# Design and analysis of sample surveys

School of Science and Informatics

Department of Mathematics, Statistics and Physical Sciences

Taita Taveta University

Dr. Noah Mutai

January 31, 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, $\theta(y)$, 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:

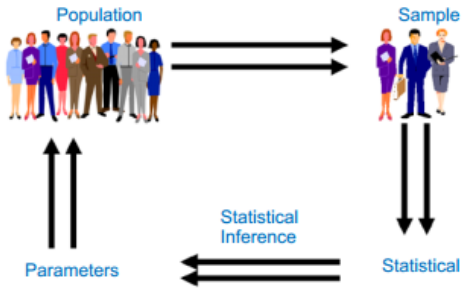


**Figure 1:** Illustration: Sampling

**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 $\pi_i$.

(ii) We select one of the $S_i$ by a process in which each $S_i$ receives its appropriate probability $\pi_i$, 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.

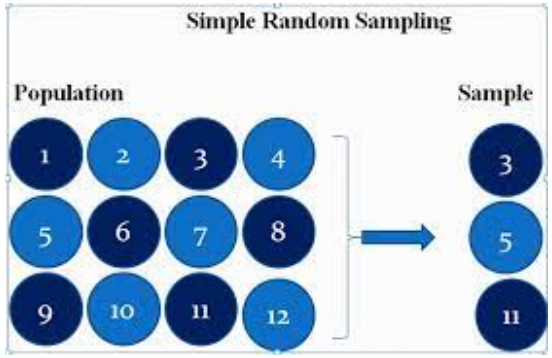


**Figure 2:** Illustration: simple random sampling

## 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 $srswr(N, n)$.

**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, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{N} t_i Y_i \Rightarrow E\left(\bar{y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{N} t_i Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} Y_i E\left(t_i\right)$. 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 $\hat{Y} = NVar\left(\bar{y}\right)$. Prove this.

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

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

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

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

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

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

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

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

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

Note:

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

□

**Corollary 2.2.1.** In $srswr\left(N, n\right)$ the sample variance is given by;

$$Var\left(\bar{y}\right) = \frac{N-1}{Nn}S_y^2; S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 \tag{2.7}$$

**Corollary 2.2.2.** In $srswr\left(N, n\right)$, $Var\left(\hat{Y}\right) = \frac{N^2\sigma^2}{n}$. As n increases, $Var\left(\bar{y}\right)$ decreases, even if $n = N$, $Var\left(\bar{y}\right)$ does not vanish. Also in $srswr$, $n$ may be arbitrarily large.

## 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.

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

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.

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

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.*

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

**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.*

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

The quantities $\pi_i$ 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; $E(a_i) = 1\times$ probability of $i^{th}$ selected unit.

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

$E(a_i, a_j) = 1\times$ 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 $Var(y) = E\left(y^2\right) - (E(y))^2$

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

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

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

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

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

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;

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

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 $\sigma^2$, $Var\left(\bar{y}\right) = \frac{\sigma^2}{n}$. 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.

## 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;

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

then we say that $(a, b)$ is a 95% confidence interval for $\theta$. It is important to realize that the $\theta$ 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.

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

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

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

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

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

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

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

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

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

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

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





$= 3302 \pm 272.05$

$\Rightarrow 3029.05 \leq Y \leq 3574.05$

## 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

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

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

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

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

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

$= \frac{PQ}{n}$

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

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

### 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, $tS\sqrt{\frac{N-n}{Nn}} = d$ where $\frac{1}{n} = \frac{1}{N} + \frac{d^2}{t^2S^2}$. Hence, $n = \frac{\left(\frac{ts}{d}\right)^2}{1+\frac{1}{N}\left(\frac{ts}{d}\right)^2}$. As a first approximation we may take $n_o = \left(\frac{ts}{d}\right)^2$. If $\frac{n_0}{N}$ is negligibly small, this may be taken as the satisfactory value of n. If not, one should calculate $n = \frac{n_0}{1+\left(\frac{n_0}{N}\right)} = n_0\left(1+\frac{n_0}{N}\right)^{-1}$. 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

$$P\left(\mid p-P \mid \leq d\right) \geq 1-\alpha \tag{2.30}$$

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

$$P\left(\mid p-P \mid \leq t\sqrt{\frac{N-n}{n\left(N-1\right)}}PQ\right)=1-\alpha \tag{2.31}$$

Equating 2.28 and 2.29 we get $t\sqrt{\frac{N-n}{n(N-1)}}PQ = d$. This gives;

$$n=\frac{\left(\frac{t^2PQ}{d^2}\right)}{1+\left(\frac{1}{N}\right)\left[\left(\frac{t^2PQ}{d^2}\right)-1\right]} \tag{2.32}$$

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 $n_0 = \frac{t^2PQ}{d^2}$. If $\frac{n_0}{N}$ is negligible, $n_0$ is a satisfactory approximation to n. If not, one should calculate $n$ as;

$$n=\frac{n_0}{1+\left[\left(\frac{n_0-1}{N}\right)\right]} \approx \frac{n_0}{1+\left(\frac{n_0}{N}\right)} \tag{2.33}$$

**Example 2.2.** 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,
$\left(\mid \bar{y}-\bar{Y} \mid \leq 0.1\bar{Y}\right) = 0.95$. 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 2.3.**

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 $\sigma^2$.

(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

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

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

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

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

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

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

□

### 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

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

This allocation arises when the $Var\left(\bar{y}_{st}\right)$is minimized subject to the constraint $\sum_{i=1}^{L} n_i$(pre-specified). 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 $Var\left(\bar{y}_{st}\right) = \sum_{i=1}^{L}\left(\frac{N_i - n_i}{N_i n_i}\right)w_i^2 S_i^2$,

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

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

(ii) under Optimum allocation

Under optimum allocation;

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

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

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

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

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

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

**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. $Var_{prop}(\bar{y}_{st})$ and $Var_{opt}(\bar{y}_{st})$

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

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

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

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

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

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

**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 Ratio and regression estimation

## 4.1 Ratio Estimation

### 4.1.1 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 4.1.** 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,

$$\hat{\bar{Y}} = \frac{\bar{y}}{\bar{x}}\bar{X} = \hat{R}\bar{X} \tag{4.1}$$

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

$$\hat{Y}_{R(tot)} = \frac{y_{tot}}{x_{tot}} X_{tot} \tag{4.2}$$

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.

### 4.1.2 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

$$\frac{4000}{400}.124 = 1240 \tag{4.3}$$

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

$$\frac{84}{240}.2700 + \frac{40}{160}.1300 = 1270 \tag{4.4}$$

This use of ratio estimation, called post-stratification.

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

**Example 4.1.** 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

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{297897.0467}{301953.7233} = 0.986565 \tag{4.5}$$

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

$$\hat{\bar{y}} = \hat{B}\bar{x} = (\hat{B})(313,343.283) = 309,133.6 \tag{4.6}$$

and

$$\hat{t}_{yr} = \hat{B}t_x = (\hat{B})(964,470,625) = 951,513,191. \tag{4.7}$$

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

### 4.1.3 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 $E\left(\hat{\bar{Y}}_R\right) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \frac{\bar{y}}{\bar{x}} \bar{X} \neq \bar{Y}$

(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,

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

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

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