)}{N-n} Cov(\bar{r}, \bar{x})$$
$$= -\frac{Nn(N-1)}{N-n} \frac{N-n}{Nn} S_{RX}$$
$$= -(N-1) S_{RX}$$

where $S_{RX} = \frac{1}{N-1} \sum_{i=1}^{N} (R_i - \bar{R})(X_i - \bar{X})$.

The following result helps is obtaining an unbiased estimator of population mean:.

Since under SRSWOR set up,

$$E(s_{xy}) = S_{xy}$$

where $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$,

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

so an unbiased estimator of the bias is $Bias(\hat{\bar{Y}}_{R0}) = -(N-1) S_{RX}$ which is obtained as follows:

$$Bias(\hat{\bar{Y}}_{R0}) = -(N-1) s_{rx}$$
$$= -\frac{N-1}{n-1} \sum_{i=1}^{n} (r_i - \bar{r})(x_i - \bar{x})$$
$$= -\frac{N-1}{n-1} \left(\sum_{i=1}^{n} r_i x_i - n \bar{r}\, \bar{x} \right)$$
$$= -\frac{N-1}{n-1} \left(\sum_{i=1}^{n} \frac{y_i}{x_i} x_i - n \bar{r}\, \bar{x} \right)$$
$$= -\frac{N-1}{n-1} (n\bar{y} - n\bar{r}\, \bar{x}).$$

So

$$Bias(Y_{R0}) = E(\hat{Y}_{R0}) - \bar{Y} = -\frac{n(N-1)}{n-1} (\bar{y} - \bar{r}\, \bar{x}).$$

Thus

$$E\left[\hat{\bar{Y}}_{R0} - Bias(\hat{\bar{Y}}_{R0}) \right] = \bar{Y}$$

or $E\left[\hat{\bar{Y}}_{R0} + \frac{n(N-1)}{n-1} (\bar{y} - \bar{r}\, \bar{x}) \right] = \bar{Y}.$

Thus

$$\hat{\bar{Y}}_{R0} + \frac{n(N-1)}{n-1} (\bar{y} - \bar{r}\, \bar{x}) = \bar{r}\bar{X} + \frac{n(N-1)}{n-1} (\bar{y} - \bar{r}\, \bar{x})$$

is an unbiased estimator of population mean.

2. Jackknife method for obtaining a ratio estimate with lower bias

Jackknife method, is used to get rid of the term of order 1/*n* from the bias of an estimator. Suppose the $E(\hat{R})$ can be expanded after ignoring finite population correction as

$$E(\hat{R}) = R + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots$$

Let $n = mg$ and the sample is divided at random ratio *g* groups, each of size *m*. Then

$$E(g\hat{R}) = gR + \frac{ga_1}{gm} + \frac{ga_2}{g^2m^2} + \ldots$$

$$= gR + \frac{a_1}{m} + \frac{a_2}{gm^2} + \ldots$$

Let $\hat{R}_i^* = \frac{\sum^* y_i}{\sum^* x_i}$ where the $\sum^*$ denote that the summation is on all values of the sample except the i^{th} group. So $\hat{R}_i^*$ is based on a simple random sample of size $m(g - 1)$, so we can express

$$E(\hat{R}_i^*) = R + \frac{a_1}{m(g-1)} + \frac{a_2}{m^2(g-1)^2} + \ldots$$

or

$$E\left[(g-1)\hat{R}_i^*\right] = (g-1)R + \frac{a_1}{m} + \frac{a_2}{m^2(g-1)} + \ldots$$

Thus

$$E\left[g\hat{R} - (g-1)\hat{R}_i^*\right] = R - \frac{a_2}{g(g-1)m^2} + \ldots$$

or

$$E\left[g\hat{R} - (g-1)\hat{R}_i^*\right] = R - \frac{a_2}{n^2}\frac{g}{g-1} + \ldots$$

Hence the bias of $\left[g\hat{R} - (g-1)\hat{R}_i^*\right]$ is of order $\frac{1}{n^2}$.

Now *g* estimates of this form can be obtained, one estimator for each group. Then the jackknife or Quenouille's estimator is the average of these of estimators

$$\hat{R}_Q = g\hat{R} - (g-1)\frac{\sum_{i=1}^{g}\hat{R}_i}{g}.$$

A large sample variance of $\hat{\bar{Y}}_{HR}$ is obtained as follows. We assume n and N are large enough so that

$\frac{n}{n-1} \cong 1$ and make $\tilde{r} \cong \bar{R}$. Then

$$\hat{\bar{Y}}_{HR} \cong (\bar{y} - \bar{R}\bar{x}).$$

Hence, large sample variance of $\hat{\bar{Y}}_{HR}$ is given by

$$Var(\hat{\bar{Y}}_{HR}) = \frac{1-f}{n}\left[S_y^2 + \bar{R}^2 S_x^2 - 2RS_{xy}\right].$$

Product method of estimation:

The ratio estimator is more efficient than the mean of a SRSWOR if $\rho > \frac{1}{2}.\frac{C_x}{C_y}$, if $R > 0,$ which is usually the case. This shows that if auxiliary information is such that $\rho < -\frac{1}{2}\frac{C_x}{C_y},$ then we cannot use the ratio method of estimation to improve the sample mean as an estimator of population mean. So there is need of another type of estimator which also makes use of information on auxiliary variable x. Product estimator is an attempt in this direction.

The product estimator of the population mean $\bar{Y}$ is defined as

$$\hat{\bar{Y}}_P = \frac{\bar{y}\,\bar{x}}{\bar{X}}.$$

We now derive the bias and variance of $\hat{\bar{Y}}_p$.

Let $\varepsilon_0 = \frac{y - Y}{\bar{Y}}$, $\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}$,

(i) Bias of $\hat{\bar{Y}}_p$.

We write $\hat{\bar{Y}}_p$ as

$$\hat{\bar{Y}}_p = \frac{\bar{y}\,\bar{x}}{\bar{X}} = \bar{Y}(1+\varepsilon_0)(1+\varepsilon_1)$$
$$= \bar{Y}(1+\varepsilon_1+\varepsilon_2+\varepsilon_1\varepsilon_0).$$

Taking expectation we obtain bias of $\hat{\bar{Y}}_p$ as

$$Bias(\hat{\bar{Y}}_p) = \frac{1}{\bar{X}} Cov(\bar{y}, \bar{x}) = \frac{f}{n\bar{X}} S_{xy},$$

which shows that bias of $\hat{\bar{Y}}_p$ decreases as n increases. Bias of $\hat{\bar{Y}}_p$ can be estimated by

$$Bias(\hat{\bar{Y}}_p) = \frac{f}{n\bar{X}} s_{xy}.$$

(ii) Variance of $\hat{\bar{Y}}_p$:

Writing $\hat{\bar{Y}}_p$ is terms of ε_0 and ε_1, we find that the variance of the product estimator $\hat{\bar{Y}}_p$ upto second order of approximation is given by

$$\begin{aligned} Var(\hat{\bar{Y}}_p) &= E(\hat{\bar{Y}}_p - \bar{Y})^2 \\ &= \bar{Y}^2 E(\varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2)^2 \\ &= \bar{Y}^2 E(\varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_1\varepsilon_2). \end{aligned}$$

Here terms in $(\varepsilon_1, \varepsilon_2)$ of degrees greater than two are assumed to be negligible. Using we find that

$$Var(\hat{\bar{Y}}_p) = \frac{f}{n}\left[S_Y^2 + R^2 S_X^2 + 2RS_{XY} \right].$$

(iii) Estimation of variance of $\hat{\bar{Y}}_p$

The variance of $\hat{\bar{Y}}_p$ can be estimated by

$$Var(\hat{\bar{Y}}_p) = \frac{f}{n}\left[s_y^2 + r^2 s_x^2 + 2rs_{xy} \right]$$

where $r = \bar{y} / \bar{x}$.

(iv) **Comparison with SRSWOR**:

From the variances of the mean of SRSWOR and the product estimator, we obtain

$$Var(\bar{y})_{SRS} - Var(\hat{\bar{Y}}_p) = -\frac{f}{n} RS_X (2\rho S_Y + RS_X),$$

which shows that $\hat{\bar{Y}}_p$ is more efficient than the simple mean $\bar{y}$ for

$$\rho < -\frac{1}{2}\frac{C_x}{C_y} \quad \text{if } R > 0$$

and for

$$\rho > -\frac{1}{2}\frac{C_x}{C_y} \quad \text{if } R < 0.$$

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 bias of $\hat{\bar{Y}}_{Ri}$ as

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

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

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

(ii) Variance of the multivariate ratio estimator:

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

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

Chapter 6

Regression Method of Estimation

The ratio method of estimation uses the auxiliary information which is correlated with the study variable to improve the precision which results in improved estimators when the regression of *y* on *x* if is linear and passes through origin. When the regression of if on *X* is linear, it is not necessary that the line should always pass through origin. Under such conditions, it is more appropriate to use the regression type estimators.

In ratio method, the conventional estimator sample mean $\bar{y}$ was improved by multiplying it by a a factor $\frac{\bar{X}}{\bar{x}}$ where $\bar{x}$ is an unbiased estimator of population mean $\bar{X}$ which is chosen as population mean of auxiliary variable. Now we consider another idea based on difference.

Consider an estimator $(\bar{x} - \bar{X})$ for which $E(\bar{x} - \bar{X}) = 0.$

Consider an improved estimator of $\bar{Y}$ as

$$\hat{\bar{Y}}^* = \bar{y} + \mu(\bar{x} - \bar{X})$$

which is an unbiased estimator of $\bar{Y}$ and μ is any constant. Now find μ such that the $Var(\hat{\bar{Y}}^*)$ is minimum

$$Var(\hat{\bar{Y}}^*) = Var(\bar{y}) + \mu^2 Var(\bar{x}) + 2\mu Cov(\bar{x}, \bar{y})$$

$$\frac{\partial Var(\bar{Y}^*)}{\partial \lambda} = 0$$

$$\Rightarrow \mu = -\frac{Cov(\bar{x}, \bar{y})}{Var(\bar{x})}$$

$$= \frac{\frac{N-n}{Nn} S_{XY}}{\frac{N-n}{Nn} S_X^2}$$

$$= \frac{S_{XY}}{S_X^2}$$

where $S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{x})(Y_i - \bar{y}), \quad S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}).$

Note that the value of regression coefficient β is a linear regression model $y = x\beta + e$ of y on x obtained by minimizing $\sum_{i=1}^{n} e_i^2$ based on n data sets $(x_i, y_i),\ i = 1,2,..,n$ is $\beta = \frac{Cov(x,y)}{Var(x)} = \frac{S_{xy}}{S_x^2}$. Thus the optimum value of μ is same as the regression coefficient of y on x with a negative sign, *i.e.*,

$$\mu = -\beta .$$

So the estimator $\hat{\bar{Y}}^*$ with optimum value of μ is

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

which is the regression estimator of $\bar{Y}$ and the procedure of estimation is regression method of estimation. The variance of $\hat{\bar{Y}}_{reg}$ *is*

$$Var(\hat{\bar{Y}}_{reg}) = V(\bar{y})[1 - \rho^2(\bar{x}, \bar{y})]$$

where $\rho(\bar{x}, \bar{y})$ is the correlation coefficient between $\bar{x}$ and $\bar{y}$. So $\hat{\bar{Y}}_{reg}$ would be efficient of $\bar{x}$ and $\bar{y}$ are highly correlated. The estimator $\hat{\bar{Y}}_{reg}$ is more efficient than $\bar{Y}$ of $\rho(\bar{x}, \bar{y}) \neq 0$ which generally holds.

Regression estimates with preassigned β :

If value of β is known as β_0, say then the regression estimator is

$$\hat{\bar{Y}}_{reg} = \bar{y} + \beta_0(\bar{X} - \bar{x}) .$$

Bias of $\bar{Y}_{reg}$:

Now, assuming that the random sample $(x_i, y_i),\ i = 1,2,..,n$ is drawn by SRSWOR,

$$\begin{aligned} E(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + \beta_0\left[\bar{X} - E(\bar{x})\right] \\ &= \bar{Y} + \beta_0\left[\bar{X} - \bar{X})\right] \\ &= \bar{Y} \end{aligned}$$

Thus $\hat{\bar{Y}}_{reg}$ is an unbiased estimator of $\bar{Y}$ when β is known.

Variance of $\hat{\bar{Y}}_{reg}$

$$
\begin{aligned}
Var(\hat{\bar{Y}}_{reg}) &= E\left[\hat{\bar{Y}}_{reg} - E(\hat{\bar{Y}}_{reg})\right]^2 \\
&= E\left[\bar{y} + \beta_0(\bar{X} - \bar{x}) - \bar{Y}\right]^2 \\
&= E\left[(\bar{y} - \bar{Y}) - \beta_0(\bar{x} - \bar{X})\right]^2 \\
&= E\left[(\bar{y} - \bar{Y})^2 + \beta_0^2(\bar{x} - \bar{X}) - 2\beta_0 E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})\right] \\
&= Var(\bar{y}) + \beta_0^2 Var(\bar{x}) - 2\beta_0 Cov(\bar{x}, \bar{y}) \\
&= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_{XY}\right] \\
&= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y\right]
\end{aligned}
$$

where

$$f = \frac{N-n}{N}$$

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

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

ρ : Correlation coefficient between X and Y.

Comparing the $Var(\hat{\bar{Y}}_{reg})$ with $Var(\bar{y})$, we note that

$$Var(\hat{\bar{Y}}_{reg}) < Var(\bar{y}).$$

If $\beta_0^2 S_X^2 - 2\beta_0 S_{XY} < 0$

or $\beta_0 S_X^2\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) < 0$

which is possible when either $\beta_0 < 0$ and $\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) > 0 \Rightarrow \frac{2S_{XY}}{S_X^2} < \beta_0 < 0$.

or $\beta_0 > 0$ and $\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) < 0 \Rightarrow 0 < \beta_0 < \frac{2S_{XY}}{S_X^2}\ \beta_0 < \frac{2S_{XY}}{S_X^2}$.

Optimal value of β

Choose β such that $Var(\hat{\bar{Y}}_{reg})$ is minimum .

So

$$\frac{\partial Var(\hat{\bar{Y}}_{reg})}{\partial \beta} = \frac{\partial}{\partial \beta}\left[S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y \right] = 0$$

$$\Rightarrow \beta = \rho \frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}.$$

The minimum value of variance of $\hat{\bar{Y}}_{reg}$ with optimum value of $\beta_{opt} = \frac{\rho S_Y}{S_X}$ is

$$Var_{min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \rho^2 \frac{S_Y^2}{S_X^2} S_X^2 - 2\rho \frac{S_Y}{S_X} \rho S_X S_Y \right]$$

$$= \frac{f}{n} S_Y^2 (1-\rho^2).$$

Since $-1 \le \rho \le 1,$ so

$$Var(\hat{\bar{Y}}_{reg}) \le Var_{SRS}(\bar{y})$$

which always holds true.

Departure from β :

If β_0 is the preassigned value of regression coefficient, then

$$Var_{\min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_Y^2 - 2\beta_0 \rho S_X S_Y \right]$$

$$= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\rho\beta_0 S_X S_Y - \rho^2 S_Y^2 + \rho^2 S_Y^2 \right]$$

$$= \frac{f}{n}\left[(1-\rho^2) S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_X \beta_{opt} + \beta_{opt}^2 S_X^2 \right]$$

$$= \frac{f}{n}\left[(1-\rho^2) S_Y^2 + (\beta_0 - \beta_{opt})^2 S_X^2 \right]$$

where $\beta_{opt} = \frac{\rho S_Y}{S_X}$.

Estimate of variance

An unbiased sample estimate of $Var(\hat{\bar{Y}}_{reg})$ is

$$Var(\hat{\bar{Y}}_{reg}) = \frac{f}{n(n-1)}\sum_{i=1}^{n}\left[(y_i-\bar{y})-\beta_0(x_i-\bar{x})\right]^2$$
$$= \frac{f}{n}\sum_{i=1}^{n}(s_y^2+\beta_0^2 s_x^2-2\beta_0 s_{xy}).$$

Regression estimates when β is computed from sample

Suppose a random sample of size n, (x_i, y_i), $i=1,2,..,n$ is drawn by SRSWOR. When β is unknown, it is estimated as

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

and then the regression estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{\beta}(\bar{X}-\bar{x}).$$

It is difficult to find the exact expressions of $E(\bar{Y}_{reg})$ and $Var(\hat{\bar{Y}}_{reg})$. So we approximate then using the same methodology as in the case of ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = \bar{Y}(1+\varepsilon_0)$$
$$\varepsilon_1 = \frac{\bar{x}-\bar{X}}{\bar{x}} \Rightarrow \bar{x} = \bar{X}(1+\varepsilon_1)$$
$$\varepsilon_2 = \frac{s_{xy}-S_{XY}}{S_{XY}} \Rightarrow s_{xy} = S_{XY}(1+\varepsilon_2)$$
$$\varepsilon_3 = \frac{s_x^2-S_x^2}{S_x^2} \Rightarrow s_x^2 = S_X^2(1+\varepsilon_3)$$

Then

$$E(\varepsilon_0) = 0$$
$$E(\varepsilon_1) = 0$$
$$E(\varepsilon_2) = 0$$
$$E(\varepsilon_3) = 0$$
$$E(\varepsilon_0^2) = \frac{f}{n} C_Y^2$$
$$E(\varepsilon_1^2) = \frac{f}{n} C_X^2$$
$$E(\varepsilon_0 \varepsilon_1) = \frac{f}{n} \rho C_X C_Y$$

and

$$\bar{Y}_{reg} = \bar{y} + \frac{s_{xy}}{s_x^2}(\bar{X} - \bar{x})$$
$$= \bar{Y}(1+\varepsilon_0) + \frac{s_{xy}(1+\varepsilon_2)}{s_x^2(1+\varepsilon_3)}(-\varepsilon_1 \bar{X})$$

The estimation error of $\hat{\bar{Y}}_{reg}$ is

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta \bar{X} \varepsilon_1 (1+\varepsilon_2)(1+\varepsilon_3)^{-1}$$

where $\beta = \frac{S_{XY}}{S_X^2}$ is the population regression coefficient.

Assuming $|\varepsilon_3| < 1$,

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) = \bar{Y}\varepsilon_0 - \beta \bar{X}(\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_3^2 - ....)$$

Retaining the terms upto second power of $\varepsilon's$ and ignoring other terms, we have

$$(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq \bar{Y}\varepsilon_0 - \beta \bar{X} \varepsilon_1 (\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_{3)}^2)$$
$$\simeq \bar{Y}\varepsilon_0 - \beta \bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)$$

Bias of $\hat{\bar{Y}}_{reg}$

Now the bias of $\hat{\bar{Y}}_{reg}$ is

$$E(\hat{\bar{Y}}_{reg} - \bar{Y}) \simeq E\left[\bar{Y}\varepsilon_0 - \beta \bar{X} \varepsilon_1 (\varepsilon_1 + \varepsilon_1\varepsilon_2)(1 - \varepsilon_3 + \varepsilon_{3)}^2)\right]$$
$$= -\frac{\beta \bar{X} f}{n}\left[\frac{\mu_{21}}{\bar{X} S_{XY}} - \frac{\mu_{30}}{\bar{X} S_X^2}\right]$$

where $f = \frac{N-n}{N}$, $(r, s)^{\text{th}}$ cross product moment is

$$\mu_{rs} = E\left[(x-\bar{X})^r (y-\bar{Y})^s\right]$$

so

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

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

So

$$Bias(\hat{\bar{Y}}_{reg}) = -\frac{\beta f}{n}\left[\frac{\mu_{21}}{S_{XY}} - \frac{\mu_{30}}{S_X^2}\right].$$

Also,

$$\begin{aligned} Bias(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + E[\hat{\beta}(\bar{X} - \bar{x})] \\ &= \bar{Y} + \bar{X}E(\hat{\beta}) - E(\hat{\beta}\bar{x}) \\ &= \bar{Y} + E(\bar{x})E(\hat{\beta}) - E(\hat{\beta}\bar{x}) \\ &= \bar{Y} - Cov(\hat{\beta}, \bar{x}) \end{aligned}$$

$$Bias(\hat{\bar{Y}}_{reg}) = E(\hat{\bar{Y}}_{reg}) - \bar{Y} = -Cov(\hat{\beta}, \bar{x})$$

MSE of $\hat{\bar{Y}}_{reg}$

To obtain the MSE of $\hat{\bar{Y}}_{reg}$, consider

$$E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \approx E\left[\varepsilon_0 \bar{Y} - \beta \bar{X}(\varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_2)\right]^2$$

Retaining the terms of $\varepsilon's$ upto the second power second and ignoring others, we have

$$\begin{aligned} E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 &\approx E\left[\varepsilon_0^2 \bar{Y}^2 + \beta^2 \bar{X}^2 \varepsilon_1^2 - 2\beta \bar{X}\bar{Y}\varepsilon_0\varepsilon_1\right] \\ &= \bar{Y}^2 E(\varepsilon_0) + \beta^2 \bar{X}^2 E(\varepsilon_1^2) - 2\beta \bar{X}\bar{Y}E(\varepsilon_0\varepsilon_1) \\ &= \frac{f}{n}\left[\bar{Y}^2 \frac{S_Y^2}{\bar{Y}^2} + \beta^2 \bar{X}^2 \frac{S_X^2}{\bar{X}^2} - 2\beta \bar{X}\bar{Y}\rho \frac{S_X S_Y}{\bar{X}\bar{Y}}\right] \end{aligned}$$

$$\begin{aligned} MSE(\hat{\bar{Y}}_{reg}) &= E(\hat{\bar{Y}}_{reg} - \bar{Y})^2 \\ &= \frac{f}{n}(S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y) \end{aligned}$$

Since $\beta = \frac{S_{XY}}{S_X^2} = \rho\frac{S_Y}{S_X}$,

so substituting it is MSE$(\hat{\bar{Y}}_{reg})$, we get

$$MSE(\bar{Y}_{reg}) = \frac{f}{n} S_Y^2 (1-\rho^2).$$

So upto second order of approximation, the regression estimator is better than the conventional sample mean estimator under SRSWOR. This is because the regression estimator uses some extra information also. Moreover, such extra information requires some extra cost also. This shows a false superiority in some sense. So the regression estimators and SRS estimates can be combined if cost aspect is also taken into consideration.

Comparison of $\hat{\bar{Y}}_{reg}$ with ratio estimate and SRS sample mean estimate

$$MSE(\hat{\bar{Y}}_{reg}) = \frac{f}{n} S_Y^2 (1-\rho^2)$$

$$MSE(\hat{\bar{Y}}_R) = \frac{f}{n} (S_Y^2 + R^2 S_X^2 - 2\rho R S_X S_Y)$$

$$Var_{SRS}(\bar{y}) = \frac{f}{n} S_Y^2.$$

(i) As $MSE(\hat{\bar{Y}}_{reg}) = Var_{SRS}(\bar{y})(1-\rho^2)$ since $\rho^2 < 1$, so $\hat{\bar{Y}}_{reg}$ is always superior to $\bar{y}$.

(ii) $\hat{\bar{Y}}_{reg}$ is better than $\hat{\bar{Y}}_R$ if $MSE(Y_{reg}) \le MSE(\hat{\bar{Y}}_R)$

or if $\frac{f}{n} S_Y^2 (1-\rho^2) \le \frac{f}{n} (S_Y^2 + R^2 S_X^2 - 2\rho R S_X S_Y)$

or if $(RS_X - \rho S_Y)^2 \ge 0$

which always holds true.

So regression estimate is always superior to ratio estimate upto second order of approximation.

Regression estimates in stratified sampling

Under the set up of stratified sampling, let the population of N sampling units is divided into k strata. The strata sizes are $N_1, N_2, .., N_k$ such that $\sum_{i=1}^{k} N_i = N$. A sample of size n_i on (x_{ij}, y_{ij}), $j = 1, 2, .., n_i$, is drawn from i^{th} strata (i = 1,2,..,k) by SRSWOR where x_{ij} and y_{ij} denotes the j^{th} unit from i^{th} strata on auxiliary and study variables, respectively.

In order to estimate the population mean, there are two approaches.

1. Separate regression estimator

- Estimate regression estimator

$$\hat{\bar{Y}}_{reg} = \bar{y} + \beta_0(\bar{X} - \bar{x})$$

from each stratum separately i.e., the regression estimate in the i^{th} stratum is

$$\hat{\bar{Y}}_{reg(i)} = \bar{y}_i + \beta_i(\bar{X}_i - \bar{x}_i).$$

- Find the stratified mean as the weighted mean of $\hat{\bar{Y}}_{reg(i)}$ $i = 1, 2, .., k$ as

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} \frac{N_i \hat{\bar{Y}}_{reg(i)}}{N}$$

$$= \sum_{i=1}^{k} [w_i \{\bar{y}_i + \beta_i(\bar{X}_i - \bar{x}_i)\}]$$

where $\beta_i = \frac{S_{ixy}}{S_{ix}^2}$, $w_i = \frac{N_i}{N}$.

In this approach , the regression estimator is separately obtained in each stratum and then combined using the philosophy of stratified sample. So $\hat{\bar{Y}}_{sreg}$ is termed as separate regression estimator,

2. Combined regression estimator

Another strategy is to estimate $\bar{x}$ and $\bar{y}$ in the $\hat{\bar{Y}}_{reg}$ as respective stratified mean. Replacing $\bar{x}$ by $\bar{x}_{st} = \sum_{i=1}^{k} w_i \bar{x}_i$ and $\bar{y}$ by $\bar{y}_{st} = \sum_{i=1}^{k} w_i \bar{y}_i$, we have

$$\hat{\bar{Y}}_{creg} = \bar{y}_{st} + \beta(\bar{X} - \bar{x}_{st}).$$

In this case, all the sample information is combined first and then implemented in regression estimator, so $\hat{\bar{Y}}_{reg}$ is termed as combined regression estimator.

Properties of separate and combined regression

In order to derive the mean and variance $\hat{\bar{Y}}_{sreg}$ and $\hat{\bar{Y}}_{creg}$, there are two cases

- when β is preassigned as β_0
- when β is estimated from the sample.

We consider here the case that β is preassigned as β_0. Other case when β is estimated as $\beta = \frac{s_{xy}}{s_x^2}$ can be dealt using the same approach based on defining various $\varepsilon's$ and using the approximation theory as in the case of $\hat{\bar{Y}}_{reg}$.

1. Separate regression estimator

Assume β is known, say β_0. Then

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} w_i[\bar{y}_i + \beta_{0i}(\bar{X}_i - \bar{x}_i)$$

$$\hat{\bar{Y}}_{sreg} = \sum_{i=1}^{k} w_i[E(\bar{y}_i) + \beta_{0i}(\bar{X}_i - E(\bar{x}_i)]$$

$$= \sum_{i=1}^{k} w_i[\bar{Y}_i + (\bar{X}_i - \bar{X}_i)]$$

$$= \bar{Y}.$$

$$Var(\hat{\bar{Y}}_{sreg}) = E\left[\hat{\bar{Y}}_{sreg} - E(\hat{\bar{Y}}_{sreg})\right]^2$$

$$= \sum_{i=1}^{k} w_i \bar{y}_i + _i + \sum_{i=1}^{k} w_i \beta_{0i}(\bar{X}_i - \bar{x}_i) - \bar{Y}]^2$$

$$= \sum_{i=1}^{k} w_i(\bar{y}_i - \bar{Y}) - \sum_{i=1}^{k} w_i \beta_{0i}(\bar{x}_i - \bar{X}_i)]^2$$

$$= \sum_{i=1}^{k} w_i^2(\bar{y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{k} w_i^2\ \beta_{0i}E(\bar{x}_i - \bar{X}_i)]^2 - \sum_{i=1}^{k} w_i^2 \beta_{0i}E(\bar{x}_i - \bar{X}_i)(\bar{y}_i - \bar{Y}_i)$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{Y}_i) + \sum_{i=1}^{k} w_i^2 \beta_{0i} Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2 \beta_{0i} Cov(\bar{x}_i, \bar{y}_i)$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i}(S_{iY}^2 + \beta_{oi}^2 S_{iX}^2 - 2\beta_{0i} S_{iXY})]$$

$Var(\hat{\bar{Y}}_{sreg})$ is minimum when $\beta_{0i} = \frac{S_{iXY}}{S_{iX}^2}$ and so substituting β_{0i}, we have

$$V_{\min}(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(S_{iY}^2 - \beta_{0i}^2 S_{iX}^2)\right]$$

where $f_i = \frac{N_i - n_i}{N_i}$.

Since SRSWOR is followed is drawing the samples from each stratum, so

$$E(s_{ix}^2) = S_{iX}^2$$
$$E(s_{iy}^2) = S_{iY}^2$$
$$E(s_{ixy}) = S_{iXY}$$

Thus an unbiased estimator of variance can be obtained by replacing S_{iX}^2 and S_{iY}^2 by their respective unbiased estimators s_{ix}^2 and s_{iy}^2 respectively as

$$Var(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 + \beta_{oi}^2 s_{ix}^2 - 2\beta_{0i} s_{ixy})\right]$$

and

$$Var_{\min}(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}(s_{iy}^2 - \beta_{oi}^2 s_{ix}^2)\right]$$

2. Combined regression estimator:

Assume β is known as β_0. Then

$$\hat{\bar{Y}}_{creg} = \sum_{i=1}^{k} w_i \bar{y}_i + \beta_0(\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i)$$

$$\hat{\bar{Y}}_{creg} = \sum_{i=1}^{k} w_i E(\bar{y}_i) + \beta_0[\bar{X} - \sum_{i=1}^{k} w_i E(\bar{x}_i)]$$

$$= \sum_{i=1}^{k} w_i \bar{Y}_i + \beta_0[\bar{X} - \sum_{i=1}^{k} w_i \bar{X}_i]$$

$$= \bar{Y} + \beta_0(\bar{X} - \bar{X})$$

$$= \bar{Y}.$$

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

$$Var(\hat{\bar{Y}}_{creg}) = E[\bar{Y}_{creg} - E(\bar{Y}_{creg})]^2$$

$$= E[\sum_{i=1}^{k} w_i \bar{y}_i + \beta_0[\bar{X} - \sum_{i=1}^{k} w_i \bar{x}_i) - \bar{Y}]^2$$

$$= E[\sum_{i=1}^{k} w_i(\bar{y}_i - \bar{Y}) -_i \beta_0 \sum_{i=1}^{k} w_i(\bar{x}_i - \bar{X}_i)]^2$$

$$= \sum_{i=1}^{k} w_i^2 Var(\bar{y}_i) + \beta_o^2\left[\sum_{i=1}^{k} w_i^2 Var(\bar{x}_i) - 2\sum_{i=1}^{k} w_i^2 \beta Cov(\bar{x}_i, \bar{y}_i)\right]$$

$$= \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i}\left[S_{iY}^2 + \beta_o^2 S_{iX}^2 - 2\beta_0 S_{iXY}\right]^2.$$

$Var(\hat{\bar{Y}}_{creg})$ is minimum when

$$\beta_0 = \frac{Cov(\bar{x}_{st}, \bar{y}_{st})}{Var(\bar{x}_{st})}$$
$$= \frac{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iXY}}{\sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} S_{iX}^2}$$

and the minimum variance is given by

$$Var_{\min}(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} (S_{iY}^2 - \beta_0^2 S_{iX}^2)$$

Since SRSWOR is followed to draw the sample from strata, so using

$$Var(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 + \beta_o^2 s_{ix}^2 - 2\beta_{0i} s_{ixy}) \right]$$

and

$$Var_{\min}(\hat{\bar{Y}}_{creg}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 - \beta_{oi}^2 s_{ix}^2) \right]$$

Comparison of $\hat{\bar{Y}}_{sreg}$ and $\hat{\bar{Y}}_{creg}$:

Note that

$$Var(\hat{\bar{Y}}_{creg}) - Var(\hat{\bar{Y}}_{sreg}) = \sum_{i=1}^{k} (\beta_{io} - \beta_0^2) \frac{w_i^2 f_i}{n_i} S_{iX}^2$$
$$= \sum_{i=1}^{k} \frac{f_i}{n_i} (\beta_{io} - \beta_0)^2 w_i^2 S_{iX}^2$$
$$\geq 0.$$

So if regression line of y on x is approximately linear and the regression coefficient do not vary much among strata, then separate regression estimate is more efficient is more efficient than combined regression estimator.

Chapter 7
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 $= \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{X_i}{M}\right)$

$$= \frac{1}{N}\left(N - \frac{N\bar{X}}{M}\right)$$
$$= 1 - \frac{\bar{X}}{M} = Q, \text{ say.}$$

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

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

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 $= 1 - \frac{\bar{X}}{M}$.

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

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

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, $i = 1, 2,\ldots,N$.

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

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

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

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

$$E(z_r) = \sum_{i=1}^{N} Z_i P_i$$

$$= \sum_{i=1}^{N} \frac{Y_i}{NP_i} P_i$$

$$= \bar{Y}.$$

Thus

$$E(\bar{z}) = \frac{1}{n} \sum_{i=1}^{n} E(z_r)$$

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

$$= \bar{Y}.$$

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

The variance of $\bar{z}$ is

$$Var(\bar{z}) = \frac{1}{n^2} Var\left(\sum_{r=1}^{n} z_r \right)$$

$$= \frac{1}{n^2} \sum_{r=1}^{n} Var(z_r) \qquad (z_r' s \text{ are independent in WR case}).$$

Now

$$Var(z_r) = E\left[z_r - E(z_r) \right]^2$$

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

$$= \sum_{i=1}^{N} \left(Z_i - \bar{Y} \right)^2 P_i$$

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

$$= \sigma_z^2 \quad \text{(say)} .$$

Thus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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
  ⋮
- U_{i-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
  ⋮
- 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 $P_i = \frac{1}{N}$.

$P_{i(2)}$ 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 $\bar{Y}$. In general, $\frac{y_i}{Np_i}(i \neq 1)$ will not provide an unbiased estimator of $\bar{Y}$.

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

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

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

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

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

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

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

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

$$= \bar{Y}$$

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

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

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

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

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

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

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

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

$= \bar{Y}.$

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

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

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

$= \dfrac{1}{4N^2}E\left[\dfrac{y_1(1+p_1)}{p_1} + \dfrac{y_2(1-p_1)}{p_2}\right]^2 - \bar{Y}^2$

$\qquad\qquad \downarrow \qquad\qquad \downarrow$

nature of variable depends only on 1^{st} draw	nature of variable depends upon 1^{st} and 2^{nd} draw

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

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

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

Using the property

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

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

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

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

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

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

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

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

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

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

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

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

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

↓ variance of WR case for $n = 2$

↓ reduction of variance in WR with varying probability

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

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

Since

$$E(z_1 z_2) = E\left[z_1 E(z_2 \mid u_1)\right]$$
$$= E\left[z_1 \bar{Y}\right]$$
$$= \bar{Y} E(z_1)$$
$$= \bar{Y}^2.$$

Consider

$$E\left[\bar{z}^2 - z_1 z_2\right] = E(\bar{z}^2) - E(z_1 z_2)$$
$$= E(\bar{z}^2) - \bar{Y}^2$$
$$= Var(\bar{z})$$

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

Alternative form

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

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

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

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

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

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

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

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

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

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

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[y_1 + y_2 + ... + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - ... - p_{r-1}) \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \sum_{j=1}^{N} {}^{*} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \sum_{j=1(\neq i(1), i(2), ..., i(r-1))}^{N} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} + \left\{ Y_{tot} - \left(Y_{i(1)} + Y_{i(2)} + ... + Y_{i(r-1)} \right) \right\} \right]$$

$$= \frac{1}{N} E_1 \left[Y_{tot} \right]$$

$$= \frac{Y_{tot}}{N}$$

$$= \bar{Y} \quad \text{for all} \quad r = 1, 2, ..., n.$$

Then

$$E(\bar{z}) = \frac{1}{n} \sum_{r=1}^{n} E(z_r)$$

$$= \frac{1}{n} \sum_{r=1}^{n} \bar{Y}$$

$$= \bar{Y}.$$

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

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

Estimate of variance:

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

Consider for $r < s$,

$$E(z_r z_s) = E\left[z_r E(z_s \mid U_1, U_2, ..., U_{s-1}) \right]$$

$$= E\left[z_r \bar{Y} \right]$$

$$= \bar{Y} E(z_r)$$

$$= \bar{Y}^2$$

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

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

Further, for $s < r$,

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

Consider

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

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

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

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

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

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

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

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 $p_s = \sum_{i=1}^{M(=n!)} p_{sio} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$

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) $Var(\hat{\theta}_u) \le Var(\hat{\theta}_0)$

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 = $n!\binom{N}{n}$

$$(i)\ E(\hat{\theta}_0) = \sum_{s=1}^{\binom{N}{n}} \sum_{i=1}^{M} z_{si} p_{si}$$

$$E(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}} \left(\sum_{i=1}^{M} z_{si} p'_{si} \right) p_s$$

$$= \sum_{s} \left(\sum_{i} z_{si} \frac{p_{si}}{p_s} \right) p_s$$

$$= \sum_{s} \sum_{i} z_{si} p_{si}$$

$$= E(\hat{\theta}_0)$$

(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

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p_{si}^{'})^2 \, p_{si}\right]$$

$$\widehat{Var}(\hat{\theta}_u) = \widehat{Var}(\hat{\theta}_0) - \sum_s \sum_i \left[\widehat{(z_{si} - \sum_i z_{si} p_{si}^{'})^2 \, p_{si}}\right]$$

$$= \sum_i p_{si}^{'} \widehat{Var}(\hat{\theta}_0) - \sum_i p_{si}^{'} \widehat{(z_{si} - \sum_i z_{si} p_{si}^{'})^2}.$$

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

$$\overline{z}(s_1^*) = \frac{1}{2N}\left[(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right]$$

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

$$\overline{z}(s_2^*) = \frac{1}{2N}\left[(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right]$$

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 $\overline{z}(s_1^*)$ and $\overline{z}(s_2^*)$ are

$$p(s_1^*) = \frac{p_i p_j}{1-p_i}$$

$$p(s_2^*) = \frac{p_j p_i}{1-p_j}$$

$$p(s) = p(s_1^*) + p(s_2^*)$$

$$= \frac{p_i p_j (2-p_i-p_j)}{(1-p_i)(1-p_j)}$$

$$p'(s_1^*) = \frac{1-p_j}{2-p_i-p_j}$$

$$p'(s_2^*) = \frac{1-p_i}{2-p_i-p_j}.$$

Murthy's unordered estimate $\bar{z}(u)$ corresponding to the Des Raj's ordered estimate is given as

$$\bar{z}(u) = \bar{z}(s_1^*)p'(s_1) + \bar{z}(s_2^*)p'(s_2)$$

$$= \frac{\bar{z}(s_1^*)p(s_1^*) + \bar{z}(s_2^*)p(s_2^*)}{p(s_1^*) + p(s_2^*)}$$

$$= \frac{\left[\frac{1}{2N}\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}\left(\frac{p_i p_j}{1-p_i}\right)\right] + \left[\frac{1}{2N}\left\{(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right\}\left(\frac{p_j p_i}{1-p_j}\right)\right]}{\frac{p_i p_j}{1-p_i} + \frac{p_j p_i}{1-p_j}}$$

$$= \frac{\frac{1}{2N}\left[\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}(1-p_j) + \left\{(1+p_j)\frac{y_j}{p_i} + (1-p_j)\frac{y_i}{p_i}\right\}(1-p_i)\right]}{(1-p_j)+(1-p_i)}$$

$$= \frac{\frac{1}{2N}\left[(1-p_j)\frac{y_i}{p_i}\{(1+p_i)+(1-p_i)\} + (1-p_i)\frac{y_j}{p_j}\{(1-p_j)+(1+p_j)\}\right]}{2-p_i-p_j}$$

$$= \frac{(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}}{N(2-p_i-p_j)}.$$

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

$$\alpha_i = \begin{cases} 1 \text{ if } Y_i \text{ is included in a sample 's' of size } n \\ 0 \text{ otherwise.} \end{cases}$$

Let $z_i = \frac{ny_i}{NE(\alpha_i)}$, $i = 1...N$ assuming $E(\alpha_i) > 0$ for all i

where

$$E(\alpha_i) = 1.P(Y_i \in s) + 0.P(Y_i \notin s)$$
$$= \pi_i$$

is the probability of including the unit *i* in the sample and is called as **inclusion probability.**

The HT estimator of $\bar{Y}$ based on $y_1, y_2, ..., y_n$ is

$$\bar{z}_n = \hat{\bar{Y}}_{HT} = \frac{1}{n}\sum_{i=1}^{n} z_i$$
$$= \frac{1}{n}\sum_{i=1}^{N} \alpha_i z_i.$$

Unbiasedness

$$E(\hat{\bar{Y}}_{HT}) = \frac{1}{n}\sum_{i=1}^{N} E(z_i\alpha_i)$$
$$= \frac{1}{n}\sum_{i=1}^{N} z_i E(\alpha_i)$$
$$= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{NE(\alpha_i)} E(\alpha_i)$$
$$= \frac{1}{n}\sum_{i=1}^{N} \frac{ny_i}{N} = \bar{Y}$$

which shows that HT estimator is an unbiased estimator of population mean.

Variance

$$V(\hat{\bar{Y}}_{HT}) = V(\bar{z}_n)$$
$$= E(\bar{z}_n^2) - \left[E(\bar{z}_n)\right]^2$$
$$= E(\bar{z}_n^2) - \bar{Y}^2.$$

Consider

$$E(\bar{z}_n^2) = \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i z_i\right]^2$$
$$= \frac{1}{n^2} E\left[\sum_{i=1}^{N} \alpha_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j z_i z_j\right]$$
$$= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2 E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_i z_j E(\alpha_i\alpha_j)\right].$$

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

$$E(\alpha_i) = E(\alpha_i^2)$$

$$E(\bar{z}_n^2) = \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\pi_i + \sum_{i(\#j)}^{N}\sum_{i=1}^{N}\pi_{ij}z_iz_j\right]$$

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

$$\begin{aligned}\bar{Y}^2 &= \left[E(\bar{z}_n)\right]^2\\ &= \frac{1}{n^2}\left[E\left(\sum_{i=1}^{N}\alpha_i z_i\right)\right]^2\\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\left[E(\alpha_i)\right]^2\right] + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N} z_iz_jE(\alpha_i)E(\alpha_j)\\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} z_i^2\pi_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_i\pi_jz_iz_j\right].\end{aligned}$$

Thus

$$\begin{aligned}Var(\hat{\bar{Y}}_{HT}) &= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_{ij}z_iz_j\right]\\ &\quad -\frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i^2 z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\pi_i\pi_j z_iz_j\right]\\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_iz_j\right]\\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)\frac{n^2y_i^2}{N^2\pi_i^2} + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)\frac{n^2y_iy_j}{N^2\pi_i\pi_j}\right]\\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)y_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_i}{\pi_i\pi_j}\right)y_iy_j\right]\end{aligned}$$

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 $E(\alpha_i\alpha_j) = P(\alpha_i = 1, \alpha_j = 1)$

$$= P(\alpha_i = 1)P(\alpha_j = 1 \mid \alpha_i = 1)$$

$$= E(\alpha_i)E(\alpha_j \mid \alpha_i = 1)$$

Therefore

$$\sum_{j(\neq i)=1}^{N}\left[E(\alpha_i\,\alpha_j)-E(\alpha_i)E(\alpha_j)\right]$$

$$=\sum_{j(\neq i)=1}^{N}\left[E(\alpha_i)E(\alpha_j\mid\alpha_i=1)-E(\alpha_i)E(\alpha_j)\right]$$

$$=E(\alpha_i)\sum_{j(\neq i)=1}^{N}\left[E(\alpha_j\mid\alpha_i=1)-E(\alpha_j)\right]$$

$$=E(\alpha_i)\left[(n-1)-(n-E(\alpha_i)\right]$$

$$=-E(\alpha_i)\left[1-E(\alpha_i)\right]$$

$$=-\pi_i(1-\pi_i) \qquad (1)$$

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

$$Var(\hat{\bar{Y}}_{HT})=\frac{1}{2n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2+\sum_{j=1}^{N}\pi_j(1-\pi_j)z_j^2-2\sum_{i\neq j=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})z_iz_j\right]$$

$$=\frac{1}{2n^2}\left[-\sum_{i=1}^{N}\left\{\sum_{j(\neq i)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_i^2\right.$$

$$\left.-\sum_{j=1}^{N}\left\{\sum_{i(\neq j)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_j^2-2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{n}\left\{E(\alpha_i)E(\alpha_j)-E(\alpha_i\alpha_j)\right\}z_iz_j\right]$$

$$=\frac{1}{2n^2}\left[\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_i^2+\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_j^2+2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_iz_j\right]\right.$$

$$=\frac{1}{2n^2}\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})(z_i^2+z_j^2-2z_iz_j)\right].$$

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

$$P_{i1} = P_i$$

$$P_{i2} = P\begin{bmatrix} y_i \text{ is not selected} \\ \text{at } 1^{st} \text{ draw} \end{bmatrix} P\begin{bmatrix} y_i \text{ is selected at } 2^{nd} \text{ draw}| \\ y_i \text{ is not selected at } 1^{st} \text{ draw} \end{bmatrix}$$

$$= \sum_{j(\neq i)=1}^{N} \frac{P_j P_i}{1 - P_j}$$

$$= \left[\sum_{j=1}^{N} \frac{P_j}{1 - P_j} - \frac{P_i}{1 - P_i} \right] P_i.$$

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

$$E(\alpha_i \alpha_j) = \text{Probability of including both } y_i \text{ and } y_j \text{ in a sample of size two}$$

$$= P_{i1} P_{j2|i} + P_{j1} P_{i2|j}$$

$$= P_i \frac{P_j}{1 - P_i} + P_j \frac{P_i}{1 - P_j}$$

$$= P_i P_j \left[\frac{1}{1 - P_i} + \frac{1}{1 - P_j} \right].$$

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.

Chapter 8

Double Sampling (Two Phase Sampling)

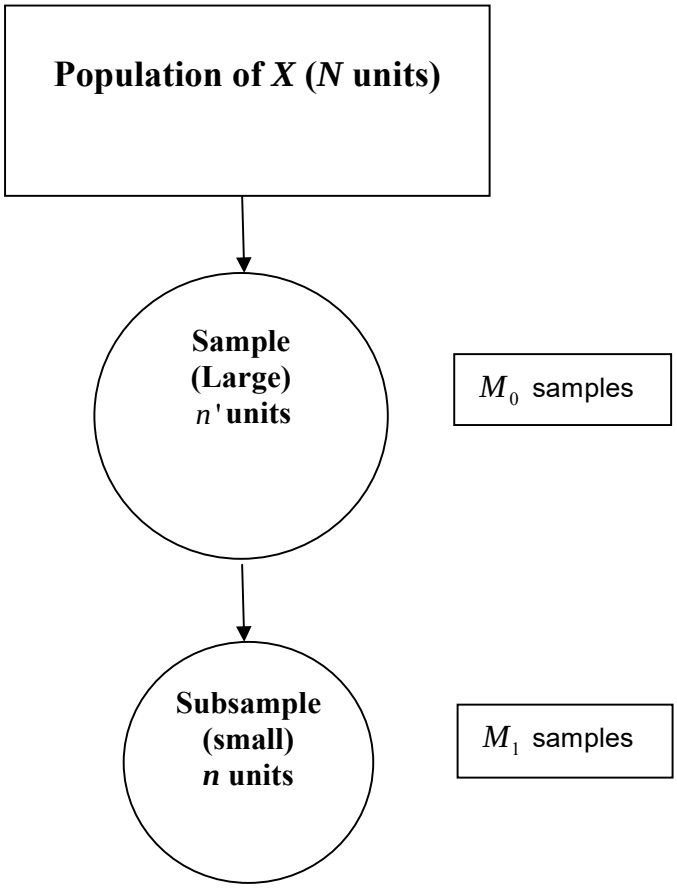
The ratio and regression methods of estimation require the knowledge of population mean of auxiliary variable $(\bar{X})$ to estimate the population mean of study variable $(\bar{Y})$. If information on the auxiliary variable is not available, then there are two options – one option is to collect a sample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



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

$$E(\tau) = \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_{ij}\tau_{ij}$$

$$= \sum_{i=1}^{M_0}\sum_{j=1}^{M_1} P_i P_{j/i}\,\tau_{ij} \quad \text{(using } P(A\cap B) = P(A)P(B/A)\text{)}$$

$$= \underbrace{\sum_{i=1}^{M_0} P_i}_{1^{st}\text{ stage}}\ \underbrace{\sum_{j=1}^{M_1} P_{j/i}\,\tau_{ij}}_{2^{nd}\text{ stage}}$$

Variance of τ

$$
\begin{aligned}
Var(\tau) &= E\left[\tau - E(\tau)\right]^2 \\
&= E\left[(\tau - E_2(\tau)) + (E_2(\tau) - E(\tau))\right]^2 \\
&= E\left[\tau - E_2(\tau)\right]^2 + \left[E_2(\tau) - E(\tau)\right]^2 + 0 \\
&= E_1E_2\left[\tau - E_2(\tau)\right]^2 + \left[E_2(\tau) - E(\tau)\right]^2 \\
&= E_1E_2\left[\tau - E_2(\tau)\right]^2 + E_1E_2\left[E_2(\tau) - E(\tau)\right]^2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad \downarrow \\
&\qquad\qquad\qquad\qquad\qquad \text{constant for } E_2 \\
&= E_1\left[V_2(\tau)\right] + E_1\left[E_2(\tau) - E_1(E_2(\tau))\right]^2 \\
&= E_1\left[V_2(\tau)\right] + V_1\left[E_2(\tau)\right]
\end{aligned}
$$

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

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

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

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

$$E(\varepsilon_1 \varepsilon_2) = \frac{1}{\bar{X}^2} E(\bar{x} - \bar{X})(\bar{x}' - \bar{X})$$

$$= \frac{1}{\bar{X}^2} E_1 \left[E_2 (\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \mid n' \right]$$

$$= \frac{1}{\bar{X}^2} E_1 \left[(\bar{x}' - \bar{X})^2 \right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2}$$

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

$$= E(\varepsilon_2^2).$$

$$E(\varepsilon_0 \varepsilon_2) = \frac{1}{\bar{X}\bar{Y}} Cov(\bar{y}, \bar{x}')$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[E(\bar{y} \mid n'), E(\bar{x}' \mid n') \right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}, \bar{x}') \mid n' \right]$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[\bar{Y}, \bar{X} \right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}', \bar{x}') \right]$$

$$= \frac{1}{\bar{X}\bar{Y}} Cov\left[(\bar{y}', \bar{x}' \right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}}$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho C_x C_y$$

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

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

$$E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2} Var(\bar{y})$$
$$= \frac{1}{\bar{Y}^2}\left[V_1\left\{E_2(\bar{y} \mid n')\right\} + E_1\left\{V_2(\bar{y}_n \mid n')\right\}\right]$$
$$= \frac{1}{\bar{Y}^2}\left[V_1(\bar{y}_n') + E_1\left\{\left(\frac{1}{n} - \frac{1}{n'}\right)s_y'^2\right\}\right]$$
$$= \frac{1}{\bar{Y}^2}\left[\left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)S_y^2\right]$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_y^2}{\bar{Y}^2}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)C_y^2$$

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

$$E(\varepsilon_1\varepsilon_2) = \frac{1}{\bar{X}^2} Cov(\bar{x}, \bar{x}')$$
$$= \frac{1}{\bar{X}^2}\left[Cov\left\{E(\bar{x} \mid n'), E(\bar{x}' \mid n')\right\} + 0\right]$$
$$= \frac{1}{\bar{X}^2} Var(\bar{X}')$$

where $Var(\bar{X}')$ 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*

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

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

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

$$E(\hat{\bar{Y}}_{Rd}) = \bar{Y}\left[1+0+0+E(\varepsilon_0\varepsilon_2)-0-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

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

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

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

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

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

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

Choice of n and n'

Write

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

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 $= \frac{C_0}{C}$.

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

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

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

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

Let

$$\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} \Rightarrow \bar{x} = (1+\varepsilon_1)\bar{X}$$

$$\varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}} \Rightarrow \bar{x}' = (1+\varepsilon_2)\bar{X}$$

$$\varepsilon_3 = \frac{s_{xy} - S_{xy}}{S_{xy}} \Rightarrow s_{xy} = (1+\varepsilon_3)S_{xy}$$

$$\varepsilon_4 = \frac{s_x^2 - S_x^2}{S_x^2} \Rightarrow s_x^2 = (1+\varepsilon_4)S_x^2$$

$$E(\varepsilon_1) = 0, E(\varepsilon_2) = 0, E(\varepsilon_3) = 0, E(\varepsilon_4) = 0$$

Define

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

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

Estimation error:

Then

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

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)

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

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:

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

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

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

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

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

Clearly, $\hat{\bar{Y}}_{regd}$ 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 $MSE\ (\hat{\bar{Y}}_{regd})$ 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 $\quad n = \sqrt{\frac{S_y^2(1-\rho^2)}{\lambda C_1}}$

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

Substituting these values in the cost function, we have

$$C_0 = C_1 n + C_2 n'$$

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

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

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

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

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

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

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

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

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 $\rho = 0, C_2 = 0$ 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

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

Double sampling for probability proportional to size estimation:

Suppose it is desired to select the sample with probability proportional to auxiliary variable x but information on x is not available. Then, in this situation, the double sampling can be used. An initial sample of size n' is selected with SRSWOR from a population of size N, and information on x is collected for this sample. Then a second sample of size n is selected with replacement and with probability proportional to x from the initial sample of size n'. Let $\bar{x}'$ denote the mean of x for the initial sample of size n', Let $\bar{x}$ and $\bar{y}$ denote means respectively of x and y for the second sample of size n. Then we have the following theorem.

Theorem:

(1) An unbiased estimator of population mean $\bar{Y}$ is given as

$$\hat{\bar{Y}} = \frac{x'_{tot}}{n'n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right),$$

where x'_{tot} denotes the total for x in the first sample.

(2) $Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \frac{(n'-1)}{N(N-1)nn'} \sum_{i=1}^{N} \frac{x_i}{X_{tot}} \left(\frac{y_i}{\frac{x_i}{X_{tot}}} - Y_{tot} \right)^2$, where X_{tot} and Y_{tot} denote the totals of x and y respectively in the population.

(3) An unbiased estimator of the variance of $\hat{\bar{Y}}$ is given by

$$\widehat{Var}(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{1}{n(n'-1)} + \left[x_{tot}' \sum_{i=1}^{n} \frac{y_i^2}{x_i} - \frac{x_{tot}'^2 (A-B)}{n'(n-1)} \right] + \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{x_{tot}' y_i}{n' x_i} - \hat{\bar{Y}} \right)$$

where $A = \left(\sum_{i=1}^{n} \frac{y_i}{x_i} \right)^2$ and $B = \sum_{i=1}^{n} \frac{y_i^2}{x_i^2}$

Proof. Before deriving the results, we first mention the following result proved in varying probability scheme sampling.

Result: In sampling with varying probability scheme for drawing a sample of size n from a population of size N and with replacement .

(i) $\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$ is an unbiased estimator of population mean $\bar{y}$ where $z_i = \frac{y_i}{Np_i}$, p_i being the probability of selection of i^{th} unit. Note that y_i and p_i can take anyone of the N values $Y_1, Y_2, ..., Y_N$ with initial probabilities $P_1, P_2, ..., P_N$, respectively.

(ii) $Var(\bar{z}) = \frac{1}{nN^2} \left[\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - N^2 \bar{Y}^2 \right] = \frac{1}{nN^2} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - \bar{Y} \right)^2 ..$

(iii) An unbiased estimator of variance of $\bar{z}$ is

$$\widehat{Var}(\bar{z}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{Np_i} - \bar{z} \right)^2 ..$$

Let E_2 denote the expectation of $\hat{\bar{Y}}$, when the first sample is fixed. The second is selected with probability proportional to x, hence using the result (i) with $P_i = \frac{x_i}{x_{tot}'}$, we find that

$$E_2\left(\frac{\hat{\bar{Y}}}{n'}\right) = E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{n'\frac{x_i}{x'_{tot}}}\right]$$
$$= E_2\left[\frac{x'_{tot}}{nn'}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)\right]$$
$$= \bar{y}'$$

where $\bar{y}'$ is the mean of y for the first sample. Hence

$$E(\hat{\bar{Y}}) = E_1\left[E_2\left(\hat{\bar{Y}} \mid n'\right)\right]$$
$$= E_1(\bar{y}_{n'})$$
$$= \hat{\bar{Y}},$$

which proves the part (1) of the theorem. Further,

$$Var(\hat{\bar{Y}}) = V_1E_2\left(\hat{\bar{Y}} \mid n'\right) + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= V_1(\bar{y}') + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + E_1V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Now, using the result (ii), we get

$$V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\sum_{i=1}^{n'}\frac{x_i}{x'_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{x'_{tot}}} - y'_{tot}\right)^2$$
$$= \frac{1}{nn'^2}\sum_{i=1}^{n'}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

and hence

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

using the probability of a specified pair of units being selected in the sample is $\frac{n'(n'-1)}{N(N-1)}$. So we can express

$$E_1 V_2\left(\hat{\bar{Y}} / n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_{tot}}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2 .$$

Substituting this in $V_2\left(\hat{\bar{Y}} \mid n'\right)$, we get

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \frac{(n'-1)}{nn'N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2 .$$

This proves the second part (2) of the theorem.

We now consider the estimation of $Var(\hat{\bar{Y}})$. Given the first sample, we obtain

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

where $p_i = \frac{x_i}{x'_{tot}}$. Also, given the first sample,

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}\right)^2\right] = V_2(\hat{\bar{Y}}) = E_2(\hat{\bar{Y}}^2) - \bar{y}'^2 .$$

Hence

$$E_2\left[\hat{\bar{Y}}^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}^2\right)^2\right] = \bar{y}'^2 .$$

Substituting $\hat{\bar{Y}} = \frac{x'_{tot}}{n'n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)$ and $p_i = \frac{x_i}{x'_{tot}}$ the expression becomes

$$E_2\left[\frac{x'^2}{nn'^2(n-1)}\left\{\left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2 - \left(\sum_{i=1}^{n}\frac{y_i^2}{x_i^2}\right)\right\}\right] = \bar{y}'^2$$

Using

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

we get

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \frac{x_{tot}^{'}}{x_i} - \frac{x_{tot}^{'2}}{nn'(n-1)}(A-B)\right] = \sum_{i=1}^{n'} y_i^2 - n'\bar{y}^{'2}$$

where $A = \left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right)^2$, and $B = \sum_{i=1}^{n} \frac{y_i^2}{x_i^2}$ which further simplifies to

$$E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = s_y^{'2},$$

where $s_y^{'2}$ is the mean sum of squares of y for the first sample. Thus, we obtain

$$E_1E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = E_1(s_y^{'2}) = S_y^2 \qquad (1)$$

which gives an unbiased estimator of S_y^2. Next, since we have

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'}\frac{(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2,$$

and from this result we obtain

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i x_{tot}^{'}}{n' x_i} - \hat{\bar{Y}}\right)^2\right] = V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Thus

$$E_1E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n' x_i} - \hat{\bar{Y}}\right)^2\right] = \frac{(n'-1)}{nn'N(n-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2 \qquad (2)$$

when gives an unbiased estimator of

$$\frac{(n'-1)}{nn'N(n-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{}{X_{tot}}} - Y_{tot}\right)^2.$$

Using (1) and (2) an unbiased estimator of the variance of $\hat{\bar{Y}}$ is obtained as

$$\widehat{Var}(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{n(n'-1)}\left[x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n' x_i} - \hat{\bar{Y}}\right)^2$$

Thus, the theorem is proved.

Chapter 9
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

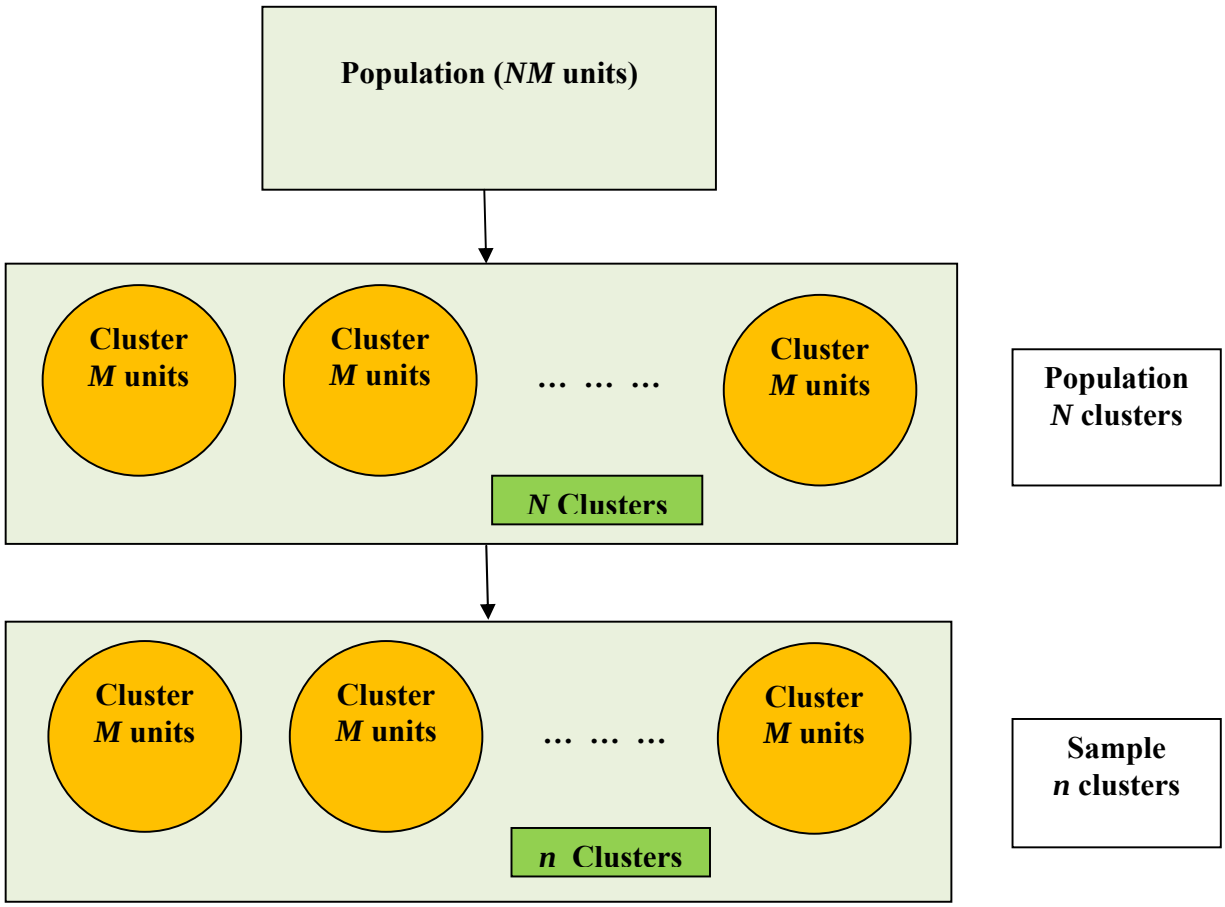
total population size = NM

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

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

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

Consider

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

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^{th} cluster.

The efficiency of cluster sampling over SRSWOR is

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

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

$$\rho = \frac{E(y_{ij}-Y)(y_{ik}-\bar{Y})}{E(y_{ij}-\bar{Y})^2}; -\frac{1}{M-1} \le \rho \le 1$$

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

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

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

Consider

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

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

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

or

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

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

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

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

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

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

- 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

  $E > 1$

  or $(M-1)\rho < 0$

  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, 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

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

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

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

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 $M(N-1) \simeq MN$ and $MN-1 \simeq MN$, 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^{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

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

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

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

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

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

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(P_i - P)^2,$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N} P_i^2 - NP^2\right]$$

$$= \frac{1}{(N-1)}\left[-\sum_{i=1}^{N} P_i(1-P_i) + \sum_{i=1}^{N} P_i - NP^2\right]$$

$$= \frac{1}{(N-1)}\left[NPQ - \sum_{i=1}^{N} P_iQ_i\right],$$

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

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

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

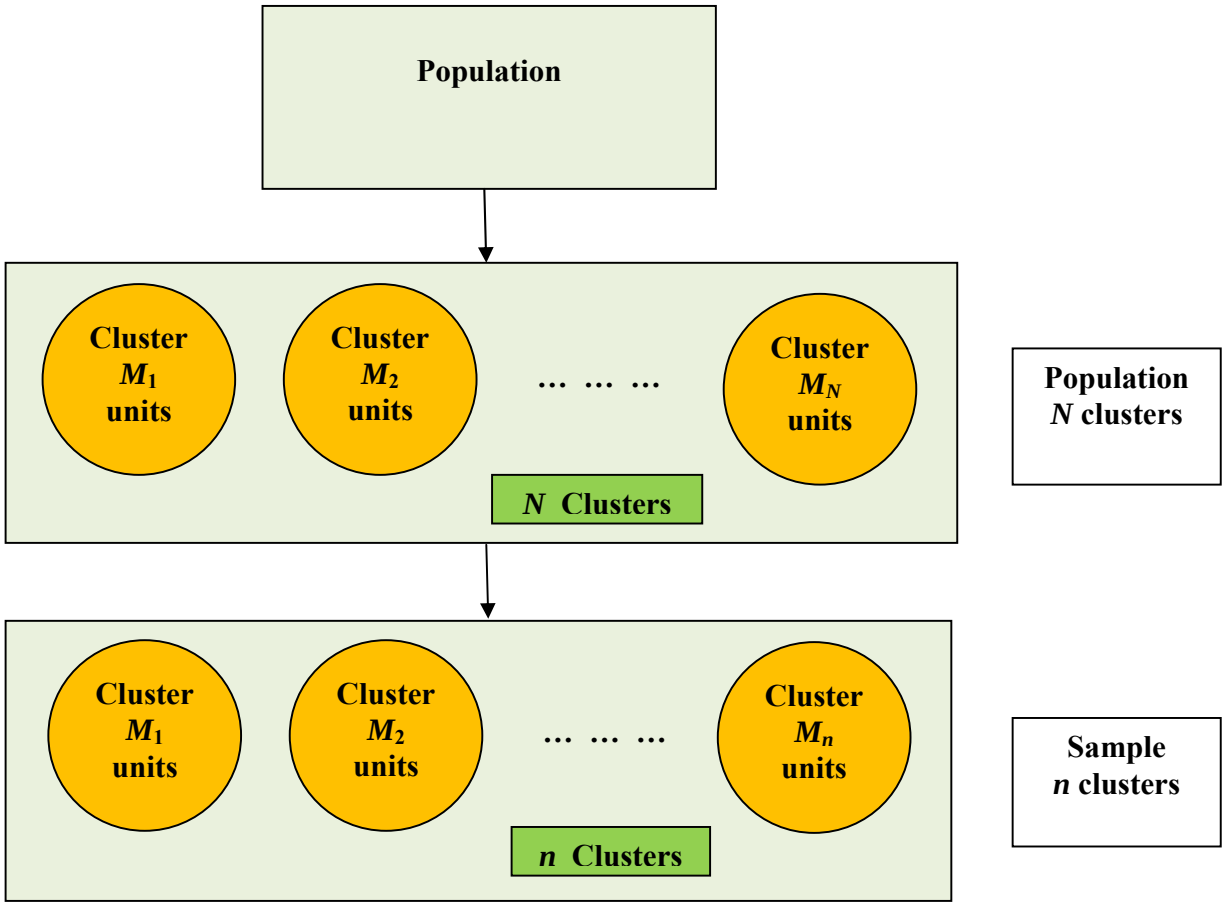
$$\bar{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} \text{ : mean of } i^{th} \text{ cluster}$$

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

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

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

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

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

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

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

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

$$Bias\left(\bar{\bar{y}}_c\right) = E\left(\bar{\bar{y}}_c\right) - \bar{Y}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i - \sum_{i=1}^{N}\left(\frac{M_i}{M_0}\right)\bar{y}_i$$

$$= -\frac{1}{M_0}\left[\sum_{i=1}^{N}M_i\bar{y}_i - \frac{M_0}{N}\sum_{i=1}^{N}\bar{y}_i\right]$$

$$= -\frac{1}{M_0}\left[\sum_{i=1}^{N}M_i\bar{y}_i - \frac{\left(\sum_{i=1}^{N}M_i\right)\left(\sum_{i=1}^{N}\bar{y}_i\right)}{N}\right]$$

$$= -\frac{1}{M_0}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y})$$

$$= -\left(\frac{N-1}{M_0}\right)S_{m\bar{y}}$$

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

The mean squared error is

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

$$= \frac{N-n}{Nn}S_b^2 + \left(\frac{N-1}{M_0}\right)^2 S_{m\bar{y}}^2$$

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 $Var\left(\bar{\bar{y}}_c\right)$ is

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

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

2. Weighted mean of cluster means

Consider the arithmetic mean based on cluster total as

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

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

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

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

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

where

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

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

$$E(s_b^{*2}) = S_b^{*2}.$$

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

Since SRSWOR is followed, so

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

Since

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

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

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

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

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

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

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

The ratio estimator based on U and V is

$$\hat{\bar{Y}}_R = \frac{\bar{u}}{\bar{v}}\bar{V}$$

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

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

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

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

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

$$= \frac{N-n}{Nn}\left(S_v^2 - \frac{S_{uv}}{\bar{U}}\right)\bar{U}$$

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

$$S_v^2 = \frac{1}{N-1}\sum_{i=1}^{N}(V_i - \bar{V})^2$$

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

$$S_{uv} = \frac{1}{N-1}\sum_{i=1}^{N}(U_i - \bar{U})(V_i - \bar{V})$$

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

$$R_{uv} = \frac{\bar{U}}{\bar{V}} = \bar{U} = \frac{1}{N\bar{M}}\sum_{i=1}^{N}M_i\bar{y}_i.$$

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,

$$MSE(\bar{y}_c^{**}) = \frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}\left(U_i - R_{uv}V_i\right)^2$$

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

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

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

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

$$= -\left(\frac{N-1}{N\bar{M}}\right) S_{m\bar{y}}$$

Since SRSWOR is used, so

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

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

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

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

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

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

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

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

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

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

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

Chapter 10

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; $i = 1, 2, ..., N;\ j = 1, 2, .., m.$

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

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

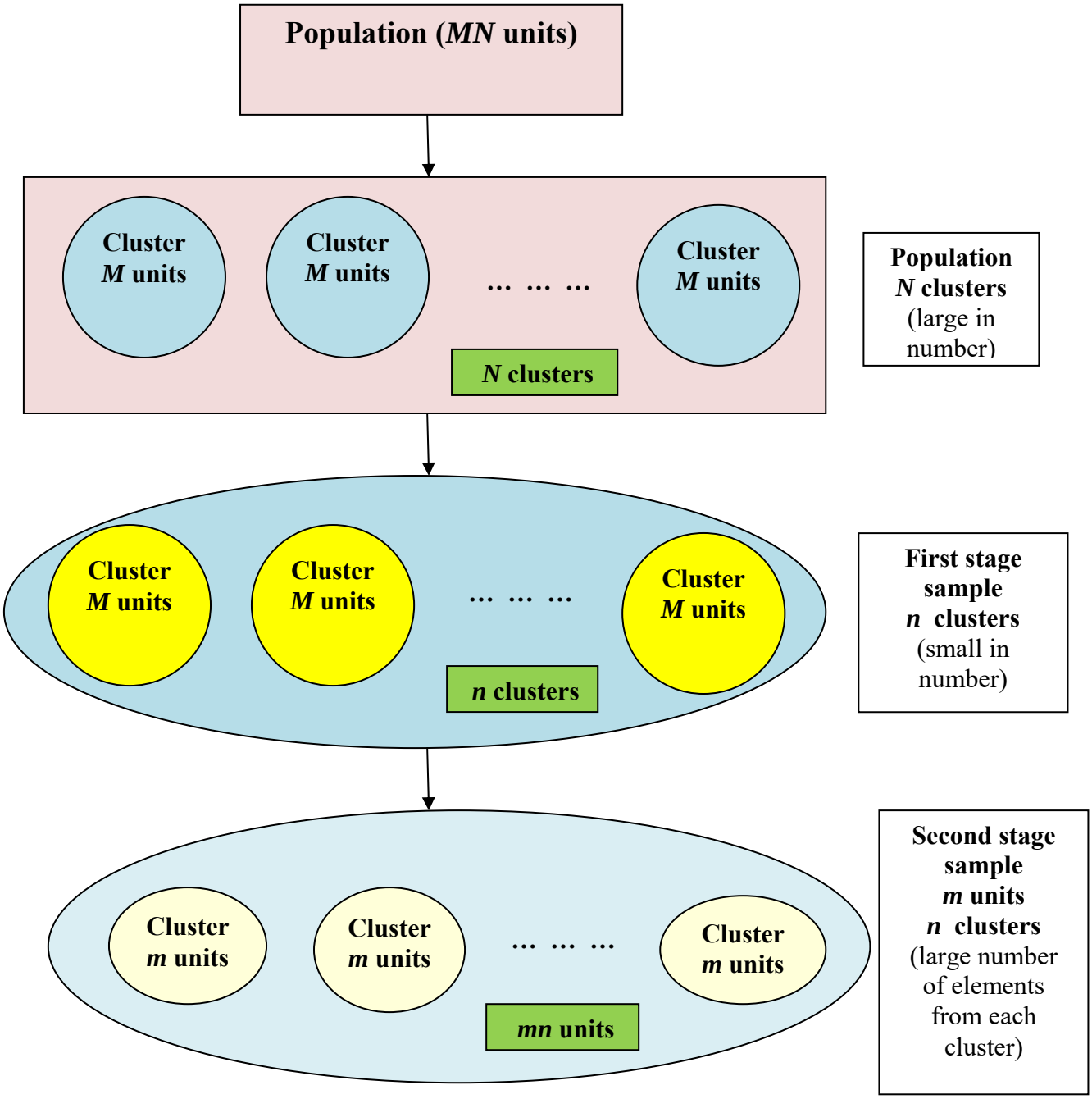
$\bar{y}_i = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$: mean per second stage unit in the i^{th} first stage unit in the sample.

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

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

$E(\hat{\theta})$	$E_1[E_2(\hat{\theta})]$
↓	↓ ↘
average over all samples	average over all 1^{st} stage samples	average over all possible 2^{nd} 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,

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

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

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

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

Variance

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

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

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

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 $\bar{y}$ 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

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

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

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

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

So

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus

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

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

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

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 $\left(S_b^2 - \frac{\overline{S}_w^2}{M}\right) < 0$ (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

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

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

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

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

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

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

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

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

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

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

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

and using the identity

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

$$(NM-1)S^2 = (N-1)MS_b^2 + N(M-1)\bar{S}_w^2 \qquad (2)$$

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

Now we need to find S_b^2 and $\bar{S}_w^2$ from (1) and (2) in terms of S^2. From (1), we have

$$\bar{S}_w^2 = -\left(\frac{MN-1}{N}\right) MS^2\rho + \left(\frac{N-1}{N}\right) MS_b^2. \qquad (3)$$

Substituting it in (2) gives

$$\begin{aligned}(NM-1)S^2 &= (N-1)MS_b^2 + N(M-1)\left[\left(\frac{N-1}{N}\right)MS_b^2 - \left(\frac{MN-1}{N}\right)MS^2\rho\right] \\ &= (N-1)MS_b^2 + (M-1)(N-1)S_b^2 - \rho M(M-1)(MN-1)S^2 \\ &= (N-1)MS_b^2[1+(M-1)] - \rho M(M-1)(MN-1)S^2 \\ &= (N-1)MS_b^2 - \rho M(M-1)(MN-1)S^2 \\ \Rightarrow S_b^2 &= \frac{(MN-1)S^2}{M^2(N-1)}\left[1+(M-1)\rho\right]\end{aligned}$$

Substituting it in (3) gives

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

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 $Var(\bar{y}_{cl})$ 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.

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

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

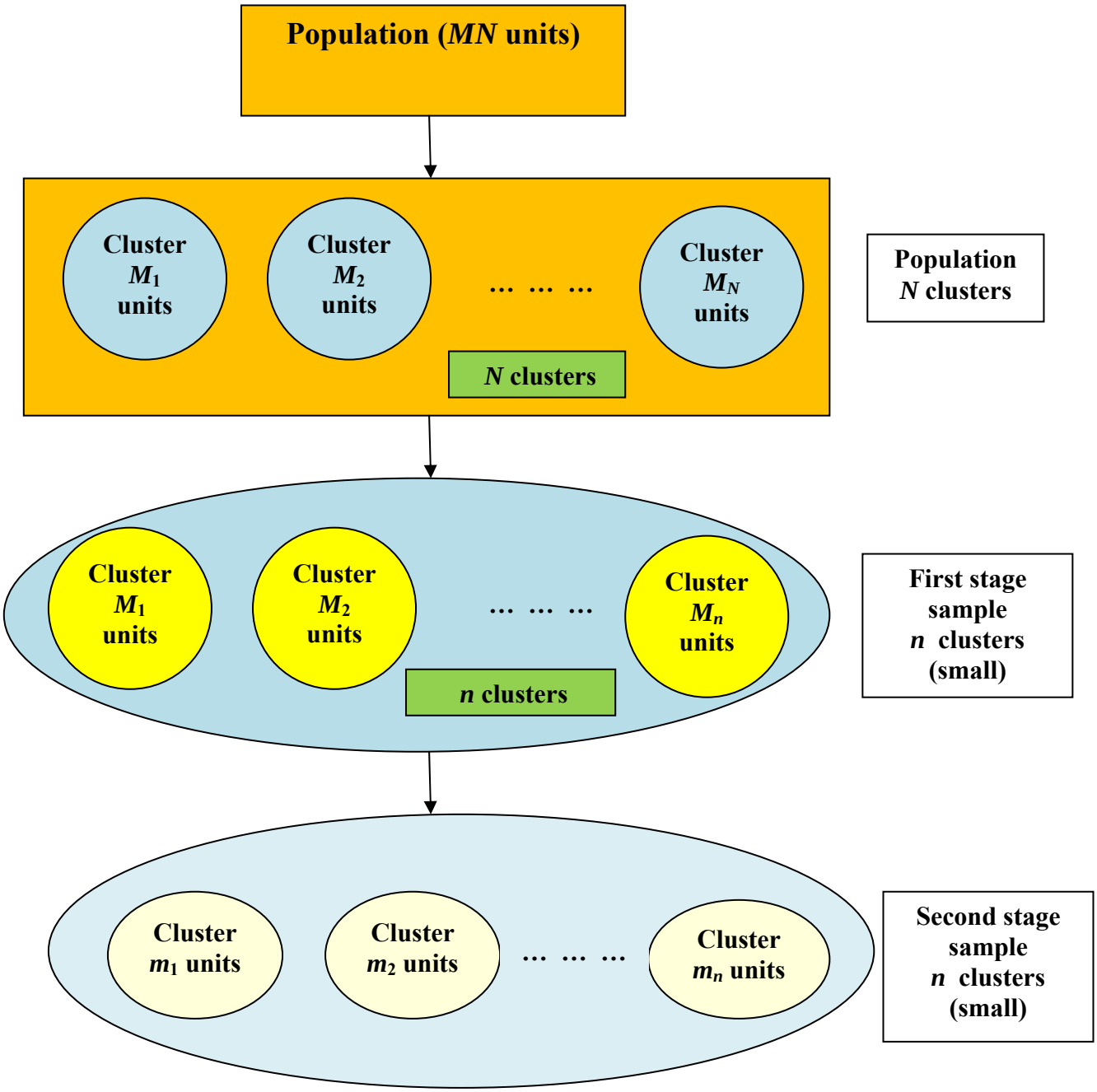
$$\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i=\bar{\bar{Y}}_N$$

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

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

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

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:

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

Bias:

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

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

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

This bias can be estimated by

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

which can be seen as follows:

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

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

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

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

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:

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

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

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

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:

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

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

Bias

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

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

Variance:

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

Mean squared error

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

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

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

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

where

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

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

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

Thus

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

Also

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

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

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