Chapter 1

1. Introduction

Sampling theory is a branch of statistics that provides a framework for making inferences about a population based on a subset of that population, called a sample.

1.2 Some Concepts

-**Statistics** is the science of data.

-**Data** are the numerical values containing some information

-**Population:** is the entire group that you want to draw conclusions about. Or it is a Collection of all the sampling units in a given region at a particular point of time or a particular period.

-**Types of statistical population**

1-Finite Population

2-Infinite Population

-**The population size** the population size, generally denoted by *N*.

can be finite or infinite (*N* is large)

-**Sample:** a sample is a group of people, objects, or items that are taken from a large population for measurement it is the specific group of individuals of collecting data from. Or a sample consists only of a portion of the population units. Or Collection of One or more sampling units selected from the population according to some specified procedure.

- Sample size

The number of individuals you should include in your sample depends on various factors, including the size and variability of the population and your research design. There are different sample size calculators and formulas depending on what our want to achieve with statistical analysis.

-Representative sample:

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

Example it is expected that a drop of blood will give the same information as all the blood in the body.

1.3Sources of data and information

1.3.1-Historical Sources

It represents data and information stored and collected by state agencies, institutions

1.3.2-Field Sources It represents data and information that can be obtained from their original sources through confrontation, correspondence, or any other method of communication Statistical tools

1.4Methods of data collection

1.4.1- Complete enumeration or census

census collects information on the whole population in this kind sample size is equal to population size i.e it is a complete count of the population For example, in every country, the census is conducted at every tenth year in which observations on all the persons staying in their country **is** collected.

1.4.2-Sampling

Why there is need of sampling? the answer is as it is too expensive and too time consuming to survey a whole population in a research study, we use sampling

Sampling is a statistical procedure that is concerned with the selection of the individual observation; it helps us to make statistical inferences about the population. and is defined as a procedure to select a sample from individual or from a large group of population for certain kind of research purpose. There are two primary technique of sampling methods that you can use in your research: Probability sampling and Non-probability sampling

-Sampling Frame The sampling frame is the actual list of individuals that the sample will be drawn from. it should include the entire target population (and who is not part of that population).

- **Sampling unit** is the individual or group that you actually choose from the sampling frame.

For example, if you are studying the opinions of college students, your sampling frame could be the enrollment records of all the colleges in your area, and your sampling unit could be a student or a class.

Example it is expected that a drop of blood will give the same information as all the blood in the body.

1.5 Advantages and disadvantages of sampling Method

1.5.1 Advantages of Sampling over Census:

1- Reduction in cost: Sampling usually results in reduction in cost in terms of money and man hours. Since in most of cases our resources are limited in terms of money and the time, sampling is more advantageous than census., sampling is best. When there is large population, sampling is the best way

2-Greater Speed (Less Time):

There is considerable saving in time and labor since only a portion of population has to be examined. At the same time, results can be obtained rapidly and analyzed much faster. Saves time and gives faster results as the sample size is smaller than the whole population

3-Greater Accuracy:

The results of a sample survey are usually much more reliable than those obtained from a complete census due to the following reasons:

➢ It is possible to determine the extent of the sampling error

- ➢ Scope of non-sampling errors is less in sampling compared with census

4-Greater Scope:

The complete census is impracticable if the survey requires a highly trained personnel and more sophisticated equipment for collection and analysis of data. It is possible to have a thorough and intensive enquiry because a more detailed information can be obtained from a small group of respondents.

1.5.2 Disadvantage of sampling Method

The main disadvantage of the sampling is chances of bias. But, seeing so many of advantages, sampling is the best way to proceed in a research.

1.6Sampling Error vs. Non-sampling Error

There are different types of errors that can occur when gathering statistical data.

1.6.1 Sampling Errors:

Sampling error is a type of error, occurs due to the sample selected does not perfectly represents the population. Cause Deviation between sample mean and population mean

1.6.2Non-Sampling Errors: are errors that result during data collection and cause the data to differ from the true values. Non-sampling errors are caused by human error, such as a mistake made in the survey process. Deficiency and analysis of data

Items	Sampling Errors	Non-Sampling Errors
Meaning	Sampling error is a type of error, occurs due to the sample selected does not perfectly represents the population.	An error occurs due to sources other than sampling, while conducting survey activities is known as non -sampling error.

Cause	Deviation between sample mean and population mean	Deficiency and analysis of data
Type	Random	Random or Non-random
Occurs	Only when sample is selected.	Both in sample and census.
Sample size	Possibility of error reduced with the increase in sample size.	It has nothing to do with the sample size.

1.7 Steps in Sampling Process

1-Defining the target population.

2-Specifying the sampling frame.

3-Selection of the sampling method.

4-Determination of sample size.

5-Collect the sample data

6-Specifying the sampling plan.

7-Analyzing the data

8-Estimation and the decision

1.8 Techniques of Sampling:

Sample surveys collect information on a fraction of total population whereas There are two primary technique of sampling methods that you can use in your research: Probability sampling and Non-probability sampling

First== probability(Random) sampling

a-Simple Probability Samples

b-Stratified Random Sampling

c-Systematic Sampling

d-Cluster Sampling

Second==Non-probability sampling(Non-randomsample or purposive sample):

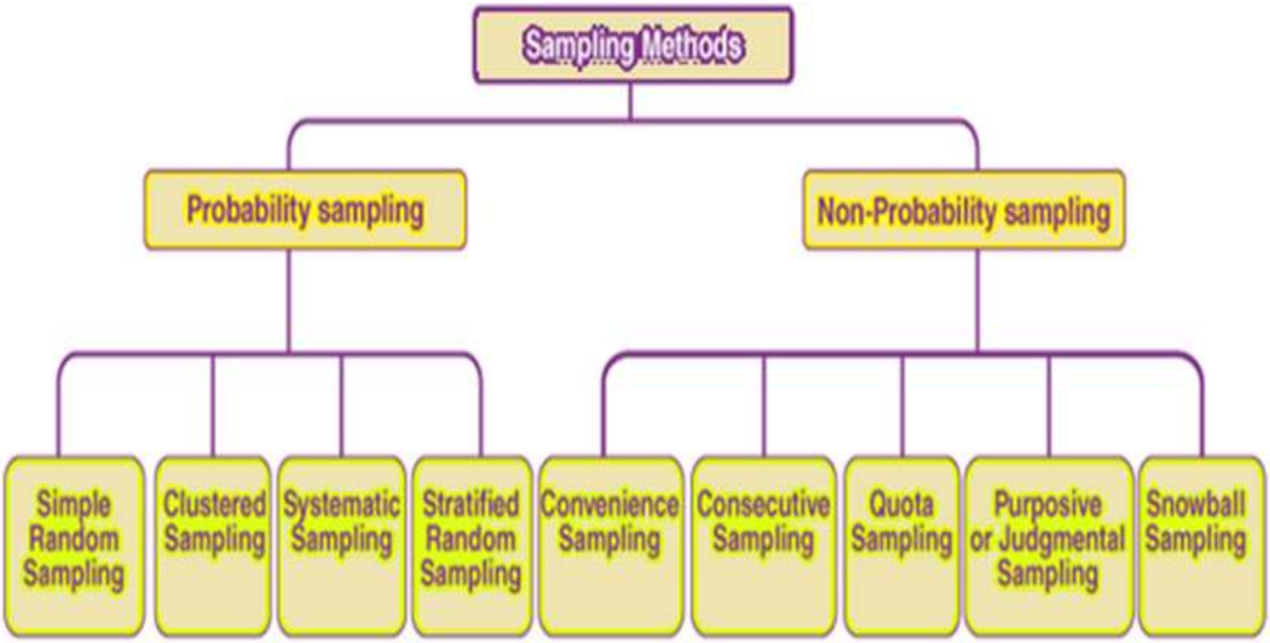
a-Convenience sampling

b-Quota sampling

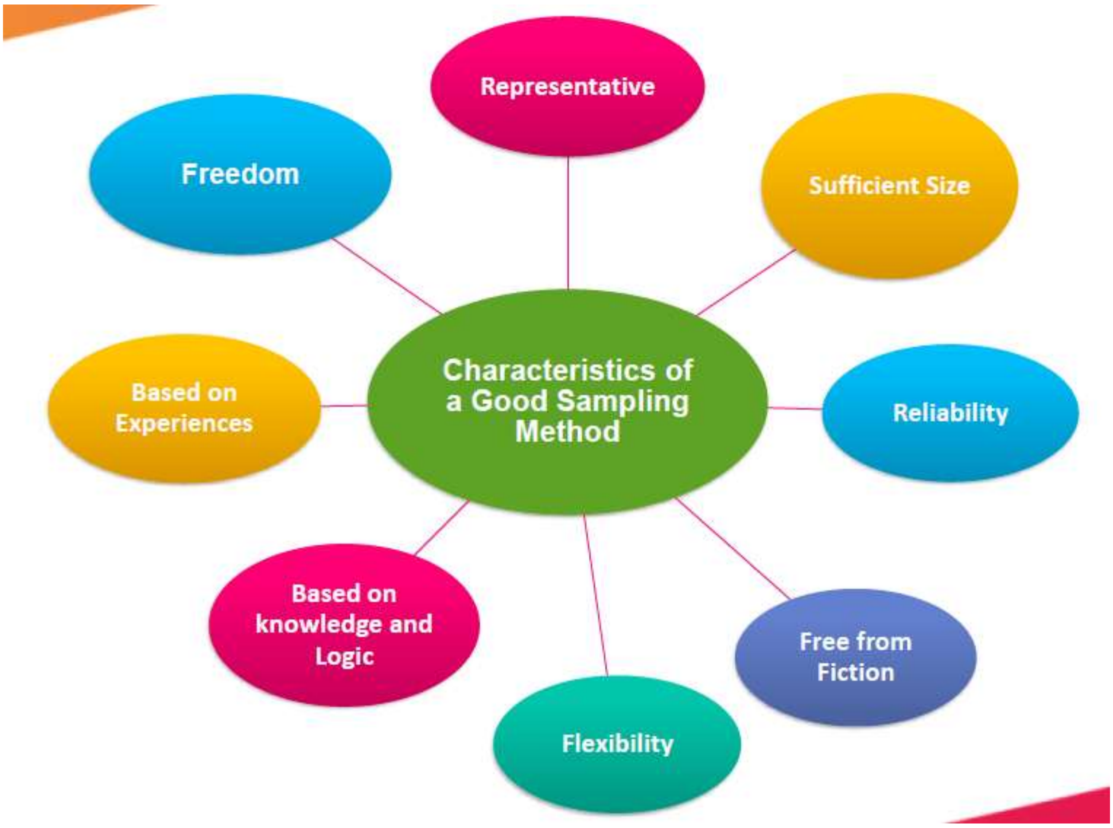
c-Self-selection (volunteer) sampling

d-Snowball sampling

e-Purposive (judgmental) sampling



Characteristics of good Sampling method



Chapter Two /First Section

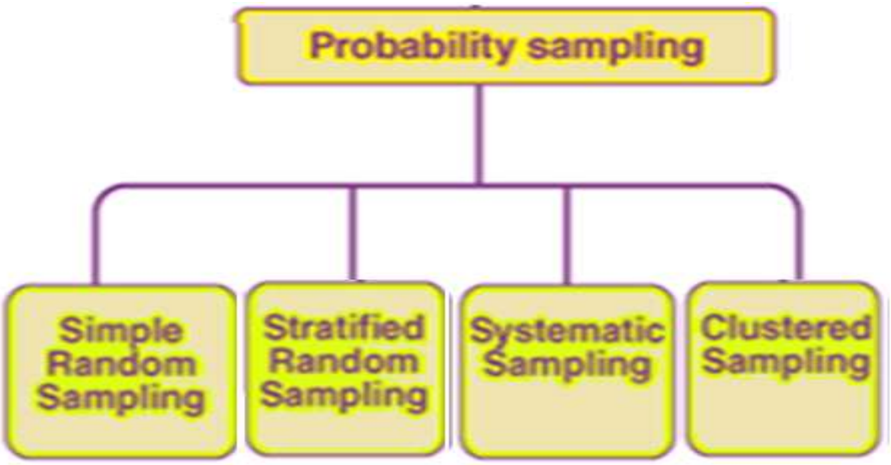
(SRS & SRSWOR)

2.1 First== probability(Random) sampling

2.1Definition of Random Sampling

Random Sampling is the simplest and most common method of selecting a sample, in which the sample is selected unit by unit, with equal probability of selection for each unit at each draw in this method, every individual has an equal chance**,** the sample chosen randomly is meant to be unbiased representation of the total population and it allows for the randomization of sample selection.

2.1.1Techniques of probability Sampling or Random Sampling:



2.1.2-Simple Random Sampling (SRS)

It is a method of selection of a sample contains n number of sampling units out of the population having N number of sampling units such that every sampling unit has an equal chance of being chosen. Here, the selection of items entirely depends on luck or probability. Therefore, this sampling technique is also a method of chance.

The important condition is the homogeneity in populations unit

When can you use random sampling?

1-If the population size is small or the size of the individual samples and their number are relatively small.

2-random sampling provides the best results since all candidates have an equal chance of being chosen.

3-It is best for population which is highly homogenous population of the study are uniformly distributed...

4-A sample chosen randomly is meant to be an unbiased representation of the total population.

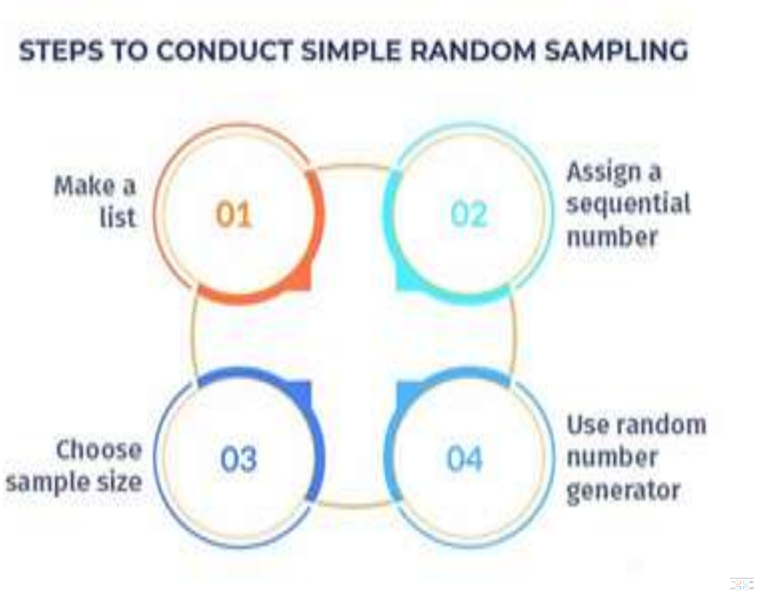
5- The sampling frame should include the whole population.

6-Simple random samples are determined by assigning sequential values to each item within a population, then randomly selecting those values.

Note:

to randomly select a percentage, is 30%, of the observations in the data set

2.1.3Steps to perform Simple Random Sampling


STEPS TO CONDUCT SIMPLE RANDOM SAMPLING
Make a list
01
02
Assign a sequential number
Choose sample size
03
04
Use random number generator

Step 1: Define the Population (Make a List)

Step 2: Assign a Sequential Number

Step 3: Choose Sample Size

Step 4: Use a Random Number

a random number is a number chosen by chance - i.e., randomly.

2.1.4Fundamental Methods to Generate a Random Variable

There several methods for determining the random **Number**s to be selected as follows:

1-Math.random(): This is a method in JavaScript that generates random numbers between 0 and 1

2- Resampling

Resampling involves the selection of randomized cases with replacement from the original data sample in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample

is the method that consists of drawing repeated samples from the original data samples. The method of Resampling is a nonparametric method of statistical inference. In other words, the method of resampling does not involve the utilization of the generic distribution tables (for example, normal distribution tables) in order to compute approximate p probability values.

There are four common ways to resample raster grids in GIS.

- Nearest Neighbor.
- Cubic Convolution.
- Bilinear.
- Majority.

3-Simulation/Game-play.

random numbers are used in simulations to model real-world phenomena and test different hypotheses. For example, scientists can use random numbers to generate different climate models and predict future weather patterns. Similarly, engineers can use random numbers to test the safety and reliability of machines and systems under different conditions.

Transform methods.

4-Random lottery. Whether by ping-pong ball or slips of paper, each population number receives an equivalent item that is stored in a box or other indistinguishable container. Then, random numbers are selected by pulling or selecting items without view from the container.

5-Physical Methods. Simple, early methods of random selection may use dice, flipping coins, or spinning wheels. Each outcome is assigned a value or outcome relating to the population.

6-Random number table. Many statistics and research books contain sample tables with randomized numbers.

7- Online random number generator.

To generate a table of random numbers, you can use a random number generator or a statistical software

Many online tools exist where the analyst inputs the population size and sample size to be selected. Random Number Generator Class : This C++ class provides many functions for generating random integers, floats, etc., as well as providing various distributions like normal, exponential, Gamma and Weibull.

8-Random numbers from Excel. Numbers can be selected in Excel using the =RANDBETWEEN formula.

Example: A cell containing =RANDBETWEEN (1,5) will selected a single random number between 1 and 5.

9-NativeCrypto : This is an Android and iOS library that provides cryptographically secure random values.

10- Choose the sampling unit whose serial number corresponds to the random number drawn from the table of random numbers.

Notes: **1- In the case of SRSWOR**, if any random number is repeated, then it is ignored, and more numbers are drawn.

2-In the case of SRSWR, all the random numbers are accepted even if repeated more than once.

2.1.5 Sample Size H.W

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

H.W

= Pre-specified variance

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

It may be noted here that can be known only when is known. This reason compels to assume that should be known. The same reasoning will also be seen in other cases. The smallest sample size needed in this case is

= Pre-specified estimation error

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

=Pre-specified width of the confidence interval

If the requirement is that the width of the confidence interval of with confidence coefficient should not exceed a pre-specified amount , then the sample size is determined such that

= Pre-specified coefficient of variation

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

= Pre-specified relative error:When is used for estimating the population mean , then the relative estimation error is defined as . If it is required that such relative estimation error should not exceed a pre-specified value with probability, then such requirement can be satisfied by expressing it like such

= Pre-specified cost

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

2.13Corollary (1)

2.14Theorem 2.2 H.W

Prove that the variance of the sample me is given by:

$\boldsymbol{V(\bar{y}) = \frac{\sigma^2}{n}(1 - f)}$ if $f = \frac{n}{N}$ is the sampling fraction

2.1.6 Types of Simple Random Sampling

There are two types of simple random sampling

1-Simple Random Sampling Without Replacement(SRSWOR)

in which a subset of the observations is selected randomly, and once an observation is selected it cannot be selected again.

Simple random sampling without replacement of size n is the probability sampling design for which a fixed number of n units are selected from a population of N units without replacement such that every possible sample of n units has equal probability of being selected.It is drawn as 1/N for the first draw, 1/ (N-1) for the second, 1/ (N-r+1) for the third, and so on. Therefore, the probability of drawing “n” units from a sample and its selection in the rth draw is n/N.

In sampling without replacement, the two sample values **aren't independent.**

How do you find probability without replacement ?

The short answer is P(A and B) = P(A) * P(B|A) or P(A and B) = P(B) * P(A|B). Without replacement means that you will have a conditional probability.

Practically, this means that the first selection has an effects on the second selection . Mathematically, this means that the covariance between the two selection not equal to zero.

the number possible samples by SRSWOR is $\binom{\boldsymbol{N}}{\boldsymbol{n}}$

The probability of each sample is $\frac{\mathbf{1}}{\binom{\boldsymbol{N}}{\boldsymbol{n}}}$

2.6.1 Probability of drawing a unit

Now if $(u_{1\,.}u_{2\,.}u_{3\,.}\cdots.u_r)$ are the r units selected in the sample, then the probability of their selection is

$$P(u_1)=\frac{N-1}{N}$$

$$P(u_2)=\frac{N-2}{N-1}$$

$$\vdots \qquad\qquad \vdots$$

$$P(u_r)=\frac{N-(N-1)}{N-r+1}=\frac{1}{N-r+1}$$

The probability of selecting a specified unit at the rth draw is:

$$P(u_{1\,.}u_{2\,.}u_{3\,.}\cdots.u_r)=P(u_1)P(u_2)P(u_3)\cdots P(u_n)$$

$$=\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\frac{N-3}{N-2}\cdots\ \cdot\frac{1}{N-r+1}=\frac{1}{N}$$

2.6.2 Probability of drawing a Sample [Simple Random Sampling Without Replacement (SRSWOR)]:

If units are selected by SRSWOR,

$$\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot\frac{N-3}{N-2}\cdots\ \cdot\cdot\frac{N-n}{N-N+1}=\frac{(N-n)!}{N!}$$

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

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

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

The total number of possible samples = $\binom{\boldsymbol{N}}{\boldsymbol{n}}$

The probability of selecting any one of these samples = $\frac{\boldsymbol{1}}{\binom{\boldsymbol{N}}{\boldsymbol{n}}}$

drawn is irrelevant =$n!\,\frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$

$$P(u_{1\,.}u_{2\,.}u_{3\,.}\cdots.u_n) = P(u_1)P(u_2)P(u_3)\cdots P(u_n) = n!\frac{(N-n)!}{N!}$$

$$P(u_{1\,.}u_{2\,.}u_{3\,.}\cdots.u_n) = P(u_1)P(u_2)P(u_3)\cdots P(u_n) = \frac{1}{\binom{\boldsymbol{N}}{\boldsymbol{n}}}$$

Example #1

There is a group of 6 people, and two will be chosen as leaders. First, let us look at the probability through SRS. Assuming that sampling is done here without replacement.

$$\text{The total number of possible samples } = \binom{\boldsymbol{N}}{\boldsymbol{n}} = \frac{\boldsymbol{N}!}{\boldsymbol{n}!\,(\boldsymbol{N}-\boldsymbol{n})!}$$

$$= \frac{\mathbf{6}!}{\mathbf{2}!\,(\mathbf{6}-\mathbf{2})!} = \frac{(\mathbf{5})(\mathbf{3})}{\mathbf{1}} = \mathbf{15}$$

$$\text{The probability of selecting any one of these samples } = \frac{\mathbf{1}}{\mathbf{15}} = \mathbf{0.06}$$

2.8. Estimation of population mean

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatteredness of the data around the central value. Among various measures of central tendency and dispersion, the popular choices are arithmetic mean and variance. So the population mean and population variability are generally measured by the arithmetic mean (or weighted arithmetic mean) and variance, respectively.

2.8.1. Estimation of population mean (SRSWOR)

Theorem 2.1

Prove that the sample mean $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ is an is an unbiased estimator of the population mean $\bar{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$.

Proof

under the two cases

First Case SRSWOR

First Method

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

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

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

since probability of selecting one sample is $\frac{1}{\binom{N}{n}}$

So

$$E(\bar{y}) = \frac{1}{n}\left[\frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} t_i\right] = \frac{1}{n}\left[\frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \left(\sum_{i=1}^{n} y_i\right)\right]$$

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

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

Now $E(\bar{y}) = \frac{1}{n}\left[\frac{1}{\binom{N}{n}}\binom{N-1}{n-1}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\left[\frac{1}{\frac{N!}{n!(N-n)!}}\frac{(N-1)!}{(n-1)!(N-1-n+1)!}\sum_{i=1}^{n} y_i\right]$

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

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

Second Method

Alternatively, the following approach can also be adapted to show the unbiasedness property. Let

$$P_j(i) = \frac{1}{N}$$

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

$$E(\bar{y}) = \frac{1}{n}\sum^{n}\left[\sum_{i=1}^{n} Y_i\, P_j(i)\right]$$

$$= \frac{1}{n}\sum^{n}\left[\sum_{i=1}^{n} Y_i \frac{1}{N}\right]$$

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

2.8.1.1 Corollary

Prove that: $\boldsymbol{E(\bar{y}) = \mu}$

1-Proof : $\boldsymbol{E\left(\frac{\sum y_i}{n}\right) = \frac{E(y_1+y_2+y_3+\cdots+y_n)}{n} = \frac{E(y_1)+E(y_2)+E(y_3)+\cdots+E(y_n)}{n}} = \frac{\mu+\mu+\mu+\cdots+\mu}{\boldsymbol{n}} = \frac{n\mu}{\boldsymbol{n}} = \mu$

Example

If we have a population contain (3,5,6,7)and we draw a sample of size 2, prove that: $\boldsymbol{E(\bar{\bar{y}})} = \mu$

Proof :

$$\mu = \frac{\sum_{ally} y_i}{n} = \frac{\sum_{ally} y_i}{n} = \frac{3+5+6+7}{4} = \frac{3+5+6+7}{4} = \frac{21}{4}$$

$$= 5 \cdot 25$$

$$number\ of \text{all possible selected samples} = \binom{N}{n} = \binom{4}{2} = \frac{4!}{2!\,(4-2)!} = 6$$

$Number\ of$ Samples	$\bar{y}$
(3.5)	4
(3.6)	$4 \cdot 5$
(3.7)	5
(5.6)	$5 \cdot 5$
(5.7)	6
(6.7)	$6 \cdot 5$
	$31 \cdot 5$

$$\boldsymbol{E(\bar{y})} = \frac{\sum_{allx} \boldsymbol{\bar{x}}}{\binom{\boldsymbol{N}}{\boldsymbol{n}}} = \frac{4 + \mathbf{4 \cdot 5} + \ 5 + \mathbf{5 \cdot 5} + \ 6 + \mathbf{6 \cdot 5}}{\binom{4}{2}} = \frac{\mathbf{31 \cdot 5}}{\mathbf{6}} = 5 \cdot 25$$

2.9. Estimation of population variance (Estimation of variance from a sample)

Since the expressions of variances of the sample mean, involve σ^2 which is based on population values, so these expressions cannot be used in real-life applications. In order to estimate the variance of $(\bar{y})$ on the basis of a sample, an estimator of (σ^2) is needed.

Consider s^2 as an estimator of $(\sigma^2\ )$ and we investigate its biasedness for s^2 in the cases of SRSWOR and SRSWR,

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

Consider

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

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

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

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

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

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

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

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

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

$$E(s^2) = \tfrac{1}{n-1} [[n\sigma^2 - nVar(\ \bar{y}\ )]] \qquad \textbf{......(1)}$$

2.9.1In the case of SRSWOR

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

From (1)

$$E(s^2) = \frac{1}{n-1}[n\sigma^2 - nVar(\,\bar{y}\,)] = \frac{n}{n-1}[\sigma^2 - Var(\,\bar{y}\,)]$$

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

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

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

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

$$E(s^2) = S^2$$

Hence

$E(s^2) = S^2 \quad \{ \quad in$ SRSWOR $\}$

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

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

2-2Chapter Two

Second Section

(SRSWR)

2.2.1 The Probability of drawing a sample [Simple Random Sampling with Replacement (SRSWR)]:

When n units are selected with SRSWR,

1 -Total number of possible samples = N^n

2-The Probability of drawing a sample = $\frac{1}{N^n}$

Proof :

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

$$P(u_1.u_2.u_3.\cdots.u_n) = P(u_1)P(u_2)P(u_3)\cdots P(u_n) = \frac{1}{N}\cdot\frac{1}{N}\cdot\frac{1}{N}\cdots\frac{1}{N}$$

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

3-Probability of drawing a unit SRSWR

$$P\left[selection\ of\ u_j\ \ at\ n^{th}\ draw\right] = \frac{1}{N}$$

In sampling with replacement, the two sample values **are independent.**

Independent Events Two events are independent if the following are true: P(A|B) = P(A) P(B) / P(B) = P(A)

P(B|A) = P(B) P(A) / P(A)= P(B)

2.2.2 Estimation of population mean SRSWR

One of the main objectives after the selection of a sample is to know about the tendency of the data to cluster around the central value and the scatteredness of the data around the central value. Among various measures of central tendency and dispersion, the popular choices are arithmetic mean and variance. So the population mean and population variability are generally measured by the arithmetic mean (or weighted arithmetic mean) and variance, respectively.

Proof:

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

$$E(\bar{y}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} E(y_i)$$

$$E(\bar{y}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} (Y_1 P_1 + Y_2 P_2 + \cdots + Y_N P_N)$$

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

$$E(\bar{y}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} \left[Y_1 \left(\frac{1}{N}\right) + Y_2 \left(\frac{1}{N}\right) + \cdots + Y_N \left(\frac{1}{N}\right) \right]$$

$$E(\bar{y}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} (\bar{Y}) = \frac{1}{n} n\, \bar{Y} = \bar{Y}$$

Where$P_i = \frac{1}{N}$ for all $i = 1.2.3. \cdots . N$ is the probability of selection of a unit. Thus$(\bar{y})$ is an unbiased estimator of the population mean$(\bar{Y})$ under SRSWR also.

2.2.3 Estimation of population variance (Estimation of variance from a sample)

Since the expressions of variances of the sample mean, involve σ^2 which is based on population values, so these expressions cannot be used in real-life applications. In order to estimate the variance of $(\bar{y})$ on the basis of a sample, an estimator of (σ^2) is needed.

Consider s^2 as an estimator of $(\sigma^2\ )$ and we investigate its biasedness for SRSWR,

unbiased estimate of is $Var(\ \bar{y}\ )$in case of SRSWR

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

Consider

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

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

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

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

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

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

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

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

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

$E(s^2) = \frac{1}{n-1}\left[[n\sigma^2 - nVar(\bar{y})]\right]$ **……(1)**

In the case of SRSWR

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

and so

$$E(s^2) = \frac{1}{n-1}[n\sigma^2 - nVar(\bar{y})]$$

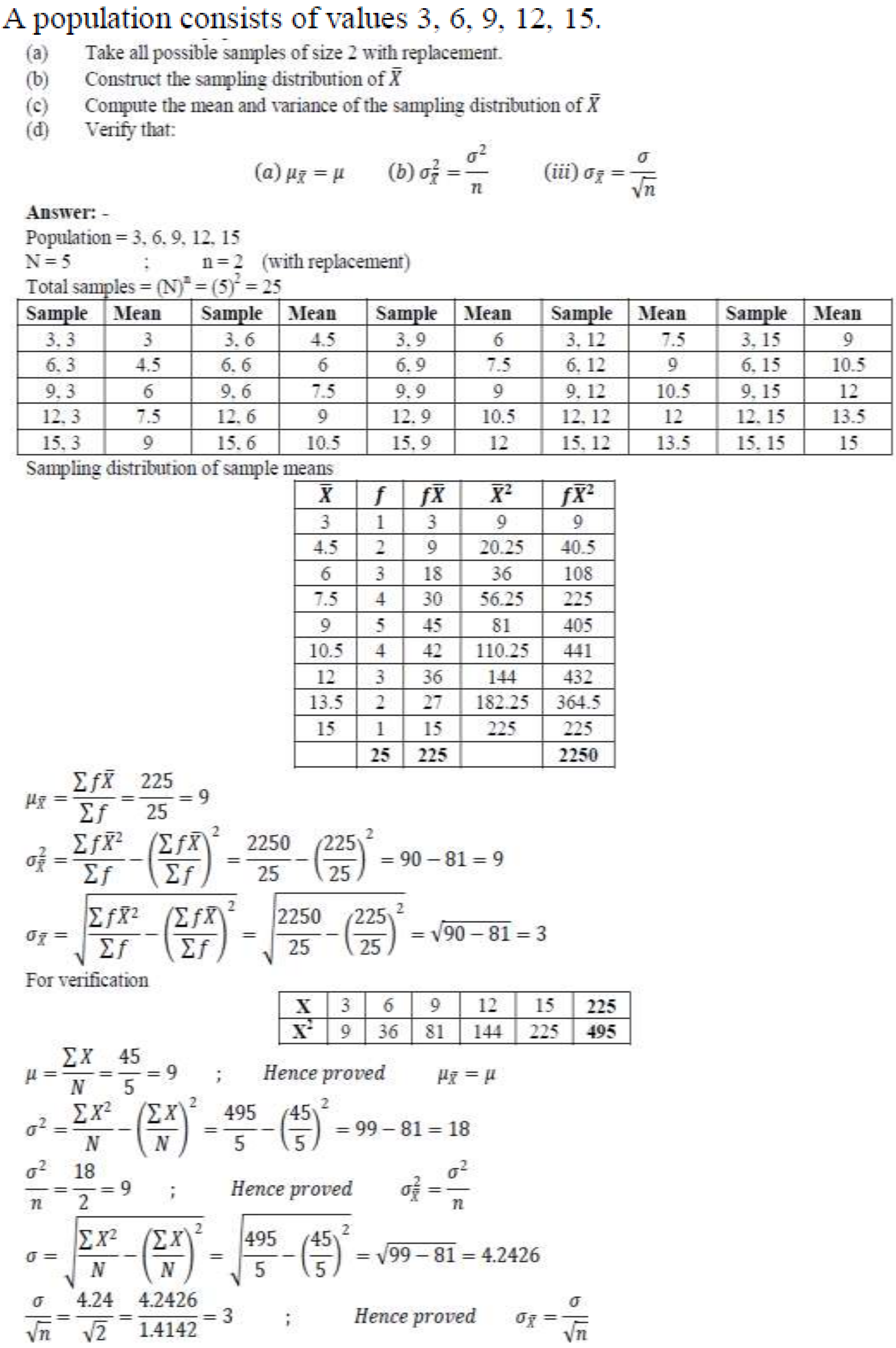
$$= \frac{n}{n-1}[\sigma^2 - Var(\bar{y})]$$

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

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

$$E(s^2) = \sigma^2$$

Example(1)



A population consists of values 3, 6, 9, 12, 15.

(a) Take all possible samples of size 2 with replacement.

(b) Construct the sampling distribution of $\bar{X}$

(c) Compute the mean and variance of the sampling distribution of $\bar{X}$

(d) Verify that:

$$(a)\ \mu_{\bar{X}} = \mu \qquad (b)\ \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \qquad (iii)\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Answer: -

Population = 3, 6, 9, 12, 15

N = 5 ; n = 2 (with replacement)

Total samples = $(N)^n = (5)^2 = 25$

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
3, 3	3	3, 6	4.5	3, 9	6	3, 12	7.5	3, 15	9
6, 3	4.5	6, 6	6	6, 9	7.5	6, 12	9	6, 15	10.5
9, 3	6	9, 6	7.5	9, 9	9	9, 12	10.5	9, 15	12
12, 3	7.5	12, 6	9	12, 9	10.5	12, 12	12	12, 15	13.5
15, 3	9	15, 6	10.5	15, 9	12	15, 12	13.5	15, 15	15

Sampling distribution of sample means

$\bar{X}$	f	$f\bar{X}$	$\bar{X}^2$	$f\bar{X}^2$
3	1	3	9	9
4.5	2	9	20.25	40.5
6	3	18	36	108
7.5	4	30	56.25	225
9	5	45	81	405
10.5	4	42	110.25	441
12	3	36	144	432
13.5	2	27	182.25	364.5
15	1	15	225	225
	25	**225**		**2250**

$$\mu_{\bar{X}} = \frac{\sum f\bar{X}}{\sum f} = \frac{225}{25} = 9$$

$$\sigma^2_{\bar{X}} = \frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2 = \frac{2250}{25} - \left(\frac{225}{25}\right)^2 = 90 - 81 = 9$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2} = \sqrt{\frac{2250}{25} - \left(\frac{225}{25}\right)^2} = \sqrt{90 - 81} = 3$$

For verification

X	3	6	9	12	15	**225**
X²	9	36	81	144	225	**495**

$$\mu = \frac{\sum X}{N} = \frac{45}{5} = 9 \quad ; \quad \textit{Hence proved} \quad \mu_{\bar{X}} = \mu$$

$$\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{495}{5} - \left(\frac{45}{5}\right)^2 = 99 - 81 = 18$$

$$\frac{\sigma^2}{n} = \frac{18}{2} = 9 \quad ; \quad \textit{Hence proved} \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{495}{5} - \left(\frac{45}{5}\right)^2} = \sqrt{99 - 81} = 4.2426$$

$$\frac{\sigma}{\sqrt{n}} = \frac{4.24}{\sqrt{2}} = \frac{4.2426}{1.4142} = 3 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Example (2)

- A population consists of values 3, 6, 9. Take all possible samples of size 3 with replacement and verify that:

$$(a)\ E(\bar{X}) = \mu \qquad (ii)\ Var(\bar{X}) = \frac{\sigma^2}{n}.\frac{N-n}{N-1} \qquad (iii)\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Answer: -

Population = 3, 6, 9

N = 3 ; n = 3 (with replacement)

Total samples = $(N)^n = (3)^3 = 27$

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
3, 3, 3	3	3, 9, 3	5	6, 6, 3	5	9, 3, 3	5	9, 9, 3	7
3, 3, 6	4	3, 9, 6	6	6, 6, 6	6	9, 3, 6	6	9, 9, 6	8
3, 3, 9	5	3, 9, 9	7	6, 6, 9	7	9, 3, 9	7	9, 9, 9	9
3, 6, 3	4	6, 3, 3	4	6, 9, 3	6	9, 6, 3	6		
3, 6, 6	5	6, 3, 6	5	6, 9, 6	7	9, 6, 6	7		
3, 6, 9	6	6, 3, 9	6	6, 9, 9	8	9, 6, 9	8		

Sampling distribution of sample means

$\bar{X}$	f	$f\bar{X}$	$\bar{X}^2$	$f\bar{X}^2$
3	1	3	0	0
4	3	12	1	3
5	6	30	4	24
6	7	42	9	90
7	6	42	16	192
8	3	24	25	300
9	1	9	36	360
	27	**162**		**1026**

$$\mu_{\bar{X}} = \frac{\sum f\bar{X}}{\sum f} = \frac{162}{27} = 6$$

$$\sigma_{\bar{X}}^2 = \frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2 = \frac{1026}{27} - \left(\frac{162}{27}\right)^2 = 38 - 36 = 2$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2} = \sqrt{\frac{1026}{27} - \left(\frac{162}{27}\right)^2} = \sqrt{38-36} = \sqrt{2} = 1.41$$

For verification

X	3	6	9	**18**
X²	9	36	81	**126**

$$\mu = \frac{\sum X}{N} = \frac{18}{3} = 6 \quad ; \quad \textit{Hence proved} \quad \mu_{\bar{X}} = \mu$$

$$\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{126}{3} - \left(\frac{18}{3}\right)^2 = 42 - 36 = 6$$

$$\frac{\sigma^2}{n} = \frac{6}{3} = 2 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{126}{3} - \left(\frac{18}{3}\right)^2} = \sqrt{42-36} = \sqrt{6} = 2.45$$

$$\frac{\sigma}{\sqrt{n}} = \frac{2.45}{\sqrt{3}} = \frac{2.45}{1.73} = 1.41 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

2.2.4Efficiency of under SRSWOR over SRSWR

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

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

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

$V(\bar{y}_{WR}) = \frac{(N-n)+(n-1)}{Nn}S^2$ **by adding(+n , -n)**

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

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

Thus

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

and so, SRSWOR is more efficient than SRSWR

2.2.5Confidence limits for the population mean

Now we construct the $100(1-\alpha)\%$ confidence interval for the population mean. Assume that the population is normally distributed $N(\mu\,.\,\sigma^2)$with mean μ and variance σ^2 then $\frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}}$ follows $N(0\,.\,1)$when $\sigma^2$is known.

If $\sigma^2$is unknown and is estimated from the sample, then $\frac{\bar{y}-\bar{Y}}{\sqrt{Var(\bar{y})}}$ follows t -distribution with$(n-1)$ degrees of freedom. When $\sigma^2$is known, then the $100(1-\alpha)\%$ confidence interval is given by

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

or $P\left[\bar{y} - Z_{\frac{\propto}{2}}\sqrt{Var(\bar{y})} < \bar{Y} < \bar{y} + Z_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\right] = 1-\propto$

and the confidence limits are

$$\left[\bar{y} - Z_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\,.\ \ \bar{y} + Z_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\right]$$

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

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

or $P\left[\bar{y} - t_{\frac{\propto}{2}}\sqrt{Var(\bar{y})} \leq \bar{Y} \leq \bar{y} + t_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\right] = 1{-}\propto$

and the confidence limits are

$$\left[\bar{y} - t_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\,.\,\bar{y} + t_{\frac{\propto}{2}}\sqrt{Var(\bar{y})}\right]$$

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

2.2.6Corollary

Prove that $\hat{Y} = N\bar{y}$ An unbiased estimate of the total population size **Y**

Proof:

$E(\hat{Y}) = EN\bar{y}$

$E(\hat{Y}) = NE\bar{y}$

$E(\hat{Y}) = N\bar{Y}$

$E(\hat{Y}) = N\frac{Y}{N} = Y$

$E(\hat{Y}) = Y$

Question 1 A population consists of ten values 2, 4, 6, 8, 10, 12. Take all possible samples of size 2 without replacement and verify that:

$$(a)\ E(\bar{X}) = \mu \qquad (ii)\ \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.\frac{N-n}{N-1} \qquad (iii)\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Answer: -

Population = 2, 4, 6, 8, 10, 12.

N = 6 ; n = 2 (without replacement)

Total samples = $^{N}C_{n} = {}^{6}C_{2} = 15$

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
2, 4	3	2, 10	6	4, 8	6	6, 8	7	8, 10	9
2, 6	4	2, 12	7	4, 10	7	6, 10	8	8, 12	10
2, 8	5	4, 6	5	4, 12	8	6, 12	9	10, 12	11

Sampling distribution of sample means

$\bar{X}$	f	$f\bar{X}$	$\bar{X}^2$	$f\bar{X}^2$
3	1	3	9	9
4	1	4	16	16
5	2	10	25	50
6	2	12	36	72
7	3	21	49	147
8	2	16	64	128
9	2	18	81	162
10	1	10	100	100
11	1	11	121	121
	15	**105**		**805**

$$\mu_{\bar{X}} = \frac{\sum f\bar{X}}{\sum f} = \frac{105}{15} = 7$$

$$\sigma_{\bar{X}}^2 = \frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2 = \frac{805}{15} - \left(\frac{105}{15}\right)^2 = 53.67 - 49 = 4.67$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2} = \sqrt{\frac{805}{15} - \left(\frac{105}{15}\right)^2} = \sqrt{53.67 - 49} = \sqrt{4.67} = 2.16$$

For verification

X	2	4	6	8	10	12	**42**
X²	4	16	36	64	100	144	**364**

$$\mu = \frac{\sum X}{N} = \frac{42}{6} = 7 \quad ; \quad \textit{Hence proved} \quad \mu_{\bar{X}} = \mu$$

$$\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{364}{6} - \left(\frac{42}{6}\right)^2 = 60.67 - 49 = 11.67$$

$$\frac{\sigma^2}{n}.\frac{N-n}{N-1} = \frac{11.67}{2}.\frac{6-2}{6-1} = \frac{11.67}{2}.\frac{4}{5} = \frac{46.68}{10} = 4.67 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{364}{6} - \left(\frac{42}{6}\right)^2} = \sqrt{60.67 - 49} = \sqrt{11.67} = 3.42$$

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3.42}{\sqrt{2}}\sqrt{\frac{6-2}{6-1}} = \frac{3.42}{\sqrt{2}}\sqrt{\frac{4}{5}} = \frac{6.84}{3.16} = 2.16 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.\frac{N-n}{N-1}$$

Question 2 : - A population consists of values 6, 12, 18, 24, 30, 36. Take all possible samples of size 3 without replacement from this population.

(a) Find the sampling distribution of $\bar{X}$

(b) Verify that:

$$(i)\ \mu_{\bar{X}} = \mu \qquad (ii)\ \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.\frac{N-n}{N-1} \qquad (iii)\ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.\sqrt{\frac{N-n}{N-1}}$$

Population = 6, 12, 18, 24, 30.

N = 5 ; n = 3 (without replacement)

Total samples = ${}^{N}C_{n} = {}^{5}C_{3} = 10$

Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean	Sample	Mean
6, 12, 18	12	6, 12, 30	16	6, 18, 30	18	12, 18, 24	18	12, 24, 30	22
6, 12, 24	14	6, 18, 24	16	6, 24, 30	20	12, 18, 30	20	18, 24, 30	24

Sampling distribution of sample means

$\bar{X}$	f	$f\bar{X}$	$\bar{X}^2$	$f\bar{X}^2$
12	1	12	144	144
14	1	14	196	196
16	2	32	256	512
18	2	36	324	648
20	2	40	400	800
22	1	22	484	484
24	1	24	576	576
	10	**180**		**3360**

$$\mu_{\bar{X}} = \frac{\sum f\bar{X}}{\sum f} = \frac{180}{10} = 18$$

$$\sigma^2_{\bar{X}} = \frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2 = \frac{3360}{10} - \left(\frac{180}{10}\right)^2 = 336 - 324 = 12$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum f\bar{X}^2}{\sum f} - \left(\frac{\sum f\bar{X}}{\sum f}\right)^2} = \sqrt{\frac{3360}{10} - \left(\frac{180}{10}\right)^2} = \sqrt{336-324} = \sqrt{12} = 3.46$$

For verification

X	6	12	18	24	30	**90**
X^2	36	144	324	576	900	**1980**

$$\mu = \frac{\sum X}{N} = \frac{90}{5} = 18 \quad ; \quad \textit{Hence proved} \quad \mu_{\bar{X}} = \mu$$

$$\sigma^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2 = \frac{1980}{5} - \left(\frac{90}{5}\right)^2 = 396 - 324 = 72$$

$$\frac{\sigma^2}{n}.\frac{N-n}{N-1} = \frac{72}{3}.\frac{5-3}{5-1} = \frac{72}{3}.\frac{2}{4} = \frac{144}{12} = 12 \quad ; \quad \textit{Hence proved} \quad \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}.\frac{N-n}{N-1}$$

$$\sigma = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\frac{1980}{5} - \left(\frac{90}{5}\right)^2} = \sqrt{396-324} = \sqrt{72} = 8.485$$

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{8.485}{\sqrt{3}}\sqrt{\frac{5-3}{5-1}} = \frac{8.485}{\sqrt{3}}\sqrt{\frac{2}{4}} = \frac{11.9996}{3.46} = 3.46 \quad ; \quad \textit{Hence proved} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.\sqrt{\frac{N-n}{N-1}}$$

Chapter Three

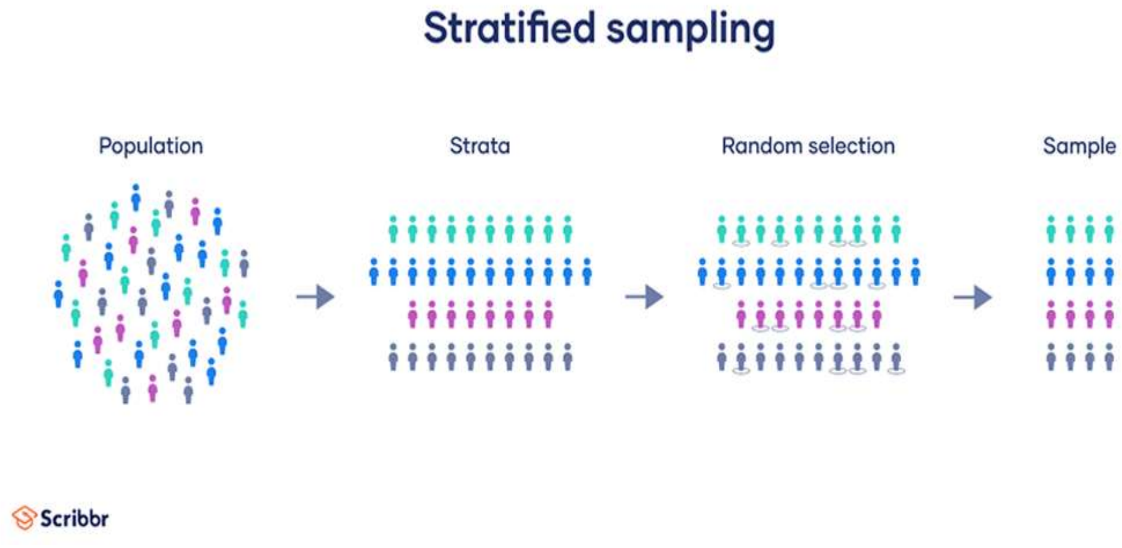
First== probability(Random) sampling

3- Stratified Random Sampling

Definition:

If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is a stratified sampling.

Stratified Random Sampling (SRS), (Stratification) is defined as the act of sorting data, people, and objects into distinct groups or layers, and It is a probability sampling technique where the total population is divided into homogenous groups these groups called 'strata'(the plural of *stratum)* and it is based on similar attributes or characteristics like race, gender, level of education, income, and more. Every member of the population studied should be in exactly one stratum. Each stratum is then sampled using another probability sampling method, such as cluster sampling or simple random sampling, allowing researchers to estimate statistical measures for each sub-population. from each group or strata, the members (sample) stratum selected randomly.



The basic idea behind the stratified sampling is to

- divide the whole heterogeneous population into smaller groups or subpopulations, such that the sampling units are homogeneous with respect to the characteristic under study within the subpopulation and

-heterogeneous with respect to the characteristic under study between/among the subpopulations. Such subpopulations are termed as **strata.**

- Treat each subpopulation as a separate population and draw a sample by SRS from each stratum.

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

Example: In order to find the average height of 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 subpopulations or strata such as

Students of class 1, 2 and 3: Stratum 1

Students of class 4, 5 and 6: Stratum 2

Students of class 7, 8 and 9: Stratum 3

Students of class 10, 11 and 12: Stratum 4

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.

3.1Types of Stratified Random Sampling

There are two types –

(a)-Proportionate stratified random sampling – in this type, the sample size is directly proportional

In this approach, each stratum sample size is directly proportional to the population size of the entire population of strata. That means each strata sample has the same sampling fraction.

Proportionate Stratified Random Sampling Formula: $n_i = (N_i / N) * n$

N = population Size

n_i = Sample size for ith stratum

N_i= Population size for ith stratum

n = Size of entire sample

Example:

If you have four strata with 500, 1000, 1500, and 2000 respective sizes, the research organization selects ½ as the sampling fraction. A researcher must choose 250, 500, 750, and 1000 members from the separate stratum.

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/2	1/2	1/2

Final Sampling Size Results	250	500	750	1000

Irrespective of the sample size of the population, the sampling fraction will remain uniform across all the strata.

2-Disproportionate Sampling:

The sampling fraction is the primary differentiating factor between proportionate and disproportionate stratified random sampling. In excessive sampling, each stratum will have a different sampling fraction. The success of this sampling method depends on the researcher’s precision at fraction allocation. If the allotted fractions aren’t accurate, the results may be biased due to the over represented or under represented strata.

Example -1-

Stratum	A	B	C	D
Population Size	500	1000	1500	2000
Sampling Fraction	1/2	1/3	1/4	1/5
Final Sampling Size Results	250	333	375	400

Example-2 –

From 1000 people, 700 males and 300 females, according to which if we want to choose 100 people, then 70 males should be selected and 30 females should be selected, and this selection will be random.

Classic stratified random sampling

Example:

Let’s say 100 (n) students of a school having 1000 (N) students were asked questions about their favorite subject. It’s a fact that the students of the 8th grade will have different subject preferences than the students of the 9th grade. For the survey to deliver precise results, the ideal manner is to divide each step into various strata. Here’s a table of the number of students in each grade:

Grade	Number of students (N_i)
5	150
6	250
7	300
8	200
9	100

Calculate the sample of each grade using the formula:

Stratified Sample (n_5) = 100 / 1000 * 150 = 15
Stratified Sample (n_6) = 100 / 1000 * 250 = 25
Stratified Sample (n_7) = 100 / 1000 * 300 = 30
Stratified Sample (n_8) = 100 / 1000 * 200 = 20
Stratified Sample (n_9) = 100 / 1000 * 100 = 10

Steps for performing a stratified random sampling:

1. Define the population and subgroups(strata).
2. Split the population into subgroups(strata).
3. Choose the sample size for subgroups (each stratum).
4. Take random samples of the subgroups (from each stratum)

3.3Procedure of stratified sampling

Divide the population of N units into k strata. Let the i^{th} stratum has N_i . $i = 1.2.3. \cdots . k$ number of units.

- Strata are constructed such that they are non-overlapping and homogeneous with respect to the characteristic under study such that $\sum_{i=1}^{k} N_i = N$

- Draw a sample of size n_i from $i^{th}(i = 1.2.3. \cdots . k)$ stratum using SRS (preferably WOR) independently from each stratum.

- All the sampling units drawn from each stratum will constitute a stratified sample of size $n = \sum_{i=1}^{k} n_i$

Notations:

We use the following symbols and notations:

N : Population size

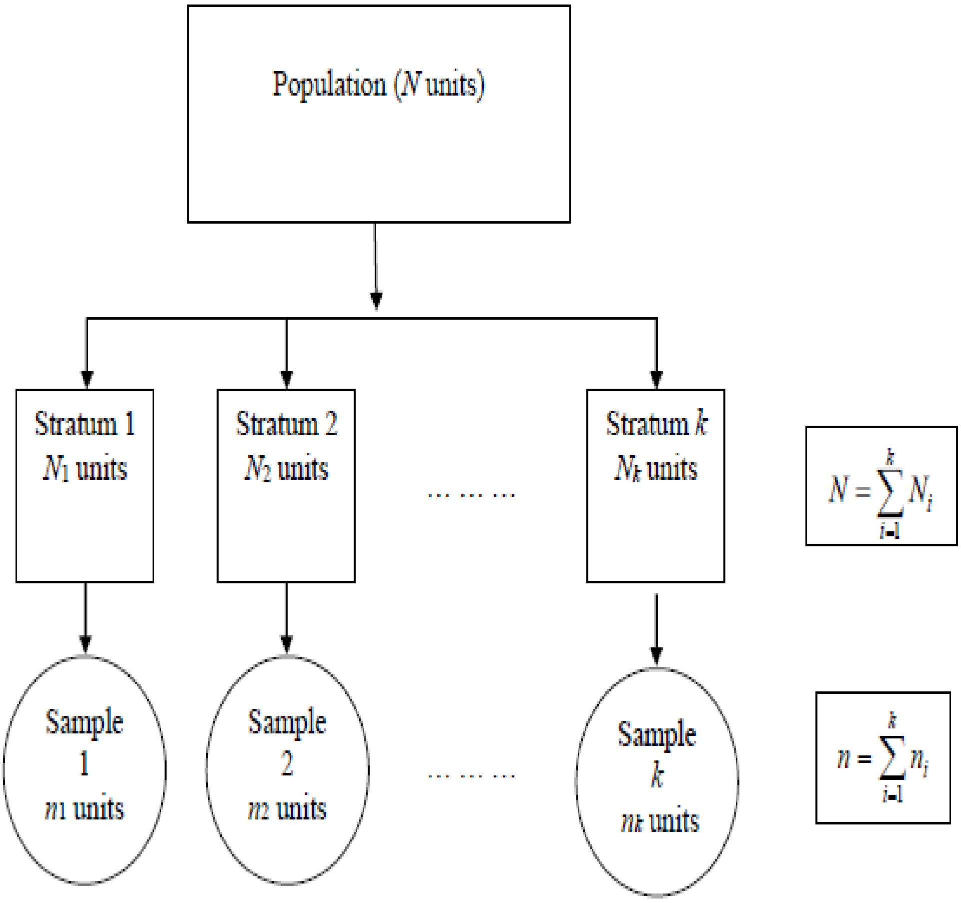
k : Number of strata

N_i : Number of sampling units in *ith* strata

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

n_i: Number of sampling units to be drawn from *ith* stratum.

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



3.4Estimation of population mean

Prove that**:**

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

Proof:

Let

Y : characteristic under study

y_{ij} value of j^{th} unit in i^{th}stratum $j \ = \ 1.2/ \ldots.ni \ \ i \ = \ 1.2.\ldots k$

$\bar{Y_\imath} = \frac{\mathbf{1}}{N_i}\sum_{\boldsymbol{j=i}}^{N_i} y_{ij}$ population mean of i^{th} stratum

$\bar{y} = \frac{\mathbf{1}}{n_i}\sum_{\boldsymbol{j=i}}^{n_i} y_{ij}$ sample mean of i^{th} stratum

$$\bar{Y} = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i\, \overline{Y_\iota} \; = \sum_{\boldsymbol{j=i}}^{k} w_i\, \overline{Y_\iota} \text{ population mean where } w_i = \frac{N_\iota}{\boldsymbol{N}}$$

First, we discuss the estimation of the population mean.

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

$\bar{Y} = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i\, \overline{Y_\iota}$ one may choose the sample mean

$$\overline{y_\iota} = \frac{\mathbf{1}}{n}\sum_{\boldsymbol{j=i}}^{k} n_i \overline{y_\iota}$$

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

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

$E(\overline{y_\iota}) = \frac{\mathbf{1}}{n}\sum_{\boldsymbol{j=i}}^{k} n_i E(\overline{y_\iota}\,) = \frac{\mathbf{1}}{n}\sum_{\boldsymbol{j=i}}^{k} n_i \overline{Y_\iota} \neq \bar{Y}$

and $\bar{y}$ turns out to be a biased estimator of $\bar{Y}$. Based on this, one can modify $\bar{y}$ so as to obtain an unbiased estimator of $\bar{Y}$.

Consider the stratum mean which is defined as the weighted arithmetic mean of strata sample means with strata sizes as weights given by

$\bar{y}_{\boldsymbol{st}} = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i\, \overline{y_\iota}$

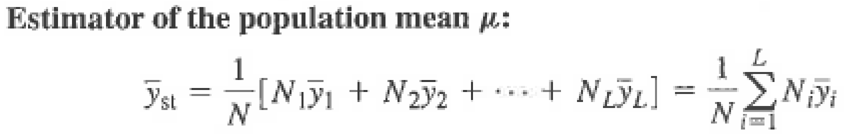
Now

$$\boldsymbol{E}(\bar{y}_{\boldsymbol{st}}) = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i E(\,\overline{y_\iota}\,)$$

$$\boldsymbol{E}(\bar{y}_{\boldsymbol{st}}) = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i \overline{Y_\iota}$$

$$\boldsymbol{E}(\bar{y}_{\boldsymbol{st}}) = \frac{\mathbf{1}}{N}\sum_{\boldsymbol{j=i}}^{k} N_i \overline{Y_\iota} = \bar{Y}$$

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



Estimator of the population mean μ:

$$\bar{y}_{\text{st}} = \frac{1}{N}[N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L] = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{y}_i$$

3.5Estimation of population variance

Since the samples have been drawn by SRSWOR, so

$$\boldsymbol{E(s^2) = S^2}$$

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

And

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

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

Note: If SRSWR is used instead of SRSWOR for drawing the samples from each stratum, then in this case

$$\bar{y} = \sum_{\boldsymbol{j=i}}^{k} w_i \bar{y}_i$$

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

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

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

Estimated variance of $\bar{y}_{st}$:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}[N_1^2\hat{V}(\bar{y}_1) + N_2^2\hat{V}(\bar{y}_2) + \cdots + N_L^2\hat{V}(\bar{y}_L)]$$

$$= \frac{1}{N^2}\left[N_1^2\left(1 - \frac{n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + \cdots + N_L^2\left(1 - \frac{n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right)\right]$$

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

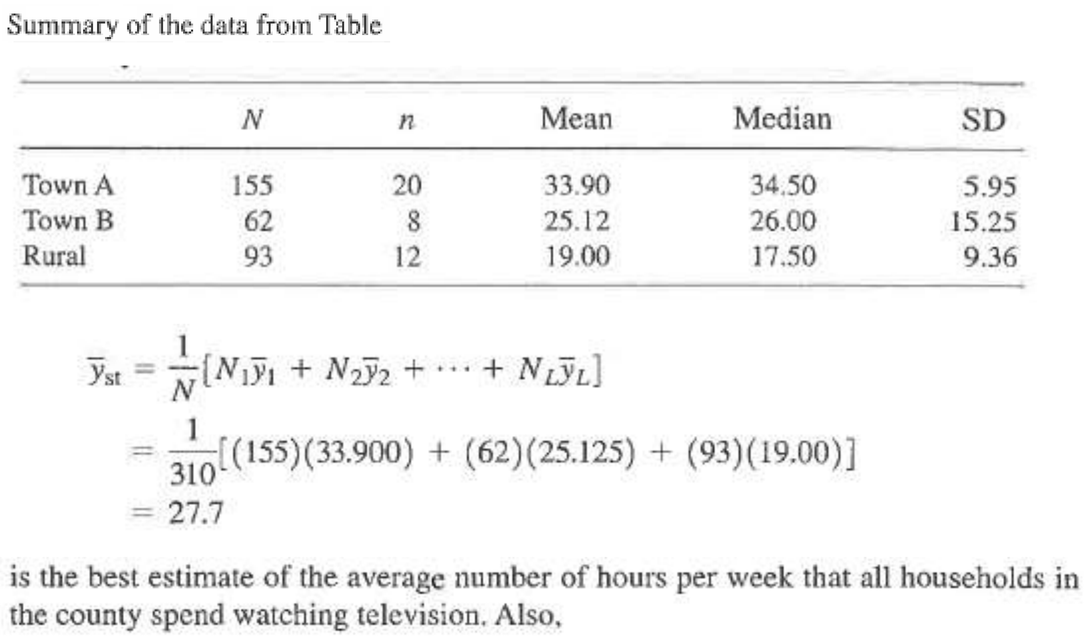
Example

Suppose the survey planned in Example is carried out. The advertising firm has enough time and money to interview $n = 40$ households and decides to select random samples of size $n_1 = 20$ from town A, $n_2 = 8$ from town B, and $n_3 = 12$ from the rural area. (We discuss the choice of sample sizes later.) The simple random samples are selected and the interviews conducted. The results, with measurements of television-viewing time in hours per week, are shown in Table below

Estimate the average television-viewing time, in hours per week, for (a) all households in the county and (b) all households in town B. In both cases, place a bound on the error of estimation.

Television-viewing time, in hours per week

Town A	Town B	Rural
35	27	8
43	15	14
36	4	12
39	41	15
28	49	30
28	25	32
29	10	21
25	30	20
38		34
27		7
26		11
32		24
29		
40		
35		
41		
37		
31		
45		
34		



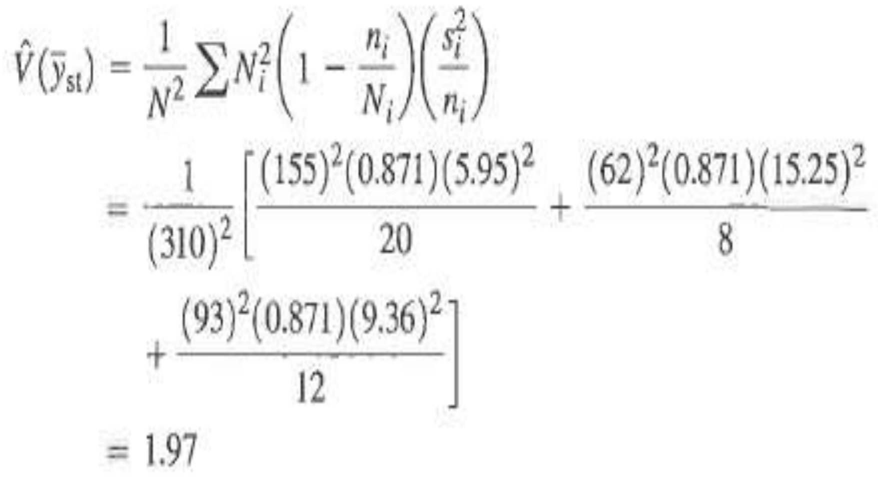
Summary of the data from Table

	N	n	Mean	Median	SD
Town A	155	20	33.90	34.50	5.95
Town B	62	8	25.12	26.00	15.25
Rural	93	12	19.00	17.50	9.36

$$\begin{aligned}\bar{y}_{\text{st}} &= \frac{1}{N}[N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L] \\ &= \frac{1}{310}[(155)(33.900) + (62)(25.125) + (93)(19.00)] \\ &= 27.7\end{aligned}$$

is the best estimate of the average number of hours per week that all households in the county spend watching television. Also,

Estimate Variance



$$\begin{aligned}\hat{V}(\bar{y}_{\text{st}}) &= \frac{1}{N^2}\sum N_i^2\left(1-\frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right) \\ &= \frac{1}{(310)^2}\left[\frac{(155)^2(0.871)(5.95)^2}{20} + \frac{(62)^2(0.871)(15.25)^2}{8} \right. \\ &\quad \left. + \frac{(93)^2(0.871)(9.36)^2}{12}\right] \\ &= 1.97\end{aligned}$$

3.6 Estimate of variance and confidence intervals

Under SRSWOR, an unbiased estimate ofS_i^2 for the i^{th} stratum

$(i \ = \ 1.2.\ldots\ldots k)$

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

In stratified sampling

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

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

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

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

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

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

Confidence limits

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

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

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

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

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

where$g_i = \frac{N_i - n_i}{n_i}$ and $Min(n_i - 1) \leq n_e \leq \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} (n_i - 1)$

assuming y_{ij} are normally distributed.

3.7Advantages of stratified sampling

1-The main advantage of this sampling is that it gives better accuracy in results as compared to other sampling methods

2-It is very easy to teach and easy to grasp by the trainees

3-Even smaller sample sizes can also give good results using strata

4-We can divide the large population into different subgroups/strata according to our need.

-5When to use stratified random sampling

6-When we want to focus on a particular strata from the given population data

7-When we want to establish relationship between two strata

8-When it is difficult to contact/access the sample population, this method is best as samples are easily involved in research with this method

9-As the elements of samples are chosen from some specific strata, the accuracy of statistical results is higher than that of simple random sampling.

What is the difference between stratified and random sampling?

The simple random sample represents the entire population. It randomly selects people from the population. A stratified, random sample on the other side, divides the population into smaller groups or strata based on common characteristics.

Chapter Four

First== probability(Random) sampling

4.1Systematic Sampling :

It is an advanced form of simple random sampling (SRS) ,

Estimators for systematic sampling and simple random sampling are identical; only the method of sample selection differs. Therefore, systematic sampling is used most often to simplify the process of selecting a sample or to ensure ideal dispersion of sample units throughout the population.

Systematic sampling is a type of probability sampling where each element in the population has a known and equal probability of being selected

Every member in the population is given a number.

The sampling interval

it is a type of probability sampling after the first (item) member is chosen, the remaining members are chosen from a given interval. selects items of data at regular intervals from a population.

To calculate the sampling interval required to select the sample data, we calculate the population size divided by the sample size.

k = Number of units in population / Number of sample units required

k =Population size/Sample size

Example.1

If the population size is 1200 and the desired sample size is 400 items of data,

Solution :

Population size =1200 items . Sample size =400 items =groups

1200 / 400 = 3 This means that every 3rd item of data in the ordered list is selected for the

1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,…,1198,1199,1200

Example.2

Choose 40 students from a College of 400 students by S.S

Solution:

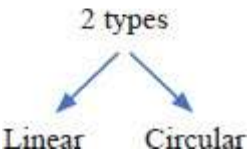
k = 400/40 = 10,

so this design would be

1 , 2 , 3 , 4 , 5 , 6 , 7, 8 , 9 , 10,11,12,13,14,15,16,17,18,19, 20 . . . ,

391 , 392 , 393 , 394, 395, 396, 397, 398 , 399 , 400 .

4.2Types of systematic sampling



4.2.1Linear systematic sampling

A list is made in a sequential manner of the whole population list. Decide the sample size and find the sampling interval by formula: $K = N/n$, where K is the K^{th} element, N is the whole population, and n = number of samples. Now, choose random number between 1 and K and then to the number what we got add K to that to get the next sample.

4.2.2Circular systematic sampling

In this, first, we will determine sample interval and then select number nearest to N/n. For example, if N = 17 and n = 4, then k is taken as 4 not 5 and then start selecting randomly between 1 to N, skip K units each time when we select the next unit until we get n units. In this type, there will be N number of samples unlike K samples in linear systematic sampling method.

4.3Advantages of systematic sampling

It is very easy to create, conduct, and analyze the sample Risk factor is very minimal. As there is even distribution of members to form a sample, systematic sampling is beneficial when there are diverse members of population.

1. 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.

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

3. 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

- Easy to implement
- Maximum dispersion of sample units throughout the population
- Requires minimum knowledge of the population

Disadvantages:

- Less protection from possible biases
- Can be imprecise and inefficient relative to other designs if the population being sampled is heterogeneous

4.4Estimation of population mean:

When *N* = *nk*

Let: y_{ij} : observation on the unit bearing the serial number

$i + (j-1)k$ in the population, $i = 1.2.\cdots.k$ $\quad j = 1.2.\cdots.n$

Suppose the drawn random number is $i \leq k$.

Sample consists of i^{th} column (in the earlier table).

Consider the sample mean given by

$\bar{y}_{sy} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$

as an estimator of the population mean given by

$\bar{Y} = \frac{1}{nk}\sum_{i=j}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i$

Probability of selecting i^{th} column as systematic sample $= \frac{1}{k}$

So $\quad E(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}\bar{y}_i = \bar{Y}$

4.5Estimation of variance

As such, there is only one cluster, so the variance in principle, cannot be estimated.

Some approximations have been suggested.

1. Treat the systematic sample as if it were a random sample. In this case, an estimate of variance is

$\widehat{Var}\bar{y}_{sy} = \left(\frac{1}{n} - \frac{1}{nk}\right)s_{wc}^2$

$s_{wc}^2 = \frac{1}{n-1}\sum_{j=0}^{n-1}\left(y_{i+jk} - \bar{y}_i\right)^2$

This estimator under-estimates the true variance

2. Use of successive differences of the values gives the estimate of variance as$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{2(n-1)}\sum_{j=0}^{n-1}(y_{i+jk} - y_{i+(j+1)k})^2$ This estimator is a biased estimator of true variance.

3. Use the balanced difference of $y_1.y_2.y_3.\cdots.y_n$ to get the estimate of variance as

$$\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{5(n-2)}\sum_{j=0}^{n-2}\left[\frac{y_i}{2} - y_{i+1} + \frac{y_{i+2}}{2}\right]^2$$

Or $\widehat{Var}(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right)\frac{1}{15(n-4)}\sum_{j=0}^{n-4}\left[\frac{y_i}{2} - y_{i+1} + y_{i+2} - y_{i+3} + \frac{y_{i+4}}{2}\right]^2$

4.6 Comparisons

Comparison of systematic sampling, stratified sampling and SRS with population with the linear trend:

We assume that the values of units in the population increase according to the linear trend.

So the values of successive units in the population increase in accordance with a linear model so that

$$y_i = a + bi,\ i = 1, 2, \ldots, N.$$

Now we determine the variances of $\bar{y}_{SRS}, \bar{y}_{sy}$ and $\bar{y}_{st}$ under this linear trend.

Under SRSWOR

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

Here $N = nk$

$$\begin{aligned}
\bar{Y} &= a + b\frac{1}{N}\sum_{i=1}^{N} i \\
&= a + b\frac{1}{N}\frac{N(N+1)}{2} \\
&= a + b\frac{N+1}{2}
\end{aligned}$$

$$\begin{aligned}
S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[a + bi - a - b\frac{N+1}{2}\right]^2 \\
&= \frac{b^2}{N-1}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)^2 \\
&= \frac{b^2}{N-1}\left[\sum_{i=1}^{N} i^2 - N\left(\frac{N+1}{2}\right)^2\right] \\
&= \frac{b^2}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right] \\
&= b^2\frac{N(N+1)}{12}
\end{aligned}$$

$$\begin{aligned}
Var(\bar{y}_{SRS}) &= \frac{nk-n}{nk.n}b^2\frac{nk(nk+1)}{12} \\
&= \frac{b^2}{12}(k-1)(nk+1).
\end{aligned}$$

4.7

Theorem: In circular systematic sampling, the sample mean is an unbiased estimator of the population mean.

Proof: If i is the number selected at random, then the circular systematic sample mean is

$$\overline{y} = \frac{1}{n}\left(\sum^{n} y\right)_i ,$$

where $\left(\sum^{n} y\right)_i$ denotes the total of y values in the i^{th} circular systematic sample, $i = 1, 2, ..., N$. We note here that in circular systematic sampling, there are N circular systematic samples, each having probability $\frac{1}{N}$ of its selection. Hence,

$$E(\overline{y}) = \sum_{i=1}^{N} \frac{1}{n}\left(\sum^{n} y\right)_i \times \frac{1}{N} = \frac{1}{Nn}\sum_{i=1}^{N}\left(\sum^{n} y\right)_i$$

Clearly, each unit of the population occurs in n of the N possible circular systematic sample means. Hence,

$$\sum_{i=1}^{N}\left(\sum^{n} y\right)_i = n\sum_{i=1}^{N} Y_i,$$

which on substitution in $E(\overline{y})$ proves the theorem.

4.8

What to do when $N \neq nk$

One of the following possible procedures may be adopted when $N \neq nk$.

(i) Drop one unit at random if the sample has $(n+1)$ units.

(ii) Eliminate some units so that $N = nk$.

(iii) Adopt circular systematic sampling scheme.

(iv) Round off the fractional interval k .

4.9

Systematic sampling when $N \neq nk$.

When N is not expressible as nk then suppose N can be expressed as

$$N = nk + p;\ p < k.$$

Then consider the following sample mean as an estimator of the population mean

$$\bar{y}_{sy} = \bar{y}_i = \begin{cases} \dfrac{1}{n+1}\displaystyle\sum_{j=1}^{n+1} y_{ij} & \text{if } i \leq p \\ \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} y_{ij} & \text{if } i > p. \end{cases}$$

In this case

$$E(\bar{y}_i) = \frac{1}{k}\left[\sum_{i=1}^{p}\left(\frac{1}{n+1}\sum_{j=1}^{n+1} y_{ij}\right) + \sum_{i=p+1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij}\right)\right]$$
$$\neq \bar{Y}.$$

So $\bar{y}_{sy}$ is a biased estimator of $\bar{Y}$.

An unbiased estimator of $\bar{Y}$ is

$$\bar{y}^*_{sy} = \frac{k}{N}\sum_{j} y_{ij}$$
$$= \frac{k}{N} C_i$$

where $C_i = n\bar{y}_i$ is the total of values of the i^{th} column.

$$E(\bar{y}^*_{sy}) = \frac{k}{N} E(C_i)$$
$$= \frac{k}{N}.\frac{1}{k}\sum_{i=1}^{k} C_i$$
$$= \bar{Y}$$

$$Var(\bar{y}^*_{sy}) = \frac{k^2}{N^2}\left(\frac{k-1}{k}\right) S_c^{*2}$$

where $S_c^{*2} = \dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k}\left(n\bar{y}_i - \frac{N\bar{Y}}{k}\right)^2.$

Chapter Five

Probability(Random) Sampling

Cluster Sampling

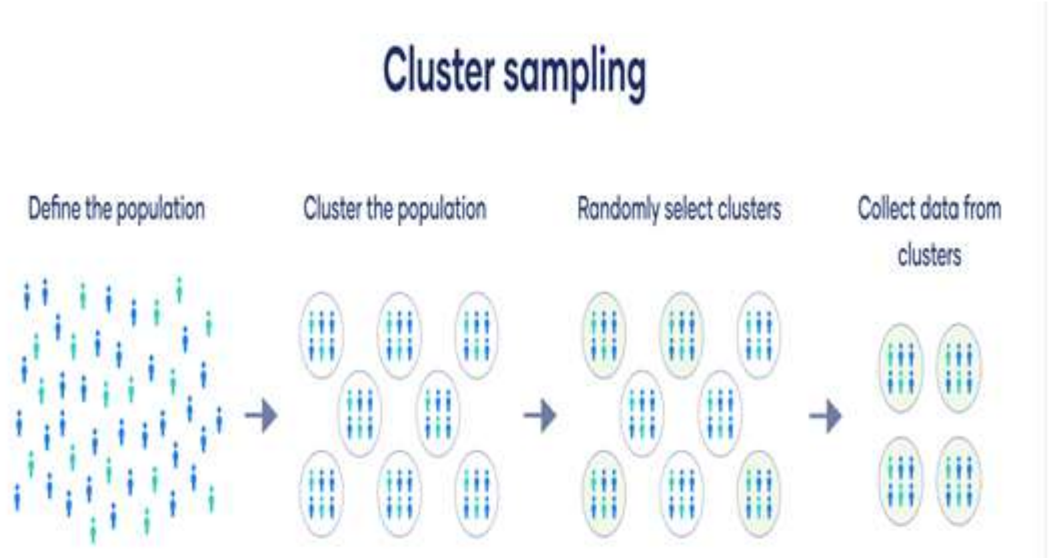
It is one of the basic assumptions in any sampling procedure that the population can be divided into a finite number of distinct and identifiable units, called sampling units. The smallest units into which the population can be divided are called elements of the population. The groups of such elements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

Cluster sampling is a probability sampling method that is often used to study large populations, particularly those that are widely geographically dispersed in which , researchers divide a population into smaller groups known as **clusters** such as districts or schools. They then randomly select among these clusters to form a sample. The clusters should ideally each be mini-representations of the population as a whole.

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

In cluster sampling

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



Examples:

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

Moreover, it is easier, faster, cheaper and convenient to collect information on clusters rather than on sampling units. In both the examples, draw a sample of clusters from houses/villages and then collect the observations on all the sampling units available in the selected clusters.

Conditions under which the cluster sampling is used: Cluster sampling is preferred when

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

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

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

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

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

Case of equal clusters

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

So total population size = NM

total sample size = nM .

Let

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

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

Estimation of population mean:

First select n clusters from N clusters by SRSWOR.

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

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

Bias:

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

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

Variance:

Variance:

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

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

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

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

Estimate of variance:

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

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

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

Comparison with SRS :

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

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

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

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

Consider

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

where

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

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

The efficiency of cluster sampling over SRSWOR is

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

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

Case of unequal clusters:

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

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

$$\begin{aligned}
M_0 &= \sum_{i=1}^{N} M_i \\
\bar{M} &= \frac{1}{N}\sum_{i=1}^{N} M_i \\
\bar{y}_i &= \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} : \text{mean of } i^{th} \text{ cluster} \\
\bar{Y} &= \frac{1}{M_0}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \\
&= \sum_{i=1}^{N}\frac{M_i}{M_0}\bar{y}_i \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{M_i}{\bar{M}}\bar{y}_i
\end{aligned}$$

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

Comparison between SRS and cluster sampling:

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

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

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

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

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

For example, in case of

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

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

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

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