Design and analysis of sample surveys

School of Science and Informatics

Department of Mathematics, Statistics and Physical Sciences

Taita Taveta University

Dr. Noah Mutai

February 19, 2023

Some important terms

- Definition — an explanation of the mathematical meaning of a word.
- Theorem — A statement that has been proven to be true.
- Proposition — A less important but nonetheless interesting true statement.
- Lemma — A true statement used in proving other true statements (that is, a less important theorem that is helpful in the proof of other results).
- Corollary — A true statement that is a simple deduction from a theorem or proposition.
- Proof — The explanation of why a statement is true.
- Conjecture — A statement believed to be true, but for which we have no proof. (a statement that is being proposed to be a true statement).
- Axiom — A basic assumption about a mathematical situation (a statement we assume to be true).

1 Introduction

Learning objectives

By the end of this lesson learners should be able to;

(a) *Define common terms used in sampling such as sample, census, target population, sampling unit, sampling frame etc.*

(b) *Identify the sampling frame for a given sampling scenario.*

(c) *State and explain why sampling is needed.*

(d) *Identify types of sampling.*

(e) *Outline steps for designing and carrying out a sample survey.*

(f) *Differentiate between a census and sample survey.*

1.1 Motivation

We all use data from samples to make decisions.

(a) Taking blood samples to test for an infection

(b) When tasting soup to correct the seasoning

(c) Deciding to buy a book after reading the first pages

(d) Choosing a major after taking first-year college classes

(e) Buying a car following a test drive

(f) Deciding to date someone we depend on few experiences with the person

In all these scenarios, we rely on a small part of information about the whole to make decisions. Therefore, our decisions are highly influenced by the quality of the data we have access to.

Definition 1.1. Sample survey, finite population sampling or survey sampling is a method of drawing an inference about the characteristics of a population or universe by **observing under only a part of the population** (Mukhopadhyay, 2008). Such methods are extensively used by government bodies throughout the world for assessing, among others, different characteristics of national economy as a required for making decisions and for the planning and projection of future economic structure. Ideally, total information about the population is obtained through census.

Definition 1.2. Census — Complete enumeration of all units in a population. In a census, every individual in the population is involved in giving out information. However, most of the times due to certain constraints to be discussed later, it is not always possible to carry out a census.

What is the interest? In a sample survey the purpose of the survey statistician is to estimate some functions of the population parameter, $\theta\left(y\right)$, say, by choosing a sample(part of the population) and by observing the values of y only on units selected in the sample. The statistician therefore want to make an inference about the population by observing only a part of it. This is essential and perhaps the only practical method of inference about the characteristics of the population since in many socioeconomic investigations the survey population may be very large, containing say hundreds or thousands of units.

Definition 1.3. Survey population — A finite(survey) population is a collection of known number N of identifiable units labeled $1, 2, 3, ..., i, ..., N$ where i stands for the label as well as the physical unit labeled i. The number N is the size of the population. The parametric functions of general interest for estimation are;

(a) Population total, $Y = \sum_{i=1}^{N} Y_i$

(b) Population mean: $\bar{Y} = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

(c) Population variance: $S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$

(d) Population coefficient of variance: $C_Y = \frac{S_Y}{\bar{Y}}$,where S_Y is the population variance and $\bar{Y}$ is the population mean.

Definition 1.4. Sample — is a part of the population/subset of the population selected for study. A sample may be drawn from a population either under with replacement (wr) or under without replacement (wor).

After a sample is selected , data are collected from the sampled units. We shall denote by y_i the value of y on the unit selected at the i^{th} draw $(i = 1, 2, ..., n)$. Thus for example if the sample is $S = \{2, 3, 2\}$,$y_1 = Y_2, y_2 = Y_3, y_3 = Y_2$ .Clearly y_i is a random variable whose possible values lie in the set $\{Y_1, Y_2, ..., Y_N\}$

For a sample s, we shall denote some statistics as follows;

(a) Sample total, $y = \sum_{i=1}^{n} y_i$

(b) Sample mean, $\bar{y} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$

(c) Sample variance, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$

(d) Sample coefficient of variation, $c_y = \frac{s_y}{\bar{y}}$,where s_y is the sample variance and $\bar{y}$ is the sample mean.

Illustration:

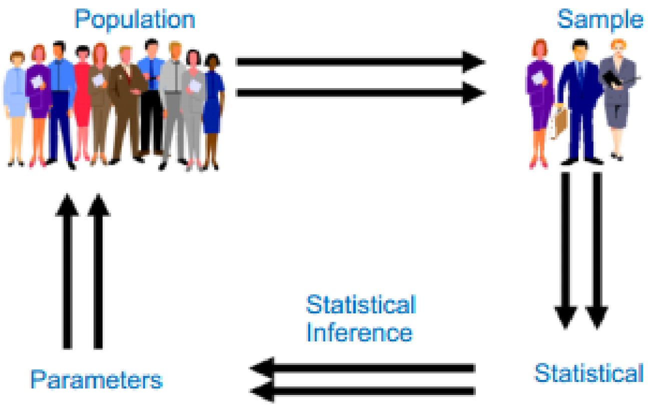


Figure 1: Illustration: Sampling

Definition 1.5. Sampling units: This refers to the individual items whose characteristics are to be measured in the sample survey.

Definition 1.6. Sampling frame: This is the list of all sampling units. It may be a list of units with identification and particulars or a map showing the boundaries of sampling unis e.g. a manufacturing firm may want to determine how popular a newly manufactured product is within the community suggests a possible frame for the survey. The firm may decide to concentrate its surveys in urban residential areas only. In this case, you have a complete list of estates in urban areas. The residents in those chosen estates will be interviewed and inferences are made.

Definition 1.7. Sampled population: It is the set of individuals in the sampling frame. Its actually the subset of the target population. Note: Sampled population is not necessarily the same as target population.

Definition 1.8. Sampling scheme: A sampling scheme is a detailed description of what data will be obtained and how this will be done.

Definition 1.9. Sampling design: A sample design is made of two elements.

(i) Sampling method: Rules and procedures by which some elements of the population are included in the sample.

(ii) Estimator: The process of calculating sample statistics is called estimator. Different sampling methods use different estimators.

1.2 Types of Sampling

(i) Probability/random sampling — statistical theory is used and the kind of inferences made are based on statistical procedures. There is some element of chance associated with selection of items into the sample.

(ii) Purposive or judgemental sampling or non probability sampling — researchers rely on their own judgment when choosing members of the population to participate in their surveys.

We shall concentrate on probability sampling in this course.

1.3 Properties of random sampling

We are able to define the set of distinct samples, $S_1, S_2, ..., S_n$, which the procedure is capable of selecting if applied to a specific population. This means that we can say precisely what sampling units belong to S_1 to S_2 and so on.

(i) Each possible sample S_i has assigned to it a known probability of selection π_i.

(ii) We select one of the S_i by a process in which each S_i receives its appropriate probability π_i, of being selected.

(iii) The method for computing the estimate from the sample must be stated and must lead to a unique estimate for any specific sample.

1.4 Types of surveys

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

(a) Demographic surveys — e.g. household surveys, family size, number of males in families, etc.

(b) Educational surveys — e.g. how many children go to school.

(c) Economic surveys — collect the economic data, e.g., data related to export and import.

(d) Employment surveys — employment related data, e.g., employment rate, labour conditions, wages, etc. in a city, state or country.

(e) Health and nutrition surveys — health and nutrition issues, e.g., number of visits to doctors, food given to children, nutritional value etc

(f) Agricultural surveys — agriculture related data to estimate, e.g., the acreage and production of crops, livestock numbers, use of fertilizers.

(g) Marketing surveys — They are conducted by major companies, manufacturers or those who provide services to consumer etc.

(h) Election surveys — conducted to study the outcome of an election or a poll.

1.5 Principal steps involved in planning and execution of 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.

12. **Pilot Survey** In planning a survey efficiently, some prior information about the population under consideration and the operational and cost aspects of of data collection will be needed. When such information is not available.

1.6 Advantages of Sampling

Sample surveys have potential advantages over complete enumeration(census). They include;

(a) **Reduced cost** — If data are secured from only a small fraction of the aggregate, expenditures may be expected to be smaller than if a complete census is attempted.

(b) **Greater speed** — For the same reason, the data can be collected and summarized more quickly with a sample than with a complete count. This may be a vital consideration when the information is urgently needed.

(c) **Greater scope** — In certain types of inquiry, highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete census may then be impracticable.

(d) **Greater accuracy** — Because personnel of higher quality can he employed and can be given intensive training, a sample may actually produce more accurate results than the kind of complete enumeration that it is feasible to take.

(e) **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.

(f) **Risk** — When a survey involves risky tests such as testing a new drug, sampling should be used.

1.7 Difference between a census and sample survey

Table 1: Difference between a census and sample survey

Parameter	Census	Sample survey
Definition	A statistical method that studies all the units or members of a population	A statistical method that studies only a representative group of the population, and not all its members.
Calculation	Total/Complete	Partial
Time involved	It is a time-consuming process	It is a quicker process.
Cost involved	It is a costly method.	It is a relatively inexpensive method
Accuracy	The results obtained are accurate.	The results are relatively inaccurate due to leaving out of items.
Reliability	Highly reliable	Low reliability for small samples.
Error	Not present	The smaller the sample size, the larger the error.
Relevance	This method is suited for heterogeneous data.	This method is suited for homogeneous data

1.8 Exercises

1. Discuss the statement: "The need to collect statistical information arises in almost every conceivable sphere of human activity."

2. Describe briefly each of the following terms:

    (a) Primary data

    (b) Secondary data

    (c) Mail inquiry

    (d) Questionnaire/schedule

    (e) Population

    (f) Census

    (g) Element

    (h) Sample

    (i) Sampling unit

    (j) Sampling frame

3. Differentiate between target and sampled population. What problem arises if two populations are not same?

4. What is the primary advantage of probability sampling over the non probability sampling? Cite three situations where non probability sampling is to be preferred

5. Assume a sample survey shall be carried out to find out about how satisfied students are with their faculty.

    (a) How would you define the population?

    (b) Would you consider a census of all students or rather a sample survey? (Why?)

    (c) How would you operationalise? being satisfied with their faculty?

    (d) What is a sampling frame and how could one be obtained in the example?

    (e) How could a random sample be obtained?

    (f) How do you consider the idea of obtaining a sample from alumni?

1.9 Solutions to exercises

1. Discuss the statement: "The need to collect statistical information arises in almost every conceivable sphere of human activity."

    The need to gather information arises in almost every conceivable sphere of human activity. Many of the questions that are subject to common conversation and controversy require numerical data for their resolution. Data resulting from the physical, chemical, and biological experiments in the form of observations are used to test different theories and hypotheses. Various social and economic investigations are carried out through the use and analysis of relevant data. The data collected and analyzed in an objective manner and presented suitably serve as basis for taking policy decisions in different fields of daily life.

The important users of statistical data, among others, include government, industry, business, research institutions, public organizations, and international agencies and organizations. To discharge its various responsibilities, the government needs variety of information regarding different sectors of economy, trade, industrial production, health and mortality, population, livestock, agriculture, forestry, environment, meteorology, and available resources. The inferences drawn from the data help in determining future needs of the nation and also in tackling social and economic problems of people. For instance, the information on cost of living for different categories of people, living in various parts of the country, is of importance in shaping its policies in respect of wages and price levels. Data on health, mortality, and population could be used for formulating policies for checking population growth. Similarly, information on forestry and environment is needed to plan strategies for a cleaner and healthier life. Agricultural production data are of immense use to the state for planning to feed the nation. In case of industry and business, the information is to be collected on labor, cost and quality of production, stock, and demand and supply positions for proper planning of production levels and sales campaigns.

2. Describe briefly each of the following terms.

   (a) Primary data — The data collected by the investigator from the original source are called primary data.

   (b) Secondary data — Already collected information.

   (c) Mail inquiry — investigator prepares a questionnaire and sends it by mail to the respondents.

   (d) Questionnaire/schedule — channel through which the needed information is elicited.

   (e) Population — total subjects under consideration.

   (f) Census — complete enumeration

   (g) Element — unit for which information is sought.

   (h) Sample — subset of the population selected for study

   (i) Sampling unit — This refers to the individual items whose characteristics are to be measured in the sample.

(j) Sampling frame — This is the list of all sampling units.

3. Differentiate between target and sampled population. What problem arises if two populations are not same?

   The target population of a survey is the population you wish to study. The sampled population is the population which you are able to observe in a sample. In an ideal world the target population and the sampled population would be the same, but often they are different.

   Sapling frame error: A sample frame error occurs when the wrong sub-population is used to select a sample.

4. What is the primary advantage of probability sampling over the non probability sampling? Cite three situations where non probability sampling is to be preferred.

   Probability gives all people a chance of being selected and makes results more likely to accurately reflect the entire population.

   Advantages:

   (i) The absence of systematic error and sampling bias.

   (ii) Higher level of reliability of research findings.

   (iii) Increased accuracy of sampling error estimation.

   (iv) The possibility to make inferences about the population.

   (v) Cost-effectiveness.

   (vi) Simple and straightforward in application.

   Non-probability sampling.

   (i) Use this type of sampling to indicate if a particular trait or characteristic exists in a population.

   (ii) Researchers widely use the non-probability sampling method when they aim at conducting qualitative research, pilot studies, or exploratory research.

   (iii) Researchers use it when they have limited time to conduct research or have budget constraints.

   (iv) When the researcher needs to observe whether a particular issue needs in-depth analysis, he applies this method.

   (v) Use it when you do not intend to generate results that will generalize the entire population

5. Assume a sample survey shall be carried out to find out about how satisfied students are with their faculty.

(a) Students enrolled in faculty . . . at a specific date.

(b) Consider size of faculty and costs of the survey/census.

(c) Consider open and closed questions, general and issue (teaching, facilities, and so on) specific opinions.

(d) Sampling frame could be a list provided by the faculty administration including, name, subject, date of enrollment, and so on.

(e) E.g. using pseudo random numbers generated with R.

(f) Alumni present a selective sub sample and results may be misleading.

2 Simple Random Sampling(SRS)

Learning objectives

By the end of this lesson learners should be able to;

(a) *Define simple random sampling.*

(b) *Draw a simple random sample using lottery, random number table and R software.*

(c) *Estimate the mean, total and variance under simple random sampling (SRS) with replacement and without replacement.*

(d) *Construct confidence intervals for the mean and total under SRS.*

(e) *Sample and estimate for proportions and percentages.*

(f) *Determine sample size for a given precision level.*

(g) *Use R to draw a simple random sample.*

2.1 Introduction

We shall consider various sampling procedures (schemes) for selection of units in the sample. Since the objective of a survey is to make inferences about the population, a procedure that provides a precise estimator of the parameter of interest is desirable. Many sampling schemes have been developed to achieve this objective. To begin with, simple random sampling, the simplest and the most basic sample selection procedure, is discussed.

Definition 2.1. Simple random sampling — The sampling procedure is known as simple random sampling if every population unit has the same chance of being selected in the sample. The sample thus obtained is termed a simple random sample.

For selecting a simple random sample in practice, units from population are drawn one by one. If the unit selected at any particular draw is replaced back in the population before the next unit is drawn, the procedure is called with replacement (WR) sampling. A set of units selected at n such draws, constitutes a simple random with replacement sample of size n. In such a selection procedure, there is a possibility of one or more population units getting selected more than once. In case, this procedure is continued till n distinct units are selected, and all repetitions are ignored, it is called simple random sampling (SRS) without replacement (WOR). This method is equivalent to the procedure, where the selected units at each draw are not replaced back in the population before executing the next draw.

Another definition of simple random sampling, both with and without replacement, could be given on the basis of probabilities associated with all possible samples that can be selected from the population.

Definition 2.2. Simple random sampling is the method of selecting the units from the population where all possible samples are equally likely to get selected.

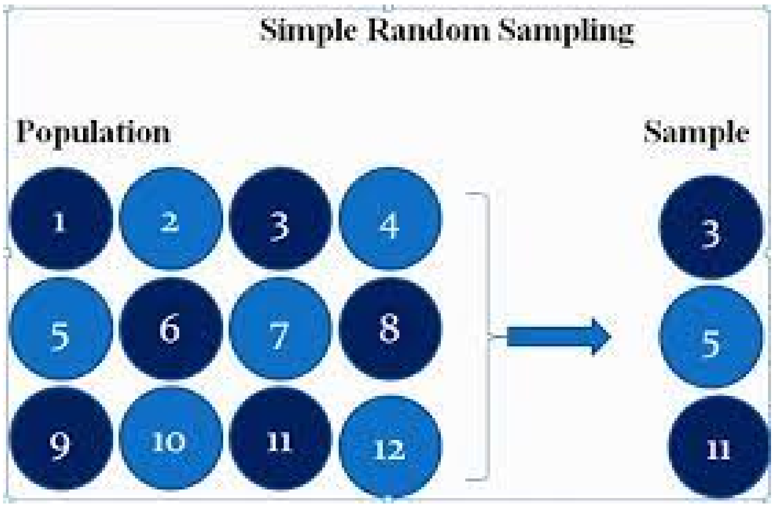


Figure 2: Illustration: simple random sampling

2.2 How to draw a simple random sample

The most commonly used procedures for selecting a simple random sample are:

1. Lottery method
2. Random number tables
3. Computer software

2.2.1 Lottery method

In this method, each unit of the population of N units is assigned a distinct identification mark (number) from 1 to N. This constitutes the population frame. Each of these numbers is then written on a different slip of paper. All the N slips of paper are identical in respect of size, color, shape, etc. Fold all these slips in an identical manner and put them in a container or drum, in which a thorough mixing of the slips is carried out before each blindfold draw. The paper slips are then drawn one by one. The units corresponding to the identification labels on the selected slips, are taken to be members of the sample.

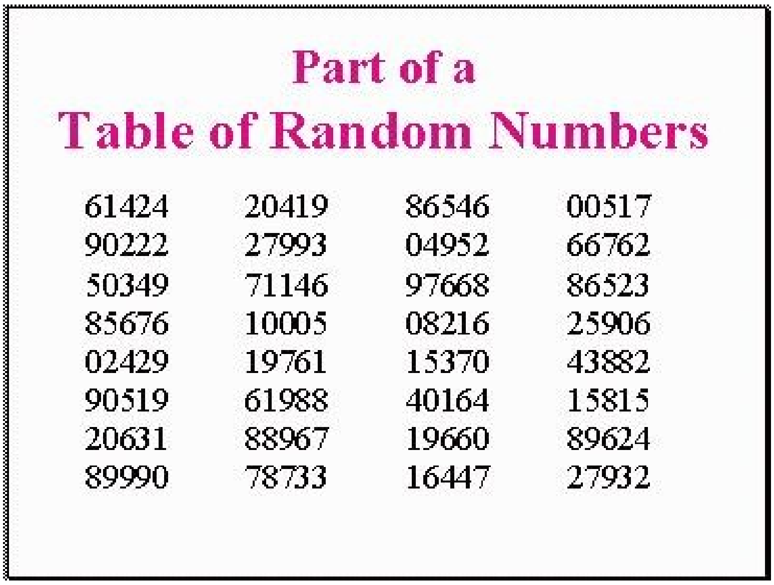
2.2.2 Random number tables

A random number table is an arrangement of ten digits from 0 to 9, occurring with equal frequencies independently of each other and without any consistently recurring trends or patterns. Several standard tables of random numbers prepared by Tippett (1927), Fisher and Yates (1938), Kendall and Smith (1939), Rand Corporation (1955), and Rao et al. (1974) are available.

Direct Approach. Again, the first step in the method is to assign serial numbers 1 to N to the N population units. If the population size N is made up of K digits, then consider K digit random numbers, either row wise or column wise, in the random number table. The sample of required size is then selected by drawing, one by one, random numbers from 1 to N, and including the units bearing these serial numbers in the sample.

This procedure may involve number of rejections of random numbers, since zero and all the numbers greater than N appearing in the table are not considered for selection. The use of random numbers has, therefore, to be modified. Two of the commonly used modified procedures are:

2.2.3 How to use a random number table.



Part of a
Table of Random Numbers

61424	20419	86546	00517
90222	27993	04952	66762
50349	71146	97668	86523
85676	10005	08216	25906
02429	19761	15370	43882
90519	61988	40164	15815
20631	88967	19660	89624
89990	78733	16447	27932

Figure 3

1. Let's assume that we have a population of 185 students and each student has been assigned a number from 1 to 185. Suppose we wish to sample 5 students (although we would normally sample more, we will use 5 for this example).

2. Since we have a population of 185 and 185 is a three digit number, we need to use the first three digits of the numbers listed on the chart.

3. We close our eyes and randomly point to a spot on the chart. For this example, we will assume that we selected 20631 in the first column.

4. We interpret that number as 206 (first three digits). Since we don't have a member of our population with that number, we go down to the next number 899 (89990). Once again we don't have someone with that number, so we continue at the top of the next column. As we work down the column, we find that the first number to match our population is 100 (actually 10005 on the chart). Student

number 100 would be in our sample. Continuing down the chart, we see that the other four subjects in our sample would be students 049, 082, 153, and 164.

5. Researchers use different techniques with these tables. Some researchers read across the table using given sets (in our examples three digit sets). For our class, we will use the technique I have described.

2.2.4 Computer software

In practice, the lottery method of selecting a random sample can be quite burdensome if done by hand. Typically, the population being studied is large and choosing a random sample by hand would be very time-consuming. Instead, there are several computer programs that can assign numbers and select n random numbers quickly and easily. Many can be found online for free.

2.3 Simple random sampling with replacement (SRSWR)

2.3.1 Definition and Estimation of Population Mean, Variance and Total

A sample is said to be selected by simple random sampling with replacement (srswr) by n draws from a population of size N if the sample is drawn by observing the following rule;

1. At each draw, each unit in the population has the same chance of being selected.

2. A unit selected at a draw is returned to the population before the next draw.

The same unit, therefore might be selected more than once. Thus the probability of getting a sample(sequence), $i = 1, 2, ..., i_n$ is;

$$P\left(\{i = 1, 2, ..., i_n\}\right) = \frac{1}{N}, ..., \frac{1}{N} = \frac{1}{N^n} \tag{2.1}$$

There are N^n possible samples(sequences) in the sample space S, for a given (N, n). A $srswr$ of n draws from a population of size N will be denoted by $srswr(N, n)$.

Example 2.1. For $N = 4$, $n = 2$ the following table shows all possible samples (sequences)

(1,1), (2,1),(3,1), (4,1)

(1,2), (2,2), (3,2), (4,2)

(1,3), (2,3), (3,3), (4,3)

(1,4), (2,4), (3,4), (4, 4)

Theorem 2.1. In $srswr(N, n)$ sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof.

$$E\left(\bar{y}\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)=y_1 \tag{2.2}$$

Since $y_1, y_2, ..., y_n$ are independently and identically distributed (iid) random variables with;

$$P\left(y_i=Y_k\right)=\frac{1}{N}, k=1,2,...,N, i=1,2,...,n. \tag{2.3}$$

Now, $E\left(y_1\right)=\frac{1}{N}\sum_{i=1}^{N}Y_k=\bar{Y}$. Hence, $E\left(\bar{y}\right)=\bar{Y}$.

Alternatively, let t_i be the number of times i occurs in the sample. Therefore, t_i follows a multinomial distribution with $E\left(t_i\right)=\frac{n}{N}$, $Var\left(t_i\right)=\frac{n}{N}\left(1-\left(\frac{1}{N}\right)\right)$,$Cov\left(t_i,t_j\right)=\frac{-n}{N^2}\left(i\neq j=1,2,...,N\right)$

(Show this)

Now, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{N}t_iY_i\Rightarrow E\left(\bar{y}\right)=E\left(\frac{1}{n}\sum_{i=1}^{N}t_iY_i\right)=\frac{1}{n}\sum_{i=1}^{n}Y_iE\left(t_i\right)$. But $E\left(t_i\right)=\frac{n}{N}$,

$\Rightarrow$

$$E\left(\bar{y}\right)=\frac{1}{n}\frac{n}{N}\sum_{i=1}^{N}Y_i=\bar{Y} \tag{2.4}$$

Hence $E\left(\bar{y}\right)=\bar{Y}$. □

Corollary 2.1.1. In $srswr\left(N,n\right)$ and unbiased estimator of Y is $\hat{Y}=N\bar{y}$.

Theorem 2.2. In $srswr(N,n)$, the sample variance is given by

$$Var\left(\bar{y}\right)=\frac{\sigma^2}{n}, \sigma^2=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right) \tag{2.5}$$

Proof. $=Var\left(\bar{y}\right)=Var\left(\frac{1}{n}\sum_{i=1}^{N}t_iY_i\right)$

$=\frac{1}{n^2}\sum_{i=1}^{N}Y_i^2Var\left(t_i\right)+\frac{1}{n^2}\sum\sum_{i\neq j}Y_iY_jCov\left(t_i,t_j\right)$

but $Var\left(t_i\right)=\frac{n}{N}\left(1-\left(\frac{1}{N}\right)\right)$

and $Cov\left(t_i,t_j\right)=\frac{-n}{N^2}\left(i\neq j=1,2,...,N\right)$

$\Rightarrow Var\left(\bar{y}\right)=\frac{1}{Nn}\left(1-\left(\frac{1}{N}\right)\right)\sum_{i=1}^{N}Y_i^2-\frac{1}{N^2n}\left[\left(\sum_{i=1}^{N}Y^i\right)^2-\sum_{i=1}^{N}Y_i^2\right]$

$=\frac{N}{N^2n}\sum_{i=1}^{N}Y_i^2-\frac{1}{N^2n}\sum_{i=1}^{N}Y_i^2+\frac{1}{N^2n}\sum_{i=1}^{N}Y_i^2-\frac{1}{N^2n}\sum_{i=1}^{n}\left(\sum_{i=1}^{N}Y_i\right)^2$

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

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

$$=\frac{\sigma^2}{n} \tag{2.6}$$

Note:

$\left(x_1+x_2\right)^2=x_1^2+2x_1x_2+x_2^2\Rightarrow\left(\sum_{i=1}^{N}x_i\right)^2=\sum_{i=1}^{N}x_i^2+\sum\sum_{i\neq j}x_ix_j$

□

Corollary 2.2.1. In $srswr\,(N,n)$, $Var\left(\hat{Y}\right) = \frac{N^2\sigma^2}{n}$. As n increases, $Var\,(\bar{y})$ decreases, even if $n = N$, $Var\,(\bar{y})$ does not vanish. Also in $srswr$, n may be arbitrarily large.

Assignment 1

1. In srswr $(N, n), n \geq 2$ and unbiased estimator of $V(\bar{y})$ is;

$$v(\bar{y}) = \frac{s^2}{n}, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{Y})^2 \tag{2.7}$$

2. The following data give the household sizes for 32 households in a village. Draw a simple random sample with replacement of 6 draws and hence obtain a estimate of the average household size along with its standard error. Obtain a 95% confidence interval for the average household size in the population. 5,3,7,11,4,6,10,9,8,12,11 10,10,11,8,7,6,8,9,4,1,5 7,7,12,8,9,10,9,7,6,8

3. For $N = 4$, $n = 2$ the following table shows all possible samples (sequences)

   (1,1), (2,1),(3,1), (4,1)

   (1,2), (2,2), (3,2), (4,2)

   (1,3), (2,3), (3,3), (4,3)

   (1,4), (2,4), (3,4), (4, 4)

   Show that $\bar{y}$, s^2 are unbiased estimators of $\bar{Y}$ and σ^2, respectively.

2.4 Simple random sampling without replacement

A sample of size n is said to be selected by simple random sampling without replacement (srswor) if the selection procedure is such that every possible sequence(sample) has the same chance of being selected. Sampling design is achieved by drawing a sample by the following draw-by-draw procedure;

1. At each draw each available unit in the population has the same chance of being selected.

2. A unit selected at a draw is removed from the population before the next draw.

If the population is of size N and we require a simple random sample without replacement of size n ,then this is chosen at random from $\begin{pmatrix} N \\ n \end{pmatrix}$ distinct sample. Each of the $\begin{pmatrix} N \\ n \end{pmatrix}$ samples has the same probability $\frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}}$ or $\begin{pmatrix} N \\ n \end{pmatrix}^{-1}$ of being selected.

Lemma 2.3. For a $srswor(N,n)$ design the probability of a specified unit being selected at any given draw is $\frac{1}{N}$ i.e.

$$P_r\left(i_k\right) = \frac{1}{N}, r = 1, 2, ..., n. \tag{2.8}$$

for any given i_k

Lemma 2.4. For a $srswor(N,n)$ the probability of two specified units being selected at any two given draws is $\frac{1}{N}\left(\frac{1}{N-1}\right)$, i.e.

$$P_{r,s}\left(i_r, i_s\right) = \frac{1}{N\left(N-1\right)}, r < s, r = 1, 2, ..., n. \tag{2.9}$$

for any given $i_r \neq i_s$

Lemma 2.5. For a $srswor(N,n)$ the probability that a specified unit is included in the sample is $\frac{n}{N}$ $i.e.$

$$P\left(i\epsilon s\right) = \pi_i\left(say\right), i = 1, 2, ..., N \tag{2.10}$$

Lemma 2.6. For a $srswor(N,n)$ the probability that any two specified units are included in the sample is $\frac{n(n-1)}{N(N-1)}$ $i.e.$

$$P\left(i\epsilon s, j\epsilon s\right) = \pi_{i,j}\left(say\right), i \neq j, i = 1, 2, ..., N \tag{2.11}$$

The quantities π_i and $\pi_{i,j}$ (as defined in 2.5 and 2.6) are respectively the inclusion probabilities of units i and (i,j) in the sample. These are called respectively, the first order and second order inclusion probabilities of a design.

2.4.1 Definition and estimation of population mean, variance and total

We consider the problem of estimating, $\bar{Y}$,Y and S^2 in srswor. Consider a population of size N and let n be the size of the simple random sample drawn from this population without replacement. Now let a_i equals 1 if the i^{th} unit is selected and 0 elsewhere, $i = 1, 2, ..., N$

Then a_i is a random variate such that; $E\left(a_i\right) = 1\times$ probability of i^{th} selected unit.

$= 1 \times \frac{n}{N} = \frac{n}{N}$ inclusion probability.

$E\left(a_i, a_j\right) = 1\times$ probability of the i^{th} and j^{th} unit selected.

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

Therefore the sample total is

$$y = \sum_{i=1}^{N} a_i Y_i = \sum_{i=1}^{n} y_i \tag{2.12}$$

The sample mean is given as;

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{N} a_i Y_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{2.13}$$

Theorem 2.7. In $srswor(N, n)$ the sample mean $\bar{y}$ is an unbiased estimator of the population mean $\bar{Y}$

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

$E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right) = \frac{1}{n}\sum_{i=1}^{N} Y_i E(a_i)$

$= \frac{1}{n}\sum_{i=1}^{N} Y_i . \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \bar{Y}$ □

Corollary 2.7.1. For $srswor(N, n), \hat{Y} = N\bar{y}$ is an unbiased estimator of the population total Y.

Theorem 2.8. In $srswor(N, n), Var(\bar{y}) = \frac{N-n}{Nn}S^2$, where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$

Proof. From $Var(y) = E\left(y^2\right) - \left(E(y)\right)^2$

it implies that $Var(\bar{y}) = E\left(\bar{y}^2\right) - \left(E(\bar{y})\right)^2$

But $E(\bar{y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right) = \frac{1}{n}\sum_{i=1}^{N} Y_i E(a_i) = \frac{1}{N}\sum_{i=1}^{N} Y_i$

Next, $E\left(\bar{y}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} a_i Y_i\right)^2$

$= E\left(\frac{1}{n^2}\sum_{i=1}^{N} a_i Y_i^2 + \frac{1}{n^2}\sum\sum_{i\neq j} a_i a_j Y_i Y_j\right)$

(Factor in expectation and use the fact that $E(a_i) = \frac{n}{N}$

and $E(a_i, a_j) = \frac{n(n-1)}{N(N-1)}$)

$= \frac{1}{nN}\sum_{i=1}^{N} Y_i^2 + \frac{n-1}{Nn(N-1)}\sum\sum_{i\neq j} Y_i Y_j$

But $\sum\sum_{i\neq j} Y_i Y_j = \left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2$

Therefore, $E\left(\bar{y}^2\right) = \frac{1}{Nn}\sum_{i=1}^{N} Y_i^2 + \frac{n(n-1)}{Nn(N-1)}\left[\left(\sum_{i=1}^{N} Y_i\right)^2 - \sum_{i=1}^{N} Y_i^2\right]$

$= \left[\frac{1}{nN} - \frac{n-1}{nN(N-1)}\right]\sum_{i=1}^{N} Y_i^2 + \frac{n-1}{Nn(N-1)}\left(\sum_{i=1}^{N} Y_i\right)^2$

Now $(x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2 \Rightarrow \left(\sum_{i=1}^{N} x_i\right)^2 = \sum_{i=1}^{N} x_i^2 + \sum\sum_{i\neq j} x_i x_j$

Therefore, $Var(\bar{y}) = \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N} Y_i^2 + \frac{n-1}{Nn(N-1)}\left(\sum_{i=1}^{N} Y_i\right)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2$

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

$= \frac{N-n}{Nn(N-1)}\sum_{i=1}^{N} Y_i^2 - \frac{N-n}{N^2n(N-1)}\left(\sum_{i=1}^{N} Y_i\right)^2$

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

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

Therefore;

$$Var(\bar{y}) = \frac{N-n}{Nn}S^2 \tag{2.14}$$

on simplification.

□

Theorem 2.9. In $srswor(N, n)$ an unbiased estimator of $Var\left(\bar{y}\right)$ is $\frac{N-n}{Nn}s^2$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$. Proof this.

Theorem 2.10. In $srswor(N, n)$ the sample variance is an unbiased estimator of the population variance i.e.

$E\left(s^2\right) = S^2$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$ and $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2$

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

$= \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 + \sum\sum_{i\neq j} y_i y_j\right)\right]$

$= \frac{1}{n-1}\left[\left(1 - \frac{1}{n}\right)\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\sum\sum_{i\neq j} y_i y_j\right]$

(open brackets and simplify)

$= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{1}{n(n-1)}\sum\sum_{i\neq j} y_i y_j$

Taking expectations on both sides we have,

$E\left(s^2\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n(n-1)}E\left(\sum\sum_{i\neq j} y_i y_j\right)$

but

$E\left(\sum_{i=1}^{n} y_i^2\right) = E\left(\sum_{i=1}^{N} a_i Y_i^2\right) = \frac{n}{N}\sum_{i=1}^{N} Y_i^2$

since $E\left(a_i\right) = \frac{n}{N}$

and $E\left(a_i a_j\right) = \frac{n(n-1)}{N(N-1)}$

and

$E\left(\sum\sum_{i\neq j} y_i y_j\right) = E\left(\sum\sum_{i\neq j} a_i a_j Y_i Y_j\right) = \frac{n}{N}\frac{n-1}{N-1}\sum\sum_{i\neq j} Y_i Y_j.$

Therefore;

$E\left(s^2\right) = \frac{1}{n}\frac{n}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{n(n-1)}\frac{n}{N}\frac{n-1}{N-1}\sum\sum_{i\neq j} Y_i Y_j$

$= \frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \frac{1}{N(N-1)}\sum\sum_{i\neq j} Y_i Y_j$

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

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

$= \frac{1}{N-1}\left[\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right] = S^2$

Hence;

$$E\left(s^2\right) = S^2 \tag{2.15}$$

□

Corollary 2.10.1. For $srswor(N, n)$, an unbiased variance estimator of Y is $Var\left(\hat{Y}\right) = \frac{N(N-n)}{n}s^2$

Proof. $Var\left(\hat{Y}\right) = Var\left(N\bar{y}\right) = N^2 Var\left(\bar{y}\right)$

$= N^2 \frac{N-n}{Nn}s^2$

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

which completes the proof. □

Corollary 2.10.2. An estimator of error of $\bar{y}$ is $\hat{\sigma}\left(\bar{y}\right) = \sqrt{\frac{N-n}{Nn}}s$. An estimator of the coefficient of variation is $c\left(\bar{y}\right) = \sqrt{\frac{N-n}{Nn}}\frac{s}{\bar{y}}$.

$c\left(\bar{y}\right)$ is a ratio estimator and biased estimator of $C\left(\bar{y}\right)$.

NOTE: The sample mean in $srswor(N, n)$ is a better estimator of $\bar{Y}$ (in the small variance sense) than sample mean in $srswr(N, n)$.

Proof. $Var\left(\bar{y}|srswr\right) - Var\left(\bar{y}|srswor\right) = \frac{n-1}{Nn}S^2 > 0$ for $n > 1$. □

In sampling from an infinite population(where each Y_i is an independently and identically distributed random variable) with variance of each random variable as σ^2, $Var\left(\bar{y}\right) = \frac{\sigma^2}{n}$. In simple random sampling with replacement, draws may be made an infinite number of times and $Var\left(\bar{y}\right) = \frac{\sigma^2}{n}$. In simple random sampling without replacement, however, $Var\left(\bar{y}\right) = \left[1 - \left(\frac{n}{N}\right)\frac{S^2}{n}\right]$.The quantity $1 - \frac{n}{N}$ appearing in the expression above is a correction factor for the finite size of the population and is called the finite population correction factor(fpc) or simply the finite multiplier. If n is very small compared to N, the fpc is close to unity and the sampling variance of $\bar{y}$ in srswor will be approximate the same as srswr. If N is very small say $N \leq 10$, then whatever n, f is not negligible and therefore there is considerable gain in using srswor over srswr.

NOTE: The finite population correction factor (fpc) is used when you **sample without replacement from more than 5% of a finite population**. It's needed because under these circumstances, the Central Limit Theorem doesn't hold and the standard error of the estimate (e.g. the mean or proportion) will be too big.

Example 2.2. For the following population, consider all possible srswor samples of size 3 and show that $\bar{y}$ and s^2 are respectively unbiased estimators of $\bar{Y}$ and S^2. Calculate the sampling variance of $\bar{y}$ and show that it agrees with the formula $\frac{N-n}{Nn}S^2$.

i: 1, 2, 3, 4, 5

Y_i: 5,8,3,11,9

2.5 Confidence intervals for population mean $\bar{Y}$ and Total Y

The sample mean $\bar{y}$ and the variance s^2 are point estimates of the unknown population mean and variance respectively. An interval estimate of unknown population parameter is a random interval constructed such that it has a given probability of including the parameters. Consider a population with unknown parameter, if one can find an interval (a, b) such that;

$$P(a \leq \theta \leq b) = 0.95 \tag{2.16}$$

then we say that (a, b) is a 95% confidence interval for θ. It is important to realize that the θ is fixed and the intervals themselves vary.

Some conditions exist under which the distribution of the sample mean in a simple random sampling tends to normal distribution. If the **sample size** is not too small and the distribution of the population from which the sample is drawn is not different from the **normal**, then in srswor, the sample mean $\bar{y}$ is approximately normal with mean $\bar{Y}$ and deviation $\frac{\sqrt{N-n}}{\sqrt{Nn}}S$ i.e.

$$\bar{y} \sim N\left(\bar{Y}, \frac{N-n}{Nn}S\right) \tag{2.17}$$

$$z = \frac{\bar{y}-\bar{Y}}{\sqrt{\frac{N-n}{Nn}S}} \sim N(0,1)$$

Hence $P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{y}-\bar{Y}}{\sqrt{\frac{N-n}{Nn}S}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$
$\Rightarrow P\left(\bar{y} - z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S \leq \bar{Y} \leq \bar{y} + z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S\right) = 1-\alpha$
where $z_{\frac{\alpha}{2}}$ is the $100\left[1-\frac{\alpha}{2}\right]\%$ point of normal distribution. Therefore;

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S, \bar{y} + z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{Nn}}S,\right) \tag{2.18}$$

is the $100\left[1-\frac{\alpha}{2}\right]\%$ confidence interval for $\bar{Y}$. For $\alpha = 0.05, 0.025, 0.01$ values for $z_{\frac{\alpha}{2}}$ are $1.96, 2.24$ and 2.58 respectively.

Example 2.3. In a private library, the books are kept on 130 shelves of similar size. The numbers of books on 15 shelves picked at random were found to be 28,23,25,3 3,31,18,22,29,30,22,26,20,21,28 and 25. Estimate the total number Y, of books in the library and calculate an approximate 95% confidence interval for Y.

Solution 1. $N = 130, n = 15, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{15}(28 + 23+, ... + 25) = 25.4$

$Y = N\bar{y}, = 130 \times 25.4 = 3302$. The 95% confidence interval is given by;

$Y \pm Nz_{0.05}\sqrt{var(\bar{y})}$

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

which is estimated by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)$

$\sum_{i=1}^{15} y_i^2 = \left(28^2 + 23^2+, ..., +25^2\right) = 9947$

$n\bar{y}^2 = 15 \times (25.4)^2$. The 95% confidence interval for Y at $\alpha = 0.05$ will be;

$\hat{Y} = Y \pm N z_{0.05}\sqrt{var\left(\bar{y}\right)}$

$= 3302 \pm 130\left(1.96 \times \sqrt{1.14}\right)$

$= 3302 \pm 272.05$

$\Rightarrow 3029.05 \leq Y \leq 3574.05$

2.6 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'. 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.

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.

2.6.1 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;

$$P = \frac{A}{N} \tag{2.19}$$

and the proportion of units in $C*$ is

$$Q = \frac{N-A}{N} = 1 - P \tag{2.20}$$

An indicator variable Y can be associated with the characteristics under study and then for $i = 1, 2, ..., N$. $Y_i = 1$ if the i^{th} unit belongs to C and 0 if the i^{th} unit belongs to $C*$. Now the population total is;

$$Y_{TOTAL} = \sum_{i=1}^{N} Y_i = A \tag{2.21}$$

and the population mean is;

$$\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N} = \frac{A}{N} = P \tag{2.22}$$

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;

$$p = \frac{a}{n} \tag{2.23}$$

which can be written as $p = \frac{a}{n} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$.

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

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

$= \frac{1}{N-1} \sum_{i=1}^{N} \left(NP - NP^2\right) = \frac{N}{N-1} PQ$

Similarly; $\sum_{i=1}^{n} y_i^2 = a = np$ and $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i^2 - n\bar{y}^2\right)$

$= \frac{1}{n-1} \sum_{i=1}^{n} \left(np - np^2\right)$

$= \frac{n}{n-1} pq$

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.

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

$E\left(p\right) = E\left(\bar{y}\right) = \bar{Y} = P$ and p is an unbiased estimator of P. Using the expression of $Var\left(\bar{y}\right)$ the variance of p can be derived as $Var\left(p\right) = Var\left(\bar{y}\right) = \frac{N-n}{Nn} S^2$

Similarly, using the estimate of variance can be derived as

$\widehat{Var}\left(p\right) = \widehat{Var}\left(\bar{y}\right) = \frac{N-n}{Nn} S^2$

$= \frac{N-n}{Nn} \cdot \frac{n}{n-1} pq$

$$= \frac{N-n}{N\left(n-1\right)} pq \tag{2.24}$$

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

$E\left(p\right) = E\left(\bar{y}\right) = \bar{Y} = P$ 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: $Var\left(p\right) = Var\left(\bar{y}\right) = \frac{N-1}{Nn} S^2$

$= \frac{N-1}{Nn} \frac{N}{N-1} PQ$

$= \frac{PQ}{n}$

$\Rightarrow \widehat{Var}\left(p\right) = \frac{n}{n-1} \cdot \frac{pq}{n}$

$$= \frac{pq}{n-1} \tag{2.25}$$

2.6.2 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 $Var\left(\hat{A}\right) = N^2 Var\left(p\right)$ and the estimate of variance is $\widehat{Var}\left(\hat{A}\right) = N^2 \widehat{Var}\left(p\right)$

2.6.3 Confidence Interval estimation for P

If N and n are large, then $\frac{p-P}{\sqrt{Var(p)}}$ approximately follows $N\left(0,1\right)$. With this approximation we can write $P\left(-z_{\frac{\alpha}{2}} \leq \frac{p-P}{\sqrt{Var(p)}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$, and the $100(1-\alpha)$ confidence interval of P is

$$p - z_{\frac{\alpha}{2}} \sqrt{Var\left(p\right)}, p + z_{\frac{\alpha}{2}} \sqrt{Var\left(p\right)} \tag{2.26}$$

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

$$p - z_{\frac{\alpha}{2}} \sqrt{Var\left(p\right)} + \frac{n}{2}, p + z_{\frac{\alpha}{2}} \sqrt{Var\left(p\right)} + \frac{n}{2} \tag{2.27}$$

2.7 Determination of sample sizes

In a field survey the statisticians would like to have a sample size that will give a desired level of precision of estimator. We note that the required precision is the difference between the estimator and the true value. This difference is denoted by d. Suppose that it is desired to find a sample size n such that the estimated value i.e. sample mean $\bar{y}$ differs from the true value (Population mean, $\bar{Y}$) by a quantity not exceeding d with a very high probability, say greater than $1-\alpha$. Hence the problem is to find n such that;

$$P\left(\mid \bar{y}-\bar{Y} \mid \leq d\right) \geq 1-\alpha \tag{2.28}$$

From srswor $\bar{y} \sim N\left(\bar{Y}, \frac{N-n}{Nn}S^2\right)$. Hence,

$$P\left(\mid \bar{y}-\bar{Y} \mid \leq S\sqrt{\frac{N-n}{Nn}}t\right) = 1-\alpha \tag{2.29}$$

where $t = z_{\frac{\alpha}{2}}$ is the 100 $\left(1-\frac{\alpha}{2}\right)$ point of normal distribution.

From equation 2.28 and 2.29,

$$tS\sqrt{\frac{N-n}{Nn}} = d \tag{2.30}$$

when

$$\frac{1}{n} = \frac{1}{N} + \frac{d^2}{t^2S^2} \tag{2.31}$$

Hence,

$$n = \frac{\left(\frac{ts}{d}\right)^2}{1+\frac{1}{N}\left(\frac{tS}{d}\right)^2} \tag{2.32}$$

As a first approximation we may take

$$n_o = \left(\frac{ts}{d}\right)^2 \tag{2.33}$$

If $\frac{n_0}{N}$ is negligibly small, this may be taken as the satisfactory value of n. If not, one should calculate

$$n = \frac{n_0}{1+\left(\frac{n_0}{N}\right)} = n_0\left((1+\frac{n_0}{N}\right)^{-1} \tag{2.34}$$

In practice one has to replace S by an advance estimate $s^{'}$ (say).

In case the problem is that of estimating a population proportion one may require to find n such that

$$P\left(\mid p-P \mid \leq d\right) \geq 1-\alpha \tag{2.35}$$

For large samples in srswor $\frac{p-P}{\sqrt{\left\{\frac{N-n}{n(N-1)}PQ\right\}}}$ is approximately a normal variable. Hence;

$$P\left(\mid p-P\mid\leq t\sqrt{\frac{N-n}{n\left(N-1\right)}}PQ\right)=1-\alpha \tag{2.36}$$

Equating 2.35 and 2.36 we get,

$$t\sqrt{\frac{N-n}{n\left(N-1\right)}PQ}=d. \tag{2.37}$$

This gives;

$$n=\frac{\left(\frac{t^2PQ}{d^2}\right)}{1+\left(\frac{1}{N}\right)\left[\left(\frac{t^2PQ}{d^2}\right)-1\right]} \tag{2.38}$$

For practical purposes, P is to be replaced by some suitable estimate p of the same. For large N a first approximation of n is

$$n_0=\frac{t^2PQ}{d^2}. \tag{2.39}$$

If $\frac{n_0}{N}$ is negligible, n_0 is a satisfactory approximation to n. If not, one should calculate n as;

$$n=\frac{n_0}{1+\left[\left(\frac{n_{0-1}}{N}\right)\right]}\approx\frac{n_0}{1+\left(\frac{n_0}{N}\right)} \tag{2.40}$$

Example 2.4. Suppose it is required to estimate the average value of output of a group of 5000 factories in a region so that the sample estimate lies within 10 of the true value with a confidence coefficient of 95%. Determine the minimum sample size required. The population coefficient of variation is known to be 60%.

Solution 2. We require n such that

$P\left(\mid \bar{y}-\bar{Y}\mid\leq 0.1\bar{Y}\right)=0.95.$

Now under normal approximation, $P|\bar{y}-\bar{Y}|\leq 1.96\sqrt{\frac{N-n}{nN}}S=0.95$

Hence, $1.96S\sqrt{\frac{N-n}{Nn}}=0.1\bar{Y}$

or $(1.96)^2\left(\frac{1}{n}-\frac{1}{N}\right)=0.01\left[\frac{\bar{Y}}{S}\right]^2=\frac{0.01}{0.36}$

Solving the above equation, we get $n=136$ (rounded off to the next integer)

Example 2.5.

Consider the population consisting of 430 units. By complete enumeration of the population it was found that $\bar{Y} = 19$, $S^2 = 85.6$ These being true population values with simple random samples, how many units must be taken to estimate $\bar{y}$ with 10% of $\bar{Y}$ a part from a chance of 1 in 20.

Solution 3. $\bar{Y} = 19, S^2 = 85.6 \Rightarrow S = \sqrt{85.6}\ N = 430,\ d = \frac{1}{20} = 0.05.\ 10\%$

of $\bar{Y} \Rightarrow d = 0.1\bar{Y} = 0.1\,(19) = 1.9.$

$n_0 = \left(\frac{ts}{d}\right)^2$

but $t = z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = 1.96.$

$\Rightarrow n_0 = \frac{(1.96)^2(85.6)}{1.9^2} = 91.09167.$

$n = n_0\left(1 + \frac{n_0}{N}\right)^{-1}, = 91.09\left[\left(1 + \frac{91.09}{430}\right)\right]^{-1} = 75.166 \simeq 75.$

2.8 Exercises

1. In a population with $N = 6$, the values of y_i are 8, 3, 1, 11, 4, and 7. Calculate the sample mean $\bar{y}$ for al1 possible simple random samples of size 2. Verify that $\bar{y}$ is an unbiased estimate of $\bar{Y}$.

2. For the same population in 1 above, calculate s^2 for al1 simple random samples of size 3, and verify that $E\left(s^2\right) = S^2$.

3. If random samples of size 2 are drawn with replacement (from this population, show by finding all possible samples that $Var\left(\bar{y}\right)$ satisfies the equation $Var\left(\bar{y}\right) = \frac{\sigma^2}{n} = \frac{S^2(N-1)}{nN}$. Give a general proof of this result.

4. A simple random sample of 30 households was drawn from a city area containing 14,848 households.The numbers of persons per household in the sample were as follows: 5,6,3,3,2,3,3,3,4,4,3,2,7,4,3,5, 4,4,3,3,4,3,3, 1,2,4,3,4,2,4. Estimate the total number of people in the area and compute the probability that this estimate is within Â±10 per cent of the true value.

5. Consider a population consisting of 6 villages, the areas (in hectares) of which are given below;

Table 2: Population of 6 villages

Village	A	B	C	D	E	F
Area	760	343	657	550	480	935

   (a) Enumerate all possible WR samples of size 3. Also, write the values of the study variable for the sampled units.

   (b) List all the WOR samples of size 4 along with their area values.

6. Among the 100 computer corporations in a region, average of the employee sizes for the largest 10 and smallest 10 corporations were known to be 300 and 100, respectively. For a sample of 20 from the remaining 80 corporations, the mean and standard deviation were 250 and 110, respectively. For the total employee size of the 80 corporations, find;

   (a) the estimate

   (b) the S.E. of the estimate

   (c) the 95% confidence limits

7. Continuing with Exercise 2, for the average and total of the 100 corporations, find;

   (a) the estimate

   (b) the S.E. of the estimate

   (c) the 95% confidence limits.

8. The height (in cm) of 6 students of M.Sc., majoring in statistics, from Punjab Agricultural University, Ludhiana was recorded during 1985. The data, so obtained, are given below:

Table 3: Heights of M.Sc. students

Student	Name	Height
1	A	168
2	B	175
3	C	185
4	D	173
5	E	171
6	F	172

Calculate;

   (a) Calculate the population mean $\bar{Y}$ and population variance σ^2.

   (b) Enumerate all possible SRS with replacement samples of size $n = 2$. Obtain sampling distribution of mean, and hence show that:

      i. $E[\bar{y}]$

      ii. $V[\bar{y}] = \frac{\sigma^2}{n}$

      iii. $E[s^2] = \sigma^2$

      iv. $E[v(\bar{y})] = V(\bar{y})$

   (c) Enumerate all possible SRS without replacement samples of size $n = 2$. Obtain sampling distribution of mean, and hence show that:

      i. $E[\bar{y}]$

ii. $V[\bar{y}] = \frac{\sigma^2}{n}$

iii. $E[s^2] = \sigma^2$

iv. $E[v(\bar{y})] = V(\bar{y})$

9. Punjab Agricultural University, Ludhiana, is interested in estimating the proportion P of teachers who consider semester system to be more suitable as compared to the trimester system of education. A with replacement simple random sample of n= 120 teachers is taken from a total of N=1200 teachers. The response is denoted by 0 if the teacher does not think the semester system suitable, and 1 if he/she does.

Table 4: Punjab Agricultural University

Teacher	1	2	3	4	5	6	...	119	120	Total
Response	1	0	1	1	0	1	...	0	1	72

(a) From the sample observations given below, estimate the proportion P along with the standard error of your estimate. Also, work out the confidence interval for P.

(b) While estimating P, the investigator feels that the tolerable error could be taken as 0.08. Do you think the sample size 120 is sufficient? If not, how many more units should be included in the sample?

3 Stratified Random Sampling

Learning objectives

By the end of this lesson learners should be able to;

(a) *Define stratified random sampling.*

(b) *Estimate the mean, total and variance under stratified random sampling.*

(c) *Allocate sample sizes under stratified random sampling.*

(d) *Construct confidence interval for mean and total under stratified random sampling.*

(e) *Apply stratified random sampling in R.*

3.1 Introduction

The objective of any sampling method is usually to estimate the unknown population parameters with the highest precision i.e. the variance of the estimators should be minimized. If the population is **heterogeneous** as will be in most situations then a sample taken via SRS might yield high levels of variability. As a result in a survey where precision is a main factor to be considered, then a strategy that addresses heterogeneity must be found.

One way of achieving higher precision is to divide the population which is originally **heterogeneous** into sub population which are to a big extent **homogeneous** with respect to survey characteristics.

In stratified random sampling, the population of N units is first divided into sub-populations $N_1, N_2, ..., N_L$ called **strata** (singular: stratum). The strata are mutually disjoint so that;

$$N_1 + N_2+, ..., +N_L = \sum_{i=1}^{L} N_i = N \tag{3.1}$$

It is important that the number of units in the stratum denoted by $N_i, i = 1, 2, ..., L$ is **known** in order to maximize the gain from stratification. After determining the strata, a sample of size $n_i, i = 1, 2, ..., L$ is drawn from each stratum. If simple random sampling procedure is used to obtain the sub-samples in each stratum then the whole procedure is called under stratified random sampling.

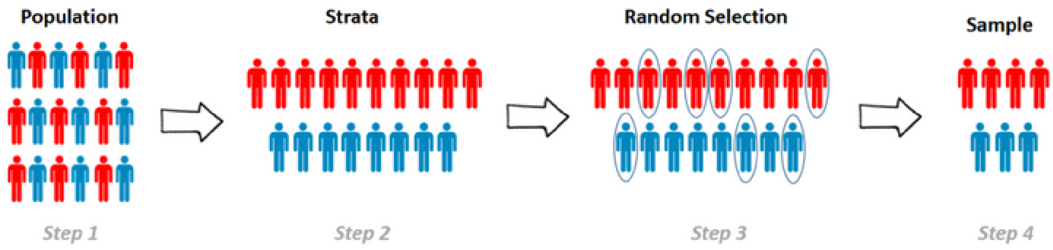


Figure 4: Illustration: stratified random sampling

The basic idea of stratification is that it may be possible to divide heterogeneous population into sub-populations which are internally homogeneous. If each sub-population is homogeneous, a precise estimate of any stratum can be obtained from a small sample of each stratum. This results in an improvement on the precision of the entire estimate.

Example 3.1. In order to find the average height of the 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 are of age around 16 years. So one can divide all the students into different sub-populations or strata such as,

Table 5: Average height of students

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

Now draw the samples by SRS from each of the strata 1, 2, 3 and 4. All the drawn samples combined together will constitute the final stratified sample for further analysis.

Notations: The following is an extension of previous notation used where the suffix i denote the stratum and j denote the j^{th} unit within the stratum.

Let Y_{ij} be the value of the characteristic y on the j^{th} unit in the i^{th} stratum in the population; y_{ij} value in the sample; $j = 1, 2, ..., N_i$ (n_i in the sample), $i = 1, 2, ..., L$

Define:

$N_i =$ Total number of units in the i^{th} stratum

$n_i =$ the number of units in the sample of the i^{th} stratum.

Note: $j = 1, 2, ..., N \rightarrow units$ in a stratum; $i = 1, 2, ..., L \rightarrow strata$

$n = \sum_{i=1}^{L} n_i =$ total sample size from all the strata

$Y_i = \sum_{i=1}^{N_i} Y_{ij} =$ population total for the i^{th} stratum.

$y_i = \sum_{i=1}^{n_i} y_{ij}$ = sample total for the i^{th} stratum.

$\bar{y}_i = \frac{y_i}{n_i}$ = sample mean for the i^{th} stratum.

$\bar{Y} = \sum_{i=1}^{L} \frac{N_i \bar{Y}_i}{N} = \frac{Y}{N}$ = overall population mean.

$S_i^2 = \frac{1}{N_i - 1} \sum_{i=1}^{N_i} \left(Y_{ij} - \bar{Y}\right)^2$ = population variance for the i^{th} stratum.

$s_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}\right)^2$ = sample variance for the i^{th} stratum.

$W_i = \frac{N_i}{N}$ = population proportion for the i^{th} stratum or stratum weight and

$f_i = \frac{n_i}{N_i}$ = sampling fraction for the i^{th} stratum.

Note: The divisor of the variance is $(N_i - 1)$

3.1.1 Estimation of Population Mean, Variance and Total

The mean of the target population is given by;

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{N_i} Y_{ij} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{Y}_i \tag{3.2}$$

where $N = N_1 + N_2 + ... + N_L$.

For the population mean per unit, the estimate used in stratified sampling is $\bar{y}_{st}$ (st for stratified), where $\bar{y}_{st} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i = \sum_{i=1}^{L} W_i \bar{y}_i$. $\left(W_i = \frac{N_i}{N}\right)$

Note: The estimate $\bar{y}_{st}$ is not in general the same as the sample mean. The sample mean $\bar{y}$ can be written as $\bar{y} = \frac{1}{n} \sum n_i \bar{y}_i$. The difference is that in $\bar{y}_{st}$ the estimates from the individual strata receive their correct weights $\frac{N_i}{N}$. It is evident that $\bar{y}$ coincides with $\bar{y}_{st}$ provided that in every stratum, $\frac{n_i}{n} = \frac{N_i}{N}$ or $\frac{n_i}{N_i} = \frac{n}{N} = f_i = f$. This means the sampling fraction is the same in all strata.

The principal properties of the estimate $\bar{y}_{st}$ are outlined in the following theorems. If simple random sample is used in each stratum then, $\bar{y}_{st}$ has the following properties.

Theorem 3.1. In stratified random sampling $\bar{y}_{st} = \sum_{i=1}^{L} \frac{N_i \bar{y}_i}{N} = \sum W_i \bar{y}_i$ is an unbiased estimator of the population mean $\bar{Y}$.

Proof. $E\left(\bar{y}_{st}\right) = \sum_{i=1}^{L} \frac{W_i E(\bar{y}_i)}{N} = \bar{Y}$ □

Theorem 3.2. In stratified random sampling using srswor in each stratum $Var\left(\bar{y}_{st}\right) = \frac{1}{N_2} \sum_{i=1}^{L} N_i^2 Var\left(\bar{y}_i\right) = \frac{1}{N^2} \sum_{i=1}^{L} \frac{N_i(N_i - n_i)}{ni} S_i^2$

Proof. $Var\left(\bar{y}_{st}\right) = Var\left(\sum_{i=1}^{L} \frac{N_i \bar{y}_i}{N}\right) = \sum_{i=1}^{L} \frac{N_i^2}{N^2} Var\left(\bar{y}_i\right)$
$= \sum_{i=1}^{L} \frac{N_i^2}{N^2} \left(\frac{N_i - n_i}{N_i}\right) \frac{S_i^2}{n_i}$.

Covariances terms vanish being independent from stratum to stratum

$\frac{1}{N^2}\sum_{i=1}^{L}\frac{N_i(N_i-n_i)}{n_i}S_i^2$ □

Corollary 3.2.1. If sampling fraction $\frac{n_i}{N_i}$ is negligibly small in each stratum, it reduces to

$Var\left(\bar{y}_{st}\right)=\frac{1}{N^2}\sum_{i=1}^{L}\frac{N_i^2S_i^2}{n_i}=\sum_{i=1}^{L}\frac{W_iS_i^2}{n_i}$

Corollary 3.2.2. If $\hat{Y}_{st}=N\bar{y}_{st}$ is the estimate of the population total Y then $Var\left(\hat{Y}_{st}\right)=\sum_{i=1}^{L}N_i\left(N_i-n_i\right)\frac{S_i^2}{n_i}$

Proof. $\hat{Y}_{st}=N\bar{y}_{st}$

$\Rightarrow Var\left(\hat{Y}_{st}\right)=Var\left(N\bar{y}_{st}\right)$

$=N^2Var\left(\bar{y}_{st}\right)$

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

$$=\sum_{i=1}^{L}N_i\left(N_i-n_i\right)\frac{S_i^2}{n_i} \tag{3.3}$$

□

3.1.2 Estimation of Variance and Confidence Intervals for the mean

In simple random sampling, the estimate of the variance of each stratum is given by $s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2$ for the i^{th} stratum.

We have found that,

$$Var\left(\bar{y}_{st}\right)=\frac{1}{N^2}\sum_{i=1}^{L}N_i\left(N_i-n_i\right)\frac{S_i^2}{n_i}. \tag{3.4}$$

In stratified random sampling, the unbiased estimate of the variance of $Var\left(\bar{y}_{st}\right)$. is given by,

$$s_{st}^2=\frac{1}{N^2}\sum_{i=1}^{L}N_i\left(N_i-n_i\right)\frac{s_i^2}{n_i}. \tag{3.5}$$

If $\bar{y}_{st}$ is normally distributed over $\bar{Y}$ then the confidence interval for $\bar{Y}$ is given by $\left[\bar{y}_{st}-z_{\frac{\alpha}{2}}S_{\bar{y}_{st}},\bar{y}_{st}+z_{\frac{\alpha}{2}}S_{\bar{y}_{st}}\right]$. Therefore,

$$\bar{y}_{st}\pm z_{\frac{\alpha}{2}}\sqrt{Var\left(\bar{y}_{st}\right)} \tag{3.6}$$

3.2 Allocation problem and choice of sample sizes is different strata

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

3.2.1 Equal allocation

Choose the sample size n to be the same for all the strata. Draw samples of equal size from each strata. Let n be the sample size and k be the number of strata, then $n_i = \frac{n}{k}$ for $i = 1, 2, ..., L$

3.2.2 Proportional Allocations

If the sample sizes in the strata are closer such that $\frac{n_i}{n} = \frac{N_i}{N} = constant$, then the stratification is defined as stratification with proportion al allocation for n_i for $i = 1, 2, ..., L$. Consider, $\bar{y}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i \bar{y}_i$, then with proportional allocation, $\bar{y}_{st} = \sum_{i=1}^{L} \frac{Nn_i}{Nn} \bar{y}_i = \frac{1}{n}\sum_{i=1}^{L} n_i \bar{y}_i$ overall mean. In this case, $\bar{y}_{st}$ coincides with $\bar{y}$ (the overall sample mean),

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

Now using proportional allocation the variance is;

$Var\left(\bar{y}_{st}\right)_{prop} = \frac{1}{N^2} N_i \left(N_i - \frac{nN_i}{N}\right) \frac{S_i^2}{\frac{nN_i}{N}}$

$= \frac{1}{N^2}\sum_{i=1}^{L} N_i \left(\frac{NN_i - nN_i}{N}\right) \frac{NS^2}{nN_i}$

$= \frac{1}{N^2}\sum_{i=1}^{L} \left(NN_i - nN_i\right) \frac{S_i^2}{n}$

$$\frac{N-n}{N^2 n}\sum_{i=1}^{L} N_i S_i^2 \tag{3.7}$$

which is the formula for the $Var\left(\bar{y}_{st}\right)$ under proportional allocation.

3.2.3 Optimal/Neymann Allocation

This allocation considers the size of strata as well as variability; $n_i \propto N_i S_i$, $n_i = C^* N_i S_i$ where C* is the constant of proportionality.

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

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

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

$n_i = \frac{n N_i S_i}{\sum_{i=1}^{L} N_i S_i}$.

This allocation arises when the $Var\left(\bar{y}_{st}\right)$is minimized subject to the constraint $\sum_{i=1}^{L} n_i$(prespecified). There are some limitations of the optimum allocation. The knowledge of $S_i, i = 1, 2, ..., L$ is needed to know n_i. If there are more than one characteristics, then they may lead to conflicting allocation.

3.2.4 Variances under different allocations

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

(i) under Proportional allocation

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

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

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

$$= \frac{N-n}{Nn} \sum_{i=1}^{L} w_i S_i^2 \qquad (3.8)$$

(ii) under Optimum allocation

Under optimum allocation;

$n_i = \frac{n N_i S_i}{\sum_{i=1}^{L} N_i S_i}$.

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

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

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

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

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

Example 3.2. A population of size 800 is divided into three strata. Their sizes and deviations are as given below. A sample of 120 is to be drawn from the population. Determine the sample size based on;

Table 6: Population

Stata	1	2	3
Size of N_i	200	300	300
Standard deviation S_i	6	8	12

(a) Proportional allocation

(b) Optimum allocation

(c) Obtain the variance of the estimates of the population mean i.e. $Var_{prop}\left(\bar{y}_{st}\right)$ and $Var_{opt}\left(\bar{y}_{st}\right)$

Solution 4. $\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = \frac{nN_i}{N}$, $n = 120 = \sum_{i=1}^{L} n_i$, $N = N_1 + N_2 + N_3 = 200 + 300 + 300 = 8000$

Therefore under proportional allocation, $n_1 = \frac{nN_1}{N} = \frac{120\times200}{800} = 30$,

$n_2 = \frac{nN_2}{N} = \frac{120\times300}{800} = 45$, $n_3 = \frac{nN_3}{N} = \frac{120\times300}{800} = 45$.

Under optimal allocation, $n_i = \frac{nN_iS_i}{\sum_{i=1}^{L} N_iS_i}$, $\sum_{i=1}^{L} N_iS_i = 200(6) + 300(8) + 300(12) = 72,000$,

$n_1 = \frac{nN_1S_1}{72,00} = \frac{120\times200\times6}{7200} = 20$, $n_2 = \frac{120\times300\times8}{7200} = 40$, $n_3 = \frac{120\times300\times12}{7200} = 60$.

$Var\left(\bar{y}_{st}\right)_{prop} = \frac{N-n}{N^2n}\sum_{i=1}^{L} N_iS_i^2$, $\sum_{i=1}^{L} N_iS_i^2 = 200\left(6^2\right) + 300\left(8^2\right) + 300\left(12^2\right) = 69,600$,

$\Rightarrow \frac{800-120}{(800)^2(120)}\left(69,600\right) = \frac{680}{64,000(120)}\left(69,600\right) = 0.61625$,

$Var\left(\bar{y}_{st}\right)_{opt} = \frac{1}{N^2}\left[\frac{1}{n}\left(\sum_{i=1}^{L} N_iS_i\right)^2 - \sum_{i=1}^{L} N_iS_i^2\right] = \frac{1}{800^2}\left[\frac{1}{120}\left(7200\right)^2 - 69,600\right] = 0.56626$.

Note: $Var\left(\bar{y}_{st}\right)_{opt} < Var\left(\bar{y}_{st}\right)_{prop}$.

Example 3.3. (a) A market researcher is allocated Ksh. 20,000 to conduct a survey by means of stratified random sampling. The population consists of stratum A of size 40,000, B of size 20,000 and C of size 10,000. The set cost of administering the survey is 200 and the cost of sampling one unit are 2.25, 4.00 and 1.00 for stratum A, B and C respectively. The deviations of observations in stratum A is thought to be twice that of stratum B and C. Find the optimum and proportional allocations, assuming that all the money is to be spent on the survey.

Solution 5. $N_1 = 40,000$, $N_2 = 20,000$, $N_3 = 10,000$, $c_0 = 200$ (fixed cost), $c = 20,000$, $c_1 = 2.25$, $c_1 = 4.0$,$c_1 = 1.0$, $A = 2S_3, B = S_3, C = S_3$.

For optimum allocation, we need to find the size of the sample in each of the stratum i.e.

$n_i = \frac{\frac{(c-c_0)N_iS_i}{\sqrt{c_i}}}{\sum_{i=1}^{3} N_iS_i\sqrt{c_i}}$.

Now, $\sum_{i=1}^{3} N_iS_i\sqrt{c_i} = 40,000\,(2S_3)\left(\sqrt{2.25}\right) + 20,000\,(S_3)\left(\sqrt{4}\right) + 10,000\,(S_3)\sqrt{1} = 170,000S_3$.

$n_1 = \frac{\frac{(20,000-200)(40,000)2S_3}{1.5}}{170,000S_3} \simeq 6211.$

$n_2 = \frac{\frac{(20,000-200)(20,000)S_3}{2}}{170,000S_3} \simeq 1164.7.$

$n_3 = \frac{\frac{(20,000-200)(10,000)S_3}{2}}{170,000S_3} \simeq 1164.7.$

Under proportional allocation,

$\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = \frac{nN_i}{N},$

$c = c_0 + \sum_{i=1}^{L} c_in_i = c_0 + \sum_{i=1}^{L} ci\frac{nN_i}{N},$

$c_0 + \frac{n}{N}\sum_{i=1}^{L} C_iN_i \Rightarrow c - c_0 = \frac{n}{N}\sum_{i=1}^{L} C_iN_i \Rightarrow n = \frac{N(c-c_0)}{\sum_{i=1}^{L} C_iN_i}, N = N_1 + N_2 + N_3 = 70,000.$

$\Rightarrow n = \frac{70,000(20,000-2,000)}{(2.25)(40,000)+4(20,000)+1(10,000)} = 77,000.$

Therefore,

$n_i = \frac{nN_i}{N} \Rightarrow n_1 = \frac{7700(40,000)}{70,000} = 4,400,$

$n_2 = \frac{7700(20,000)}{70,000} = 2200,$

$n_3 = \frac{7700(10,000)}{70,000} = 1,100.$

3.3 Exercises

(a) Given a population $U = 1, 2, 3, 4$ and $y_1 = y_2 = 0, y_3 = 1, y_4 = -1$, the values taken by the characteristic y.

  i. Calculate the variance of the mean estimator for a simple random design without replacement of size $n = 2$.

  ii. Calculate the variance of the mean estimator for a stratified random design for which only one unit is selected per stratum and the strata are given by $U_1 = 1, 2$ and$U_2 = 3, 4$.

(b) A sample of 30 students is to be drawn from a population of 300 students belonging to two colleges A and B. The means and deviations of their marks are given below. Use the information to confirm that Neyman allocation scheme is a more efficient scheme when compared to proportional allocation.

(c) A stratified population has 5 strata. The stratum sizes N_i and means $\bar{Y}_i$ and S_i^2 of some variable Y are as follows.

Table 7: A sample of 30 students

	Number of students	Mean	SD
College A	200	30	10
College B	100	60	40

Table 8: Stratified population

$Stratum$	N_i	$\bar{Y}_i$	S_i^2
1	117	7.3	1.31
2	98	6.9	2.03
3	74	11.2	1.13
4	41	9.1	1.96
5	45	9.6	1.74

i. Calculate the overall population mean and variance.

ii. For a stratified simple random sample of size 80, determine the appropriate stratum sample sizes under proportional allocation and Neyman allocation.

(d) Among the 7500 employees of a company, we wish to know the proportion P of them that owns at least one vehicle. For each individual in the sampling frame, we have the value of his income. We then decide to construct three strata in the population: individuals with low income (stratum 1), with medium income (stratum 2), and with high income (stratum 3). We denote:

N_h = the stratum size h,

n_h = the sample size in stratum h (simple random sampling),

p_h = the estimator of the proportion of individuals in stratum h owning at least one vehicle.

The results are given in Table 9

Table 9: Employees according to income

	h=1	h=2	h=3
Nh	3500	2000	2000
nh	500	300	200
ph	0.13	0.45	0.5

i. What estimator $\hat{P}$ of P do you propose? What can we say about its bias?

ii. Calculate the accuracy of $\hat{P}$, and give a 95% confidence interval for P.

iii. Do you consider the stratification criteria to be adequate?

4 Cluster sampling

4.1 Introduction

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;

(a) Divide the whole population into clusters according to some well defined rule.

(b) Treat the clusters as sampling units.

(c) Choose a sample of clusters according to some procedure.

(d) Carry out a complete enumeration of the selected clusters, i.e., collect information on all the sampling units available in selected clusters.

Example 4.1. 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.

Example 4.2. 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.

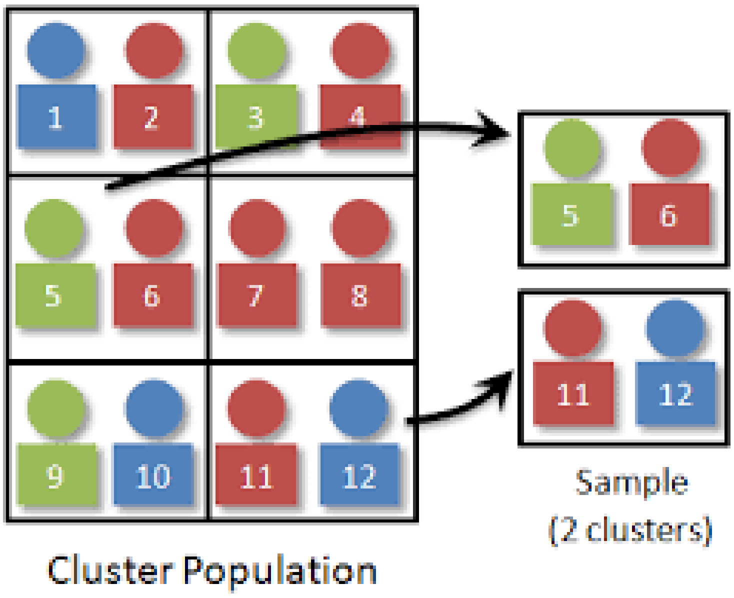


Figure 5: Illustration: cluster sampling

Conditions under which the cluster sampling is used

Cluster sampling is preferred when;

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

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

(c) 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.

4.2 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} be the 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.

4.2.1 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; $E\left(\bar{y}_{cl}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(\bar{y}i\right)$

or $\frac{1}{n}\sum_{i=1}^{n} \bar{Y}$ (since SRS is used) $= \bar{Y}$

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 the case of $\bar{y}_{cl}$,the sampling units are $\bar{y}_1, \bar{y}_2, ..., \bar{y}_n$.

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

$E\left(\bar{y}_{cl}\right) = E\left(\bar{y}_{cl} - \bar{Y}\right)^2 = \frac{N-n}{Nn}S_b^2$

where $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{y}_i - \bar{Y}\right)^2$ 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 $\widehat{Var}\left(\bar{y}_{cl}\right) = \frac{N-n}{Nn}s_b^2$

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

4.2.2 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. Suppose that n clusters are selected with SRSWOR and all the elements in these selected clusters are surveyed. Assume that $M_i, (i = 1, 2, ..., N)$ are known. Based on this scheme, several estimators can be obtained to estimate the population mean. We consider one types of such estimators.

Let there be N clusters and M_i, be the size of the i^th cluster, let,
$M_0 = \sum_{i=1}^{N} M_i$
$\bar{M} = \frac{t}{N}\sum_{i=1}^{N} M_i$
$\bar{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$: mean for the i_{th} cluster.
$\bar{Y} = \frac{1}{M_0}\sum_{i}^{N}\sum_{j}^{M_i} y_{ij}$
$= \sum_{i=1}^{N} \frac{M_i}{M_0}\bar{y}_i$
=$\frac{1}{N}\sum_{i=1}^{N} \frac{M_i}{M_0}\bar{y}_i$

4.2.3 Mean of cluster means

Consider the simple arithmetic mean of the cluster means as

$\bar{\bar{y}}_c = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$
$E[\bar{\bar{y}}_c] = \frac{1}{N}\sum_{i=1}^{N} \bar{y}_i$
$\neq \bar{Y}$ where $\bar{Y} = \sum_{i=1}^{N} \frac{M_i}{M_0}\bar{y}_i$.
The bias of $\bar{\bar{y}}_c$ is
$Bias(\bar{\bar{y}}_c) = E[\bar{\bar{y}}_c - \bar{Y}] = -\frac{1}{M_0}[\sum_{i=1}^{N} M_i\bar{y}_i - \frac{M_0}{N}\sum_{i}^{N} \bar{y}_i]$
$= \frac{1}{M_0}[\sum_{i=1}^{N} M_i\bar{y}_i - \frac{(\sum_{i=1}^{N} M_i)(\sum_{i=1}^{N} \bar{y}_i)}{N}]$
$= -\frac{1}{M_0}(M_i - \bar{M})(\bar{y}_i - \bar{Y})$
$= -(\frac{N-1}{M_0})S_{m\bar{y}}$
Bias $\bar{\bar{y}}_c = 0$ if M_i and $\bar{y}_i$ are uncorrelated.

The mean squared error is

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

where

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

$S^2_{m\bar{y}} = \frac{1}{N-1}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{Y})$

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

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

where;

$s^2_b = \frac{n-1}{\sum_1 n-1}(\bar{y}_c - \bar{\bar{y}}_c)^2$

4.3 Exercises

(a) On a micro-computer hard disk, we count 400 files, each one consisting of exactly 50 records. To estimate the average number of characters per record, we decide to sample using simple random sampling 80 files, then 5 records in each file. We denote: $m = 80$ and $\bar{n} = 5$. After sampling we find: the sample variance of the estimators for the total number of characters per file, which is $s^2_T = 905000$; the mean of the m sample variances $s^2_{2,i}$ is equal to 805, where $s^2_{2,i}$ represents the variance for the number of characters per record in file i.

(i) How do we estimate without bias the mean number $\bar{Y}$ of characters per record?

(ii) How do we estimate without bias the accuracy of the previous estimator?

(iii) Give a 95% confidence interval for $\bar{Y}$.

(b) The objective is to estimate the mean income of households in a district of a city consisting of 60 blocks of houses (of variable size). For this, we select three blocks using simple random sampling without replacement and we interview all households which live there. Furthermore, we know that 5000 households reside in this district. The result of the survey is given in the following table.

i. Estimate the mean income and the total income of the households in the district using the Horvitz-Thompson estimator.

ii. Estimate without bias the variance of the Horvitz-Thompson mean estimator.

iii. Estimate the mean income of the households in the district using the Hájek ratio, and compare with the estimation from 1. Was the direction of the change predictable?

(c) Consider a simple random sample of clusters. We suppose that all clusters are of the same size. Recall the expression of the Horvitz-Thompson estimator. Give an expression of its variance as a function of the inter-cluster population variance.

5 Systematic random sampling

5.1 Introduction

The systematic sampling technique is operationally more convenient than the simple random sampling. It also ensures at the same time that each unit has equal probability of inclusion in the sample. In this method of sampling, the first unit is selected with the help of random numbers and the remaining units are selected automatically according to a predetermined pattern. This method is known as systematic sampling.

(a) Suppose the N units in the population are numbered 1 to N in some order.

(b) Suppose further that N is expressible as a product of two integers n and k, so that $N = nk$.

(c) To draw a sample of size n, select a random number between 1 and k. Suppose it is r. Select the first unit whose serial number is r.

(d) This first unit is called a **random start.** Select every k^{th} unit after i^{th} unit.

(e) Sample will contain $r, r + k, r + 2k, ..., r + (n - 1)\, k$ serial number units.

(f) So first unit is selected at random and other units are selected systematically. This systematic sample is called k^{th} systematic sample and k is termed as **sampling interval**.

(g) This is also known as linear systematic sampling.

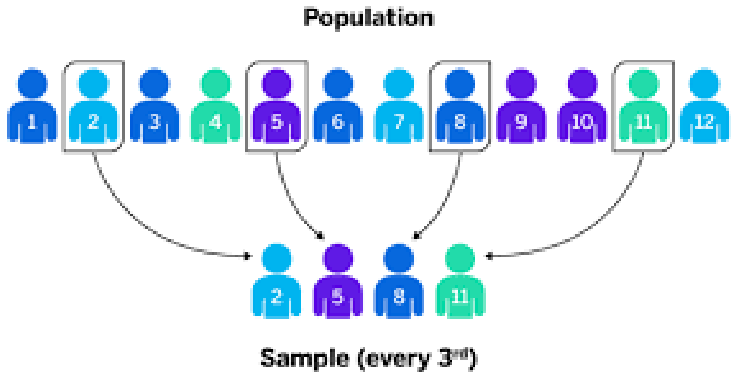


Figure 6: Illustration: Systematic random sampling

Example 5.1. Let $N = 50$ and $n = 5$. So $k = 10$. Suppose first selected number between 1 and 10 is 3.

Solution 6. Then systematic sample consists of units with following serial number 3, 13, 23, 33, 43.

Example 5.2. An insurance company's claims, in dollars, for one day are 400,600,570,960, 780, 800, 460, 650,440, 530, 470, 810, 625, 510, and 700. List all possible systematic samples of size 3, that can be drawn from this set of claims using linear systematic sampling. Also, obtain corresponding sample means.

Solution 7. Here the population size $N = 15$, and the size of the sample to be selected is $n = 3$. The sampling interval k will thus be $15/3 = 5$. The random number r to be selected from 1 to k can, therefore, take any value in the closed interval $[1, 5]$. Each random start from 1 to 5 will yield corresponding systematic sample. In all, there will be $k = 5$ possible samples. These are given below in Table 10 along with their means.

Table 10: Systematic sample

Random start	Serial number of sample units	y values of sampled units	Sample mean
1	1,6,11	400,800,470	556.67
2	2,7,12	600,460,810	623.33
3	3,8,13	570,650,625	615.00
4	4,9,14	960,440,510	636.67
5	5,10,15	780,530,700	670.00

5.1.1 Advantages of systematic sampling:

(a) It is easier to draw a sample and often easier to execute it without mistakes. This is more advantageous when the drawing is done in fields and offices as there may be substantial saving in time.

(b) The cost is low and the selection of units is simple. Much less training is needed for surveyors to collect units through systematic sampling.

(c) The systematic sample is spread more evenly over the population. So no large part will fail to be represented in the sample. The sample is evenly spread and cross section is better. Systematic sampling fails in case of too many blanks.

5.2 Estimation of Population Mean, Variance and Total

5.2.1 Estimation of population mean

When $N = nk$. Let y_{ij} be observation on the unit bearing the serial number $i + (j-1)\,k$ in the population, $i = 1, 2, ..., k$, $j = 1, 2, ..., n$. Suppose the drawn random number is $i \leq k$. Sample consists of ith column (in earlier table).

Consider the sample mean given by; $\bar{y}_{sys} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$ as an estimator of the population mean given by $\bar{Y} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{nk}\sum_{i=1}^{k}\bar{y}_i$. The probability of selecting i^{th} column as systematic sample $= \frac{1}{k}$. So, $E\left(\bar{y}_{sys}\right) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i = \bar{Y}$. Therefore, $\bar{y}_{sys}$ is an unbiased estimator of $\bar{Y}$. Further, $Var\left(\bar{y}_{sy}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$.

Consider $\left(N-1\right)S^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{Y}\right)^2$

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

$=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2 + n\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$

$= k\left(n-1\right)S_{wsy}^2 + n\sum_{i=1}^{k}\left(\bar{y}_i - \bar{Y}\right)^2$

where $S_{wsy}^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_i\right)^2$ is the variation among the units that lie within the same systematic sample.

Thus;

$Var\left(\bar{y}_{sys}\right) = \frac{N-1}{N}S^2 - \frac{k(n-1)}{N}S_{wsy}^2$

$= \frac{N-1}{N}S^2 - \frac{n-1}{n}S_{wsy}^2$ where $\frac{N-1}{N}S^2$ is the variation as a whole while $\frac{n-1}{n}S_{wsy}^2$ is the pooled within variation of the k^{th} systematic sample with $N = nk$.

This expression indicates that when the within variation is large, then $Var\left(\bar{y}_i\right)$ becomes smaller. Thus higher heterogeneity makes the estimator more efficient and higher heterogeneity is well expected in systematic sample.

Example 5.3. On a particular day, 162 boats had gone to sea from the coast for fishing. It was desired to estimate the total catch of fish at the end of the day. As it was not possible to weigh the catch for all the 162 boats, it was decided to weigh fish for only 15 boats selected using circular systematic sampling. Discuss the selection procedure, and obtain the estimate of total catch of fish using data on the 15 sample boats given in the table below.

Table: Catch of fish (in quintals) for 15 selected boats

Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish	Serial No. of boat	Catch of fish
73	5.614	128	9.225	21	8.460
84	8.202	139	6.640	32	10.850
95	6.115	150	7.350	43	6.970
106	9.765	161	5.843	54	5.524
117	8.550	10	6.875	65	7.847

Solution 8. In this case, we have $N = 162$ and $n = 15$. Since $N/n = 162/15 = 10.8$ is not a whole number, the value of sampling interval k is taken as 11, an integer nearest to 10.8, and circular systematic sampling is used for selection of boats. If the selected random number r, $1 \leq$ r $\leq$ 162, is 73 , then the boats bearing serial numbers $73, 84, \ldots, 65$ will be included in the

sample. The serial numbers of selected boats, along with the corresponding catch of fish, are presented in table 6.4. We now proceed to estimate the total catch of fish using (6.1). This estimate is

$$\begin{aligned}\hat{\mathrm{Y}}_{\mathrm{sy}} &= \mathrm{N}\overline{\mathrm{y}}_{\mathrm{sy}} = \frac{\mathrm{N}}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y}_{\mathrm{i}} \\ &= \frac{162}{15}(5.614 + 8.202 + \ldots + 7.847) \\ &= \frac{(162)(113.83)}{15} \\ &= 1229.364\end{aligned}$$

The estimate of variance $\mathrm{V}\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right)$ is then computed by using the expression (6.4). Thus,

$$\begin{aligned}v\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right) &= \mathrm{N}^2\mathrm{v}\left(\bar{y}_{\mathrm{sy}}\right) = \frac{\mathrm{N}(\mathrm{N}-\mathrm{n})}{2\mathrm{n}(\mathrm{n}-1)}\sum_{\mathrm{i}=1}^{\mathrm{n}-1}\left(\mathrm{y}_{\mathrm{i}+1} - \mathrm{y}_{\mathrm{i}}\right)^2 \\ &= \frac{162(162-15)}{2(15)(14)}\left[(8.202-5.614)^2 + (6.115-8.202)^2 + \ldots + (7.847-5.524)^2\right] \\ &= \frac{(162)(162-15)(67.596)}{2(15)(14)} \\ &= 3832.693\end{aligned}$$

The confidence interval, for the total catch of fish for 162 boats, can then be calculated from

$$\begin{aligned}&\hat{Y}_{\mathrm{sy}} \pm 2\sqrt{\mathrm{v}\left(\hat{\mathrm{Y}}_{\mathrm{sy}}\right)} \\ =&1229.364 \pm 2\sqrt{3832.693} \\ =&1105.547, 1353.181\end{aligned}$$

Thus, the estimate of total catch of fish obtained from a single sample is 1229.364 quintals. The confidence limits, obtained above, indicate that the total catch from all the 162 boats is likely to fall in the interval [1105.547, 1353.181] quintals

Example 5.4. A population is comprised of 6 households with respective sizes $2, 4, 3, 9, 1$ and 2 (the size x_k of household k is the number of people included). We select 3 households without replacement, with a probability proportional to its size.

(a) Give, in fractional form, the inclusion probabilities of the 6 households in the sampling frame (note: we may recalculate certain probabilities).

(b) Carry out the sampling using a systematic method.

(c) Using the sample obtained in 2., give an estimation for the mean size $\bar{X}$ of households; was the result predictable?

Solution 9. (a) For all $k : \pi_k = 3\frac{x_k}{X}$, with $X = 21$ Therefore

$$\pi_k = \frac{x_k}{7}, k \in U.$$

A problem arises for unit 4 because $\pi_4 > 1$. We assign the value 1 to π_4 and for the other units we recalculate the $\pi_k, k \neq 4$, according to:

$$\pi_k = 2\frac{x_k}{X - 9} = 2\frac{x_k}{12} = \frac{x_k}{6}.$$

Finally, the inclusion probabilities are presented in Table 9a. We can verify that

$$\sum_{k=1}^{6} \pi_k = 3$$

k	1	2	3	4	5	6
π_k	$1/3$	$2/3$	$1/2$	1	$1/6$	$1/3$

(b) We select a random number between 0 and 1, and we are interested in the cumulative probabilities presented in Table 11. We advance in this list using a sampling interval of 1. In each case, we obtain in fine three distinct individuals (including household 4).

Table 11: Cumulative inclusion probabilities

k	1	2	3	4	5	6
$\sum_{j \leq k}$	1/3	1	1 1/2	572	2 2/3	3

(c) We have

$\hat{\hat{X}} = \frac{1}{6}\sum_{k \epsilon S} \frac{x_k}{\pi_k}[\frac{xk_1}{1} + \frac{x_k 2}{xk_2/6} + \frac{xk_3}{x_k 3/6}]$ with $k_1 = 4$ (household 4 is definitely chosen) and k_2 and k_3 being the other two selected households

$\hat{\hat{X}} = \frac{1}{6}[9 + 6 + 6] = 3.5 = \bar{X}$

This result was obvious, as x_k and π_k are perfectly proportional, by construct (we have a null variance, thus a 'perfect' estimator for the estimation of the mean size $\bar{X}$).

5.3 Exercises

(a) Describe the circular systematic sample.

(b) Show how to estimate the mean in systematic sampling when $n \neq k$.

(c) A census was conducted in a community. In addition to obtaining the usual population information the surveys questioned the occupants of every 20^{th} household to determine how long they have occupied their present homes. The results are summarized as follows; $n = 115$, $\sum y_i^2 = 2011.15$, $\sum \bar{y}_i = 407.1$, $N = 2300$, $k = 20$. Use this results to estimate the average amount of time people have lived in their present homes and place a bound on the error of estimation.

(d) Out of 24 villages in an area, two linear systematic samples of 4 villages each were selected. The total area under wheat is given in the table below.

i. Estimate the total area under wheat.

ii. Estimate the variance of the sample mean and place an upper bound on the error of estimation.

(e) In a small municipality, we listed six businesses for which total sales (variable xk) are respectively 40, 10, 8, 1, 0.5 and 0.5 million Euros. With the aim of estimating total paid employment, select three businesses at random and without replacement, with unequal probabilities according to total sales, using systematic sampling (by justifying your process). To do this, we use the following result for a uniform random variable between [0, 1]: 0.83021. What happens if we modify the order of the list?

(f) Consider a population of 5 units. We want to select using systematic sampling with unequal probabilities a sample of two units with inclusion probabilities proportional to the following values of X_i,

$1, 1, 6, 6, 6.$

i. Calculate the first-order inclusion probabilities.

ii. Considering the two units where the value of X_i is 1, calculate their second-order inclusion probabilities for every possible permutation of the list. What is the outcome?

6 Ratio and regression estimation

6.1 Ratio Estimation

6.1.1 Introduction

An important objective in 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 is made through the ratio method of estimation to obtain an **improved estimator** of population mean and total. 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 observations 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 $x_i^{'}s$ from;

(a) some earlier completed census

(b) some earlier surveys

(c) 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 in previous year.

Theorem 6.1. 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,

$$\hat{\bar{Y}} = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{R}\bar{X} \tag{6.1}$$

assuming that the population mean $\bar{X}$ is known.

The ratio estimator of the population total $Y_{tot} = \sum_{i=1}^{N} Y_i$ is

$$\hat{Y}_{R(tot)} = \frac{y_{tot}}{x_{tot}} X_{tot} \tag{6.2}$$

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

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 $\frac{\bar{y}}{\bar{x}}$) vary little from sample to sample and the ratio estimate will be of high precision.

6.1.2 Why use ratio estimation?

(a) Sometimes, we simply want to estimate a ratio e.g. change of liabilities to assets, the ratio of the number of fish caught to the number of hours spent fishing, or the per-capita income of household members in a country.

(b) Sometimes we want to estimate a population total, but the population size N is unknown. Then we cannot use the estimator $\hat{Y}_y = N\bar{y}$ as in SRS. But we know that $N = \frac{t_x}{\bar{x}}$ and can estimate N by $\frac{t_x}{\bar{x}}$. We thus use another measure of size, tx, instead of the population count N.

(c) Ratio estimation is often used to increase the precision of estimated means and totals.

(d) Ratio estimation is used to adjust estimates from the sample so that they reflect demographic totals. An SRS of 400 students taken at a university with 4,000 students may contain 240 women and 160 men, with 84 of the sampled women and 40 of the sampled men planning to follow careers in teaching. Using only the information from the SRS, you would estimate that

$$\frac{4000}{400}.124 = 1240 \tag{6.3}$$

students plan to be teachers. Knowing that the college has 2,700 women and 1,300 men, a better estimate of the number of students planning teaching careers might be

$$\frac{84}{240}.2700 + \frac{40}{160}.1300 = 1270 \tag{6.4}$$

This use of ratio estimation, called post-stratification.

(e) Ratio estimation may be used to adjust for nonresponse, as will be discussed later.

Example 6.1. U.S. Census of Agriculture, a SRS of 300 of the 3,078 counties. For this example, suppose we know the population totals for 1987, but have 1992 information only for the SRS of 300 counties. When the same quantity is measured at different times, the response of interest at an earlier time often makes an excellent auxiliary variable. Let

y_i = total acreage of farms in county i in 1992

x_i = total acreage of farms in county i in 1987.

In 1987 a total of $t_x = 964,470,625$ acres were devoted to farms in the United States. The average acres of farms per county for the population is $\bar{y} = \frac{964,470,625}{3078} = 313,343.3$

The estimated ratio is

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{297897.0467}{301953.7233} = 0.986565 \tag{6.5}$$

and the ratio estimators of $\bar{y}$ and t_y are:

$$\hat{\bar{y}} = \hat{B}\bar{x} = (\hat{B})(313,343.283) = 309,133.6 \tag{6.6}$$

and

$$\hat{t}_{yr} = \hat{B}t_x = (\hat{B})(964,470,625) = 951,513,191. \tag{6.7}$$

Note that y for these data is 297,897.0, so $\hat{t}_ySRS = (3078)\bar{y} = 916,927,110$

6.1.3 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 $E\left(\hat{\bar{Y}}_R\right) = \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} \sum_{i=1}^{\begin{pmatrix} N \\ n \end{pmatrix}} \frac{\bar{y}}{\bar{x}}\bar{X} \neq \bar{Y}$

(in general). 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 $\varepsilon_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = (1 - \varepsilon_0)\bar{Y}$,

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

Since SRSWOR is being followed, so ; $E(\varepsilon_0) = 0$, $E(\varepsilon_1) = 0$,

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

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

where $f = \frac{N-n}{N}$, $S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N \left(Y_i - \bar{Y}\right)^2$

and $C_Y = \frac{S_Y}{\bar{Y}}$ is the coefficient of variation related to Y.

Similarly,

$E\left(\varepsilon_1^2\right) = \frac{1}{n} C_X^2$,

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

$= \frac{1}{\bar{X}\bar{Y}} \cdot \frac{f}{n} S_{XY} = \frac{1}{\bar{X}\bar{Y}} \frac{f}{n} \rho S_X S_Y = \frac{f}{n} \rho \frac{S_X}{\bar{X}} \frac{S_Y}{\bar{Y}} = \frac{f}{n} \rho C_X C_Y$

where $C_X = \frac{S_X}{\bar{X}}$ is the coefficient of variation related to X

and ρ is the population correlation coefficient between X and Y.

Writing $\hat{\bar{Y}}_R$ in terms of $\varepsilon_i^{'}s$ 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 $|\frac{\bar{x}-\bar{X}}{\bar{X}}| < 1$ i.e., 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 is fairly large. With this assumption,

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

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

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

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 $\hat{\bar{Y}}_R - \bar{Y} = \bar{Y}\left({}_0 - \varepsilon_1\right)$ and $E\left(\hat{\bar{Y}}_R - \bar{Y}\right) = 0$. So the ratio estimator is an unbiased estimator of population mean up 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 than assuming that powers more than one are negligibly small), then the estimation error of $\hat{\bar{Y}}_R$ can be approximated as

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

Then the bias of $\hat{\bar{Y}}_R$ is given by

$E\left(\hat{\bar{Y}}_R - \bar{Y}\right) = \bar{Y}\left(0 - 0 + \frac{f}{n}C_X^2 - \frac{f}{n}\rho C_X C_Y\right)$ upto the 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. $Bias\left(\hat{\bar{Y}}\right) = 0$. If $E\left(\varepsilon_1^2 - \varepsilon_0\varepsilon_1\right) = 0$

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

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

or if $Var(\bar{x}) - \frac{Cov(\bar{x},\bar{y})}{R} = 0$

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

$MSE\left(\hat{\bar{Y}}_R\right) = E\left(\hat{\bar{Y}}_R - \bar{Y}\right)^2$

$= E\left(\bar{Y}^2\left(\varepsilon_0 - \varepsilon_1 + \varepsilon_1^2 - \varepsilon_1\varepsilon_0 + ...\right)^2\right)$

$= E\left(\bar{Y}^2\left(\varepsilon_0^2 + \varepsilon_1^2 - 2\varepsilon_0\varepsilon_1\right)\right)$

Under the assumption $|\varepsilon_1| < 1$, and the terms of ε_0 and ε_1 involving powers more than two are negligible small, $MSE\left(\hat{\bar{Y}}_R\right) = \bar{Y}^2\left[\frac{f}{n}C_X^2 + \frac{f}{n}C_Y^2 - \frac{2f}{n}\rho C_X C_Y\right] = \frac{\bar{Y}^2 f}{n}\left[C_x^2 + C_Y^2 - 2\varrho C_X C_Y\right]$ up to the second order of approximation.

6.2 Regression Estimation

6.2.1 Introduction

The ratio method of estimation uses the **auxiliary information** which is correlated with the study variable to improve the precision which results in the improved estimators when the regression of Y on X is linear and passes through origin. When the regression of Y 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 estimator to estimate the population means.

In ratio method, the conventional estimator sample mean $\bar{y}$ was improved by multiplying it by 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 $\left(\bar{x} - \bar{X}\right)$ for which $E\left(\bar{x} - \bar{X}\right) = 0$. Consider an improved estimator of $\bar{Y}$ as $\hat{\bar{Y}}^* = \bar{y} + \mu\left(\bar{x} - \bar{X}\right)$ which is an unbiased estimator of $\bar{Y}$ and μ is any constant. Now find μ such that the $Var\left(\hat{\bar{Y}}^*\right)$ is minimum.

$Var\left(\hat{\bar{Y}}^*\right) = Vay\left(\bar{y}\right) + \mu^2 Var\left(\bar{x}\right) + 2\mu Cov\left(\bar{x}, \bar{y}\right)$

$\frac{\partial\left(\hat{\bar{Y}}^*\right)}{\partial\mu} = 0 \Rightarrow \mu = -\frac{Cov(\bar{x},\bar{y})}{Var(\bar{x})}$

$= -\frac{\frac{N-n}{Nn}S_{XY}}{\frac{N-n}{Nn}S_X^2} = -\frac{S_{XY}}{S_X^2}$

where $S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right), S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$

Consider a linear regression model $y = x\beta + e$ where y is the dependent variable, x is the independent variable and e is the random error component which takes care of the difference arising due to lack of exact relationship between x and y. Note that the value of regression coefficient β in 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 $\left(x_i, y_i\right), i = 1, 2, ..., n$ 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 the optimum value of μ is $\hat{\bar{Y}}_{reg} = \bar{y} + \beta\left(\bar{X} - \bar{x}\right)$ which is the regression estimator of $\bar{Y}$ and the procedure of estimation is called as under the regression method of estimation.

The variance of $\hat{\bar{Y}}_{reg}$ is $Var\left(\hat{\bar{Y}}_{reg}\right) = V\left(\bar{y}\right)\left[1-\rho^2\left(\bar{x},\bar{y}\right)\right]$ where $\rho\left(\bar{x},\bar{y}\right)$ is the correlation coefficient between $\bar{x}$ and $\bar{y}$. So $\hat{\bar{Y}}_{reg}$ would be efficient if x and y are highly correlated. The estimator $\hat{\bar{Y}}_{reg}$ is more efficient than $\bar{Y}$ if $\rho\left(\bar{x},\bar{y}\right) \neq 0$ which generally holds.

6.2.2 Estimate of variance

An unbiased sample estimate of $Var\left(\hat{\bar{Y}}_{reg}\right)$ is

$\widehat{Var}\left(\hat{\bar{Y}}_{reg}\right) = \frac{f}{n(n-1)}\sum_{i=1}^{n}\left[\left(y_i-\bar{y}\right)-\beta_0\left(x_i-\bar{x}\right)\right]^2$

$= \frac{f}{n}\sum_{i=1}^{n}\left(s_y^2+\beta_0^2 s_x^2-2\beta_0 s_{xy}\right)$

6.3 Regression estimates when β is computed from sample

Suppose a random sample of size n on paired observations on $\left(x_i,y_i\right), i=1,2,...,n$ 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 given by $\hat{\bar{Y}}_{reg}=\bar{y}+\beta\left(\bar{X}-\bar{x}\right)$. It is difficult to find the exact expressions of $E\left(\bar{Y}_{reg}\right)$ and $Var\left(\hat{\bar{Y}}_{reg}\right)$. So we approximate them using the same methodology as in the case of ratio method of estimation.

Let

$\varepsilon_0=\frac{\bar{y}-\bar{Y}}{\bar{Y}} \Rightarrow \bar{y}=\bar{Y}\left(1+\epsilon_0\right)$

$\varepsilon_1=\frac{\bar{x}-\bar{X}}{\bar{x}} \Rightarrow \bar{x}=\bar{X}\left(1+\epsilon_1\right)$

$\varepsilon_2=\frac{s_{xy}-S_{XY}}{S_{XY}} \Rightarrow s_{xy}=S_{XY}\left(1+\epsilon_2\right)$

$\varepsilon_3=\frac{s_x^2-S_X^2}{S_X^2} \Rightarrow s_x^2=S_X^2\left(1+\epsilon_3\right)$

then $E\left(\varepsilon_0\right)=0$, $E\left(\varepsilon_1\right)=0$, $E\left(\varepsilon_2\right)=0$, $E\left(\varepsilon_3\right)=0$, $E\left(\varepsilon_0^2\right)=\frac{f}{n}C_X^2$, $E\left(\varepsilon_0,\varepsilon_1\right)=\frac{f}{n}\rho C_X C_Y$ and $\bar{Y}_{reg}=\bar{y}+\frac{s_{xy}}{s_x^2}\left(\bar{X}-\bar{x}\right)\bar{Y}\left(1+\varepsilon_0\right)+\frac{S_{XY}(1+\varepsilon_2)}{S_X^2(1+\varepsilon_3)}\left(-\varepsilon_1\bar{X}\right)$.

The estimation error of $\bar{Y}_{reg}$ is

$\left(\hat{\bar{Y}}_{reg}-\bar{Y}\right)=\bar{Y}\varepsilon_0-\beta\bar{X}\varepsilon_1\left(1+\varepsilon_2\right)\left(1+\varepsilon_3\right)^{-1}$ where $\beta=\frac{S_{XY}}{S_X^2}$

is the population regression coefficient. Assuming $\mid \varepsilon_3 \mid < 1$,

$\left(\hat{\bar{Y}}_{reg}-\bar{Y}\right)\simeq\bar{Y}\varepsilon_0-\beta\bar{X}\left(\varepsilon_1+\varepsilon_1\varepsilon_2\right)\left(1-\varepsilon_3+\varepsilon_3^2\right)$

$$\simeq\bar{Y}\varepsilon_0-\beta\bar{X}\left(\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2\right) \qquad (6.8)$$

6.3.1 Bias of $\hat{\bar{Y}}_{reg}$

Now the bias of $\hat{\bar{Y}}_{reg}$ up to the second order of approximation is $E\left(\hat{\bar{Y}}_{reg}-\bar{Y}\right)^2 \approx E\left[\bar{Y}\varepsilon_0-\beta\bar{X}\left(\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2\right)\right]^2$

Retaining the terms of $\varepsilon_i^{'}s$ up to the second power second and ignoring others, we have
$= E\left(\hat{\bar{Y}}_{reg} - \bar{Y}\right)^2 \approx E\left[\varepsilon_0^2\bar{Y}^2 + \beta^2\bar{X}^2\varepsilon_1^2 - 2\beta\bar{X}\bar{Y}\varepsilon_0\varepsilon_1\right]$
$= \bar{Y}^2E\left(\varepsilon_0^2\right) + \beta^2\bar{X}^2E\left(\varepsilon_1^2\right) - 2\beta\bar{X}\bar{Y}E\left(\varepsilon_0\varepsilon_1\right)$
$= \frac{f}{n}\left[\bar{Y}^2\frac{S_Y^2}{\bar{Y}^2} + \beta^2\bar{X}^2\frac{S_X^2}{\bar{X}^2} - 2\beta\bar{X}\bar{Y}\rho\frac{S_XS_Y}{\bar{X}\bar{Y}}\right]$
$MSE\left(\hat{\bar{Y}}_{reg}\right) = E\left(\hat{\bar{Y}}_{reg} - \bar{Y}\right)^2 = \frac{f}{n}\left(S_Y^2 + \beta^2S_X^2 - 2\beta\rho S_XS_Y\right).$
Since $\beta = \frac{S_{XY}}{S_X^2} = \rho\frac{S_Y}{S_X}$, so substituting it in $MSE\left(\hat{\bar{Y}}_{reg}\right)$, we get.

$$MSE\left(\hat{\bar{Y}}_{reg}\right) = \frac{f}{n}S_Y^2\left(1 - \rho^2\right). \tag{6.9}$$

So up to 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.

6.3.2 Exercises

(a) The number of inhabitants(1000's) in each of of a simple random sample of 49 cities drawn from a population of 196 large cities is as given below. $\sum y_i = 6262$, $\sum x_i = 5054$. The true total number of inhabitants in the 196 cities in 1929, x is assumed to be known. Its value is 22,919. Estimate the total number of inhabitants in the 196 cities in 1980. (i) Using the ratio estimator (ii) Using the sample estimator.

(b) In a study to estimate the total sugar content of a truck load of oranges a random sample of n=10 oranges was juiced and weighed as shown in the table below. The total weight of all the oranges obtained by first weighing the truck loaded and then unloaded was found to be 1800 pounds. Estimate the total sugar content of oranges and place a bound on the error of estimation.

(c) A company wish to estimate the average amount of money i.e. $\bar{Y}$ paid to employees for medical expenses during the first three months of the calendar year.Average quarterly reports available in the fiscal report of the previous year. A random sample of 100 employees records are taken from the population of 1000 employees. The sample results are summarized below. Use the data to estimate the population mean and place a bound on the error of estimation.
$n = 100,$
$N = 1000,$
$\sum y_i = 1750$

The total for the corresponding quarter corresponding to the previous year is $\sum_{i=1}^{100} x_i = 1200$

Population total X

for the corresponding quarter of the previous year $X = 12500$,

$\sum y_i^2 = 31650$,

$\sum x_i^2 = 15620$,

$\sum x_i y_i = 22059.35$

7 Introduction to complex surveys and weighting

7.1 What is a complex survey?

If selection of final units of observation is accomplished through a **series of stages** in which larger units are first selected then the sample survey is said to be **complex (or multistage)**. Generally, complex sample surveys employ simple random sampling at each stage in a series of stages to culminate in the final sample of observation units at the desired level.

7.2 Steps in conducting/analyzing complex survey samples

(a) Define problem/state objectives

  i. Specify the problem to be addressed.

  ii. Objectives support/ seek to answer problems stated

  iii. Problems can be highly variable, even within one survey

(b) Understand the sample design

  i. Understand how the survey was conducted, the intended population of interest and the levels at which data are available?

  ii. If a problem seeks to understand trends over time and the survey is cross-sectional, then the utility of the survey data in meeting estimation objectives will be limited.

(c) Select the sample (or obtain the data)

  i. There are many types of random sampling designs that can be used at various stages of a complex, multistage sampling design: Simple Random Sampling (with/without replacement), Systematic Sampling, Stratified Sampling, Cluster Sampling

(d) Process and analyze data

(e) Interpret and evaluate the results of the analysis

(f) Report results from survey data

7.3 Sampling weights

Sampling weights are the number of individuals in the population **each respondent in the sample is representing**. Sampling weights are needed to correct for **imperfections in the sample** that might lead to **bias** and other departures between the sample and the reference population.

A sample weight is the inverse of the probability of selection. For example, if my simple random sample is one tenth of the population size (i.e. my sampling fraction is $\frac{1}{10}$), then each respondent in the sample is representing 10 people in the population.

Sampling weights compensate for:

(a) Unequal probabilities of selection.

(b) Unit non-response.

(c) To adjust the weighted sample distribution for key variables of interest (for example, age, race, and sex) to make it conform to a known population distribution.

Example 7.1. How weights work.

Table 12: Example on sampling weights

Score	Weight
4	1
2	2
1	4
5	1
2	2

Simple mean $= \frac{(4+2+1+5+2)}{5} = 2.8$

Weighted mean $= \frac{(4x1)+(2x2)+(1x4)+(5x1)+(2x2)}{10} = 2.1$

Weights are frequencies of each observation in the population.

Types of weights

(a) Raw or base weights

(b) Relative or normalized weights

(c) Design effect adjusted weights

(a) Raw weights or base weights

Raw weights sum to the population size. They are the inverse of the probability of selection:

$$w_i = \frac{1}{p_i} \tag{7.1}$$

For example, if the probability of selection of a unit is $\frac{1}{50}$, its raw weight is 50.

Raw weights for multi-stage sampling — The raw weight is the inverse of the product of the probabilities of selection at each stage.

$$w_i = \frac{1}{p_{1i} \cdot p_{2i}} \tag{7.2}$$

where,

p_{1i} = probability of selection at stage 1

p_{2i} = probability of selection at stage 2.

For example, if the probability of selection at stage 1 (schools) is $\frac{1}{5}$, and the probability of selection at stage 2 (teachers) is $\frac{1}{20}$, the final probability of selection is $\frac{1}{5} \cdot \frac{1}{20} = \frac{1}{100}$ and the raw weight is 100.

Problems of using raw weights

(a) Estimates of means, proportions and standard errors obtained using raw weights will be based on the population size, not the sample size. The means and proportion estimates will be correct, but the test statistics will have too much power.

(b) **Solution:** Convert raw weights to normalized weights.

(b) Normalized or relative weights

Normalized weights sum to the sample size. With normalized weights in the analyses, the estimates of means, and proportions are correct. The estimates of standard errors are correct given a simple random sample or stratified sample.

When a cluster or multi-stage sample is used, the estimation of standard errors will not be correct using only case weights. Special procedures such as Taylor series approximation, bootstrapping or design effects need to be used to obtain correct standard errors.

Converting a raw weight to a normalized weights There are two ways of converting a raw weight to a normalized way:

(a) Dividing the raw weights by the mean of the raw weights:

$$w_N = \frac{w_i}{\bar{w}} \tag{7.3}$$

(b) Multiplying the raw weight by the overall sampling fraction:

$$w_N = w_i \frac{n}{N} \tag{7.4}$$

Adjusting for unit non-response

(a) Calculate the weight for non-response, which is the inverse of the subgroup response rates.

$$w_{nr} = \frac{S_s}{n_s} \tag{7.5}$$

where S_s is the number of sampled cases for the subgroup and S_s is the number of response obtained for the subgroup.

(b) Calculate the non-response adjusted weight: It is the product of the original weight and the weight for non-response.

$$w_a = \frac{w_i}{w_{nr}} \tag{7.6}$$

Example 7.2. Weighting for unequal probabilities of selection: An epsem sample of 5 households is selected from 250. One adult is selected at random in each sampled household. The monthly income (y_{ij}) and the level of education ($z_{ij} = 1$, if secondary or higher; 0 otherwise) of the j^{th} sampled adult in the i^{th} household are recorded. Let M_i denote the number of adults in household i. Then, the overall probability of selection of a sampled adult is given by:

$$p_{ij} = p_i . p_{ji} = \frac{5}{250} . \frac{1}{M_i} = \frac{1}{50} . \frac{1}{M_i} \tag{7.7}$$

Therefore, the weight of a sampled adult is given by:

$$w_i = \frac{1}{p_{ij}} = 50.M_i \tag{7.8}$$

To illustrate the estimation procedure, let us assume a first-stage sample of 5 households with data obtained from the single sampled adult for each household as given in the table below:

Table 13: Example: weighting for unequal probabilities.

Sampled Households	M_i	w_i	y_ij	z_ij	$w_i.y_{ij}$	$w_i.z_{ij}$	$w_i.z_{ij}.y_{ij}]$
1	3	150	70	1	10500	150	10500
2	1	50	30	0	1500	0	0
3	3	150	90	1	13500	150	13500
4	5	250	50	1	12500	250	12500
5	4	200	60	0	12000	0	0
16	800	800	300	3	50000	550	36500

Estimates of various characteristics can then be obtained from the above table as follows:

(a) The estimate of monthly income is

$$\bar{y_i} = \frac{\sum w_i.y_{ij}}{\sum w_i} = \frac{50000}{800} = 62.5 \tag{7.9}$$

If weights were not used, this estimate would be $\frac{300}{5} = 60\%$.

(b) The estimate of the proportion of people with secondary or higher education is

$$\bar{y_i} = \frac{\sum w_i.z_{ij}}{\sum w_i} = \frac{550}{800} = 0.6975. \tag{7.10}$$

If weights are not used, this estimate would be $\frac{3}{5}$ or 0.60 or 60%.

(c) The estimate of the total number of people with secondary or higher education is

$$\hat{t} = \sum w_i.z_{ij} = 550 \tag{7.11}$$

(d) The estimate of the mean monthly income of adults with secondary or higher education is

$$\bar{y_w} = \frac{\sum w_i.z_{ij}.y_{ij}}{\sum w_i.z_{ij}} = \frac{36500}{550} = 66.36 \tag{7.12}$$

(e) Note that for estimating totals, sampled elements need to be weighted by the reciprocal of their selection probabilities. For estimating means and proportions, the weights need only be; proportional to the reciprocals of the selection probabilities. Thus, in the preceding example, the weights w_i are proportional to $M_i(w_i = 50 * M_i)$. If we use M_i as the weights, then the estimate of the proportion with secondary or higher education is

$$\hat{p} = \frac{M_i.z_{ij}}{\sum M_i} = \frac{(3.1) + (1.0) + (3.1) + (5.1) + (4.0)}{3 + 1 + 3 + 5 + 4} = \frac{11}{16} = 0.6875 \tag{7.13}$$

as before. However, the estimate of the total number of adults with secondary or higher education is

$$\hat{p_s} = 50 \sum M_i.z_{ij} = 50.11 = 550 \tag{7.14}$$

7.3.1 Self weighting samples

When the weights of all sampled units are the same, the sample is referred to as **self weighting.** Samples are rarely self-weighting at the national level for several reasons.

(a) First, sampling units are selected with unequal probabilities of selection. Indeed, even though the PSUs are often selected with probability proportional to size, and households selected at an appropriate rate within PSUs to yield a self-weighting design, this may be nullified by the selection of one person for interview in each sampled household.

(b) Second, the selected sample often has deficiencies including non-response and under-coverage.

(c) Third, the need for precise estimates for domains and special sub-populations often requires oversampling these domains.

7.3.2 The adjustment of sample weights for non-response

It is rarely the case that all desired information is obtained from all sampled units in surveys. For instance, some households may provide no data at all while other households may provide only partial data, that is, data on some but not all questions in the survey. The former type of non-response is called unit or total non-response, while the latter is called item nonresponse. If there are any systematic differences between the respondents and non-respondents, then naïve estimates based solely on the respondents will be biased. It is important to keep survey non-response as low as possible, in order to reduce the possibility that the survey estimates could be biased in some way by failing to include (or including a disproportionately small percentage of) a particular portion of the population. For example, persons who live in urban areas and have relatively high incomes might be less likely to participate in a multipurpose survey that includes income modules. Failure to include a large segment of this portion of the population could affect national estimates of average household income, educational attainment, literacy, etc.

7.3.3 Reducing non-response bias in household surveys

The size of the non-response bias for a sample mean, for instance, is a function of two factors:

- The proportion of the population that does not respond. - The size of the difference in population means between respondent and non respondent

Reducing the bias due to non-response therefore requires that either the non-response rate be small, or that there are small differences between responding and non-responding households and persons. With proper record keeping of every sampled unit that is selected for the survey, it is possible to estimate directly from the survey data, the non-response rate for the entire sample and for sub-domains of interest. Furthermore, special carefully designed studies can be carried out to evaluate the differences between respondents and non-respondents (Groves and Couper, 1998).

7.3.4 Compensating for non-response bias

A number of techniques can be employed to reduce the potential for non-response bias in household surveys. The standard method of compensating for partial or item non-response is **imputation**.

For unit or total non-response, there are three basic procedures for compensation:

(a) Non-response adjustment of the weights.

(b) Drawing a larger sample than needed and creating a reserve sample from which replacements are selected in case of non-response.

(c) Substitution, the process of replacing a non-responding household with another household that was not sampled which is in close proximity to the non-responding household with respect to the characteristic of interest.

Example 7.3. A stratified multi-stage sample of 1000 households is selected from two regions (North and South) of a country. Households in the North are sampled at a rate of $\frac{1}{100}$ and those in the south at a rate of $\frac{1}{200}$. Response rates in urban areas are lower that those in rural areas. Let n_h denote the number of households sampled in stratum h, let r_h denote the number of eligible households that responded to the survey, and let th denote the number of responding households with access to primary health care. Then, the non-response adjusted weight for the households in stratum h is given by:

$$w_h = w_{1h} * w_2 h \tag{7.15}$$

where $w_{2h} = \frac{n_h}{r_h}$. Assume that the stratum-level data are as given in the following table:

Table 14: 1000 households is selected from two regions (North and South)

Stratum	n_h	r_h	t_h	w_{1h}	w_{2h}	w_h	$w_h r_h$	$w_h t_h$
North-Urban	100	80	70	100	1.25	125	10000	8750
North-Rural	300	120	100	100	2.5	250	30000	25000
South-Urban	200	170	150	200	1.18	236	40120	35400
South-Rural	400	360	180	200	1.11	222	79920	39960
	1000	730	500				160400	109110

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p_s} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109,110}{160,040} = 0.682 \tag{7.16}$$

The estimated number of households with access to primary health care is

$$\hat{p_t} = \sum w_h t_h = 109,110 = 68.2\% \tag{7.17}$$

Note that the unweighted estimated proportion of households with access to primary health care, using only the respondent data is

$$\hat{p_{nw}} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0.685 \tag{7.18}$$

and the estimated proportion using the initial weights without non-response adjustment

$$\hat{p_1} = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83000}{126000} = 0.659 \tag{7.19}$$

7.3.5 The adjustment of sample weights for non-coverage

Non-coverage refers to the failure of the sampling frame to cover all of the target population and thus some sampling units have no probability of selection into the sample selected for the household survey. This is just one of many possible deficiencies of sampling frames used to select samples for surveys in developing countries.

Most household surveys in developing countries are based on stratified multi-stage area probability designs.

(a) The first-stage units, or primary sampling units, are usually geographic area units.

(b) At the second stage, a list of households or dwelling units is created, from which the sample of households is selected.

(c) At the last stage, a list of house members or residents is created, from which the sample of persons is selected.

Thus non-coverage may occur at three levels: the PSU level, the household level, and the person level. There are several procedures for handling the problem of non-coverage in household surveys. These include:

(a) Improved field procedures such as the use of multiple frames and improved listing procedures.

(b) Compensating for the non-coverage through a **statistical adjustment of the weights.**

If reliable control totals are available for the entire population and for specified subgroups of the population, one could attempt to adjust the weights of the sample units in such a way as to make the sum of weights match the control totals within the specified subgroups. The subgroups are called **post-strata**, and the statistical adjustment procedure is called **post-stratification**. This procedure simultaneously compensates for non-response and non-coverage.

Example 7.4. In the preceding example, suppose that the number of households is known to be 45,025 in the North and 115,800 in the South. Suppose further that the weighted sample totals are respectively 40,000 and 120,040.

(a) Step 1: Compute the post-stratification factors.

For the North region, we have: $w_{3h} = \frac{45025}{40000} = 1.126$

For the South region, we have: $w_{3h} = \frac{115800}{120040} = 0.965$

(b) Step 2: Compute final, adjusted weight: $w_f = w_h.w_{3h}$ The numerical results are summarized in the following table:

Table 15: Example

Stratum	r_h	t_h	w_h	w_f	$w_f.r_h$	$w_f.t_h$
North-Urban	80	70	125	140.75	11256	9849
North-Rural	120	100	250	281.4	33768	28140
South-Urban	170	150	236	227.77	38709	34155
South-Rural	360	180	222	214.2	77112	38556
Total	730	500			160845	110700

Therefore, the estimated proportion of households with access to primary health care is:

$$\hat{p_f} = \frac{\sum w_f t_h}{\sum w_f r_h} = \frac{110700}{160845} = 0.688 \tag{7.20}$$

NOTE: Check sample design, stratification and sampling weights for Demographic and Health Survey (DHS) `https://dhsprogram.com/data/Guide-to-DHS-Statistics/Analyzing_DHS_Data.ht`

8 Introduction to non-response and missing data analysis

8.1 Introduction

In almost every survey, some of the persons, households, and other types of units selected into the sample are not contacted. Persons away from home on business or vacation, wrong addresses and telephone numbers, households without telephones or with unlisted numbers, inability of the interviewers to reach households in remote places, and similar reasons contribute to the non contacts. Even if they are contacted, some of the units may not respond to one or more characteristics of the survey.

Some factors affecting non-response Lohr (2010)

(a) Survey design:

   (i) Survey content — A survey on drug use or financial matters may have a large number of refusals. Sometimes the response rate can be increased for surveys involving sensitive subjects by careful ordering of the questions, or by using a self-administered questionnaire on the computer to protect the respondents' privacy.

   (ii) Questionnaire design — Question wording has a large effect on the responses received; it can also affect whether a person responds to the survey or to a particular item on the questionnaire. A well-designed questionnaire form may increase data accuracy and reduce item non-response.

   (iii) Survey introduction — The survey introduction, sometimes sent in an advance letter telling a household or business that they have been selected to participate in the survey, provides the first contact between the interviewer and potential respondent. A good introduction, giving the recipient motivation to respond, can increase response rates dramatically.

   (iv) Sampling frame — Some sampling frames have more detailed or accurate information than others. A sampling frame of addresses for a household survey may also contain telephone numbers for some of the addresses; these, if accurate, could be used for following up with nonrespondents.

   (v) Sampling design — The survey design affects the response rate.

   (vi) Survey mode — Household surveys conducted in person (sometimes called face-to-face surveys) have in general had higher response rates than surveys conducted by telephone, mail, e-mail, or internet. In a high quality in-person survey, interviewers are sent to the households selected in a probability sample of areas or addresses.

   (vii) Follow-up — Almost all high-quality household surveys follow up with households or persons who do not respond to the initial survey request. In some surveys, more than 30 attempts are made to reach a household or person selected to be in the sample.

(b) Interviewers — Some interviewers are better at this than others. Experience, workload, training, attitudes and motivation, speech patterns, persistence, and personal characteristics can all affect the response rate achieved by an interviewer (Lohr, 2010).

(c) Respondent

   (i) Availability and access — Some calling periods, times of day, or seasons of the year may yield higher response rates than others.

   (ii) Burden — ersons who respond to a survey are doing you an immense favor, and the survey should be as nonintrusive as possible.

   (iii) Motivation — Much research has been done on how to motivate sampled persons to respond to the survey. Incentives, financial or otherwise, may increase the response rate

We explore two non-response types.

(a) Unit-non-response — entire observation unit is missing.

(b) Item-non-response — measurements are present for the observation unit but at least one item is missing.

Example 8.1. In a survey of persons, unit non-response means that the person provides no information for the survey; item non-response means that the person does not respond to a particular item on the questionnaire. Unit non-response can arise for a variety of reasons:

(a) The interviewer may not be able to contact the household.

(b) Someone may be ill and cannot respond to the survey.

(c) A person who is contacted may refuse to participate in the survey.

Item non-response often occurs because someone declines to answer a question (a person may skip the question about income, for example) or does not finish the survey.

8.2 Missing data mechanism

Rubin (1976) classified missing data problems into three categories.

(a) Missing Complete At Random (MCAR)

(b) Missing Complete At Random (MAR)

(c) Missing Complete At Random (MNAR)

In his theory every data point has some **likelihood** of being missing. The process that governs these probabilities is called the **missing data mechanism** or **response mechanism**. The model for the process is called the **missing data model or response model.**

(a) **Missing Completely At Random (MCAR):** If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This implies that causes of the missing data are unrelated to the data. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. While convenient, MCAR is often unrealistic for the data at hand.

(b) **Missing Not At Random (MAR):** If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. Modern missing data methods generally start from the MAR assumption.

(c) **Missing Complete At Random (MNAR):** If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also

find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. MNAR is the most complex case. Strategies to handle MNAR are to

find more data about the causes for the missingness.

8.3 Approaches for dealing with missing data

(a) Ignore the non-response (not recommended, but unfortunately common in practice) — Non-response increases the variance of estimators (because the sample size is smaller than anticipated) but the main concern is potential bias.

(b) Prevent it. Design the survey so that non-response is low. This is by far the best method.

A common feature of poor surveys is a lack of time spent on design and non-response follow-up. Many persons new to surveys (and some, unfortunately, not new) simply jump in and start collecting data without considering potential problems in the data collection process; they send questionnaires to everyone in the target population and analyze those that are returned. It is not surprising that such surveys have poor response rates.

A researcher who knows the target population well will be able to anticipate some of the reasons for non-response and prevent some of it. Most investigators, however, do not know as much about reasons for non-response as they think they do. They need to discover why the non-response occurs and resolve as many of the problems as possible before commencing the survey.

(c) Take a representative sub-sample of the non-respondents; use that sub-sample to make inferences about the other non-respondents.

(d) Use a statistical model to predict values for the non-respondents

   i. Listwise deletion
   ii. Pairwise deletion
   iii. Mean imputation
   iv. Regression imputation
   v. Stochastic regression imputation
   vi. LOCF and BOFC
   vii. Indicator method

8.3.1 Listwise deletion

Complete case analysis (listwise deletion) is the default way of handling incomplete data in many statistical packages, including SPSS, SAS and Stata. The function na.omit() does the same in S-PLUS and R. The procedure eliminates all cases with one or more missing values on the analysis variables.

The major advantage of complete case analysis is convenience. If the data are MCAR, listwise deletion produces unbiased estimates of means, variances and regression weights. Under MCAR, listwise deletion produces standard errors and significance levels that are correct for the reduced subset of data, but that are often larger relative to all available data. A disadvantage of listwise deletion is that it is potentially wasteful

8.3.2 Pairwise deletion

Pairwise deletion, also known as available-case analysis, attempts to remedy the data loss problem of listwise deletion. The method calculates the means and (co)variances on all observed data. Thus, the mean of variable X is based on all cases with observed data on X, the mean of variable Y uses all cases with observed Y -values, and so on. For the correlation and covariance, all data are taken on which both X and Y have non-missing scores. Subsequently, the matrix of summary statistics are fed into a program for regression analysis, factor analysis or other modeling procedures.

8.3.3 Mean imputation

A quick fix for the missing data is to replace them by the mean. We may use the mode for categorical data. Mean imputation distorts the distribution in several ways. Mean imputation is a fast and simple fix for the missing data. However, it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Mean imputation should perhaps only be used as a rapid fix when a handful of values are missing, and it should be avoided in general.

8.3.4 Regression imputation

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. The first step involves building a model from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data. Regression imputation yields unbiased estimates of the means under MCAR, just like mean imputation, and of the regression weights of the imputation model if the explanatory variables are complete. Moreover, the regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model.

8.3.5 Stochastic regression imputation

Stochastic regression imputation is a refinement of regression imputation that adds noise to the predictions. This will have a downward effect on the correlation. Stochastic regression imputation is an important step forward. In particular it preserves not only the regression weights, but also the correlation between variables.

8.3.6 LOCF and BOFC

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) require longitudinal data. The idea is to take the last observed value as a replacement for the missing data. LOCF is convenient because it generates a complete dataset. The method is used in clinical trials.

8.3.7 Indicator method

Suppose that we want to

fit a regression, but there are missing values in one of the explanatory variables. The indicator method replaces each missing value by a zero and extends the regression model by the response indicator. The procedure is applied to each incomplete variable. The user analyzes the extended model instead of the original. This method is popular in public health and epidemiology. An advantage is that the indicator method retains the full dataset.

8.3.8 Multiple Imputation

Multiple imputation creates $m > 1$ complete datasets. Each of these datasets is analyzed by standard analysis software. The m results are pooled into a

final point estimate plus standard error by simple pooling rules (Rubin's rules").

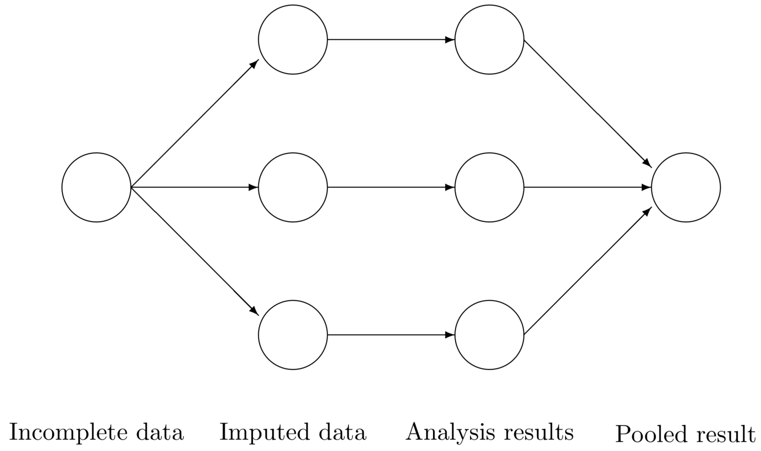


Figure 7: Scheme of main steps in multiple imputation Van Buuren (2018)

.

9 Sources of Errors in Surveys

9.1 Introduction

It is a general assumption in the sampling theory that the **true value** of each unit in the population can be obtained and tabulated without any errors. In practice, this assumption may be violated due to several reasons and practical constraints. This results in errors in the observations as well as in the tabulation. Such errors which are due to the factors other than sampling are called non-sampling errors.

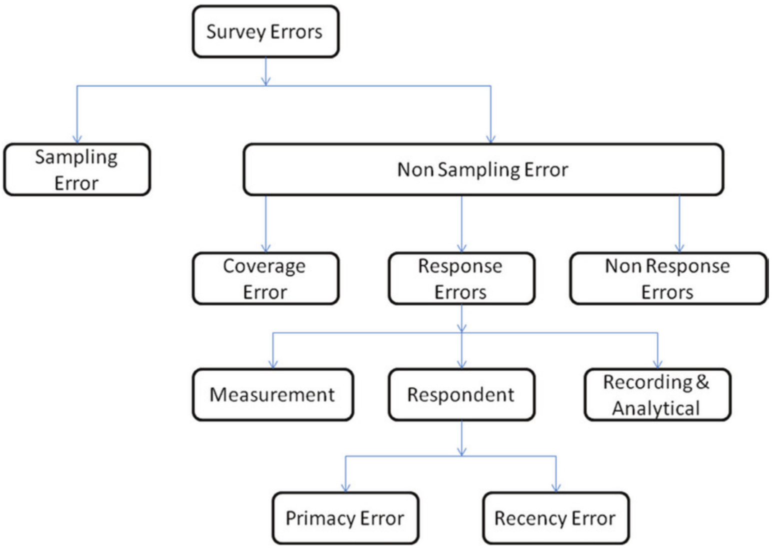


Figure 8

The non-sampling errors are **unavoidable in census and surveys.** The data collected by complete enumeration in census is free from sampling error but would not remain free from non-sampling errors. The data collected through sample surveys can have both sampling errors as well as non-sampling errors. The non-sampling errors arise because of the factors other than the inductive process of inferring about the population from a sample.

In general, the sampling errors decrease as the sample size increases whereas non-sampling error increases as the sample size increases. In some situations, the non-sampling errors may be large and deserve greater attention than the sampling error. In any survey, it is assumed that the value of the characteristic to be measured has been defined precisely for every population unit. Such a value exists and is unique. This is called the true value of the characteristic for the population value.

In practical applications, data collected on the selected units are called survey values and they differ from the true values. Such difference between the true and observed values is termed as the observational error or response error. Such an error arises mainly from the lack of precision in measurement technique s and variability in the performance of the investigators.

9.2 Non-Sampling Errors

Non sampling errors can occur at every stage of planning and execution of survey or census. It occurs at planning stage, field work stage as well as at tabulation and computation stage.

The main sources of the non sampling errors are;

(a) lack of proper specification of the domain of study and scope of investigation,

(b) incomplete coverage of the population or sample,

(c) faulty definition,

(d) defective methods of data collection and tabulation errors.

More specifically, one or more of the following reasons may give rise to non-sampling errors or indicate its presence:

(a) The data specification may be inadequate and inconsistent with the objectives of the survey or census.

(b) Due to imprecise definition of the boundaries of area units, incomplete or wrong identification of units, faulty methods of enumeration etc, the data may be duplicated or may be omitted.

(c) The methods of interview and observation collection may be inaccurate or inappropriate.

(d) The questionnaire, definitions and instructions may be ambiguous.

(e) The investigators may be inexperienced or not trained properly.

(f) The recall errors may pose difficulty in reporting the true data.

(g) The scrutiny of data is not adequate.

(h) The coding, tabulation etc. of the data may be erroneous.

(i) There can be errors in presenting and printing the tabulated results, graphs etc.

(j) In a sample survey, the non-sampling errors arise due to defective frames and faulty selection of sampling units.

These sources are not exhaustive but surely indicate the possible source of errors. Non-sampling errors may be broadly classified into three categories.

(a) Specification errors: These errors occur at planning stage due to various reasons, e.g., inadequate and inconsistent specification of data with respect to the objectives of surveys/ census, omission or duplication of units due to imprecise definitions, faulty method of enumeration/interview/ambiguous schedules etc.

(b) Ascertainment errors: These errors occur at field stage due to various reasons e.g., lack of trained and experienced investigations, recall errors and other type of errors in data collection, lack of adequate inspection and lack of supervision of primary staff etc.

(c) Tabulation errors: These errors occur at tabulation stage due to various reasons, e.g., inadequate scrutiny of data, errors in processing the data, errors in publishing the tabulated results, graphs etc.

Ascertainment errors may be further sub-divided into

(i) under Coverage errors owing to over-enumeration or under-enumeration of the population or the sample, resulting from duplication or omission of units and from the non-response.

(ii) under Content errors relating to the wrong entries due to the errors on the part of investigators and respondents.

Same division can be made in the case of tabulation error also. There is a possibility of missing data or repetition of data at tabulation stage which gives rise to coverage errors and also of errors in coding, calculations etc. which gives rise to content errors.

Treatment of non-sampling errors: Some conceptual background is needed for the mathematical treatment of non-sampling errors.

Total error: Difference between the sample survey estimate and the parametric true value being estimated is termed as total error.

9.3 Sampling errors

If complete accuracy can be ensured in the procedures such as determination, identification and observation of sample units and the tabulation of collected data, then the total error would consist only of the error due to sampling, termed as bold sampling error.

Measure of sampling error is mean squared error (MSE).

The MSE is the difference between the estimator and the true value and has two components:

(a) square of sampling bias.

(b) sampling variance.

If the results are also subjected to the non-sampling errors, then the total error would have both sampling and non-sampling error.

Total bias: The difference between the expected value and the true value of the estimator is termed as total bias. This consists of sampling bias and non-sampling bias.

Non-sampling bias: For the sake of simplicity, assume that the two following steps are involved in the randomization:

(i) for selecting the sample of units and (ii) for selecting the survey personnel.

10 Organization of National surveys, and the Kenya Bureau of Statistics(K.N.B.S)

10.1 Introduction

The Statistics Act 2006 specifically mandates KNBS to: Enumerate Act as the principal agency of the government for collecting, analyzing and disseminating statistical data in Kenya

(a) Act as custodian of official statistics.

(b) Conduct the Population and Housing Census every ten years, and such other censuses and surveys as the Board may determine;

(c) Maintain a comprehensive and reliable national socio-economic database.

(d) Establish and promote the use of best practices and methods in the production and dissemination of statistical information across the National Statistical System (NSS); and

(e) Plan, authorize, coordinate and supervise all official statistical programmes undertaken within the national statistical system.

NOTE: Check `www.knbs.org` for more information.

References

Lohr, S. L. (2010). Sampling: design and analysis (advanced series). Brooks/Cole Cengage Learning.

Mukhopadhyay, P. (2008). Theory and methods of survey sampling. PHI Learning Pvt. Ltd.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3), 581–592.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.