SAMPLING THEORY

M.Sc., STATISTICS First Year

Semester – I, Paper-IV

Lesson Writers

Dr. N. Viswam
Department of Statistics
Hindu College,
Guntur.

Dr. B. Guravaiah
Assistant Professor,
Department of Mathematics &
Statistics. Vignan's Foundations
for Science, Technology &
Research, Vadlamudi, Guntur

Dr. U. Ramkiran
Tech. Asst.
Department of Statistics
Acharya Nagarjuna University

Lesson Writer & Editor
Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

Director, I/c
Prof. V.VENKATESWARLU
MA., M.P.S., M.S.W., M.Phil., Ph.D.
CENTRE FOR DISTANCE EDUCATION
ACHARAYANAGARJUNAUNIVERSITY
NAGARJUNANAGAR – 522510
Ph:0863-2346222,2346208,
0863-2346259(Study Material)
Website: www.anucde.info
e-mail:anucdedirector@gmail.com

M.Sc., STATISTICS - Sampling Theory

First Edition 2025

No. of Copies :

This book is exclusively prepared for the use of students of M.SC.(Statistics) Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited Circulation only.

Published by:

Prof. V.VENKATESWARLU,

Director, I/C

Centre for Distance Education, Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A^{+}' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the doorstep of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K.GangadharaRao

M.Tech.,Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University

M.Sc. – Statistics Syllabus

SEMESTER-I

104ST24: SAMPLING THEORY

UNIT-I:
Systematic Sampling: Allocation problem in stratified sampling, gain in precision due to stratification, estimation of sample size with continuous data, stratified sampling for proportions. Methods of populations with linear trend. Yates end correction, Modified systematic sampling, balanced systematic sampling, centrally located sampling, circular systematic sampling.

UNIT-II:
Varying probability and Cluster sampling: Cluster sampling with equal and unequal cluster sizes, optimum cluster size for fixed cost. PPS sampling with and without replacements procedures of selection of a sample, estimator of population total and its sampling variance in PPS with replacement, Des Raj and Murthy's estimator (for sample sizi trvoy, Horvitz- Thomson estimator, Grundy's estimator, Midzuno-sen sampling scheme.

UNIT-III:
Two-stage sampling: Two-stage sampling with equal number of second stage units, estimation of population mean, its variance and estimation of variance. Double sampling (twophase sampling) for stratification, variance of the estimated mean, optimum allocation in double sampling.

UNIT-IV:
Multiphase Sampling: Introduction, Double sampling for Difference estimation. Double sampling for ratio estimation. Double sampling for regression estimator, Optimum allocation varying probability sampling. Non sampling errors: Sources and types of non Sampling errors, Non response errors, techniques for adjustment of non response, Hansen and Hurwitz Technique, Deming's Model.

UNIT-V:
Ratio Estimator: Introduction, Bias and Mean square error, Estimation of variance, confidence interval, comparison with mean per unit estimator, Ratio estimator in stratified random sampling. Difference estimator and Regression estimator: Introduction, Difference estimator, Difference estimator in stratified sampling. Regression estimator, Comparison of regression estimator with mean per unit estimator and ratio estimator. Regression estimator instratified sampling**.**

BOOKS FOR STUDY:

1) Sampling techniques by W.G. Cochran, John Wiley
2) Sampling theory by Singh & Chaudhary
3) Sampling Theory, Narosa Publication by Des Raj and Chandok (199g)
4) Sampling Theory and Methods, Narosa publishers by S. Sampath (2001)
5) Theory and Analysis of Sample Survey Designs, F.S. Chaudhary: New Age International Publishers, Delhi.

BOOKS FOR REFERENCES:

l) Sampling Theory & Methods by M.N. Murthy.
2) Sampling theory of surveys with Applications: P.V.Sukhatme & B.V. Sukhatme.
3) Theory and methods of survey sampling. Mukhopadhyay (l988).

CODE:104ST24

M.Sc DEGREE EXAMINATION

First Semester

Statistics :: Paper IV- Sampling Theory

Time: Three hours Maximum:70 Marks

Answer ONE question from each unit (5x14=70)

UNIT – I

1. (a) Discuss the optimum allocation in stratified sampling. Obtain the variance of an estimate of population proportion with stratified random sampling.
   (b) Define systematic sampling. If a population consists of a linear trend, then prove that $Var(\bar{y}_{st}) \le Var(\bar{y}_{sys}) \le Var(\bar{y}_n)_R$

(or)

2. (a) Discuss the estimation of gain in precision due to stratification.
   (b) Explain circular systematic sampling. Obtain the method of estimation of sample size with continuous data.

UNIT – II

3. (a) What is cluster sampling? Obtain an unbiased estimator of population total based on cluster sampling, with clusters of equal size, and derive an expression for the sampling variance of this estimator.
   (b) How do you determine the optimum cluster size so as to minimize the variance for a fixed cost.

(or)

4. (a) Explain PPS sampling with replacement (wr-pps). Obtain an unbiased estimator of the population total and variance of the estimator under wr-pps. Also derive the estimator for the variance.
   (b) Define Horvitz – Thompson estimator of the population mean and derive the variance of this estimator.

UNIT – III

5. (a) Derive an expression for estimating the variance of population mean in two stage sampling where SRSWOR is adopted at both stages.
   (b) Obtain an estimator for the population mean under double sampling with SRSWR at the first stage and SRSWR at the second stage.

(or)

6. (a) Obtain the variance of an estimate for the population mean under double sampling with SRSWR at the first stage and SRSWR at the second stage.
   (b) Discuss the problem of optimal allocation in double sampling.

UNIT – IV

7. (a) What is double sampling? In case of double sampling for difference estimation, propose an estimator for the population mean and derive its variance, stating the necessary assumptions, If any.
   (b) Distinguish between multistage sampling and multiphase sampling.

(or)

8. (a) What are the various sources and types of non-sampling errors. Explain in detail?
(b) Briefly explain the concepts i) Hansen and ii) Hurwitz Technique and Deming's Model.

UNIT – V

9. (a) Derive the bias and mean square error of regression estimator of the population total assuming SRSWR for the units.
(b) Explain difference estimation. Define a separate difference estimator for population mean and obtain its variance.

(or)

10. (a) Explain ratio estimation. Obtain the variances of ratio estimates in stratified sampling.
(b) Obtain the several of the regression estimation. Obtain the variance of regression coefficient with pre-assigned value.

CONTENTS

LESSON-1

SYSTEMATIC SAMPLING – AN OVERVIEW

OBJECTIVES:

After study the lesson, the student is able to:

- Visualize the reasons for studying Systematic Sampling
- To identify the principal elements of Population
- Simple Random Sampling with Replacement / Without Replacement
- Variance of Systematic Sample mean in terms of Stratification
- Comparison between Systematic Sampling and Stratified Sampling
- To select representative sample from a population.

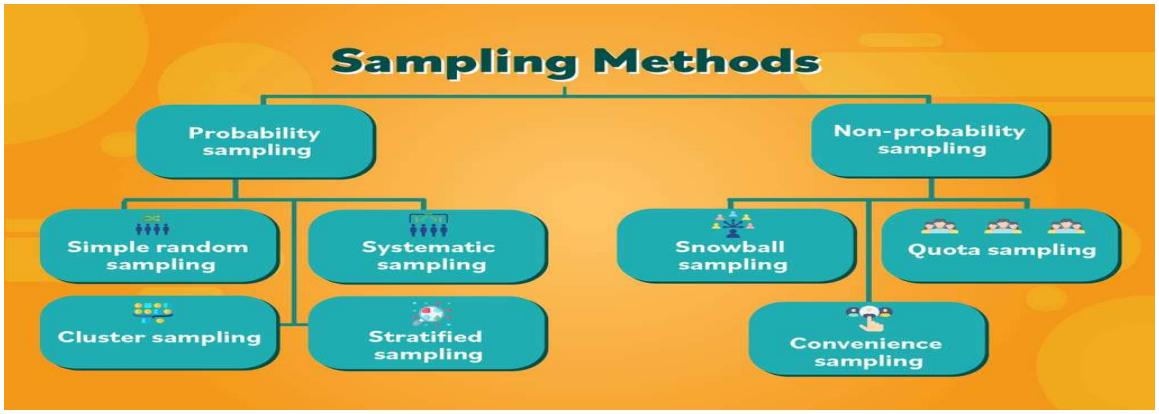
STRUCTURE:

1.1 INTRODUCTION:

Statistics is the science of data.

Data are the numerical values containing some information.

Statistical tools can be used on a data set to draw statistical inferences. These statistical inferences are in turn used for various purposes. For example, government uses such data for policy formulation for the welfare of the people, marketing companies use the data from consumer surveys to improve the company and to provide better services to the customer, etc. Such data is obtained through sample surveys. Sample surveys are conducted throughout the world by governmental as well as non-governmental agencies. For example, "National Sample Survey Organization (NSSO)" conducts surveys in India, "Statistics Canada" conducts surveys in Canada, agencies of United Nations like "World Health Organization (WHO), "Food and Agricultural Organization (FAO)" etc. conduct surveys in different countries.



❖ **POPULATION:**

Collection of all the sampling units in a given region at a particular point of time or a particular period is called the population. For example, if the medical facilities in a hospital are to be surveyed through the patients, then the total number of patients registered in the hospital during the time period of survey will the population. Similarly, if the production of wheat in a district is to be studied, then all the fields cultivating wheat in that district will be constitute the population. The total number of sampling units in the population is the population size, denoted generally by *N*. The population size can be finite or infinite (*N* is large).

❖ **CENSUS:**

The complete count of population is called census. The observations on all the sampling units in the population are collected in the census. For example, in India, the census is conducted at every tenth year in which observations on all the persons staying in India is collected.

❖ **SAMPLE:**

One or more sampling units are selected from the population according to some specified procedure. A sample consists only of a portion of the population units. Such a collection of units is called the sample.

Population

Total : $Y = \sum_{i=1}^{N} Y_i = Y_1 + Y_2 + Y_3 \ldots + Y_N$

Mean $\bar{Y} = (Y_1 + Y_2 + Y_3 \ldots + Y_N) / N = \sum_{i=1}^{N} Y_i \Big/ N$

Sample: Total: $\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 \ldots + y_n$,Then $\bar{y} = (y_1 + y_2 + y_3 \ldots + y_n) / n = \sum_{i=1}^{n} y_i / n$

Result-1

SRSWR: Show that sample mean is an unbiased estimate of the population mean.

$E(\bar{y}_n) = \bar{Y}_N$

Proof: $E\left(\bar{y}_n\right)=E\left[\frac{1}{n}\sum_{i=1}^{N}a_iY_i\right]=\frac{1}{n}\sum_{i=1}^{N}E(a_i)Y_i$, $\because$ ($E(a_i) = n/N$)

Hence $E\left(\bar{y}_n\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{n}{N}Y_i=\frac{1}{N}\bar{Y}_N$.

Result-2

SRSWOR: Show that $V\left(\bar{y}_n\right)=(1-f)\frac{S^2}{n}$ .where $f=\frac{n}{N}$ is called the sampling fraction and $(1-f)$ is called finite population correction (F.P.C).

Where $S^2=\frac{\sum_{i=1}^{N}(Y_i-\bar{Y}_N)^2}{N-1}$ = Population scatterness of the observation from the mean values.

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

Proof. We have $V\left(\bar{y}_n\right)=E\left(\bar{y}^2_n\right)-\left[E\left(\bar{y}_n\right)\right]^2$,$[\because E(\bar{y}_n)=\bar{Y}_N]$

$=E\left(\bar{y}_n\right)^2-\left(\bar{Y}_N\right)^2 \rightarrow (1)$

Consider, $E\left(\bar{y}_n\right)^2=E\left[\frac{1}{n}\sum_{i=1}^{n}y_i\right]^2=\frac{1}{n^2}E\left[\sum_{i=1}^{n}y_i\right]^2=\frac{1}{n^2}E\left[\sum_{i=1}^{n}y_i^2+\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}y_iy_j\right]\rightarrow(2)$

In equation (2), Consider, $E\left[\sum_{i=1}^{n}y_i^2\right]=E\left[\sum_{i=1}^{n}a_iY_i^2\right]=\frac{n}{N}\sum_{i=1}^{N}Y_i^2$ $\because$ ($E(a_i) = n/N$)

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

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

$\therefore E\left[\sum_{i=1}^{n}y_i^2\right]=\frac{n}{N}\left[(N-1)S^2+N\bar{Y}_N^2\right]$

$=\frac{n(N-1)}{N}S^2+\frac{n\cdot N}{N}\bar{Y}_N^2=\frac{n(N-1)}{N}S^2+n\bar{Y}_N^2\rightarrow(i)$

In equation (2) consider,

$E\left[\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n}y_iy_j\right]=E\left[\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jY_iY_j\right]$

$\therefore=\frac{n(n-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}Y_iY_j$

$\because$($E(a_ia_j) = n(n-1)/N(N-1)$)

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

$=\sum_{i=1}^{N}Y_i^2-N\bar{Y}_N^2$

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

$= \left(N\overline{Y}_N\right)^2 - \left[(N-1)S^2 + N\overline{Y}_N^2\right]$

$= N^2\overline{Y}_N^2 - (N-1)S^2 - N\overline{Y}_N^2 \because S^2 = \sum_{i=1}^{N}\left(Y_i - \overline{Y}_N\right)^2$

$\therefore E\left[\sum_{i=1}^{n}\sum_{\substack{j=1, \\ i\neq j}}^{n} y_i y_j\right] = \frac{n(n-1)}{N(N-1)}\left[N^2\overline{Y}_N^2 - (N-1)S^2 - N\overline{Y}_N^2\right] \rightarrow (ii)$

From equation (2)

$E\left(\overline{y}_n^2\right) = \frac{1}{n^2}\left[\frac{n}{N}(N-1)S^2 + n\overline{Y}_N^2 + \frac{n(n-1)}{N(N-1)}\left\{N^2\overline{Y}_N^2 - (N-1)S^2 - N\overline{Y}_N^2\right\}\right]$

$= \frac{N-1}{Nn}S^2 + \frac{\overline{Y}^2}{N} + \frac{(n-1)}{nN(N-1)}N^2\overline{Y}_N^2 - \frac{n(n-1)(N-1)}{n^2N(N-1)}S^2 - \frac{n(n-1)N}{nN(N-1)}\overline{Y}_N^2$

$= \frac{S^2}{nN}[N-1-(n-1)] + \frac{\overline{Y}_N^2}{n} + \frac{N(n-1)}{n(N-1)}\overline{Y}_N^2 - \frac{(n-1)\overline{Y}_N^2}{n(N-1)}$

$= \frac{S^2}{nN}[N-n] + \overline{Y}_N^2\left[\frac{1}{n} + \frac{N(n-1)}{n(N-1)} - \frac{(n-1)}{n(N-1)}\right]$

$E\left(\overline{y}_n^2\right) = \frac{S^2}{nN}[N-n] + \overline{Y}_N^2\left[\frac{N-1+Nn-N-n+1}{n(N-1)}\right]$

$E\left(\overline{y}_n^2\right) = \frac{S^2}{nN}(N-n) + \overline{Y}_N^2\left[\frac{Nn-n}{n(N-1)}\right] = \frac{N-n}{nN}S^2 + \overline{Y}_N^2\frac{n(N-1)}{n(N-1)}$

$E\left(\overline{y}_n^2\right) = \frac{N-n}{nN}S^2 + \overline{Y}_N^2 \rightarrow (3)$

Substitute equation (3) in equation (1) we get from equation (1)

$V\left(\overline{y}_n\right) = E\left(\overline{y}_n^2\right) - \overline{Y}_N^2 \Rightarrow \frac{N-n}{nN}S^2 + \overline{Y}_N^2 - \overline{Y}_N^2 \Rightarrow \frac{N-n}{nN}S^2$

$\therefore V\left(\overline{y}_n\right) = \left(\frac{N-n}{nN}\right)S^2 = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = (1-f)\frac{S^2}{n}\left[\because f = \frac{n}{N}\right]$

Also we can write $V\left(\overline{y}_n\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2$

Result-3:

Standard error $\left(\overline{y}\right) = \sqrt{\text{var}\left(\overline{y}\right)}$

$S.E\left(\overline{y}\right) = \sqrt{(1-f)\frac{S^2}{n}} = \frac{S}{\sqrt{n}}\sqrt{(1-f)}$

Result-4:

Unbiased estimate of $V(\bar{y})$ is $v(\bar{y})$

Where $v(\bar{y}) = (1-f)\frac{s^2}{n}$, where $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$\hat{Y}$ = Estimate of the total population.

Result-5:

$V(\hat{Y}) = V(N\bar{y}) = N^2 V(\bar{y})[\because \hat{Y} = N\bar{y}]$

We know that $V(\bar{y}) = \frac{S^2}{n}(1-f)$

Variance of $\hat{Y} = N\bar{y}$ is an estimate of the population total Y is

$V(\hat{Y}) = E((\hat{Y} - \bar{Y})^2 = \frac{N^2S^2}{n}\left(\frac{N-n}{N}\right) = \frac{N^2S^2}{n}(1-f)$

Result-6: Unbiased estimates of the variance of $\bar{y}$ and $\hat{Y} = N\bar{y}$ are

$$s^2_{\hat{y}} = V(s_{\hat{y}}) = v(\hat{y}) = N^2\frac{s^2}{n}(1-f) \Rightarrow s_{\hat{y}} = Ns\sqrt{(1-f)/n}$$

Result-7: Variance of the sample estimates,

Population proportion is unknown $P = \frac{A}{N}; p = \frac{a}{n}$; $\bar{y} = \frac{\sum_i^n y_i}{n} = \frac{a}{n} = p$

P: population proportion and p: sample proportion

i. $V(p) = \frac{N-n}{N-1}\frac{PQ}{n}$

ii. $v(\hat{p}) = v(p) = \frac{N-n}{(n-1)}\frac{pq}{N}$

Proof. I. We have $V(p) = V(\bar{y}_n)_{SRSWOR} = (\frac{N-n}{nN})S^2$

$= \frac{N-n}{nN}\cdot\frac{NPQ}{N-1}$

$V(p) = \frac{N-n}{N-1}\cdot\frac{PQ}{n}$

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

$\bar{Y}_N = \frac{NP}{N} = P$, $\sum_{i=1}^{N} Y_i^2 = A = NP$,

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

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

$= NP - NP^2$

$S^2 = NP(1-P)/(N-1) = NPQ/(N-1)$

II. We have $E(s^2) = S^2 = E(\frac{N-n}{Nn}s^2) = \frac{N-n}{Nn}S^2$

$= E(\frac{N-n}{Nn}\cdot\frac{npq}{n-1}) = Var(p) = E(\frac{N-n}{N}\cdot\frac{pq}{n-1}) = Var(p)$

Hence $v(p) = \frac{(N-n)pq}{N(n-1)}$, provides an unbiased estimate of Var(p).

Result-8: If $(X_i, Y_i) \forall i = 1,2,\cdots,n$ are the pair of the variable defined for every i^{th} unit of the population and $(\bar{x}_n, \bar{y}_n)$ are the corresponding sample mean of SRSWOR of size 'n'.

Then prove that $\text{cov}(\bar{x}_n, \bar{y}_n) = \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)$

Proof: Let $U_i = X_i + Y_i$, $\forall i = 1,2,\cdots,N$

$$\frac{1}{N}\sum_{i=1}^{N}U_i = \frac{1}{N}\sum_{i=1}^{N}X_i + \frac{1}{N}\sum_{i=1}^{N}Y_i$$

$$\bar{U}_N = \bar{X}_N + \bar{Y}_N$$

Similarly $\bar{u}_n = \bar{x}_n + \bar{y}_n$

Consider $V(\bar{u}_n) = E\left[\bar{u}_n - E(\bar{u}_n)\right]^2$

$V(\bar{u}_n) = E[\bar{u}_n - \bar{U}_N]^2$

R.H.S:$E(\bar{u}_n - \bar{U}_N)^2 = E[(\bar{x}_n + \bar{y}_n) - (\bar{X}_N + \bar{Y}_N)]^2$

$$= E\left[(\bar{x}_n - \bar{X}_N) + (\bar{y}_n - \bar{Y}_N)\right]^2$$

$$= E\left[\bar{x}_n - \bar{X}_N\right]^2 + E\left[\bar{y}_n - \bar{Y}_N\right]^2 + 2E\left[(\bar{x}_n - \bar{X}_N)(\bar{y}_n - \bar{Y}_N)\right]$$

$$= E[\bar{x}_n - E(\bar{x}_n)]^2 + E[\bar{y}_n - E(\bar{y}_n)]^2 + 2E[\{\bar{x}_n - E(\bar{x}_n)\}\{\bar{y}_n - E(\bar{y}_n)\}]$$

$$= v(\bar{x}_n) + v(\bar{y}_n) + 2\,\text{cov}(\bar{x}_n, \bar{y}_n) \rightarrow (2)$$

L.H.S: $V(\bar{u}_n)_{SRSWOR} = \frac{N-n}{nN}s^2$

$$= \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}(U_i - \bar{U}_N)^2$$

$$= \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}\left[(X_i + Y_i) - (\bar{X}_N + \bar{Y}_N)\right]^2$$

$$= \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}\left[(X_i - \bar{X}_N) + (Y_i - \bar{Y}_N)\right]^2$$

$$V(\bar{u}_n) = \frac{1}{N-1}\frac{N-n}{nN}\left[\sum_{i=1}^{N}(X_i - \bar{X}_N)^2 + \sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2 + 2\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)\right]$$

$$= \frac{N-n}{nN}\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)^2 + \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2 + 2\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)\right]$$

$$= \frac{N-n}{nN}\left[s_X^2 + s_Y^2 + 2\sum_{i=1}^{N}\frac{(X_i - \bar{X}_N)(Y_i - \bar{Y}_N)}{N-1}\right]$$

$$= \frac{N-n}{nN}s_X^2 + \frac{N-n}{nN}s_Y^2 + \frac{2}{N-1}\frac{N-n}{nN}\sum_{i=1}^{N}(X_i - \bar{X}_N)(Y_i - \bar{Y}_N) \rightarrow (3)$$

From (2) & (3)

$$\cancel{\therefore} \operatorname{cov}\left(\bar{x}_n, \bar{y}_n\right) = \frac{\cancel{1}}{N-1}\frac{N-n}{nN}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)\left(Y_i - \bar{Y}_N\right)$$

$$\therefore \operatorname{cov}\left(\bar{x}_n, \bar{y}_n\right) = \frac{N-n}{nN}\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)\left(Y_i - \bar{Y}_N\right)$$

1.2 DESCRIPTION OF SYSTEMATIC SAMPLING:

A Sampling Technique in which only the first unit is selected with the help of random numbers and the rest get selected automatically according to some pre-designed pattern is known as "Systematic Random Sampling".

Suppose N units of the population are numbered from 1 to N in some order. Let N= nk, where 'n' is the sample size and 'k' is an integer and a random number less than or equal to 'k' be selected and every kth unit thereafter.

For instance, if k=15 and if the first unit drawn is no.10, the subsequent units are no's 10,25,40,55,70 and so on. This type is called as every kth Systematic Sampling and such a procedure termed "Linear Systematic Sampling". If N≠nk, and every k^{th} unit be included in a circular manner till the whole list is exhausted. It will be called "Circular Systematic Sampling".

1.3 ADVANTAGES OVER SIMPLE RANDOM SAMPLING:

(A)

1 .It is easier to draw a sample and often easier to execute without mistake.
2. This is a particular advantage when the drawing is done in an office then maybe a substantial saving in time.
3. For instance, if the units are described on cards that are all of the same size and lie in a file drawer, a card can be drawn out every inch along file as measured by ruler (scale).
4. This operation is speedy where as SRSING would be slow.

(B)

1. Institutively Systematic sampling sums likely to be more precise than SRSING.
2. In effect if stratified the population into n-strata with consists of the first k-units, and second k-units and so on.
3. We might therefore expect the SYSING to be about as precise as the corresponding Stratified Random Sampling with one unit per stratum.
4. The difference is that with the SYS the units occur at the same relative position, in the stratum where as with the stratified random sample the position in the stratum is determined separately by randomization within the each stratum.

[The position is differ from stratum to stratum in stratified random sampling]. N= n k

1111 111 11111 - - - - - - - - - - 1

k k k k

(C)

1. Systematic Sampling useful in forest surveys for estimating the volume of Timber.
2. In fisheries for estimating the total catch of fish.
3. In milk yield surveys for estimating of the location yield. (We use Systematic Sampling in agriculture field)

Notations: Let y_{ij} denote the j^{th} number of the i^{th} systematic sample, so that

j=1,…..n; i=1,…k

The mean of the i^{th} sample is denoted by:- $\bar{y}_i = \sum_{j=1}^{n} y_{ij} / n$

The mean of a systematic sample is $\bar{y}_{sy} = \sum_{i=1}^{k} \bar{y}_i / k$

Is an unbiased estimator of the population mean $\bar{Y}$.

1.4 VARIANCE OF SYSTEMATIC SAMPLING:

THEOREM :

The variance of the mean of a systematic sample is :

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

where

$$S_{wsy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \qquad (1)$$

Is the variance among units that lie within the same systematic sample

S_{wsy}^2

PROOF:-

By usual identity of ANOVA

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

The cross product term is ‘zero’ since

Within each sample. The variance of $\bar{y}_{st}$ by definition

$$Var(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2 \qquad (3)$$

K* $Var(\bar{y}_{sy}) = \sum_{i=1}^{k}(\bar{y}_i - \bar{Y})^2$ from equation (1)

Hence, (N-1)S^2=nk* $Var(\bar{y}_{sy})$ +k(n-1)*S^2_{wsy}

nk* $Var(\bar{y}_{sy})$ =(N-1)S^2 -k(n-1)*S^2_{wsy}

divide with nk on both sides

$$Var(\bar{y}_{sy}) = \frac{N-1}{N}S^2 - \frac{n-1}{n}S^2_{wsy}$$

1.5 SUMMARY:

Systematic sampling is a practical and efficient method for selecting samples from large, ordered populations. While it offers numerous advantages over simple random sampling, such as simplicity and improved coverage, care must be taken to ensure that the population's ordering does not introduce bias. When properly applied, systematic sampling can yield highly reliable and representative results with reduced variance.

1.6 KEY WORDS:

- Sampling
- Systematic Sampling
- Random Start
- Sampling Interval (k)
- Simple Random Sampling
- Variance
- Bias
- Efficiency

1.7 SELF ASSESSMENT QUESTIONS:

1. What is the key difference between systematic sampling and simple random sampling?
2. How is the sampling interval (k) determined in systematic sampling?

3. List two advantages of systematic sampling over simple random sampling.
4. In what situations might systematic sampling yield a higher variance than simple random sampling?
5. Explain with an example how a hidden pattern in data could affect systematic sampling results.
6. Derive the formula for the variance of the sample mean in systematic sampling.
7. If a population has 500 elements and a sample of 50 is needed, what is the sampling interval? Illustrate how you would select the sample.
8. Why is systematic sampling considered more practical in field surveys?

1.8 SUGGESTED READINGS:

1. Cochran, W.G. (1977). Sampling Techniques. *3rd* ed. Wiley.
2. Kish, L. (1965). Survey Sampling. Wiley.
3. Lohr, S.L. (2010*).* Sampling: Design and Analysis. 2nd ed. Brooks/Cole.
4. Sukhatme, P.V. et al. (1984). Sampling Theory of Surveys with Applications. Iowa State University Press.
5. Levy, P.S., & Lemeshow, S. (2013). Sampling of Populations: Methods and Applications (4th ed.). Wiley.

Prof. G. V. S. R. Anjaneyulu

LESSON- 2

STRATIFIED RANDOM SAMPLING

OBJECTIVES:

By the end of this lesson, learners will be able to:

- Understand the concept of stratified random sampling**,** including its purpose and advantages over simple random sampling.
- Illustrate stratified sampling through practical examples to comprehend its application in real-world scenarios.
- Familiarize with key notations and terminology used in stratified sampling methodology.
- Explain the step-by-step procedure involved in conducting stratified random sampling.
- Differentiate between stratified sampling and cluster sampling schemes, highlighting their respective use cases.
- Identify potential issues in the estimation of parameters when using stratified sampling.
- Calculate the population mean and its variance using appropriate estimation techniques in the context of stratified sampling.

STRUCTURE:

2.1 INTRODUCTION:

An important objective in any estimation problem is to obtain an estimator of a population parameter that can take care of the salient features of the population. If the population is

homogeneous with respect to the characteristic under study, then the method of simple random sampling will yield a homogeneous sample, and in turn, the sample mean will serve as a good estimator of the population mean. Thus, if the population is homogeneous with respect to the characteristic under study, then the sample drawn through simple random sampling is expected to provide a representative sample. Moreover, the variance of the sample mean not only depends on the sample size and sampling fraction but also on the population variance. To increase the precision of an estimator, we need to use a sampling scheme that can reduce the heterogeneity in the population. If the population is heterogeneous with respect to the characteristic under study, then one such sampling procedure is stratified sampling.

The basic idea behind 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].

2.2 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 classes 1, 2, and 3: Stratum 1

Students of classes 4, 5, and 6: Stratum 2

Students of classes 7, 8, and 9: Stratum 3

Students of classes 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.

2.3 NOTATIONS:

We use the following symbols and notations:

N : Population size

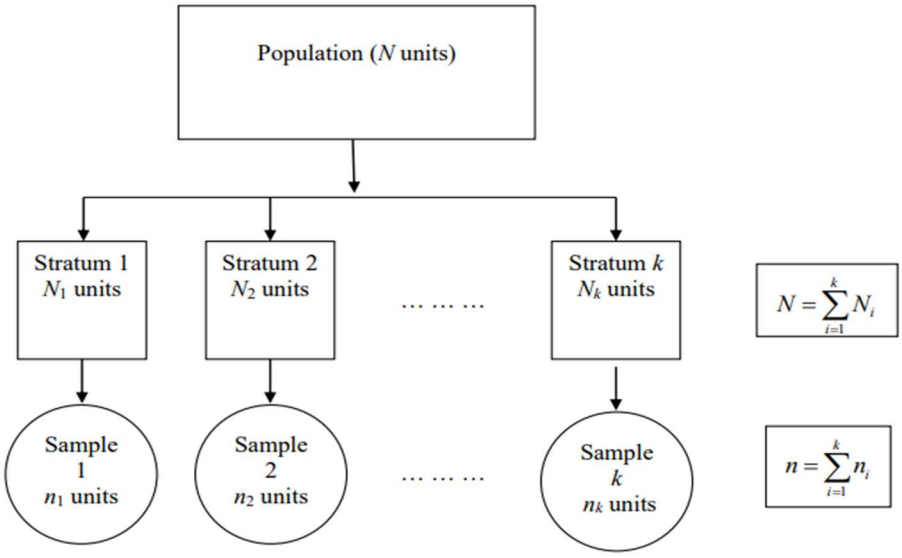
k : Number of strata

N_i : Number of sampling units in i[th] strata

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

n_i : Number of sampling units to be drawn from i[th] stratum

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



2.4 PROCEDURE OF STRATIFIED SAMPLING:

Divide the population of N units into k strata. Let the i[th] stratum has $N_i, i = 1,2,....,k$ 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] $(1,2,....,k)$ stratum using SRS (preferably WOR) independently from each stratum.

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

2.5 DIFFERENCE BETWEEN STRATIFIED AND CLUSTER SAMPLING SCHEMES:

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

- within homogeneous and
- among heterogeneous.

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

- within heterogeneous and
- among homogeneous.

[Note: We discuss the cluster sampling later.]

2.6 ISSUES IN THE ESTIMATION OF PARAMETERS IN STRATIFIED SAMPLING:

Divide the population of N units into k strata. Let the ith stratum has $N_i, i = 1,2,....,k$ number of units.

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

We now consider the estimation of population mean and population variance from a stratified sample.

2.7 ESTIMATION OF POPULATION MEAN AND ITS VARIANCE:

Let

Y : Characteristic under study,

y_i : Value of jth unit in the ith stratum $j = 1,2,....., n,$ i = 1,2,.....,k,

$\overline{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$: Population mean of ith stratum

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: Sample mean of ith stratum

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

Estimation of Population Mean:

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

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

as a possible estimator of Y.

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

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

Thus

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

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

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

Now

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

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

Variance of $\overline{y}_{st}$:

$$Var\left(\overline{y}_{st}\right)=\sum_{i=1}^{k}w_i^2 Var\left(\overline{y}_i\right)+\sum_{i(\neq j)=1}^{k}\sum_{j=1}^{n_i}w_i w_j Cov\left(\overline{y}_i,\overline{y}_j\right).$$

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

$$Cov\left(\overline{y}_i,\overline{y}_j\right)=0, i\neq j$$

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

Where

$$S_i^2=\frac{1}{N_i-1}\sum_{J=1}^{N_i}\left(Y_{ij}-\overline{Y}_i\right)^2$$

Thus,

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

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

Observe that $Var\left(\overline{y}_{st}\right)$ is small when S_i^2 is small. This observation suggests how to construct the strata. If S_i^2 is small for all $i=1,2,.....,k,$ then $Var\left(\overline{y}_{st}\right)$ will also be small.

The total variation in the population is fixed and can be orthogonally partitioned into between and within strata variations, i.e.,

Total variation = Between strata variation + Within strata variation $\left(S_i^2\right)$

Since S_i^2 is small, so obviously "Between strata variation" has to be large. That is why it was mentioned earlier that the strata are to be constructed such that they are within homogeneous, i.e., S_i^2 is small and among heterogeneous ("Between strata variation" is large).

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

Estimate of Variance:

Since the samples have been drawn by SRSWOR, so

$$E\left(s_i^2\right)=S_i^2$$

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

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

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

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

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

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

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

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

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

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

Advantages of stratified sampling:

1. Data of known precision may be required for certain parts of the population. This can be accomplished with a more careful investigation of a few strata.

Example: To know the direct impact of the hike in petrol prices, the population can be divided into strata, such as lower income group, middle-income group, and higher income group. Obviously, the higher-income group is more affected than the lower-income group. So, a more careful investigation can be conducted in the higher-income group strata.

2. Sampling problems may differ in different parts of the population.

Example: To study the consumption pattern of households, the people living in houses, hotels, hospitals, prisons, etc., are to be treated differently.

3. Administrative convenience can be exercised in stratified sampling.

Example: In taking a sample of villages from a big state, it is more administratively convenient to consider the districts as strata so that the administrative set up at the district

level may be used for this purpose. Such administrative convenience and the convenience of organizing fieldwork are important aspects of national-level surveys.

4. Full cross-section of the population can be obtained through stratified sampling. It may be possible in SRS that some large part of the population may remain unrepresented. Stratified sampling enables one to draw a sample representing different population segments to any desired extent. The desired degree of representation of some specified parts of the population is also possible.

5. Substantial gain in efficiency is achieved if the strata are formed intelligently.

6. In the case of a skewed population, the use of stratification is of importance since a larger weight may have to be given for the few extremely large units, which in turn reduces the sampling variability.

7. When estimates are required for the population and the subpopulations, then stratified sampling is helpful.

8. When the sampling frame for subpopulations is more easily available than the sampling frame for the whole population, then stratified sampling is helpful.

9. If the population is large, it is convenient to sample separately from the strata rather than the entire population.

10. The population mean or population total can be estimated with higher precision by suitably providing the weights to the estimates obtained from each stratum.

2.8 SUMMARY:

- Stratification is the process of grouping heterogeneous members of the population into relatively homogeneous subgroups.
- The allocation of sample sizes in different strata can be done either by proportional allocation.
- The allocation of sample sizes in different strata is said to be proportional if the sampling fraction f_i is constant for all strata.
- The allocation of sample sizes in different strata is said to be optimal if it holds any one of the characteristics (i) $V(\bar{y}_{st})$ is minimum for a specified size n. (ii) $V(\bar{y}_{st})$ is minimum for s specified cost C and (iii) Cost C is minimum for a specified $V(\bar{y}_{st})$.
- If intelligently used, stratification always results in a smaller variance for the estimated mean or total than is given by a comparable simple random sampling.

- Stratified random sampling with proportional allocation gives a more precise estimate of the population mean as compared with that of simple random sampling. We observe that grater the difference between stratum means, greater would be the gain in precision in Stratified random sampling with proportional allocation over simple random sampling.
- Stratified random sampling with proportional allocation gives a more precise estimate of the population mean as compared with that of simple random sampling. We observe that grater the difference between stratum standard deviations, would be the gain in precision of optimum allocation over simple random sampling.
- Stratified random sampling is an effective method to obtain accurate and representative estimates from a population by dividing it into meaningful strata. It ensures better precision than simple or cluster sampling, especially when strata differ significantly. However, its success depends on proper stratification, appropriate allocation of sample sizes, and careful handling of practical challenges. When implemented correctly, it offers high-quality statistical estimates with minimized sampling error.

2.9 KEY WORDS:

- Stratified Sampling
- Stratum
- Sample Allocation
- Population Mean
- Sampling Variance
- Homogeneous Groups
- Estimation
- Sampling Scheme

2.10 SELF ASSESSMENT QUESTIONS:

1. What is the key advantage of stratified sampling over simple random sampling?
2. Define stratum and give an example.
3. How is the sample mean estimated in stratified sampling?
4. What are two major allocation strategies used in stratified sampling?
5. Explain one disadvantage of stratified sampling.
6. Differentiate between stratified and cluster sampling with an example.
7. How do overlapping strata affect the quality of stratified sampling?

8. Given $N = 1000,\ N_1 = 400,\ N_2 = 600,\ n_1 = 40,\ n_2 = 60,$ and $\bar{y}_1$=50, $\bar{y}_2$=70 calculate $\bar{y}_{st}$.
9. Why is proportional allocation commonly used?
10. What assumptions are made while estimating variance in stratified sampling?

2.11 SUGGESTED READINGS:

1. Cochran, W.G. (1977). Sampling Techniques. *3rd* ed. Wiley.
2. Theory and methods of survey sampling. Parimal Mukhopadhyay(1988).
3. Sukhatme, P.V. et al. (1984). Sampling Theory of Surveys with Applications. Iowa State University Press.
4. Levy, P.S., & Lemeshow, S. (2013). Sampling of Populations: Methods and Applications (4th ed.). Wiley.
5. Sampling Theory & Methods by M.N. Murthy.

Prof. G. V. S. R. Anjaneyulu

LESSON-3

SAMPLING PROPORTIONS & ESTIMATION OF SAMPLE SIZE

OBJECTIVES:

By the end of this lesson, learners will be able to:

- Understand the Concept of Stratified Sampling for Proportions Learn how stratification is applied in estimating population proportions and why it improves accuracy over simple random sampling (SRS).
- Evaluate the Gain in Precision Quantify the increase in precision due to stratification and understand conditions under which stratification offers maximum benefit.
- Solve Allocation Problems Analyze how to allocate samples optimally across strata under different constraints such as cost and population variability.
- Determine Sample Sizes Across Strata Decide on the number of samples to be drawn from each stratum based on proportional, equal, or optimal allocation principles.
- Compare Variances under Various Allocation Methods Study the behaviour of variances under proportional, optimal, and other allocation strategies.
- Compare Stratified Sampling with SRS Examine how stratified sampling improves estimation accuracy over SRS by comparing variances under different scenarios.
- Apply Stratified Sampling Techniques in Practical Surveys Integrate all above concepts to design efficient and statistically sound survey strategies using stratified random sampling.

STRUCTURE:

3.1 INTRODUCTION:

Stratified sampling is a powerful and widely used technique in survey sampling that enhances the precision of estimates by dividing the population into distinct subgroups, or *strata*, and then sampling from each stratum. This method is especially beneficial when the population exhibits considerable heterogeneity, but within each stratum there is more homogeneity.

In the context of estimating proportions, stratified sampling ensures that each subgroup is appropriately represented, which can lead to more reliable and precise estimates than simple random sampling (SRS), especially when the proportions vary significantly across strata. This unit explores the application of stratified sampling to proportions, with particular attention to the estimation of gains in precision, allocation of sample sizes, and the comparison of variances under different allocation strategies. We will also examine the optimal and proportional allocation methods and their impact on the efficiency of the sampling design.

3.2 STRATIFIED SAMPLING FOR PROPORTIONS:

If we wish to estimate the proportion of units in the population that fall into some defined class C, the ideal stratification is attained if we can place in the first stratum every unit that falls in C ,and in the 2nd stratum every unit that does not fall in C.

Let $P_h = \frac{A_h}{N_h}, p_h = \frac{a_h}{n_h}$ be the proportions of units in C, in the hth stratum and in the sample from that stratum respectively. For the proportion in the whole population, the estimate appropriate to stratified random sampling is

$$p_{st} = \sum_h \frac{N_h p_h}{N}$$

$$\bar{y}_{st} = \sum_h w_h \bar{y}_h = \frac{\sum N_h \bar{y}_h}{N}$$

$\bar{y}_h$ = sample mean from h^{th} stratum

p_h = sample proportion from h^{th} stratum.

Theorem: With stratified random sampling, the variance of p_{st} is

$$V(p_{st}) = \frac{1}{N^2} \sum_h \frac{N_h^2 (N_h - n_h)}{N_h - 1} \cdot \frac{P_h Q_h}{n_h} \quad , Q_h = 1 - P_h$$

Proof: we have $V\left(\bar{y}_{st}\right)=\frac{1}{N^2}\sum_h N_h\left(N_h-n_h\right)\frac{S_h^2}{n_h}\rightarrow(1)$

Let y_{hi} be a variate which has the value '1' when the unit is in C, and zero otherwise.

For this variate , $S_h^2=\frac{N_h}{N_h-1}P_hQ_h$.

Substitute this value of S_h^2 in equation (1),we get

$$V(p_{st})=\frac{1}{N^2}\sum_h\frac{N_h(N_h-n_h)}{n_h}\cdot\frac{N_h}{N_h-1}P_hQ_h$$

$$\therefore V\left(p_{st}\right)=\frac{1}{N^2}\sum_h\frac{N_h^2\left(N_h-n_h\right)}{N_h-1}\cdot\frac{P_hQ_h}{n_h}\text{ ------(2)}$$

Note: In nearly all applications, even if the FPC is not ignored ,terms in $\frac{1}{N_h}$ will be negligible, and the slightly simpler formula,

$$V(p_{st})=\frac{1}{N^2}\sum_h N_{h(N_h-n_h)}\cdot\frac{P_hQ_h}{n_h}\cdot\frac{N_h}{N_h-1}$$

Dividing both Numerator and Denominator by N_h

$$V(p_{st})=\frac{1}{N^2}\sum_h\frac{N_h(N_h-n_h)}{n_h}\cdot P_hQ_h\cdot\frac{1}{1-\frac{1}{N_h}}$$

$$=\frac{1}{N^2}\sum_h N_h(N_h-n_h).\frac{P_hQ_h}{n_h}.$$

$$V\left(p_{st}\right)=\frac{1}{N^2}\sum_h N_h^2\frac{\left(N_h-n_h\right)}{N_h}\cdot\frac{P_hQ_h}{n_h}$$

$$V\left(p_{st}\right)=\sum_h w_h^2\frac{P_hQ_h}{n_h}\left(1-f_h\right)\rightarrow(3)$$

Corollary-1:If FPC is ignored i.e.,$1-f_h\cong1-0=1$

$$V\left(p_{st}\right)=\sum_h\frac{w_h^2P_hQ_h}{n_h}$$

Corollary-2: With proportional allocation

$$V\left(p_{st}\right)=\sum_h w_h^2\frac{P_hQ_h}{n_h}(1-f)=\sum_h w_hP_hQ_h(1-f)\frac{N_h}{Nn_h}\quad\left(\because W_h=\frac{N_h}{N}\right)$$

$$=\sum_h w_hP_hQ_h\left(\frac{1-f}{n}\right)\qquad\left[\because n=\frac{Nn_h}{N_h}\right]$$

$$\therefore V\left(p_{st}\right)=\frac{1-f}{n}\sum_h w_hP_hQ_h$$

3.3 ESTIMATION OF GAIN IN PRECISION DUE TO STRATIFICATION:

In comparing the precision (variance) of stratified with unstratified random sampling, it was assumed that the values of stratum means $(\overline{Y}_h)$ and stratum variance (S_h^2) were known. Many times stratum means and stratum variances are unknown, What is available only a stratified sample and the problem is to estimate the gain in precision due to stratification.

An estimate of the variance of the estimate from Simple Random Sampling is obtained from stratified sample and a comparison can be made with a situation in which no stratification can be made.

The true variance of the mean of a SRS is $V(\bar{y}) = \frac{N-n}{nN} s^2 = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi}-\overline{Y})^2}{N-1}\left(\frac{1-f}{n}\right)$

$$= \frac{(1-f)}{n}\frac{\sum_{h}^{L}\sum_{i}^{N_h}\left\{(y_{hi}-\overline{Y}_h)+(\overline{Y}_h-\overline{Y})\right\}^2}{N-1}$$

$$V(\bar{y}) = \frac{(1-f)}{n}\left[\frac{\sum_h\sum_i(y_{hi}-\overline{Y}_h)^2+\sum_h\sum_i(\overline{Y}_h-\overline{Y})^2+2\sum_h\sum_i(y_{hi}-\overline{Y}_h)(\overline{Y}_h-\overline{Y})}{N-1}\right]$$

The cross product term vanishes since $\sum_i(y_{hi}-\overline{Y}_h)=0$ in all strata

$$V(\bar{y}) = \frac{1-f}{n}\left[\frac{\sum_h(N_h-1)S_h^2+\sum_h N_h(\overline{Y}_h-\overline{Y})^2}{(N-1)}\right] \rightarrow (1) \quad \because S_h^2 = \frac{\sum(y_{hi}-\overline{Y}_h)^2}{N_h-1}$$

In the first term $\left[\sum_h(N_h-1)S_h^2\right]$ of the equation(1), we need only put s_h^2 for S_h^2

The second term $\left[\sum_h N_h(\overline{Y}_h-\overline{Y})^2\right]$ requires investigation because

$\sum_h N_h(\bar{y}_h-\bar{y}_{st})^2$ is not an unbiased estimate of $\sum_h N_h(\overline{Y}_h-\overline{Y})^2$

[The relevant result for $\sum_h N_h(\overline{Y}_h-\overline{Y})^2$ is stated in theorem-5A.1]

Theorem-5A.1: in stratified random sampling

$$E\left[\sum_h N_h(\bar{y}_h-\bar{y}_{st})^2\right] = \sum_h N_h(\overline{Y}_h-\overline{Y})^2 + \frac{\sum_h S_h^2(N_h-n_h)}{n_h}\left(1-\frac{N_h}{N}\right)$$

$$= \sum_h N_h \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_h \frac{N_h S_h^2}{n_h}(1-f_h)(1-W_h)$$

Proof: we may write

$$\sum_h N_h \left(\overline{y}_h - \overline{y}_{st}\right)^2 = \sum_h N_h \left[\left(\overline{Y}_h - \overline{Y}\right) + \left(\overline{y}_h - \overline{Y}_h\right) - \left(\overline{y}_{st} - \overline{Y}\right) \right]^2$$

$$= \sum_h N_h [(\bar{Y}_h - \bar{Y})^2 + (\bar{y}_h - \bar{Y}_h)^2 + (\bar{y}_{st} - \bar{Y})^2 + 2(\bar{Y}_h - \bar{Y})(\bar{y}_{st} - \bar{Y}) - 2(\bar{Y}_h - \bar{Y})(\bar{y}_{st} - \bar{Y}) - 2(\bar{y}_h - \bar{Y}_h)(\bar{y}_{st} - \bar{Y})]$$

We now expand and take the average overall possible samples. It may be verified that the average of each of two cross product terms involving $\left(\overline{Y}_h - \overline{Y}\right)$ vanishes (becomes zero)i.e., $\left(\overline{Y}_h - \overline{Y}\right) = 0$.

This gives

$$\sum_h N_h \left(\overline{y}_h - \overline{y}_{st}\right)^2 = \sum_h N_h \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)^2 + \sum_h N_h \left(\overline{y}_{st} - \overline{Y}\right)^2 - 2\sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)\left(\overline{y}_{st} - \overline{Y}\right)$$

Taking expectation on both sides

$$E\left[\sum_h N_h \left(\overline{y}_h - \overline{y}_{st}\right)^2\right] = \sum_h N_h \left(\overline{Y}_h - \overline{Y}\right)^2 + E\left[\sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)^2\right] + E\left[\sum_h N_h \left(\overline{y}_{st} - \overline{Y}\right)^2\right] - 2E\left[\sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)\left(\overline{y}_{st} - \overline{Y}\right)\right]$$

but $2E\left[\sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)\left(\overline{y}_{st} - \overline{Y}\right)\right] = 2NE\left[\left(\overline{y}_{st} - \overline{Y}\right)^2\right]$ [from the-5.2]

$$2E[\textstyle\sum_h N_h (\bar{y}_{st} - \bar{Y})^2] - 2E[\sum_h N_h (\bar{y}_h - \bar{Y}_h)(\bar{y}_{st} - \bar{Y})]$$

$$= \textstyle\sum_h N_h E(\bar{y}_{st} - \bar{Y})^2 - 2E(\bar{y}_{st} - \bar{Y})^2$$

$$= NV\left(\overline{y}_{st}\right) - 2NV\left(\overline{y}_{st}\right) = -NV\left(\overline{y}_{st}\right)$$

$$= -N \cdot \frac{1}{N^2} \sum_h N_h \left(N_h - n_h\right) \frac{S_h^2}{n_h} \quad \text{[from the-5.3]}$$

$$= -\sum_h N_h \left[\frac{N_h - n_h}{N}\right] \frac{S_h^2}{n_h}$$

$$E\left[\sum_h N_h \left(\overline{y}_h - \overline{Y}_h\right)^2\right] = \sum_h N_h E\left[\left(\overline{y}_h - \overline{Y}_h\right)^2\right] = \sum_h N_h \frac{\left(N_h - n_h\right)}{N_h} \frac{S_h^2}{n_h}$$

Because with in each stratum $\overline{y}_h$ is the mean of SRS , Hence

$$E\left[\sum_h N_h \left(\overline{y}_h - \overline{y}_{st}\right)^2\right] = \sum_h N_h \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_h \frac{\left(N_h - n_h\right)}{n_h} S_h^2 - \sum_h N_h \left[\frac{N_h - n_h}{N}\right] \frac{S_h^2}{n_h}$$

$$= \sum_h N_h \left(\overline{Y}_h - \overline{Y} \right)^2 + \sum_h S_h^2 \left(\frac{N_h - n_h}{n_h} \right) \left(1 - \frac{N_h}{N} \right)$$

$$= \sum_h N_h \left(\overline{Y}_h - \overline{Y} \right)^2 + \sum_h \frac{N_h S_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) \left(1 - \frac{N_h}{N} \right)$$

$$= \sum_h N_h \left(\overline{Y}_h - \overline{Y} \right)^2 + \sum_h \frac{N_h S_h^2}{n_h} (1 - f_h)(1 - W_h)$$

Corollary: An unbiased estimator of $= \sum_h N_h \left(\overline{y}_h - \overline{Y} \right)^2$ is

$\sum_h N_h (\bar{y}_h - \bar{y}_{st})^2 - \sum_h \frac{N_h s_h^2}{n_h}(1-f_h)(1-W_h) \longrightarrow (2)$

Substituting equation (2) in (1) we obtained an unbiased estimator of $V\left(\overline{y}\right)$ i.e.,

$$v\left(\overline{y}\right) = \frac{N-n}{n(N-1)} \left[\sum_h W_h s_h^2 - \sum_h \frac{W_h s_h^2}{n_h} + \sum_h \frac{W^2_h s_h^2}{n_h} - \sum_h \frac{W_h s_h^2}{N} + \sum_h W_h \overline{y}_h^2 - \left(\sum_h W_h \overline{y}_h \right)^2 \right] \rightarrow (3)$$

In nearly all applications simplification can be utilized

(i) N>50 . The fourth term inside the bracket of equation (3) be ignored.

$$\because v\left(\overline{y}\right) = \frac{N-n}{n(N-1)} \left[\sum_h W_h s_h^2 - \sum_h \frac{W_h s_h^2}{n_h} + \sum_h \frac{W^2_h S_h^2}{n_h} + \sum_h W_h \overline{y}_h^2 - \left(\sum_h W_h \overline{y}_h \right)^2 \right] \rightarrow (4)$$

If the sample allocation is large enough in each stratum, i.e., All $n_h > 50$. The second and third terms inside the bracket of equation (4) may be dropped

$$\because v\left(\overline{y}\right) = \frac{N-n}{n(N-1)} \left[\sum_h W_h s_h^2 + \sum_h W_h \overline{y}_h^2 - \left(\sum_h W_h \overline{y}_h \right)^2 \right] \rightarrow (5)$$

Relative precision of method-1 to method-2 is $= \dfrac{\text{precision of estimate in method - 1}}{\text{precision of estimate in method - 2}}$

Suppose, V_1 : variance in method 1, V_2 : variance in method 2

$$R.P = \frac{\left(1 / V_1 \right)}{\left(1 / V_2 \right)} = \frac{V_2}{V_1}$$

The estimate of the relative gain in precision due to stratification is thus obtained by

$$\frac{v(\bar{y}_{sR}) - v(\bar{y}_{st})}{v(\bar{y}_{st})}$$

3.4 ALLOCATION PROBLEM:

Question: How do you choose the sample sizes $n_1, n_2, ..., n_k$ so that the available resources are used effectively?

There are two aspects of choosing the sample sizes:

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

(ii) Maximize the precision for a given cost.

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

Based on different ideas, some allocation procedures are as follows:

1. **Equal allocation:**

   Choose the sample size i n to be the same for all the strata.

   Draw samples of equal size from each stratum.

   Let n be the sample size and k be the number of strata, then

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

2. **Proportional allocation:**

   For fixed k, select n_1 such that it is proportional to stratum size N_i, i.e.,

   $n_i \propto N_i$

   $or\ \mathrm{n}_i = \mathrm{C}N_i$

   Where C is the constant of proportionality.

   $\sum_{i=1}^{k} \mathrm{n}_i = \sum_{i=1}^{k} CN_i$

   $or\ \mathrm{n} = \mathrm{CN}$

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

Thus $\mathrm{n}_i = \left(\frac{n}{N}\right) N_i.$

Such allocation arises from considerations like operational convenience.

3. **Neyman or optimum allocation:**

   This allocation considers the size of strata as well as variability

   $n_i \propto N_i S_i$

   $n_i = \mathrm{C}^* N_i S_i$

Where C^* is the constant of proportionality.

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

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

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

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

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

3.5 CHOICE OF SAMPLE SIZES BASED ON DIFFERENT STRATAS:

Choice of sample size based on the cost of the survey and variability

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

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

where

C: total cost

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

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

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

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

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

Thus ϕ is minimum when

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

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

How to determine λ ?

There are two ways to determine λ.

(i) Minimize variability for a fixed cost.

(ii) Minimize cost for given variability.

We consider both cases.

How to determine λ ?

There are two ways to determine λ.

(i) Minimize variability for a fixed cost.

(ii) Minimize cost for given variability.

We consider both cases.

(i) Minimize variability for fixed cost

Let $C = C_0^*$ be the pre-specified cost which is fixed.

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

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

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

Substituting λ in the expression for $n_i = \frac{1}{\lambda} \frac{w_i S_i}{\sqrt{C_i}}$, the optimum n_i is obtained as

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

The required sample size to estimate $\bar{Y}$ such that the variance is minimum for the given cost $C = C_0^*$ is $n = \sum_{i=1}^{k} n_i^*$.

(ii) Minimize cost for a given variability

Let $V = V_0$ be the pre-specified variance. Now determine n_i such that

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

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

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

or $$\lambda = \frac{V_0 + \sum_{i=1}^{k} \frac{w_i^2 S_i^2}{N_i}}{\sum_{i=1}^{k} w_i S_i \sqrt{C_i}}$$ (after substituting $n_i = \frac{1}{\lambda}\frac{w_i S_i}{\sqrt{C_i}}$).

Thus the optimum n_i is

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

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

Sample size under proportional allocation for fixed cost and for fixed variance

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

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

So $C_0 = n \sum_{i=1}^{k} w_i C_i$ or $n = \frac{C_0}{\sum_{i=1}^{k} w_i C_i}$. Thus $n_i = \frac{C_o w_i}{\sum w_i C_i}$.

The required sample size to estimate $\bar{Y}$ in this case is $n = \sum_{i=1}^{k} n_i$.

(ii) If variance $= V_0$ is fixed, then

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

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

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

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

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

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

3.6 VARIANCES UNDER DIFFERENT ALLOCATIONS:

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

(i) Proportional allocation

Under proportional allocation

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

and

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

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

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

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

(ii) Optimum allocation

Under optimum allocation

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

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

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

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

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

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

3.7 COMPARISON OF VARIANCES OF THE SAMPLE MEAN UNDER SRS WITH STRATIFIED MEAN UNDER PROPORTIONAL AND OPTIMAL ALLOCATION:

(a) Proportional allocation:

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

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

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

Consider

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

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

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

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

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

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

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

Thus

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

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

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

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

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

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

(b) Optimum allocation

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

Consider

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

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

$$= \frac{1}{n}\sum_{i=1}^{k}w_iS_i^2 - \frac{1}{n}\bar{S}^2$$

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

where $\bar{S} = \sum_{i=1}^{k}w_iS_i$ and the larger gain in efficiency is achieved when S_i differs from $\bar{S}$ more.

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

Combining the results in (a) and (b), we have $Var_{opt}(\bar{y}_{st}) \leq Var_{prop}(\bar{y}_{st}) \leq Var_{SRS}(\bar{y})$.

3.8 SUMMARY AND CONCLUSIONS:

In this unit, we explored the key concepts and applications of stratified sampling, an essential technique in survey sampling designed to improve precision by dividing the population into homogeneous subgroups (strata).

- Introduction provided a foundation for understanding the motivation and rationale behind stratification—mainly to reduce sampling error and ensure representative sampling across key subgroups.
- Stratified Sampling for Proportions, we examined how to estimate population proportions using stratified sampling, including formulas for weighted means and variance.
- Estimation of the Gain in Precision highlighted the advantage of stratified sampling over simple random sampling (SRS), demonstrating how stratification can yield lower variances and more reliable estimates, especially when there is heterogeneity across strata.
- The Allocation Problem addressed how to distribute the total sample size across different strata optimally. We reviewed proportional, equal, and Neyman (optimal) allocation methods, each with different criteria for balancing precision and cost.
- hoice of Sample Sizes focused on determining sample sizes in each stratum, considering population size, variability, and resource constraints.
- In Variances under Different Allocations, we compared how the allocation strategy influences the variance of estimators. Optimal allocation generally results in the smallest variance, followed by proportional and then equal allocation.
- Finally, Comparison of Variances showed how stratified sampling, especially under optimal or proportional allocation, typically outperforms SRS in terms of efficiency and accuracy of the sample mean.

CONCLUSION:

Stratified sampling is a powerful and flexible technique, particularly effective when the population is heterogeneous. By thoughtfully dividing the population and choosing appropriate allocation strategies, researchers can significantly enhance the precision of their estimates. The choice of allocation method and sample size per stratum plays a crucial role in minimizing variance and ensuring cost-effective sampling. Overall, stratified sampling offers a methodological advantage in both theory and practical survey applications.

3.9 KEY WORDS:

- Stratified Sampling
- Stratum/Strata
- Proportion Estimation
- Precision Gain
- Between-stratum Variance
- Within-stratum Variance
- Optimal Allocation
- Proportional Allocation
- Neyman Allocation
- Sample Size Determination
- Population Mean
- Variance Estimation
- Simple Random Sampling (SRS)
- Cost-effective Sampling
- Homogeneity within Strata
- Heterogeneity between Strata
- Sampling Efficiency
- Allocation Problem
- Sampling Design
- Estimation Accuracy

3.10 SELF ASSESSMENT QUESTIONS:

1. What are the main advantages of stratified sampling over simple random sampling?
2. How do you estimate a population proportion using stratified sampling?
3. What is meant by "gain in precision" due to stratification? How is it quantified?
4. Describe the Neyman allocation and how it differs from proportional allocation.
5. How does the choice of sample sizes affect the precision of a stratified estimator?
6. What are the variances under proportional and optimal allocation?
7. When would stratified sampling provide significantly better estimates than SRS?
8. In what situations would you prefer proportional allocation over Neyman allocation?
9. What factors should be considered when deciding the allocation of sample sizes in different strata?
10. Give an example where stratified sampling leads to a more precise result than SRS.

3.11 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
3. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
4. Singh, D. and Chaudhary, F.S. (1986) *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.
5. Kish, L. (1965) *Survey Sampling*, Wiley.
6. Mukhopadhyay, P. (2008) *Theory and Methods of Survey Sampling*, PHI Learning.
7. T.W. Anderson (2003) *An Introduction to Multivariate Statistical Analysis.*

Prof. G. V. S. R. Anjaneyulu

LESSON- 4

METHODS OF POPULATION WITH LINEAR TREND

OBJECTIVES:

By the end of this lesson, learners will be able to:

Understand Linear Trends in Populations:

- Identify the presence and nature of linear trends in finite populations.

Introduce Systematic Sampling Techniques:

- Understand the limitations of simple systematic sampling when linear trends are present.

Learn Yates End Correction:

- Study Yates' approach to correct bias in systematic sampling under linear trend conditions.

Explore Modified Systematic Sampling:

- Learn the method of constructing modified systematic samples that reduce trend-induced bias.

Understand Balanced Systematic Sampling:

- Examine the concept of balancing sample selection to account for population trends.

Study Centrally Located Sampling:

- Understand the concept of sampling units near the center of each sampling interval.

Understand Circular Systematic Sampling:

- Learn how circular systematic sampling addresses edge effects and trend biases.

Compare Sampling Methods:

- Compare the efficiencies, variances, and biases of various sampling techniques in the presence of linear trends.

Apply Techniques to Practical Situations:

- Use real-world examples to demonstrate the effectiveness of each method.

STRUCTURE:

4.1 INTRODUCTION:

In survey sampling, **systematic sampling** is a widely used technique due to its simplicity and practical applicability. However, its effectiveness significantly diminishes when the population exhibits a **linear trend**—that is, a consistent increase or decrease in the values across units ordered in a particular fashion (e.g., time, geography, or rank).

In such cases, simple systematic sampling can introduce bias and lead to inefficient estimates**.** This has motivated the development of modified sampling strategies that specifically address the shortcomings of systematic sampling in the presence of trends.

To mitigate the effect of linear trends and enhance the precision of estimates, several improved methods have been proposed:

- Yates End Correction attempts to reduce bias by adjusting for the ends of the population.
- Modified Systematic Sampling restructures the selection intervals to counteract trend effects.
- Balanced Systematic Sampling ensures that samples are selected in a way that balances the trend influence.
- Centrally Located Sampling focuses on choosing units near the center of each interval to minimize variability.
- Circular Systematic Sampling addresses edge bias by wrapping the population around in a circular manner.

These methods aim to retain the operational ease of systematic sampling while significantly improving accuracy and reliability when linear trends are present in the population.

4.2 YATES END CORRECTION:

If the linear trend is present in the population, LSS estimator for $\bar{\boldsymbol{Y}}$ can be improved by giving the weights:

$$\frac{\mathbf{1}}{\mathbf{n}}+\frac{(\mathbf{2r-k-1})}{\mathbf{2(n-1)k}}\ \mathbf{and}\ \frac{\mathbf{1}}{\mathbf{n}}-\frac{(\mathbf{2r-k-1})}{\mathbf{2(n-1)k}}$$

To the first and the last units in the sample respectively instead of the usual weight of 1/n. These weights have been determined such that when applied to the specular linear population considered. The estimator turns out to be $\bar{\mathbf{Y}}$ giving rise to zero variance. For, letting the weights for the first and last units to be $\left(\frac{1}{n}+x\right)$ and $\left(\frac{1}{n}-x\right)$ respectively we get the estimate for the r[th] systematic sample is

$$\bar{\boldsymbol{y}}_{\boldsymbol{r}}=\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{j=0}}^{\boldsymbol{n-1}}\{\boldsymbol{a}+\boldsymbol{b}(\boldsymbol{r}+\boldsymbol{jh})\}+\boldsymbol{x}(\boldsymbol{a}+\boldsymbol{br})-\boldsymbol{x}[\boldsymbol{a}+\boldsymbol{b}(\boldsymbol{r}+(\boldsymbol{n}-\mathbf{1})\boldsymbol{k})]$$

$$= a+b\ (r+\frac{\boldsymbol{n-1}}{\boldsymbol{n2}}\boldsymbol{k})\text{-}xb(n\text{-}1)k$$

Equating this to the population mean a+b (N+1) and solving the x, we get

$$X=\frac{2r-k-1}{2(n-1)k}$$

These corrections to the estimator based on the systematic sample drawn from a population exhibiting a linear trend in the population values are called End corrections, invented by Yates in 1948. It may be pointed out that these end corrections may make the estimator slightly biased through the variance is likely to be reduced.

4.3 MODIFIED SYSTEMATIC SAMPLING:

Suppose, we have a population of size N, the units of which are denoted by $\{U_1, U_2, U_3, ..., U_N\}$. To select a sample of size n from this population, we will arrange N units into $k_1 = L/n$ (where L is the least common multiple of N and n) groups, each containing $s = N/k_1$ elements. The partitioning of groups is shown in Table 1. A set of $m = L/N$ groups from these k_1 groups are selected using simple random sampling without replacement to get a sample of size $ms = n$.

Table 1: Labels of population units arranged in MSSM.

		Labels of Sample units			
Groups	G_1	U_1	U_{k_1+1} . .		$U_{(s-1)k_1+1}$
	G_2	U_2	U_{k_1+2} . .		$U_{(s-1)k_1+2}$
	G_3	U_3	U_{k_1+3} . .		$U_{(s-1)k_1+3}$
	G_i	U_i	U_{k_1+i} . .		$U_{(s-1)k_1+i}$
	G_{k_1}	U_{k_1}	U_{2k_1} . .		$U_{sk_1=N}$

Thus sample units with random starts $r_i (i = 1, 2, ..., m)$ selected from 1 to k_1 correspond to the following labels:

(2.1) $\quad r_i + (j-1)k_1, \qquad i = 1, 2, ..., m \quad \text{and} \quad j = 1, 2, ..., s.$

Consider the mean estimator

$$\bar{y}_{MSSM} = \frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s} y_{r_ij} = \frac{1}{m}\sum_{i=1}^{m}\Big(\frac{1}{s}\sum_{j=1}^{s} y_{r_ij}\Big).$$

where y_{r_ij} is the value of the jth unit of the ith random group.

Taking expectation on both sides, we get:

$$\begin{aligned} E\left(\bar{y}_{MSSM}\right) &= \frac{1}{m}\sum_{i=1}^{m} E\Big(\frac{1}{s}\sum_{j=1}^{s} y_{r_ij}\Big) \\ &= \frac{1}{m}\sum_{i=1}^{m}\frac{1}{k_1}\sum_{i=1}^{k_1}\Big(\frac{1}{s}\sum_{j=1}^{s} y_{ij}\Big) = \frac{1}{sk_1}\sum_{i=1}^{k_1}\sum_{j=1}^{s} y_{ij} = \mu, \end{aligned}$$

where y_{ij} is the value of the jth unit of the ith group and μ is the population mean.

The variance of $\bar{y}_{MSSM}$ is given by

$$V\left(\bar{y}_{MSSM}\right) = E\left(\bar{y}_{MSSM} - \mu\right)^2 = \frac{1}{m^2} E\left[\sum_{i=1}^{m}\left(\bar{y}_{r_i.} - \mu\right)\right]^2,$$

where $\bar{y}_{r_i.}$ is the mean of ith random group.

After simplification, we have:

$$V\left(\bar{y}_{MSSM}\right) = \frac{1}{mk_1}\frac{(k_1 - m)}{(k_1 - 1)}\sum_{i=1}^{k_1}\left(\bar{y}_{i.} - \mu\right)^2, \tag{2.2}$$

where $\bar{y}_{i.}$ is the mean of ith group.

Further, it can be observed that in a situation when MSSM becomes LSS, the variance expression given in Equation (2.2) reduces to variance of LSS, i.e.,

$V\left(\bar{y}_{MSSM}\right) = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}_{i.} - \mu\right)^2 = V\left(\bar{y}_{LSS}\right).$

Similarly, in the case when MSSM becomes SRS, $V\left(\bar{y}_{MSSM}\right)$ reduces to variance of SRS without replacement, i.e.,

$$V\left(\bar{y}_{MSSM}\right) = \frac{(N-n)}{nN}\frac{1}{(N-1)}\sum_{i=1}^{N}\left(y_i - \mu\right)^2 = V\left(\bar{y}_{SRSWOR}\right).$$

The alternative expressions for $V\left(\bar{y}_{MSSM}\right)$ have been presented in Theorems 2.1, 2.2 and 2.3:

Theorem 2.1. *The variance of sample mean under MSSM is:*

$$V(\bar{y}_{MSSM}) = \frac{1}{mN}\frac{(k_1-m)}{(k_1-1)}\Big[(N-1)S^2 - k_1(s-1)S_{wg}^2\Big],$$

where $S^2 = \frac{1}{N-1}\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2$, *and* $S_{wg}^2 = \frac{1}{k_1(s-1)}\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\bar{y}_i)^2$ *is the variance among the units that lie within the same group.*

Proof: From analysis of variance, we have:

$$\sum_{i=1}^{N}(y_i-\mu)^2 = s\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 + \sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\bar{y}_i)^2, \quad \text{or}$$

$$(N-1)S^2 = s\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 + k_1(s-1)S_{wg}^2.$$

Thus

$$V(\bar{y}_{MSSM}) = \frac{1}{mN}\frac{(k_1-m)}{(k_1-1)}\Big[(N-1)S^2 - k_1(s-1)S_{wg}^2\Big]. \tag{2.3}$$ □

Theorem 2.2. *The variance of sample mean under MSSM is:*

$$V(\bar{y}_{MSSM}) = \frac{1}{n}\left(\frac{k_1-m}{k_1-1}\right)\left(\frac{N-1}{N}\right)S^2\Big[1+(s-1)\rho_w\Big],$$

where

$$\rho_w = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{s}\sum_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\mu)(y_{ij'}-\mu)/s(s-1)k_1}{\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2/sk_1}.$$

Proof: Note that

$$\begin{aligned}
V(\bar{y}_{MSSM}) &= \frac{1}{mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2 \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\sum_{j=1}^{s}(y_{ij}-\mu)\Big]^2 \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\Big[\sum_{i=1}^{k_1}\sum_{j=1}^{s}(y_{ij}-\mu)^2 + \sum_{i=1}^{k_1}\sum_{j\neq 1}^{s}(y_{ij}-\mu)(y_{iu}-\mu)\Big] \\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\Big[(sk_1-1)S^2 + (sk_1-1)(s-1)S^2\rho_w\Big].
\end{aligned}$$

Hence

$$V(\bar{y}_{MSSM}) = \frac{1}{n}\frac{(k_1-m)}{(k_1-1)}\frac{(N-1)}{N}S^2\Big[1+(s-1)\rho_w\Big], \tag{2.4}$$

where ρ_w is the intraclass correlation between the pairs of units that are in the same group. □

Theorem 2.3. *The variance of $\bar{y}_{MSSM}$ is:*

$$V(\bar{y}_{MSSM}) = \frac{(k_1-m)}{mN}S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big],$$

where

$$S^2_{wst} = \frac{1}{s(k_1-1)}\sum_{j=1}^{s}\sum_{i=1}^{k_1}(y_{ij}-\bar{y}_{.j})^2$$

and

$$\rho_{wst} = \frac{\sum_{i=1}^{k_1}\sum_{j=1}^{s}\sum_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\bar{y}_j)(y_{ij'}-\bar{y}_{j'})}{s(s-1)(k_1-1)S^2_{wst}}.$$

Proof: Note that

$$\begin{aligned}
V(\bar{y}_{MSSM}) &= \frac{1}{mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}(\bar{y}_i-\mu)^2\\
&= \frac{1}{mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\frac{1}{s}\sum_{j=1}^{s}y_{ij}-\frac{1}{s}\sum_{j=1}^{s}\bar{y}_j\Big]^2\\
&= \frac{1}{s^2mk_1}\frac{(k_1-m)}{(k_1-1)}\sum_{i=1}^{k_1}\Big[\sum_{j=1}^{s}(y_{ij}-\bar{y}_j)\Big]^2\\
&= \frac{1}{smN}\frac{(k_1-m)}{(k_1-1)}\Bigg[\sum_{j=1}^{s}\sum_{i=1}^{k_1}(y_{ij}-\bar{y}_j)^2+\sum_{i=1}^{k_1}\sum_{j=1}^{s}\sum_{\substack{j'=1\\ j'\neq j}}^{s}(y_{ij}-\bar{y}_j)(y_{ij'}-\bar{y}_{j'})\Bigg]\\
&= \frac{1}{smN}\frac{(k_1-m)}{(k_1-1)}s(k_1-1)S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big].
\end{aligned}$$

Hence

$$V(\bar{y}_{MSSM}) = \left(\frac{k_1-m}{mN}\right)S^2_{wst}\Big[1+(s-1)\rho_{wst}\Big]. \tag{2.5}$$

□

4.4 BALANCED SYSTEMATIC SAMPLING:

If one wishes to draw a sample of size n from a population of size N, then a simple way to do this would be to randomly select a unit and then to select subsequent units at equally spaced intervals, until a sample of size n is achieved. More specifically, if one randomly selects a unit from the first $k = N/n$ units and every kth unit thereafter, then this sampling design is known as linear systematic sampling (LSS), provided that k is an integer (Cochran, 1977). LSS is advantageous over simple random sampling without replacement (SRS) and stratified random sampling (STR) (based on the random selection of one unit per stratum from n strata, each of size k), owing to its convenience and operational simplicity when implemented.

Consider a finite population $U = (U_1, ..., U_N)$ of size N and let y_q be the value of the study variable of the qth unit of population U, for $q \in \{1, ..., N\}$. Accordingly, the population mean $\overline{Y} = \sum_{q=1}^{N} y_q/N$ is estimated from the sample mean $\bar{y}$. Suppose a population that exhibits linear trend, given by the model A

$$y_q = a + bq + e_q, \qquad q = 1, ..., N \tag{1}$$

where a and b are constants and the e_q's denote the random errors which follows Cochran's (1946) super-population model, i.e. if the function $\mathscr{E}$ denotes the average of all potential finite populations that can be drawn from model A, then

$$\mathscr{E}(e_q) = 0, \qquad \mathscr{E}(e_q^2) = \sigma^2, \qquad \mathscr{E}(e_q e_z) = 0 \, (q \neq z).$$

By using (1), the population mean is given by

$$\overline{Y} = \frac{1}{N}\sum_{q=1}^{N} y_q = \frac{1}{N}\sum_{q=1}^{N} a + \frac{b}{N}\sum_{q=1}^{N} q + \frac{1}{N}\sum_{q=1}^{N} e_q = a + \frac{b(N+1)}{2} + \bar{\bar{e}},$$

where $\bar{\bar{e}} = \sum_{q=1}^{N} e_q/N$ denotes the mean random error of the population. Now, let $\bar{y}_{\text{LSS}}$, $\bar{y}_{\text{SRS}}$, and $\bar{y}_{\text{STR}}$, denote the sample means when conducting LSS, SRS and STR, respectively. Thus, when estimating $\overline{Y}$ under model A, we note that the expected mean square errors (MSEs) of $\bar{y}_{\text{LSS}}$, $\bar{y}_{\text{SRS}}$, and $\bar{y}_{\text{STR}}$, are respectively given by

$$M_{\text{LSS}} = \sigma_e^2 + \frac{b^2(k^2-1)}{12}, \tag{2}$$

$$M_{\text{SRS}} = \sigma_e^2 + \frac{b^2(N+1)(k-1)}{12}, \tag{3}$$

and

$$M_{\text{STR}} = \sigma_e^2 + \frac{b^2(k^2-1)}{12n}, \tag{4}$$

where $\sigma_e^2 = \sigma^2(1/n - 1/N)$ represents the minimum expected error variance component, while the second terms on the right hand side represent the linear trend components (Bellhouse, 1988). By comparing Equations (2) through to (4), we obtain

$$M_{\text{STR}} \leq M_{\text{LSS}} \leq M_{\text{SRS}}. \tag{5}$$

Accordingly, some authors have suggested modified LSS designs to remove the linear trend component in Equation (2) and thus improve efficiency. Yates (1948) proposed a corrected estimator which uses the LSS design and is termed as the *Yates' end corrections* (YEC) estimator. This

estimate is obtained by applying appropriate weights to the first and the last sampling units. *Centered systematic sampling* (CESS) was first discussed by Madow (1953), where the centrally located linear systematic sample is selected and thus no randomization is required. The centered systematic sample mean is subject to bias, since certain population units have no chance of being selected for the sample (Murthy, 1967). A balanced arrangement reverses the order, with respect to the population unit indices, of every alternative set of k population units. Sethi (1965) considered the application of LSS on this arrangement and this design was later named as *balanced systematic sampling* (BSS) by Murthy (1967, p. 165). Singh, Jindal and Garg (1968) suggested the application of LSS on a modified arrangement, where a subset of units from the end of the population is reversed, with respect to their population unit indices. This sampling design is known as *modified systematic sampling* (MSS). Denote $\bar{y}_{\text{YEC}}$, $\bar{y}_{\text{CESS}}$, $\bar{y}_{\text{BSS}}$, and $\bar{y}_{\text{MSS}}$, as the sample means associated with YEC, CESS, BSS and MSS, respectively. When estimating $\overline{Y}$ under model A, the expected MSEs of these sample means are respectively given by

$$M_{\text{YEC}} = \sigma_e^2 + \frac{\sigma^2\left(k^2-1\right)}{6(n-1)^2k^2}, \tag{6}$$

$$M_{\text{CESS}} = \begin{cases} \sigma_e^2, & \text{if } k \text{ is odd} \\ \sigma_e^2 + b^2/4, & \text{if } k \text{ is even} \end{cases} \tag{7}$$

and

$$M_{\text{BSS}} = M_{\text{MSS}} = \begin{cases} \sigma_e^2, & \text{if } n \text{ is even} \\ \sigma_e^2 + b^2(k^2-1)/12n^2, & \text{if } n \text{ is odd} \end{cases} \tag{8}$$

(Fountain and Pathak, 1989). By referring to Equations (6) to (8), we note that: (*i*) while there is a complete removal of the linear trend component in M_{YEC}, there is a larger error variance component, owing to the uneven weighting of the sampling units; (*ii*) the linear trend component in M_{CESS} is only eliminated when k is odd; and (*iii*) both M_{BSS} and M_{MSS} are equivalent, with the linear trend components being removed only for the case when n is even. Good reviews for these designs are provided by Bellhouse and Rao (1975), Cochran (1977), Fountain and Pathak (1989), Singh (2003) and the corresponding references cited therein.

More recent optimal modified LSS designs for linear trend populations have been suggested by Subramani (2000, 2009, 2010) and Khan, Shabbir and Gupta (in press), while Mukerjee and Sengupta (1990) proposed optimal design-unbiased strategies to estimate $\overline{Y}$. As in the case of the earlier designs, these recent solutions are based on certain assumptions and/or are optimal for linear trend populations under certain conditions.

In the present paper, a modified LSS design, termed as *balanced modified systematic sampling* (BMSS), is proposed. In Section 2, a discussion on the methodology of BMSS is provided. For Section 3, the expected MSE of the BMSS sample mean, is compared to that of M_{LSS}, M_{SRS}, M_{STR}, M_{YEC}, M_{CESS}, M_{BSS} and M_{MSS}. As a result, BMSS is only optimal for the case when $n/2$ is an even integer. A BMSS with end corrections (BMSSEC) estimator is thus constructed, so as to remove the linear trend component in the corresponding expected MSE for the other cases of n. A numerical example on a hypothetical population is then considered in Section 4, before carrying out

a simulation study in Section 5. Note that k is assumed to be an integer throughout this paper, i.e. assuming that N is an exact multiple of n, so that sampling is conducted linearly.

2. Balanced Modified Systematic Sampling (BMSS)

A modified arrangement used for BMSS is defined as follows: (a) if n is even, then the order of every alternative set of k population units is reversed, before reversing the order of the first/last $n/2$ sets of k population units; and (b) if n is odd, then the order of every alternative set of k population units is reversed, before reversing the order of the last $(n-1)/2$ sets of k population units. LSS is then applied to this modified arrangement, so as to select the required sample. Note that different arrangements, before applying LSS, will result in different compositions of samples and this paper deals with a specific arrangement, as explained above. By reversing the order of $n/2$ (or $(n-1)/2$) sets of k population units, a balancing effect is obtained which is optimal for populations exhibiting linear trend. Note that MSS reverses the order of the last $n/2$ (or $(n-1)/2$) sets of k population units, without alternating the order of each set, while BSS alternates the order of each set, without reversing the order of the last $n/2$ (or $(n-1)/2$) sets of k population units. Thus, the ordering of BMSS is a mixture of both, the MSS and BSS orderings. Moreover, BMSS reduces to LSS when $n=2$ and we will thus assume that $n>2$.

The above-mentioned design is equivalent to selecting sampling units according to the following indices:

(A) if $n/2$ is an even integer, then

$$i+2jk, \qquad 2(j+1)k-i+1, \qquad \text{for } j=0,\ldots,(n-4)/4$$

and

$$N+i-k-2jk, \qquad N-i-k-2jk+1, \qquad \text{for } j=0,\ldots,(n-4)/4;$$

(B) if $n/2$ is an odd integer, then

$$i+2jk, \qquad N+i-k-2jk, \qquad \text{for } j=0,\ldots,(n-2)/4$$

and

$$2(j+1)k-i+1, \qquad N-i-k-2jk+1, \qquad \text{for } j=0,\ldots,(n-6)/4;$$

(C) if $n=3$, then

$$i, \qquad 2k-i+1 \qquad \text{and} \qquad N-i+1;$$

(D) if $n\neq 3$ and $(n+1)/2$ is an even integer, then

$$i+2jk, \qquad 2(j+1)k-i+1, \qquad N-i-2jk+1, \qquad \text{for } j=0,\ldots,(n-3)/4$$

and

$$N+i-2(j+1)k, \qquad \text{for } j=0,\ldots,(n-7)/4;$$

(E) if $(n+1)/2$ is an odd integer, then

$$i+2jk, \quad 2(j+1)k-i+1, \quad N-i-2jk+1, \quad N+i-2(j+1)k,$$
$$\text{for } j=0,\ldots,(n-5)/4 \quad \text{and} \quad i+(n-1)k/2.$$

Note that Cases (A) and (B) are sub-cases of n being even, while Cases (C) to (E) are sub-cases of $n>1$ being odd.

The ith ($i \in \{1,\ldots,k\}$) sample mean, denoted by $\bar{y}_{BMSS}$, is obtained by using the above sampling unit indices for the respective cases, e.g. if we consider Case (A), then the sample mean is given as

$$\bar{y}_{BMSS} = \frac{1}{n}\sum_{j=0}^{(n-4)/4}\left(y_{i+2jk}+y_{2(j+1)k-i+1}+y_{N+i-k-2jk}+y_{N-i-k-2jk+1}\right).$$

Note that $\bar{y}_{BMSS}$ is design-unbiased, since BMSS is viewed as an arrangement of units before applying LSS.

4.5 CENTRALLY LOCATED SAMPLING:

Another situation where the estimator in the case of the hypothetical linear population equals. $\bar{Y}$ is obtained by considering only the systematic sample with the start $\frac{(k+1)}{2}$ if k is add or the two systematic samples with starts $\frac{k}{2}$ and $\frac{(k+1)}{2}$ if k is even for, if k is add, substituting r=$\frac{(k+1)}{2}$ in $\bar{y}_r$, we get

$$\bar{y}_r = a + b\left(\frac{(k+1)}{2}, \frac{(n-1)}{2}k\right)$$
$$= a+b\,.\frac{(N+1)}{2}$$

And if k is even substituting $\frac{k}{2}$ and $\frac{(k+2)}{2}$ for r, we get

$$\bar{y}_{\frac{k}{2}} = a + \frac{1}{2}\,bN \quad \text{and} \quad \bar{y}_{\frac{(k+1)}{2}} = a + \frac{1}{2}\,b\,(N+2)$$

The mean of which is $\bar{Y}$. Hence, it may be desirable to consider only the systematic sample with $\frac{(k+1)}{2}$ as the random start if k is odd and the systematic samples with $\frac{k}{2}(or)\ \frac{(k+1)}{2}$ as random starts for selection with probability $\frac{1}{2}$ if k is even, when even there is a linear trend. Present in the population such a sample is known Centrally located sample.

But in practice, it is not advisable to use such a sample, especially when one is in doubt about the presence of present linear trend in the arrangement used, since it is not a valid sample due to certain units not getting any chance at all of being included in the sample and hence it is subject to bias and it is not possible to estimate the error involved in the estimator.

4.6 CIRCULAR SYSTEMATIC SAMPLING:

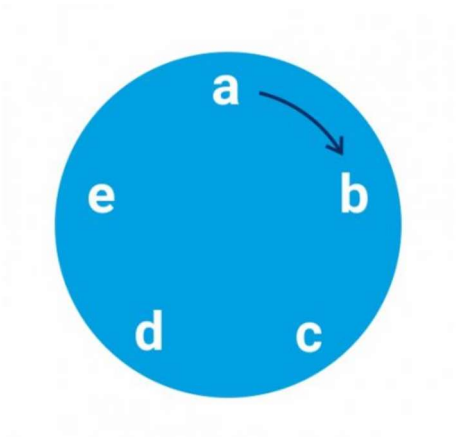
This is when a sample starts again at the same point after ending. This means that once the sampling interval reaches the last member of the population, it wraps around to the beginning and continues the selection process. Circular systematic sampling is often used in situations where the population exhibits cyclical patterns or where there is no clear starting or ending point. For example, if researchers are studying tree growth in a forest, they could use circular systematic sampling by selecting trees at regular intervals along a circular path, ensuring comprehensive coverage of the forest area.

Circular systematic sampling is like regular systematic sampling. However, rather than stopping at the end of the population list, you start over and continue sampling using your numerical interval until you've sampled every individual in the population. Researchers use this approach in cases where k isn't an integer.

Example 1: We have a population of 14 individuals numbered from 1 to 14. We want to select a sample of 4 individuals using circular systematic sampling.

1. Calculate the sampling interval: k = 14/4 = 3 (choose the closest integer to N/n)
2. Start randomly between 1 to 14: Let's say we randomly start at individual number 4.
3. Create samples by skipping through k units: We select individuals 4, 7, and 11.
4. Repeat until you select members of the entire population: Since we have only two individuals in our sample, the process ends here. However, we would continue until all 14 individuals are sampled, resulting in 14 samples, if we wanted to sample the entire population

In circular systematic sampling, a sample starts again from the same point once again after ending; thus, the name.
For example, if N = 7 and n = 2, k=3.5. There are two probable ways to form sample:



1. If we consider k=3, the samples will be – ad, be, ca, db and ec.
2. If we consider k=4, the samples will be – ae, ba, cb, dc and ed.
   How is a circular systematic sample selected?

- Calculate sampling interval (k) = N/n. (If N = 11 and n = 2, then k is taken as 5 and not 6)

- Start randomly between 1 to N
- Create samples by skipping through k units every time until you select members of the entire population.
- In the case of this method, there will be N number of samples, unlike k samples in the linear systematic sampling method.

Difference between linear systematic sampling and circular systematic sampling:
Here is the difference between linear and circular.

Linear	Circular
Create samples = k (sampling interval)	Create samples = N (total population)
The start and endpoints of this sample are distinct.	It restarts from the start point once the entire population is considered.
All sample units should be arranged in a linear manner prior to selection.	Elements will be arranged in a circular manner.

4.7 SUMMARY AND CONCLUSION:

In populations exhibiting a **linear trend**, conventional systematic sampling methods can lead to **biased estimates** and **inefficient results**, as the sample may not adequately reflect the trend in the data. To address these challenges, several refined sampling techniques have been developed.

- **Yates End Correction** modifies the estimation procedure to account for bias at the beginning and end of the population sequence.
- **Modified Systematic Sampling** alters the sample selection pattern to minimize the correlation between units introduced by the trend.
- **Balanced Systematic Sampling** ensures that units are selected in a way that balances the rising and falling values across the trend, reducing variance.
- **Centrally Located Sampling** selects units near the center of each interval, minimizing the potential bias caused by edge values within intervals.
- **Circular Systematic Sampling** removes boundary effects by treating the population as circular, thus equalizing the chance of selection for all units.

These advanced methods retain the **simplicity and practicality** of systematic sampling while significantly **improving the efficiency and accuracy** of estimates when the population is not homogeneous but follows a linear trend.

CONCLUSION:

Understanding and choosing appropriate sampling strategies is crucial when dealing with populations with linear trends**.** The use of modified techniques such as Yates End Correction or Balanced Systematic Sampling helps to reduce bias and improve precision**,** ensuring more reliable and valid statistical inferences**.** Researchers and practitioners should be equipped to recognize trend patterns in data and apply these alternative methods to enhance the quality of survey results.

4.8 KEY WORDS:

- Linear Trend
- Systematic Sampling
- Yates End Correction
- Modified Systematic Sampling
- Balanced Systematic Sampling
- Centrally Located Sampling
- Circular Systematic Sampling
- Trend-Induced Bias
- Sampling Interval
- Population Ordering
- Sampling Efficiency
- Sampling Variance
- Edge Effect
- Sampling Design
- Unbiased Estimation

4.9 SELF- ASSESSMENT QUESTIONS:

1. What is a linear trend in the context of a population? Why does it pose a problem in systematic sampling?
2. Explain the concept of Yates End Correction**.** How does it help in reducing bias?
3. How does Modified Systematic Sampling differ from ordinary systematic sampling?
4. What is the main idea behind Balanced Systematic Sampling**?** In what situations is it most effective?
5. Describe Centrally Located Sampling**.** Why is it useful in the presence of a trend?
6. What is the key advantage of using Circular Systematic Sampling over linear systematic sampling?
7. When should one prefer circular systematic sampling over centrally located sampling**?**
8. How does population ordering affect the efficiency of systematic sampling techniques**?**

4.10 SUGGESTED READINGS:

1. Cochran, W.G. (1977). Sampling Techniques. *3rd* ed. Wiley.
2. Theory and methods of survey sampling. Parimal Mukhopadhyay (1988).
3. Sukhatme, P.V. et al. (1984). Sampling Theory of Surveys with Applications. Iowa State University Press.
4. Levy, P.S., & Lemeshow, S. (2013). Sampling of Populations: Methods and Applications (4th ed.). Wiley.
5. Des Raj and Chandhok, P. (1998)**.** *Sample Survey Theory*. Narosa Publishing House
6. Murthy, M.N. (1967)**.** *Sampling Theory and Methods*, Statistical Publishing Society
7. Singh, D. and Chaudhary, F.S. (1986)**.** *Theory and Analysis of Sample Survey Designs,* New Age International Publishers

Prof. G. V. S. R. Anjaneyulu

LESSON - 5
CLUSTER SAMPLING

OBJECTIVES:

By the end of this module, learners will be able to:

- Understand the fundamentals of cluster sampling**,** including when and why it is used as an alternative to simple or stratified random sampling.
- Differentiate between equal and unequal cluster sizes**,** and comprehend their implications on sampling design and estimation.
- Describe and apply methods of selecting clusters with equal probability (SRS of clusters) and with varying probabilities (Probability Proportional to Size – PPS).
- Calculate unbiased estimators of population totals and means using cluster sampling with both equal and unequal cluster sizes.
- Evaluate and compare the variances of estimates obtained through cluster sampling under different designs.
- Analyze the efficiency and cost-effectiveness of cluster sampling relative to other sampling techniques in large-scale surveys.
- Implement practical procedures for drawing samples from populations using cluster sampling in real-world scenarios.
- Recognize the challenges and limitations of cluster sampling, including intra-cluster correlation and design effects.

STRUCTURE:

5.1 INTRODUCTION:

In statistics cluster sampling is a sampling plan used when mutually homogeneous yet internally heterogeneous groupings are evident in a statistical population. It is often used in marketing research.

In this sampling plan, the total population is divided into these groups (known as clusters) and a simple random sample of the groups is selected. The elements in each cluster are then

sampled. If all elements in each sampled cluster are sampled, then this is referred to as a "one-stage" cluster sampling plan. If a simple random subsample of elements is selected within each of these groups, this is referred to as a "two-stage" cluster sampling plan. A common motivation for cluster sampling is to reduce the total number of interviews and costs given the desired accuracy. For a fixed sample size, the expected random error is smaller when most of the variation in the population is present internally within the groups, and not between the groups.

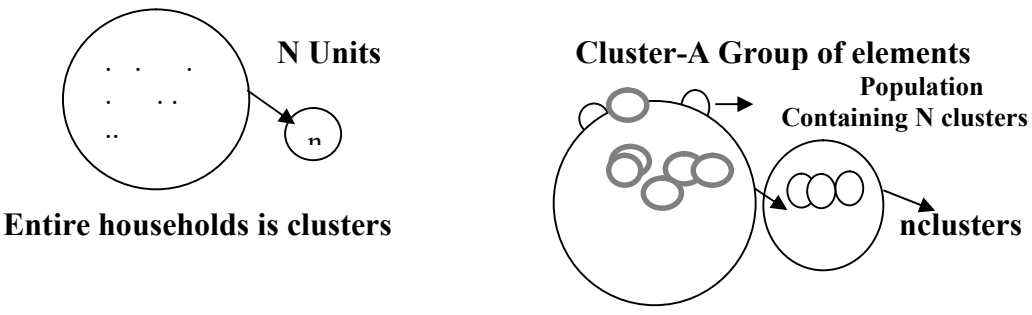
5.2 VARYING PROBABILITY SAMPLING:

The simple random sampling scheme provides a random sample where every unit in the population has an equal probability of selection. Under certain circumstances, more efficient estimators are obtained by assigning unequal probabilities of selection to the units in the population. This type of sampling is known as a varying probability sampling scheme.

If Y is the variable under study and X is an auxiliary variable related to Y, then in the most commonly used varying probability scheme, the units are selected with probability proportional to the value of X, called as size. This is termed as probability proportional to a given measure of size (pps) sampling. If the sampling units vary considerably in size, then SRS does not take into account the possible importance of the larger units in the population. A large unit, i.e., a unit with a large value of Y contributes more to the population total than the units with smaller values, so it is natural to expect that a selection scheme that assigns more probability of inclusion in a sample to the larger units than to the smaller units would provide more efficient estimators than the estimators which provide equal probability to all the units. This is accomplished through pps sampling.

Note that the “size” is the value of auxiliary variable X and not the value of study variable Y. For example, in an agriculture survey, the yield depends on the area under cultivation. So, bigger areas are likely to have a larger population, and they will contribute more to wards the population total, so the value of the area can be considered as the size of the auxiliary variable. Also, the cultivated area for a previous period can also be taken as the size while estimating the yield of the crop. Similarly, in an industrial survey, the number of workers in a factory can be considered as the measure of size when studying the industrial output from the respective factory.

5.3 CLUSTER SAMPLING:



4each+every are events **N-Clusters, M-elements** **nM Elements**

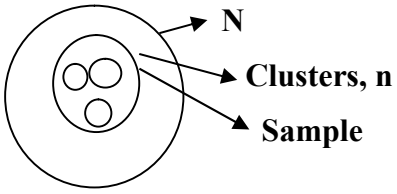
Ex: wards: n×1=50 (clusters)

Households, M=200 ,n=5×200=1000

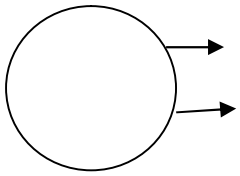
NM elements are there in the population n clusters

1. For artificial clusters has equal size.

2. For natural clusters has an unequal size



- In SRS



Prepare a sampling (for all the units in the population)frame for all at 1,2,…10,000, and we draw 1000 households using random number table, where as in cluster sampling we don't use sampling frame for entire ' N', and we prepare sampling frame only for ' n'

Cluster may be in equal or unequal size:

EX:- If the population is apple tree then each branch is a cluster and fruits are elements.

Population (N) > sample unit (1) > cluster (n) > element (M).

Q:1)How to draw a sample unit from cluster.

DESCRIPTION: The smallest unit into which the population can be divided is called an element of the population. A group of such elements is known as a cluster. When the sampling unit is a cluster, the procedure is called cluster sampling. Hence cluster sampling consists in forming suitable clusters of elements and surveying all the elements in a sample of clusters selected according to an appropriate sampling scheme.

i) Cluster may be of equal size and ii) Cluster may be of unequal size.

5.3.1 Examples of Sampling of Equal and Unequal Clusters:

There are no. of situations where it is convenient to take certain naturally formed groups of elements as clusters and in such cases the cluster size would in generally vary from cluster to cluster for instance, households, which are groups of persons, and villages or urban blocks, which are groups of households and persons, are usually considered as clusters for purpose of sampling because of operational convenience. Though the size of natural clusters such as villages, branches of trees (clusters of leaves, flowers, fruits (elements)etc). Usually varies over cluster, it is possible to have equal clusters when clusters are artificially formed.

For instance, in a crop survey, we may consider clusters of two or more plots or other area units of a given size and shape as clusters and in a house hold survey two or more neighbouring household may be grouped to form clusters. Similarly in a production process in an industry the no. of items produced at regular intervals of time may be the same and the

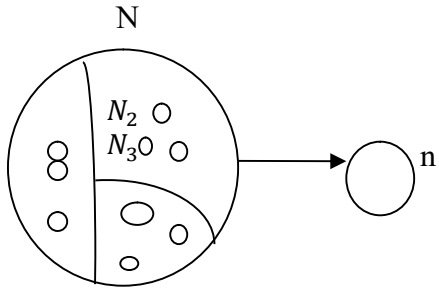
production at different intervals of time can be considered to constitute the clusters.

The rational choice between the two types of clusters may be made by the familiar principle of selecting the cluster that gives the smaller variance for a given cost or the minimum cost for a prescribed variance. When a list of individual houses is available, economic considerations may point to the choice of a larger cluster. For a given size of sample, a small cluster usually gives more precise results than a larger cluster. When cost is balanced against precision the larger cluster may prove superior.

Clusters are generally made up of neither elements and therefore the elements within a cluster tend to have similar characteristics.

After dividing the population into specified clusters the required no. of clusters can be selected and all the elements in selected clusters are enumerated. Various sampling procedures, e.g., Simple Random, Stratified (or) Systematic Sampling procedure can be applied to cluster sampling by treating the clusters themselves as sampling unit.

- Advantages of cluster sampling:-
  1) Collection of data for neighbouring elements in easier, cheaper, faster and operationally more convenient than observing units spread over a season.
  2) It is less costly than Simple Random Sampling due to the saving of time in Journeys, Identification, Contacts….etc.
  3) When the sampling frame of elements may not be readily available



Cluster (N) = $N_1 + N_2 + N_3$

Notations in the case of equal cluster size:-

N= No. of clusters in the population

n= No. of clusters in the sample

M= No. of elements in the cluster

Let $y_{ij=}$ observed values of j^{th} element in the i^{th} cluster ($i = 1,2, \dots, N; j = 1,2, \dots, M$).

$y_i = i^{th}$ Cluster, total $= \sum_{j=1}^{M} y_{ij} = y_{i1} + y_{i2} + \cdots + y_{iM}$

$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$ = Mean per i^{th} cluster in the population.

$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N} y_i}{NM}$ = Mean per element in the population, $\bar{\bar{Y}} = \bar{Y}/M$

$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$ = Mean per cluster in the sample

$\bar{\bar{y}} = \frac{\bar{y}}{M} =$ Mean per element in the sample

$S^2 = \frac{\sum_{i=1}^{N}\sum_{i=1}^{M}(y_{ij}-\bar{\bar{Y}})^2}{NM \quad 1}$ = variance among elements in the population.

ρ: Intra cluster correlation coefficient between elements within clusters.

$$\rho = \frac{E(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})}{E(y_{ij}-\bar{\bar{Y}})^2} = \frac{covariance}{variance}$$

$$\rho = \frac{\frac{\sum_i\sum_j(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})}{\frac{\cancel{NM}(M-1)}{2}}}{\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{Y}})^2}{\cancel{NM}}} = \frac{2\sum_i\sum_j(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})}{\frac{(M-1)}{(NM-1)S^2}}$$

$$\rho = \frac{2\sum_i\sum_{j<k}(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})}{(M-1)(NM-1)S^2} \qquad [\because S^2 = \frac{\sum i\sum j(y_{ij}-\bar{\bar{Y}})^2}{NM-1}]$$

THEOREM $-$ 2.2.1: $-$

A simple random sample of 'n' cluster, each containing M elements, is drawn from N clusters in the population. Then the sample mean per element $\bar{\bar{y}}$ is an unbiased estimate of $\bar{\bar{Y}}$ with variance is

$$V(\bar{\bar{y}}) = \frac{(1-f)}{n}\frac{(NM-1)}{M^2(N-1)}S^2[1+(M-1)\rho]$$

where ρ is the intra cluster correlation coefficient.

PROOF:-

Let y_i denote the total for the i^{th} cluster and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

$\therefore \bar{y}\ is$ an unbiased estimate of $\bar{Y}$, with variance

$$V(\bar{y}) = \frac{(1-f)}{n}\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1}$$

$$\bar{y} = M\bar{\bar{y}}\ and\ \bar{Y} = M\bar{\bar{Y}}$$

But $\frac{\bar{y}}{M} = \bar{\bar{y}}\ and\ \frac{\bar{Y}}{M} = \bar{\bar{Y}}$. Hence $\bar{\bar{y}}$ is an unbiased estimate of $\bar{\bar{Y}}$ with variance (M constant)

$$V(\bar{\bar{y}}) = V\left(\frac{\bar{y}}{M}\right) = \frac{1}{M^2}V(\bar{y})$$

$= \frac{(1-f)}{nM^2}\frac{\sum_{i=1}^{N}(y_i-\bar{Y})^2}{N-1} \rightarrow (1)$, Since $\frac{\bar{Y}}{M} = \bar{\bar{Y}}$.

But $(y_i-\bar{Y}) = (y_{i1}-\bar{\bar{Y}}) + (y_{i2}-\bar{\bar{Y}}) + \cdots + (y_{iM}-\bar{\bar{Y}})$

Square and sum over all N clusters

$$\sum_{i=1}^{N}(y_i-\bar{Y})^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{Y}})^2 + 2\sum_{i=1}^{N}\sum_{j<k}^{M}(y_{ij}-\bar{\bar{Y}})(y_{ik}-\bar{\bar{Y}})$$

$$\left[\because S^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{\bar{Y}})^2}{NM-1}\right]$$

$=(NM-1)s^2+(M-1)(NM-1)\rho s^2$

$= (NM-1))s^2[1+(M-1)\rho] \rightarrow (2)$

Substituting equ (2) in equ (1) we get

$$V(\bar{\bar{y}}) = \frac{(1-f)}{nM^2(N-1)}(NM-1)s^2[1+(M-1)\rho]$$

$$V(\bar{\bar{y}}) = \frac{(1-f)}{n}\frac{(NM-1)}{M^2(N-1)}s^2[1+(M-1)\rho]$$

NOTE:- If M =1, it gives the sampling variance of a SRSing of nM elements taken individually.

Both the procedures are equally good in this situations.

If M>1and ρ is positive cluster sampling with give a higher variance than the mean per element (i.e., SRS).

If ρ is negative cluster sampling is more precise.

QUESTION:- Describe the method of cluster sampling then define the mean of element (or) unit element per population and obtain variance $[\because (NM-1)s^2]$.

COROLLARY:- When N is large, prove that

$$V(\bar{\bar{y}}) = \frac{(1-f)}{nM}s^2[1+(M-1)]\rho$$

substitute $S^2=\frac{N}{N-1}\sigma^2$ we get ,$[\because N\sigma^2 = (N-1)S^2]$

$$= \frac{(N-n)}{(N-1)nM}\sigma^2[1+(M-1)\rho]$$

PROOF:- Let $V(\bar{\bar{y}}) = \frac{(1-f)}{n}\frac{(NM-1)}{M^2(N-1)}s^2[1+(M-1)\rho]$.

When N is large $\frac{1}{NM}$ *and* $\frac{1}{N}$ are considered negligible.

$$\therefore V(\bar{\bar{y}}) = \frac{(1-f)}{nM}s^2[1+(M-1)\rho]\frac{(NM-1)}{M(N-1)}$$

Dividing with NM both in Numerator and Denominator of R.H.S then we get

$$\therefore V(\bar{\bar{y}}) = \frac{(1-f)}{nM}s^2[1+(M-1)\rho]\frac{1-\frac{1}{NM}}{1-\frac{1}{N}}$$

$\Longrightarrow V(\bar{\bar{y}}) = \frac{(1-f)}{nM}s^2[1+(M-1)\rho].1 \qquad [\because \frac{1-0}{1-0}]$

From the definitions of $s^2$and σ^2

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

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

$\therefore V(\bar{\bar{y}}) = \frac{(1-f)}{nM}(\frac{N}{N-1})\sigma^2[1+(M-1)\rho] \qquad [\because f = \frac{n}{N}]$

$$V(\bar{\bar{y}}) = \frac{1-\frac{n}{N}}{nM}(\frac{N}{N-1})\sigma^2[1+(M-1)\rho]$$

$$\therefore V(\bar{\bar{y}}) = \frac{(N-n)}{nM(N-1)}\sigma^2[1+(M-1)\rho]$$

5.4 SUMMARY:

Summary and Conclusion:

- **Varying Probability Sampling** improves precision by giving higher selection chances to more important or larger units.
- **Cluster Sampling** reduces costs and increases feasibility when population elements are naturally grouped.
- While equal-sized clusters simplify the analysis, real-world applications often involve unequal clusters, requiring appropriate estimation techniques.
- Both methods are valuable in large-scale surveys and form the backbone of sampling strategies in fields like social sciences, economics, and public health.

5.5 KEY WORDS:

- Varying Probability Sampling
- PPS Sampling
- Horvitz-Thompson Estimator
- Cluster Sampling
- Equal Cluster Sizes
- Unequal Cluster Sizes.

5.6 SELF- ASSESSMENT QUESTIONS:

1. Define varying probability sampling and give an example.
2. What are the advantages of using PPS sampling?
3. Distinguish between equal and unequal cluster sampling.
4. In what situations is cluster sampling more suitable than simple random sampling?
5. How does the choice of cluster size affect the precision of estimates?
6. Explain cluster sampling technique with equal and unequal cluster sizes.
7. Describe the sampling technique with varying probabilities without replacement.

5.7 SUGGESTED READINGS:

1. Cochran, W.G. (1977) – *Sampling Techniques*, 3rd Edition, Wiley Eastern. Hansen, Hurwitz, and Madow (1953) – *Sample Survey Methods and Theory*, Wiley.
2. Des Raj and Chandhok, P. (1998) – *Sample Survey Theory*, Narosa Publishing House. Sarndal, C.E., Swensson, B., and Wretman, J. (1992) – *Model Assisted Survey Sampling*, Springer.
3. Singh, D. and Chaudhary, F.S. (1986) – *Theory and Analysis of Sample Survey Designs*, New Age International.
4. Murthy, M.N. (1967) – *Sampling Theory and Methods*, Statistical Publishing Society.

LESSON- 6

OPTIMUM CLUSTER SIZE

OBJECTIVES:

Upon completion of this unit, learners will be able to:

1. **Understand the Concept of Optimum Cluster Size**
   - Define and determine the optimum cluster size under a fixed survey cost constraint.
   - Analyze the trade-offs between cost and precision in cluster sampling.
   - Apply mathematical expressions to identify the optimal number of elements per cluster for cost-efficient data collection.
2. **Comprehend PPS (Probability Proportional to Size) Sampling Techniques**
   - Understand the principles of PPS sampling with and without replacement.
   - Differentiate between sampling with and without replacement in terms of procedure, estimator properties, and efficiency.
3. **Learn the Procedures for Selecting a Sample in PPS Sampling**
   - Describe and perform PPS sample selection using various methods such as cumulative total method and systematic PPS.
   - Implement appropriate procedures for both with-replacement and without-replacement designs.
4. **Estimation and Variance in PPS Sampling**
   - Derive and apply unbiased estimators for the population total under PPS sampling with replacement.
   - Compute the sampling variance of the estimator under PPS with replacement.
   - Compare efficiency and reliability of estimators under PPSWR and PPSWOR schemes.
5. **Apply Theoretical Knowledge to Practical Survey Design**
   - Integrate concepts of optimum cluster size and PPS sampling into the design of practical, cost-effective survey plans.
   - Evaluate the advantages and limitations of PPS methods in real-world sampling situations.

STRUCTURE:

6.1 INTRODUCTION:

The optimum number of clusters can be defined as follow: Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters. For each k, calculate the total within-cluster sum of square (wss). Plot the curve of wss according to the number of clusters k.

These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

In addition to *elbow*, *silhouette* and *gap statistic* methods, there are more than thirty other indices and methods that have been published for identifying the optimal number of clusters. We'll provide R codes for computing all these 30 indices in order to decide the best number of clusters using the "majority rule".

6.2 OPTIMUM CLUSTER SIZE FOR FIXED COST:

In cluster sampling, the sampling variance increases with cluster size and decreases with increasing no. of sample clusters, while the cost decreases with increasing cluster size and increases with no. of sample clusters. Hence ,in practice it is necessary to strike a balance between these two opposing points of view by finding the optimum values for the cluster size (M) and the no. of sample clusters(n), which would minimize the sampling variance for a fixed cost (or) alternatively minimize the cost for a specified sampling variance.

(1) $V \uparrow M \uparrow \qquad V \downarrow n \uparrow$

(2) $C \downarrow M \uparrow \qquad C \uparrow n \uparrow$

QUESTION:- In cluster sampling of equal cluster sizes, derive the expressions for the optimum values of M and n (cluster size and no. of clusters) for a fixed cost function:-

ANS :- ANOVA for the whole population (on element basis)

Source of variation	Degrees of freedom	Mean square
Between clusters	(N-1)	S_b^2
Between elements within clusters	N(M-1)	S_w^2
Between elements in the population	NM-1	$S^2 = \frac{(N-1)S_b^2 + N(M-1)S_w^2}{NM-1}$

In several agricultural surveys, S_w^2 appeared to be related to M by the empirical formula $S_w^2 = AM^g (g > 0)$, where "A" and "g" are constants that do not depend on M. In this formula S_w^2 increases steadily as M increases. Usually "g" is small.

From the ANOVA table, we find

$$s_b^2 = \frac{(NM-1)S^2 - N(M-1)S_w^2}{(N-1)}$$

$$= \frac{(NM-1)S^2 - N(M-1)AM^g}{(N-1)} [\because S_w^2 = AM^g]$$

Dividing 'N' on both Numerator and Denominator

$$s_b^2 = \frac{\left(M - \frac{1}{N}\right)S^2 - (M-1)AM^g}{1 - \frac{1}{N}}$$

When N is large, ignore $\frac{1}{N}$,

$$\therefore s_b^2 = MS^2 - (M-1)AM^g \longrightarrow (1)$$

In an extensive survey the nature of the field costs plays a large part in determining the optimum cluster. Two components of field costs are distinguished. The component c_1Mn comprises costs that vary directly with the total no. of elements (farms). Thus c_1 contains the cost of interview and the cost of travel from farm to farm with in cluster. The 2^{nd} component, $c_2\sqrt{n}$, measures the cost of travel between clusters. The total field cost is therefore

$$C = c_1Mn + c_2\sqrt{n} \longrightarrow (2) \qquad \boxed{\frac{\circ\longrightarrow\circ\longrightarrow\circ\ldots\circ}{c_1mn+c_2\sqrt{n}=C}}$$

Assuming SRSing and ignoring the fpc, the variance of the mean per element $\bar{\bar{y}}$ is $V(\bar{\bar{y}}) = \frac{S_b^2}{nM}$ [∵From Theorem-9.2 &S_b^2]

From equation (1) ,this equals

$$V(\bar{\bar{y}}) = \frac{MS^2-(M-1)AM^g}{nM}$$

Dividing both Nr and Dr with M

$$= \frac{\frac{1}{M}[MS^2 - (M-1)AM^g]}{\frac{nM}{M}}$$

$V = V(\bar{y}) = \frac{S^2-(M-1)AM^{g-1}}{n} \longrightarrow (3)$ (dv/dn= -k/(n^2) = -(k/n) 1/n= -v/n)

To determine the optimum size of cluster, we find M, and incidentally n, to minimize V for fixed C

This gives,

$$\frac{2c_1M\sqrt{n}}{c_2} = (1 + \frac{4Cc_1M}{c_2^2})^{\frac{1}{2}} - 1 \longrightarrow (4)$$

From equation (2)

$$ax^2 + bx + c = 0; \qquad c_1Mn + c_2\sqrt{n} - C = 0$$

$$a = c_1M, \qquad b = c_2, \qquad c = -C \therefore x = \frac{-b \pm \sqrt{b^2-4ac}}{2a} \quad ,$$

$$\therefore \sqrt{n} = \frac{-c_2 \pm \sqrt{c_2^2+4c_1MC}}{2c_1M}$$

$$2c_1M\sqrt{n} = -c_2 \pm (c_2^2 + 4c_1MC)^{\frac{1}{2}}$$

We take some positive quantity

$$= -c_2 + c_2\left[1 + \frac{4c_1MC}{c_2^2}\right]^{\frac{1}{2}} = c_2\left[-1 + (1 + \frac{4c_1MC}{c_2^2})^{\frac{1}{2}}\right]$$

$$\frac{2c_1M\sqrt{n}}{c_2} = (1 + \frac{4c_1MC}{c_2^2})^{\frac{1}{2}} - 1 \longrightarrow (4)$$

The equation to be minimized is

$$C+\lambda V = c_1Mn + c_2\sqrt{n} + \lambda V \longrightarrow (5) \quad (\because \text{Adding } `\lambda V' \text{ on both sides})$$

Differentiating and noting that $\frac{\partial v}{\partial n} = \frac{-V}{n}$.

We obtain the equation

for n : $c_1M + \frac{1}{2}c_2n^{\frac{-1}{2}} = -\lambda\frac{\partial V}{\partial n} = \lambda\frac{V}{n} \longrightarrow (6)$

and for M: $c_1n = -\lambda\frac{\partial V}{\partial M} \longrightarrow (7)$

Dividing (7) by (6) to eliminate λ . This gives

$$\lambda\frac{\partial V}{\partial M}.\frac{\cancel{n}}{\cancel{\lambda}\,\mathrm{v}} = \frac{-c_1\cancel{n}}{c_1M + \frac{1}{2}c_2n^{\frac{-1}{2}}}$$

Multiply with M on both sides

$$\frac{M\partial V}{V\partial M} = \frac{-c_1M}{c_1M + \frac{1}{2}c_2n^{\frac{-1}{2}}} = \frac{-1}{1 + \frac{c_2}{2c_1M\sqrt{n}}} \longrightarrow (8)$$

If we substitute for $\sqrt{n}$ from eq (4) we obtain after some manipulation

$$\frac{M\partial V}{V\partial M} = (1 + \frac{4Cc_1M}{c_2^2})^{-\frac{1}{2}} - 1 \longrightarrow (9)$$

By writing out the left side of equ (9) in full and changing signs on both sides we find,

$$V=\frac{S^2-(M-1)AM^{g-1}}{n} = \frac{S^2-AM^g+AM^{g-1}}{n}$$

$$\frac{\partial V}{\partial M} = \frac{1}{n}[-AgM^{g-1} + A(g-1)M^{g-2}]$$

$$\frac{M}{V}\frac{\partial V}{\partial M} = \frac{\frac{M}{n}[-AgM^{g-1} + A(g-1)M^{g-2}]}{\frac{S^2-(M-1)AM^{g-1}}{n}}$$

$$-\frac{M}{V}\frac{\partial V}{\partial M} = \frac{AM^{g-1}[gM-(g-1])}{S^2-(M-1)AM^{g-1}}$$

$$\frac{AM^{g-1}[gM-(g-1)]}{S^2-(M-1)AM^{g-1}} = 1 - \left(1 + \frac{4Cc_1M}{c_2^2}\right)^{-1/2} \rightarrow (10)$$

This equ (10) gives optimum 'M' by iterative method (it is difficult to get on explicit expression for M).

Equation (4) can be written as

$$\frac{2c_1M\sqrt{n}}{c_2} = \frac{(c_2^2 + 4c_1CM)^{\frac{1}{2}}}{c_2} - 1$$

$$\frac{2c_1M\sqrt{n}}{\cancel{c_2}} = \frac{(c_2^2 + 4c_1CM)^{\frac{1}{2}} - c_2}{\cancel{c_2}}$$

$$2c_1M\sqrt{n} = (c_2^2 + 4c_1CM)^{\frac{1}{2}} - c_2$$

$$\left(2c_1M\sqrt{n}\right)^2 = \left((c_2^2 + 4c_1CM)^{\frac{1}{2}} - c_2\right)^2$$

$$n = \left[\frac{(c_2^2 + 4c_1CM)^{\frac{1}{2}} - c_2}{2c_1M}\right]^2 \rightarrow (11)$$

On substituting the value of "M" obtained from (10) in (11) , we can obtain the optimum value of 'n'.

UNEQUAL CLUSTSER SAMPLING:-

Suppose there are N clusters in the population. Let the i^{th} cluster consists of M_i elements $(i = 1,2, \dots, N)$ and $\sum_{i=1}^{N} M_i = M_0$

The population mean for element $\bar{\bar{Y}}$ is defined by

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{Mi} y_{ij}}{\sum_{i=1}^{N} M_i} = \frac{\sum_{i=1}^{N} M_i\, \bar{y}_{i.}}{M_0}$$

Where $\bar{y}_{i.} = \frac{\sum_{j=1}^{Mi} y_{ij}}{M_i}$ is the mean for elements of the i^{th} cluster.

The sample mean for element is given by

$\bar{y}_n' = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i} \rightarrow (1)$ where $\bar{y}_n'$ is not unbiased.

THEOREM- 2.2.2 :-

The estimator of the mean $\bar{\bar{Y}}$ is given by equation (1) which is a weighted mean of the cluster means and a ratio of two random variables is not unbiased. Its sampling variance is given by

$V(\bar{y}_n') = \frac{(1-f)}{n} s_b'^2$.

Where $s_b'^2 = \frac{\sum_{i=1}^{M} M_i^2 (\bar{y}_{i.} - \bar{\bar{Y}})^2}{\bar{M}^2 (N-1)}$

and $\bar{M} = \frac{M_0}{N} = \frac{\sum_{i=1}^{N} M_i}{N}$.

PROOF :- From Theorem 6.1 of ratio estimation, we know that

$V(\hat{R}) = \frac{(1-f)}{n\bar{X}^2} \frac{[\sum_{i=1}^{N} (y_i - Rx_i)^2]}{(N-1)}$.

Where $\hat{R} = \frac{\bar{y}}{\bar{x}}$, $R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} = \frac{\bar{Y}}{\bar{X}}$

The estimator $\bar{y}_n'$ is given by replace in x_i by M_i and y_i by $M_i \bar{y}_{i.}$ in ratio estimator mean $\hat{R}$.

We have seen that the ratio estimate is not unbiased.

Substituting $x_i = M_i, y_i = M_i \bar{y}_{i.}$

Then $\hat{R} = \frac{\sum_{i=1}^{n} M_i \bar{y}_{i.}}{\sum_{i=1}^{n} M_i} = \bar{y}_n'$ (from definition)

$$R = \frac{\sum_{i=1}^{N} M_i \bar{y}_{i.}}{\sum_{i=1}^{N} M_i} = \bar{\bar{Y}} \quad (by \text{ definition})$$

$\bar{X}^2 = (\frac{\sum_{i=1}^{N} x_i}{N})^2 = (\frac{\sum_{i=1}^{N} M_i}{N})^2 = \bar{M}^2$ (by given statement)

$$V(\bar{y}_n') = \frac{(1-f)}{n\bar{M}^2(N-1)} \left[\sum_{i=1}^{N} (M_i \bar{y}_{i.} - \bar{\bar{Y}} M_i)^2\right]$$

$$= \frac{(1-f)}{n\bar{M}^2(N-1)} \sum_{i=1}^{N} M_i^2 (\bar{y}_{i.} - \bar{\bar{Y}})^2$$

$$V(\bar{y}_n') = \frac{(1-f)}{n} (S_b')^2 \quad \text{(given in the stratum)}$$

$\bar{M}$= Average cluster size in the population

COROLLARY:- An unbiased estimator of $V(\bar{y}_n')$is given by

$v(\bar{y}_n') = \frac{(1-f)}{n}(s_b')^2$, Where $(s_b')^2 = \frac{\sum_{i=1}^{n} M_i^2(\bar{y}_{i.}-\bar{y}_n')^2}{(\bar{M}')^2(n-1)}$And $\bar{M}' = \frac{\sum_{i=1}^{n} M_i}{n}$

$\overline{M}' \rightarrow$ Average cluster size in the sample.

DIFFERENCE OF STRATA AND CLUSTSER:-

Strata and cluster are both non-overlapping sub-sets of the population

(1) All strata are represented in the sample; but only a subset of clusters are in the sample.

(2) Cluster sampling gives less precision than either SRS or STRS

Ex:- sometimes the cost per sample points less for cluster sampling than for other sampling methods . Given a fixed budget, the researcher may be able to use a bigger sample with cluster sampling than with the other methods, when the increased sample size is sufficient to offset the loss in precession, cluster sampling may be the best choice.

DISADVANTAGES:-

It gives higher sampling error, which can be expressed in the so-called "design effect".

Definition:- The ratio between the no. of subjects in the cluster study and the no. of subjects in an equally reliable, randomly sampled un clustered study.

6.3 PROBABILITY PROPORTIONAL SIZE SAMPLING (OR) PPS SAMPLING:

We are drawing a sample from population probability is proportional to the size of the sample. PPS-probability proportional size sampling-size is important.

In SRSing the selection probabilities were equal for all units of the population. When ever the units vary in size SRSing is not an appropriate procedure as no importance is given to the size of the unit such ancillary information about the size of the units can be utilized in selecting the sample so as to get more efficient estimators of the population parameter's. one such method is to assign unequal probability of selection to different units in the population depending on their sizes with orchards (garden of fruit trees) having varying no's of fruit trees, it may be desirable to provide Sampling scheme in which orchards are selected with probability proportional to the number of trees in the orchards. When units vary in their sizes and the variate under study is highly correlated with the size of the unit, the probability of selection may be assigned in proportion to the size of the unit, (the probability). This type of sampling procedure where the probability of selection is proportional to the size of the units known as probability proportional size sampling abbreviated as PPS sampling.

QUESTION:- What is PPS Sampling and describe PPS Sampling?

There is a basic difference between SRSing and PPS Sampling procedures. In SRSing the Probability of drawing any specified unit at any given draw is the same, while in PPS Sampling it differs, from draw to draw. The theory of PPS Sampling is consequently more complex than that of SRSing.

6.4 PPS SAMPING WITH REPLACEMENT :-

There are two methods of selection:

(1) Cumulative total method

(2) Lahiri's method

1. CUMULATIVE TOTAL METHOD:-

Let the size of the i^{th} unit be $X_i(i = 1,2, \dots, N)$ the total being $= \sum_{i=1}^{N} X_i$. We associate the no's 1toX_1 with the first unit, the no's $(X_1 + 1)$ to $(X_1 + X_2)$ with second unit and so on. A number 'k' is choose at random from 1toX and the unit which this number is associated is selected clearly the i^{th} unit in the population is being selected with a probability proportional to X_i. If a sample of size, 'n' is required the procedure is repeated n-times with replacement of the units selected. This procedure of selection is known as cumulative total method for the method needs cumulation of the unit sizes.

The main difficulty in this procedure is the compulsion to complete successive cumulative totals, which becomes time consuming and costly when the population size, 'N' is large.

Example:- A village has 10-holdings consisting of 50,30,45,25,40,26,44,35,28 and 27 fields respectively select a sample of 4-holdings with the replacement method and which Probability proportional to the no. of fields in the holding.

The first step in the selection of holdings is to cumulative totally as shown below:

S. No. HOLDINGS	SIZE(X_i)	CUMMULATIVE SIZE	NUMBERS ASSOCIATED
1	50	50	1-50
2	30	80	51-80
3	45	125	81-125
4	25	150	126-150
5	40	190	151-190
6	26	216	191-216
7	44	260	217-260
8	35	295	261-295
9	28	323	296-323
10	27	350	324-350

272-8, 326-10, 165-5, 094-3 (these 4 no's are selected only by randomly and using with replacement)

To select a holding, a random number not exceeding 350 is drawn with the help of a random number table. Suppose the random number thus selected in 272. It can be seen from the cumulative totals that the number is associated with the group 261-295 i.e., the 8^{th} holding is selected corresponding to the random no.272.

Similarly, we select 3 more random numbers. Suppose these no's are 326,165,and 094 then the holdings selected corresponding to these random no's are 326,165, and 094 then the holdings selected corresponding to these random no's of the$10^{th}, 5^{th}, and\ 3^{rd}$ respectively.

Hence a sample of 4 holding selected with Probability Proportional to size will contain the $8^{th}, 10^{th}, 5^{th}, and\ 3^{rd}$ holdings.

QUESTION:- Explain cumulative method with give illustrative example?

2. LAHIRI'S METHOD [FOR PPS SAMPLING WITH REPLACEMENT]:

Lahiri in 1951 suggested an alternative procedure in which cumulations are avoided completely. It consists in selecting a number at random between 1and N and noting down the unit with the corresponding serial number, provisionally. Another random no. is then chosen b/w I and M, where M is the maximum size of the N units of the population.

If the second random no is smaller than the size of the unit provisionally selected, the units selected into the sample. If not, the entire procedure is repeated until a unit is finally selected for selecting a sample of n-units, the procedure is to be repeated until n-units are selected.

Example:

S. No of holdings	1	2	3	4	5	6	7	8	9	10
size(X_i)	50	30	45	25	40	26	44	35	28	27

In this case N=10,M=50. First we have to select a random number which is not greater than 10 and a 2nd random number which is not greater than 50. Referring to the random number table, the pairs is (10,13). Hence the 10th unit is selected in the sample. Similarly choosing other pairs we can have (4,26),(5,35),(7,26). The pair (4, 26) is rejected as 26 is greater than the size value (25) and so another pair is drawn which turns out to be (8, 16) hence the sample will consists of the holdings with serial no's 10, 5, 7 and 8.

6.6 ESTIMATION IN PPS SAMPLING WITH REPLACEMENT:

1	2	.	.	.	N
u_1	u_2	.	.	.	u_N
y_1	y_2	.	.	.	y_N

Consider a population of N-units and let y_i be the value of the characteristic under study for the unit u_i of the population($i = 1,2, \dots, N$). Suppose further that $p_i = \frac{X_i}{X}$ (=size of the unit) be the Probability that the unit u_i is selected in a sample of, '1' such that $\sum_{i=1}^{N} p_i = 1$. Let n independent selections be made with the replacement method and the value of y_i for each selected unit be observed further, let (y_i, p_i) be the value and Probability of selection of the ith unit of the sample. It can be seen that the random variables $\frac{y_i}{p_i}(i = 1,2, \dots, n)$ are independently and identically distributed. If $p_i = \frac{1}{N}$, it gives rise to a SRS. This shows that SRSing is a particular case of PPS Sampling.

Theorem-2.2.3: In PPS sampling with replacement, an unbiased estimator of the population total Y is given by $\hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)$ with its Sampling variance $V(\hat{Y}_{PPS}) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2$

Proof: Let us define random variates $z_i = \left(\frac{y_i}{p_i}\right)$ and $(i = 1,2,\ldots,n)$

Which are independently and identically distributed

Hence $E(z_i) = \sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i}\right) = Y$.

Now, let us consider $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ ($\because$ Simple average of z_i)

Since $E(\hat{Y}_{PPS}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right) = E(\bar{z})$[$\because$ By statement $\hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)$]

$$\therefore E(\hat{Y}_{PPS}) = E(\bar{z}) = \sum_{i=1}^{n}\frac{1}{n}E(Z_i) = Y$$

$\therefore \hat{Y}_{PPS}$ is an unbiased estimator of Y.

$$V(\hat{Y}_{PPS}) = V(\bar{z}) = V\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right)$$

$= \frac{1}{n^2}\sum_{i=1}^{n} V(z_i) = \frac{1}{n^2} n \sum_{i=1}^{N} p_i (z_i - Y)^2$, since $E(z_i) = Y$

$$V(\hat{Y}_{PPS}) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2$$

Corollary: An unbiased estimator of the population mean $\bar{Y}$ is given by $\hat{\bar{Y}}_{PPS} = \frac{1}{nN}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)$, with its Sampling variance is $V\left(\hat{\bar{Y}}_{PPS}\right) = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{Np_i} - \bar{Y}\right)^2$.

Proof: Let $V\left(\hat{\bar{Y}}_{PPs}\right) = \frac{1}{N^2}V(\hat{Y}_{PPS})$

$= \frac{1}{N^2}\frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - Y\right)^2 = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{Np_i} - \frac{Y}{N}\right)^2 = \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{Np_i} - \bar{Y}\right)^2$.

Theorem-2.2.4: In PPS Sampling with replacement an unbiased estimator of $V(\hat{Y}_{PPS})$ is given by

$v(\hat{Y}_{PPS}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2$,for n>1

$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 + \sum_{i=1}^{n}(\hat{Y}_{PPS})^2 - 2\sum_{=1}^{n}\frac{y_i}{p_i}\hat{Y}_{PPS}\right]$

$= \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 + n\hat{Y}_{PPS}^2 - 2n\hat{Y}_{PPS}^2\right]$ $\left[\because \hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}\right]$

$v(\hat{Y}_{PPS}) = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\hat{Y}_{PPS}^2\right] \longrightarrow (1)$

Proof: By the usual algebraic identity

$$\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2 = \sum_{i=1}^{n}\left[\left(\frac{y_i}{p_i} - Y\right) - (\hat{Y}_{PPS} - Y)\right]^2$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 + \sum_{i=1}^{n}(\hat{Y}_{PPS} - Y)^2 - 2\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)(\hat{Y}_{PPS} - Y)$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 + n\left(\hat{Y}_{PPS} - Y\right)^2 - 2\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)\left(\hat{Y}_{PPS} - Y\right)$$

Since $\hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right) \Rightarrow n\hat{Y}_{PPS} = \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)$

Then $\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right) = n\left(\hat{Y}_{PPS} - Y\right)\left[\because \sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right) - nY = n\hat{Y}_{PPS} - nY = n\left(\hat{Y}_{PPS} - Y\right)\right]$

$$\therefore \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right) = \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 + n\left(\hat{Y}_{PPS} - Y\right)^2 - 2n\left(\hat{Y}_{PPS} - Y\right)^2$$

$$\therefore \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2 = \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 - n\left(\hat{Y}_{PPS} - Y\right)^2 \rightarrow (2)$$

From equation (1)

$$n(n-1)v\left(\hat{Y}_{PPS}\right) = \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2$$

From (2), we have

$$n(n-1)v\left(\hat{Y}_{PPS}\right) = \sum_{i=1}^{n}\left(\frac{y_i}{p_i} - Y\right)^2 - n\left(\hat{Y}_{PPS} - Y\right)^2 \rightarrow (3)$$

Taking expectation on both sides of equation (3)

$$n(n-1)E\left[v\left(\hat{Y}_{PPS}\right)\right] = E\left[\sum_{i=1}^{N} t_i\left(\frac{y_i}{p_i} - Y\right)^2 - n\left(\hat{Y}_{PPS} - Y\right)^2\right]$$

$$n(n-1)E\left[v\left(\hat{Y}_{PPS}\right)\right] = E\left[\sum_{i=1}^{N} t_i\left(\frac{y_i}{p_i} - Y\right)^2 - nV\left(\hat{Y}_{PPS}\right)\right]$$

$$\left[\because E\left(\hat{Y}_{PPS} - Y\right)^2 = V\left(\hat{Y}_{PPS}\right)\right]$$

Since $E\left[\hat{Y}_{PPS} - Y\right]^2 = V\left(\hat{Y}_{PPS}\right)$ and $E(t_i) = np_i \sim B(0,1) = np$

$$n(n-1)E\left[v\left(\hat{Y}_{PPS}\right)\right] = n\sum_{i=1}^{N} p_i\left(\frac{y_i}{p_i} - Y\right)^2 - nV\left(\hat{Y}_{PPS}\right)$$

$n(n-1)E\left[v\left(\hat{Y}_{PPS}\right)\right] = n.nV\left(\hat{Y}_{PPS}\right) - nV\left(\hat{Y}_{PPS}\right)$ [from theorem-5.3.1]

$$n(n-1)E\left[v\left(\hat{Y}_{PPS}\right)\right] = n^2V\left(\hat{Y}_{PPS}\right) - nV\left(\hat{Y}_{PPS}\right)$$

$$\cancel{n(n-1)}E\left[v\left(\hat{Y}_{PPS}\right)\right] = \cancel{n(n-1)}V\left(\hat{Y}_{PPS}\right)$$

$\therefore v\left(\hat{Y}_{PPS}\right)$ is an unbiased estimate of $V\left(\hat{Y}_{PPS}\right)$

$$\therefore E\left[v\left(\hat{Y}_{PPS}\right)\right] = V\left(\hat{Y}_{PPS}\right)$$

$$\therefore v\left(\hat{Y}_{PPS}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2$$

Corollary: An unbiased estimator $of\ v\left(\hat{\bar{Y}}_{PPS}\right)$ is given by

$$v\left(\hat{\bar{Y}}_{PPS}\right) = \frac{1}{N^2}v\left(\hat{Y}_{PPS}\right)$$

$$= \frac{1}{N^2}\cdot\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{PPS}\right)^2$$

$$= \frac{1}{N^2}\frac{1}{n(n-1)}\left[\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\hat{Y}_{PPS}^2\right]$$

$$v\left(\hat{\bar{Y}}_{PPS}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\frac{y_i}{Np_i} - \hat{\bar{Y}}_{PPS}\right]^2 \; [\because From\ Th - 5.3.2]$$

PPS SAMPLING WITHOUT REPLACEMENT:

In general of the sampling scheme is to select a PPS sample of size unity and remove the selected unit from the population. From the remaining units, another PPS sample of size one is taken as before and the selected unit removed from the population. This process is repeated until, 'n' selections are made.

Suppose n, units are selected one by one, with Probability proportional to size measure x, at each draw, without replacing the units selected in the previous draws. The Probability of selection at the first draw for the j^{th} unit is given by

$p_j = {}^{X_j}/_{X}, j = 1,2,\dots,N,$ where X=$\sum_{j=1}^{N} X_j$

Similarly, the Probability that the ith unit is selected at 2nd draw is given by $\frac{p_i}{j} = \frac{p_i}{(1-p_j)}, i \neq j$, and so on.

This set up of Sampling comprises an ordered set of Sample values $(y_1, y_2, \dots, y_n)$ with Probabilities $(p_1, p_2, \dots, p_n)$

Examples: In a village, there are 8 orchards (garden tree fruits) with 50,30,25,40,26,44,20 and 35 trees, respectively . Select a sample of 2 orchards with Probability proportional to the number of trees in the orchard and WOR.

Sol: By using Lahiri's method of selection

S.No of house holdings	Orchard number	1	2	3	4	5	6	7	8
Size(X_i)	Number of trees	50	30	25	40	26	44	20	35

i. For selecting a pair of random no's$(i,j), (i \leq 8, j \leq 50)$, using the random table. The random pair is (5,17). Here (5,17)- 5th orchard is selected in the sample

ii. For selecting this orchard after deleting of 5th orchard.

S. No of house holdings	Orchard number	1	2	3	4	5	6	7
Size(X_i)	Number of trees	50	30	25	40	26	46	35

As a pair of random no's$(i,j)/(i \leq 7, j \leq 50)$ then using random no's the selected pair is (6,18). Here (6,18)-6th orchard is selected in the sample. Thus, the sample selected consists of the units at serial no's 5 and 7 of the original list with the no. of trees being 26&20, respectively.

In WOR PPS has two methods:

i. Narains scheme of sample selection.

ii. Sen-Midzuno method.

i. **Narain's scheme of sample selection**: This scheme was introduced by Narain's (1951). The scheme consists of constructing revised probabilities of selection $p_i'(i = 1,2, \dots, N)$ such that the inclusion probabilities π_i are proportional to the original probabilities of selection $p_i(i = 1,2, \dots, N)$ and sampling is done without replacement.

Here the inclusion probability's π_i are given by $\pi_i = np_i, i = 1,2, \dots, N$.

The probability selection of second draw is p_i' is the similar to 1st draw. Let us consider the simple case N=4 and n=2. The problem is to evaluate π_{ij} given $p_i(i,j = 1,2,3,4)$. The relationship $\sum_{j \neq i} \pi_{ij} = \pi_i$ provides a system of four linear equations with size unknowns. For this problem choose the values of two arbitrary parameters that restriction to positive values. The computations become tedious for n greater than 2.

ii. **Sen –Midzuno Method:** This method was suggested by midzuo (1952) and independently by Sen (1952). It consists in selecting the first unit with PPS and the remaining (n-1) units from (N-1) units of the population by SRSING WOR.

In this procedure, the inclusion probability's for individual and pairwise units are given by

$$\pi_i = p_i + (1 - p_i)\left(\frac{n-1}{N-1}\right) \text{ for } i = 1,2, \dots, N.$$

$$= [(N-1)p_i + n - 1 - np_i + p_i]\frac{1}{(N-1)}$$

$$= [Np_i - \cancel{p_i} + n - 1 - np_i + \cancel{p_i}]\frac{1}{(N-1)}$$

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

And $\pi_{ij} = p_i\left(\frac{n-1}{N-1}\right) + p_j\left(\frac{n-1}{N-1}\right) + \left(1 - p_i - p_j\right)\frac{(n-1)(n-2)}{(N-1)(N-2)}, for\ i \neq j = 1,2, \dots, N$

$$= \frac{(n-1)}{(N-2)}\left[p_i\frac{(N-2)}{(N-1)} + \frac{p_j(N-2)}{(N-1)} + \left(1 - p_i - p_j\right)\frac{(n-2)}{(N-1)}\right]$$

$$= \frac{(n-1)}{(N-2)}\left[\frac{(N-2)}{(N-1)}\left(p_i + p_j\right) + \left(1 - p_i - p_j\right)\frac{(n-2)}{(N-1)}\right]$$

$$= \frac{(n-1)}{(N-2)}\left[\frac{Np_i + Np_j - \cancel{2p_i} - \cancel{2p_j} + n - np_i - np_j - 2 + \cancel{2p_i} + \cancel{2p_j}}{(N-1)}\right]$$

$$= \frac{(n-1)}{(N-2)}\left[\frac{(N-n)p_i + (N-n)p_j + (n-2)}{(N-1)}\right]$$

$$= \frac{(n-1)}{(N-2)}\left[\frac{(N-n)}{(N-1)}\left(p_i + p_j\right) + \frac{(n-2)}{(N-1)}\right]$$

By extension of the above argument, we can have $y_i, y_j, \dots, y_q$, a sample of n-units. The Probability of including these n-units in the sample is given by

$\pi_{ij \dots q} = \frac{1}{\binom{N-1}{n-1}}\left(p_i + p_j + \cdots + p_q\right)$.

6.6 SUMMARY AND CONCLUSION:

- Reviewed key concepts of optimum cluster size and PPS sampling
- Highlighted procedures for selecting samples in PPS
- Compared with and without replacement methods
- Discussed estimators and their variances
- Emphasized practical application in surveys with cost constraints and unequal unit sizes.

6.7 KEY WORDS:

- **Cluster Sampling** – A sampling technique where the population is divided into groups (clusters), and a sample of clusters is selected to represent the population.
- **Optimum Cluster Size** – The ideal number of elements in a cluster that minimizes the total survey cost for a fixed budget while maintaining acceptable precision.
- **Fixed Cost Sampling** – Sampling design that considers a budget constraint where the total cost of sampling cannot exceed a given limit.
- **Intra-cluster Correlation** – The similarity among units within a cluster, which affects the efficiency of cluster sampling.
- **Probability Proportional to Size (PPS)** – A sampling method where the selection probability of each unit is proportional to a known size measure (e.g., population, area, revenue).
- **PPS With Replacement (PPSWR)** – A method in which units can be selected multiple times during sampling, allowing for simple estimation formulas.
- **PPS Without Replacement (PPSWOR)** – A sampling approach in which each unit is selected only once, requiring more complex estimation techniques.
- **Cumulative Total Method** – A technique used in PPS sampling where cumulative sizes are used to facilitate the selection of units.
- **Lahiri's Method** – A method for PPS sampling that involves random selection of a unit and acceptance based on a probability condition.
- **Systematic PPS Sampling** – A method where a random start is chosen, followed by systematic steps using size measures to select units.
- **Horvitz-Thompson Estimator** – An unbiased estimator used in unequal probability sampling, especially in PPSWR.
- **Estimator of Population Total** – A statistical estimate of the total value of a variable in the population based on the sample data.
- **Sampling Variance** – The variability in an estimator due to the randomness of sample selection, particularly important in PPS designs.
- **Efficiency** – The degree to which a sampling design reduces variance for a given cost or sample size.

6.8 SELF-ASSESSMENT QUESTIONS:

1. What are the advantages of using cluster sampling in survey research?
2. Explain the concept of unequal probability sampling.
3. What factors influence the determination of an optimum cluster size under a fixed cost?
4. Derive the expression for optimum cluster size given a total cost constraint.
5. What is the basic principle behind PPS sampling?
6. Give examples of size measures that can be used in PPS sampling.
7. Distinguish between PPS sampling with replacement (PPSWR) and without replacement (PPSWOR).
8. What are the advantages and disadvantages of PPSWR compared to PPSWOR?
9. Explain the cumulative total method for selecting a PPS sample.
10. Describe Lahiri's method. What are its limitations?

6.9 SUGGESTED READINGS:

1. Cochran, W.G. (1977) – *Sampling Techniques*, 3rd Edition, Wiley Eastern. Hansen, Hurwitz, and Madow (1953) – *Sample Survey Methods and Theory*, Wiley.
2. Des Raj and Chandhok, P. (1998) – *Sample Survey Theory*, Narosa Publishing House. Sarndal, C.E., Swensson, B., and Wretman, J. (1992) – *Model Assisted Survey Sampling*, Springer.
3. Singh, D. and Chaudhary, F.S. (1986) – *Theory and Analysis of Sample Survey Designs*, New Age International.
4. Murthy, M.N. (1967) – *Sampling Theory and Methods*, Statistical Publishing Society.

Dr. B.Guravaiah

LESSON- 7

DES RAJ, MURTHY'S ESTIMATOR

OBJECTIVES:

After completing this unit, the learner will be able to:

- Understand the concept of PPS sampling and its importance in survey sampling when the units vary in size.
- Explain the need for unequal probability sampling and distinguish between sampling with and without replacement.
- Describe and apply Des Raj's estimator for estimating the population total or mean using PPS sampling with a sample size of two.
- Explain Murthy's estimator as a symmetric alternative to Des Raj's estimator and understand its derivation.
- Compare the efficiency and bias of Des Raj and Murthy's estimators under different sampling scenarios.
- Perform calculations and estimations using these estimators through worked examples.
- Evaluate the applicability of Des Raj and Murthy's methods in practical survey designs.

STRUCTURE:

7.1 INTRODUCTION:

Des Raj and Murthy's estimators are two well-known unequal probability sampling estimators used for estimating the population total or mean when the sample size is two and units are selected with probability proportional to size (PPS).

These estimators were developed as part of efforts to improve estimation efficiency in survey sampling, especially when:

- The population units have known and unequal sizes (e.g., villages with different populations).
- Simple random sampling may not be efficient due to heterogeneity in sizes.

7.2 ORDERED ESTIMATES:

To overcome the difficulty of changing expectations with each draw, associate a new variate with each draw such that its expectation is equal to the population value of the variate under study. Such estimators take into account the order of the draw. They are called the ordered estimates. The order of the values obtained at the previous draw will affect the unbiasedness of the population mean.

7.3 UNORDERED ESTIMATES:

Corresponding to any ordered estimator, there exist an unordered estimator which does not depend on the order in which the units are drawn and has smaller variance than the ordered estimator. Unordered Estimator: In case of sampling WOR from a population of size N, there are unordered sample(s) of size n.

7.4 DES RAJ'S & MURTHY'S ESTIMATOR (SAMPLE SIZE TWO):

Des Raj ordered estimator

Case 1: Case of two draws:

Let y_1 and y_2 denote the values of units $U_{i(1)}$ and $U_{i(2)}$ drawn at the first and second draws respectively. Note that anyone out of the N units can be the first unit or second unit, so we use the notations $U_{i(1)}$ and $U_{i(2)}$ instead of U_1 and U_2. Also note that y_1 and y_2 are not the values of the first two units in the population. Further, let p_1 and p_2 denote the initial probabilities of selection of $U_{i(1)}$ and $U_{i(2)}$, respectively.

Consider the estimators

$$z_1 = \frac{y_1}{Np_1}$$

$$z_2 = \frac{1}{N}\left[y_1 + \frac{y_2}{p_2/(1-p_1)}\right]$$

$$= \frac{1}{N}\left[y_1 + y_2\frac{(1-p_1)}{p_2}\right]$$

$$\bar{z} = \frac{z_1 + z_2}{2}.$$

Note that $\frac{p_2}{1-p_1}$ is the probability $P(U_{i(2)} \mid U_{i(1)})$.

Estimation of Population Mean:

First, we show that $\bar{z}$ is an unbiased estimator of $\bar{Y}$.

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

Note that $\sum_{i=1}^{N} P_i = 1$.

Consider

$$E(z_1) = \frac{1}{N} E\left(\frac{y_1}{p_1}\right) \quad \left(\text{Note that } \frac{y_1}{p_1} \text{ can take any one of out of the } N \text{ values } \frac{Y_1}{P_1}, \frac{Y_2}{P_2}, \ldots, \frac{Y_N}{P_N}\right)$$

$$= \frac{1}{N}\left[\frac{Y_1}{P_1}P_1 + \frac{Y_2}{P_2}P_2 + \ldots + \frac{Y_N}{P_N}P_N\right]$$

$$= \bar{Y}$$

$$E(z_2) = \frac{1}{N} E\left[y_1 + y_2\frac{(1-p_1)}{p_2}\right]$$

$$= \frac{1}{N}\left[E(y_1) + E_1\left\{E_2\left(y_2\frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right)\right\}\right] \quad \text{(Using } E(Y) = E_X[E_Y(Y \mid X)]\text{.}$$

where E_2 is the conditional expectation after fixing the unit $U_{i(1)}$ selected in the first draw.

Since $\frac{y_2}{p_2}$ can take any one of the $(N-1)$ values (except the value selected in the first draw) $\frac{Y_j}{P_j}$ with probability $\frac{P_j}{1-P_1}$, so

$$E_2\left[y_2\frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right] = (1-P_1)E_2\left[\frac{y_2}{p_2}\middle| U_{i(1)}\right] = (1-P_1)\sum\nolimits_j^{*}\left[\frac{Y_j}{P_j}.\frac{P_j}{1-P_1}\right].$$

where the summation is taken over all the values of Y except the value y_1 which is selected at the first draw. So

$$E_2\left[y_2\frac{(1-P_1)}{p_2}\middle| U_{i(1)}\right] = \sum\nolimits_j^{*} Y_j = Y_{tot} - y_1.$$

Substituting it in $E(z_2)$, we have

$$E(z_2) = \frac{1}{N}\left[E(y_1) + E_1(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N}\left[E(y_1) + E(Y_{tot} - y_1)\right]$$

$$= \frac{1}{N}E(Y_{tot}) = \frac{Y_{tot}}{N} = \bar{Y}.$$

Thus

$$E(\bar{z}) = \frac{E(z_1)+E(z_2)}{2}$$
$$= \frac{\bar{Y}+\bar{Y}}{2}$$
$$= \bar{Y}.$$

Variance:

The variance of $\bar{z}$ for the case of two draws is given as

$$Var(\bar{z}) = \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\left[\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i}-Y_{tot}\right)^2\right] - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i}-Y_{tot}\right)^2$$

Proof: Before starting the proof, we note the following property

$$\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i\left[\sum_{j=1}^{N} b_j - b_i\right]$$

which is used in the proof.

The variance of $\bar{z}$ is

$$Var(\bar{z}) = E(\bar{z}^2) - \left[E(\bar{z})\right]^2$$
$$= E\left[\frac{1}{2N}\left\{\frac{y_1}{p_1}+y_1+\frac{y_2(1-p_1)}{p_2}\right\}\right]^2 - \bar{Y}^2$$
$$= \frac{1}{4N^2}E\left[\frac{y_1(1+p_1)}{p_1}+\frac{y_2(1-p_1)}{p_2}\right]^2 - \bar{Y}^2$$

p_1 ↓ nature of variable depends only on 1st draw

p_2 ↓ nature of variable depends upon 1st and 2nd draw

$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i(1+P_i)}{P_i}+\frac{Y_j(1-P_i)}{P_j}\right\}^2\frac{P_iP_j}{1-P_i}\right] - \bar{Y}^2$$
$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i^2}\frac{PP_j}{1-P_i}+\frac{Y_j^2(1-P_i)^2}{P_j^2}\frac{P_iP_j}{1-P_i}+2Y_iY_j\frac{(1-P_i^2)}{P_iP_j}\frac{P_iP_j}{1-P_i}\right\}\right] - \bar{Y}^2$$
$$= \frac{1}{4N^2}\left[\sum_{i\neq j=1}^{N}\left\{\frac{Y_i^2(1+P_i)^2}{P_i}\frac{P_j}{1-P_i}+\frac{Y_j^2(1-P_i)^2}{P_j}\frac{P_i}{1-P_i}+2Y_iY_j(1+P_i)\right\} - \bar{Y}^2.\right.$$

Using the property

$$\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i \left[\sum_{j=1}^{N} b_j - b_i\right], \text{ we can write}$$

$$Var(\bar{z}) = \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2(1+P_i)^2}{P_i(1-P_i)}\left\{\sum_{j=1}^{N}P_j - P_i\right\} + \sum_{i=1}^{N}P_i(1-P_i)\left\{\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i}\right\} + 2\sum_{i=1}^{N}Y_i(1+P_i)(\sum_{j=1}^{N}Y_j - Y_i)\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i}(1+P_i^2+2P_i) + \sum_{i=1}^{N}P_i(1-P_i)\left\{\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \frac{Y_i^2}{P_i}\right\} + 2\sum_{i=1}^{N}Y_i(1+P_i)(\sum_{j=1}^{N}Y_j - Y_i)\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} + \sum_{i=1}^{N}Y_i^2P + 2\sum_{i=1}^{N}Y_i^2 + \sum_{i=1}^{N}P_i\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \sum_{i=1}^{N}Y_i^2 - \sum_{i=1}^{N}P_i^2\sum_{j=1}^{N}\frac{Y_j^2}{P_j}\right.$$

$$\left. + \sum_i P_iY_i^2 + 2\sum_{i=1}^{N}Y_i\sum_{j=1}^{N}Y_j - 2\sum_{i=1}^{N}Y_i^2P_i + 2\sum_{i=1}^{N}Y_iP_i\sum_{j=1}^{N}Y_j - 2\sum_{i=1}^{N}Y_i^2\right] - \bar{Y}^2$$

$$= \frac{1}{4N^2}\left[2\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - \sum_{i=1}^{N}P_i^2\sum_{j=1}^{N}\frac{Y_j^2}{P_j} - \sum_{i=1}^{N}Y_i^2 + 2Y_{tot}^2 + 2Y_{tot}\sum_{i=1}^{N}Y_iP_i\right] - \bar{Y}^2$$

$$= 2\left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{4N^2}\left(\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - Y_{tot}^2 + Y_{tot}^2\right) - \frac{1}{4N^2}\left[\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i + 4N^2\bar{Y}^2\right]$$

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}(\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i - 2Y_{tot}^2 + 4Y_{tot}^2)$$

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

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}(\sum_{i=1}^{N}Y_i^2 - 2Y_{tot}\sum_{i=1}^{N}Y_iP_i + 2Y_{tot}^2 - 2Y_{tot}^2 + \sum_i P_i^2Y_{tot}^2)$$

$$= \left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\frac{1}{2N^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}\left(Y_i^2 - 2Y_{tot}Y_iP_i + P_i^2Y_{tot}^2\right)$$

$$= \frac{1}{2N^2}\left(1-\frac{1}{2}\sum_{i=1}^{N}P_i^2\right)\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

$$(\bar{z}) = \frac{1}{2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{NP_i} - \bar{Y}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\sum_{i=1}^{N}\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2 - \frac{1}{4N^2}\sum_{i=1}^{N}P_i^2\left(\frac{Y_i}{P_i} - Y_{tot}\right)^2$$

↓ variance of WR case for $n = 2$

↓ reduction of variance in WR with varying probability

Estimation of $Var(\bar{z})$

$$Var(\bar{z}) = E(\bar{z}^2) - (E(\bar{z}))^2$$
$$= E(\bar{z}^2) - \bar{Y}^2$$

Since

$$E(z_1 z_2) = E\left[z_1 E(z_2 \mid u_1)\right]$$
$$= E\left[z_1 \bar{Y}\right]$$
$$= \bar{Y} E(z_1)$$
$$= \bar{Y}^2.$$

Consider

$$E\left[\bar{z}^2 - z_1 z_2\right] = E(\bar{z}^2) - E(z_1 z_2)$$
$$= E(\bar{z}^2) - \bar{Y}^2$$
$$= Var(\bar{z})$$

$\Rightarrow Var(\bar{z}) = \bar{z}^2 - z_1 z_2$ is an unbiased estimator of $Var(\bar{z})$

Alternative form

$$Var(\bar{z}) = \bar{z}^2 - z_1 z_2$$
$$= \left(\frac{z_1 + z_2}{2}\right)^2 - z_1 z_2$$
$$= \frac{(z_1 - z_2)^2}{4}$$
$$= \frac{1}{4}\left[\frac{y_1}{Np_1} - \frac{y_1}{N} - \frac{y_2}{N}\frac{1-p_1}{p_2}\right]^2$$
$$= \frac{1}{4N^2}\left[(1-p_1)\frac{y_1}{p_1} - \frac{y_2(1-p_1)}{p_2}\right]^2$$
$$= \frac{(1-p_1)^2}{4N^2}\left(\frac{y_1}{p_1} - \frac{y_2}{p_2}\right)^2.$$

Case 2: General Case

Let $(U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)})$ be the units selected in the order in which they are drawn in n draws where $U_{i(r)}$ denotes that the i^{th} unit is drawn at the r^{th} draw. Let $(y_1, y_2, .., y_r, ..., y_n)$ and $(p_1, p_2, ..., p_r, ..., p_n)$ be the values of the study variable and corresponding initial probabilities of selection, respectively. Further, let $P_{i(1)}, P_{i(2)}, ..., P_{i(r)}, ..., P_{i(n)}$ be the initial probabilities of $U_{i(1)}, U_{i(2)}, ..., U_{i(r)}, ..., U_{i(n)}$, respectively.

Further, let

$$z_1 = \frac{y_1}{Np_1}$$

$$z_r = \frac{1}{N}\left[y_1 + y_2 + \ldots + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1})\right] \text{ for } r = 2,3,\ldots,n.$$

Consider $\bar{z} = \frac{1}{n}\sum_{r=1}^{n} z_r$ as an estimator of population mean $\bar{Y}$.

We already have shown in case 1 that $E(z_1) = \bar{Y}$.

Now we consider $E(z_r), r = 2,3,\ldots,n$. We can write

$$E(z_r) = \frac{1}{N}E_1E_2\left[z_r \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)}\right]$$

where E_2 is the conditional expectation after fixing the units $U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)}$ drawn in the first $(r - 1)$ draws.

Consider

$$E\left[\frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1})\right] = E_1E_2\left[\frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)}\right]$$

$$= E_1\left[(1 - P_{i(1)} - P_{i(2)} \ldots - P_{i(r-1)})E_2\left(\frac{y_r}{p_r} \middle| U_{i(1)}, U_{i(2)}, \ldots, U_{i(r-1)}\right)\right].$$

Since conditionally $\frac{y_r}{p_r}$ can take any one of the N-$(r$-$1)$ values $\frac{Y_j}{P_j}, j = 1,2,\ldots,N$ with probabilities

$\frac{P_j}{1 - P_{i(1)} - P_{i(2)} \ldots - P_{i(r-1)}}$, so

$$E\left[\frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1})\right] = E_1\left[(1 - P_{i(1)} - P_{i(2)} \ldots - P_{i(r-1)}) \sum_{j=1}^{N}{}^{*} \frac{Y_j}{P_j} \cdot \frac{P_j}{(1 - P_{i(1)} - P_{i(2)} \ldots - P_{i(r-1)})}\right]$$

$$= E_1\left[\sum_{j=1}^{N}{}^{*} Y_j\right]$$

where $\sum_{j=1}^{N}{}^{*}$ denotes that the summation is taken over all the values of y except the y values selected in the first $(r$ -1) draws like as $\sum_{j=1(\neq i(1), i(2), \ldots, i(r-1))}^{N}$, i.e., except the values $y_1, y_2, \ldots, y_{r-1}$ which are selected in the first $(r$-1) draws.

Thus now we can express

$$E(z_r) = \frac{1}{N} E_1 E_2 \left[y_1 + y_2 + \ldots + y_{r-1} + \frac{y_r}{p_r}(1 - p_1 - \ldots - p_{r-1}) \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1}^{N}{}^{*} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \sum_{j=1(\neq i(1), i(2), \ldots, i(r-1))}^{N} Y_j \right]$$

$$= \frac{1}{N} E_1 \left[Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} + \left\{ Y_{tot} - \left(Y_{i(1)} + Y_{i(2)} + \ldots + Y_{i(r-1)} \right) \right\} \right]$$

$$= \frac{1}{N} E_1 \left[Y_{tot} \right]$$

$$= \frac{Y_{tot}}{N}$$

$$= \bar{Y} \quad \text{for all} \quad r = 1, 2, \ldots, n.$$

Then

$$E(\bar{z}) = \frac{1}{n} \sum_{r=1}^{n} E(z_r)$$

$$= \frac{1}{n} \sum_{r=1}^{n} \bar{Y}$$

$$= \bar{Y}.$$

Thus $\bar{z}$ is an unbiased estimator of population mean $\bar{Y}$.

The expression for variance of $\bar{z}$ in general, the case is complex, but its estimate is simple.

Estimate of variance:

$Var(\bar{z}) = E(\bar{z}^2) - \bar{Y}^2$.

Consider for $r < s$,

$$E(z_r z_s) = E\left[z_r E(z_s \mid U_1, U_2, \ldots, U_{s-1}) \right]$$

$$= E\left[z_r \bar{Y} \right]$$

$$= \bar{Y} E(z_r)$$

$$= \bar{Y}^2$$

because for $r < s$, z_r will not contribute

and similarly for $s < r, z_s$ will not contribute in the expectation.

Further, for $s < r$,

$$\begin{aligned} E(z_r z_s) &= E\left[z_s E(z_r \mid U_1, U_2, \ldots, U_{r-1})\right] \\ &= E\left[z_s \bar{Y}\right] \\ &= \bar{Y} E(z_s) \\ &= \bar{Y}^2. \end{aligned}$$

Consider

$$\begin{aligned} E\left[\frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s\right] &= \frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} E(z_r z_s) \\ &= \frac{1}{n(n-1)} n(n-1) \bar{Y}^2 \\ &= \bar{Y}^2. \end{aligned}$$

Substituting $\bar{Y}^2$ in $Var(\bar{z})$, we get

$$\begin{aligned} Var(\bar{z}) &= E(\bar{z}^2) - \bar{Y}^2 \\ &= E(\bar{z}^2) - E\left[\frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s\right] \end{aligned}$$

$$\Rightarrow Var(\bar{z}) = \bar{z}^2 - \frac{1}{n(n-1)} \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s$$

Using $\left(\sum_{r=1}^{n} z_r\right)^2 = \sum_{r=1}^{n} z_r^2 + \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s$

$$\Rightarrow \sum_{r(\neq s)=1}^{n} \sum_{s=1}^{n} z_r z_s = n^2 \bar{z}^2 - \sum_{r=1}^{n} z_r^2,$$

The expression of $Var(\bar{z})$ can be further simplified as

$$\begin{aligned} Var(\bar{z}) &= \bar{z}^2 - \frac{1}{n(n-1)}\left[n^2 \bar{z}^2 - \sum_{r=1}^{n} z_r^2\right] \\ &= \frac{1}{n(n-1)}\left[\sum_{r=1}^{n} z_r^2 - n\bar{z}^2\right] \\ &= \frac{1}{n(n-1)} \sum_{r=1}^{n} (z_r - \bar{z})^2. \end{aligned}$$

Unordered estimator:

In an ordered estimator, the order in which the units are drawn is considered. Corresp ordered estimator, there exists an unordered estimator that does not depend on the order units are drawn and has a smaller variance than the ordered estimator.

In the case of sampling WOR from a population of size N, there are $\binom{N}{n}$ unordered san n. Corresponding to any unordered sample(s) of size n units, there are $n!$ ordered samples. For example, for $n=2$ if the units are u_1 and u_2, then

- there are 2! ordered samples - (u_1,u_2) and (u_2,u_1)
- there is one unordered sample (u_1,u_2).

Moreover,

$$\begin{pmatrix}\text{Probability of unordered}\\ \text{sample } (u_1,u_2)\end{pmatrix}=\begin{pmatrix}\text{Probability of ordered}\\ \text{sample } (u_1,u_2)\end{pmatrix}+\begin{pmatrix}\text{Probability of ordered}\\ \text{sample } (u_2,u_1)\end{pmatrix}$$

For $n=3$, there are three units u_1,u_2,u_3 and

-there are the following 3! = 6 ordered samples:

$$(u_1,u_2,u_3),(u_1,u_3,u_2),(u_2,u_1,u_3),(u_2,u_3,u_1),(u_3,u_1,u_2),(u_3,u_2,u_1)$$

- there is one unordered sample (u_1,u_2,u_3).

Moreover,

Probability of unordered sample

= Sum of probability of ordered sample, i.e.

$$P(u_1,u_2,u_3)+P(u_1,u_3,u_2)+P(u_2,u_1,u_3)+P(u_2,u_3,u_1)+P(u_3,u_1,u_2)+P(u_3,u_2,u_1),$$

Let $z_{si}, s=1,2,..,\binom{N}{n}, i=1,2,...,n!(=M)$ be an estimator of population parameter θ based on the ordered sample s_i. Consider a scheme of selection in which the probability of selecting the ordered sample (s_i) is p_{si}. The probability of getting the unordered sample(s) is the sum of the probabilities, i.e.,

$$p_s=\sum_{i=1}^{M}p_{si}.$$

For a population of size N with units denoted as 1, 2,…, N, the samples of size n are $n-$ tuples. In the n^{th} draw, the sample space will consist of $N(N-1)...(N-n+1)$ unordered sample points.

$$p_{sio} = P[\text{selection of any ordered sample}] = \frac{1}{N(N-1)...(N-n+1)}$$

$$p_{siu} = P[\text{selection of any unordered sample}] = \frac{n!}{N(N-1)...(N-n+1)} = n!P\begin{bmatrix}\text{selection of any}\\ \text{ordered sample}\end{bmatrix}$$

then $p_s = \sum_{i=1}^{M(=n!)} p_{sio} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}$.

Theorem: If $\hat{\theta}_0 = z_{si}$, $s = 1,2,...,\binom{N}{n}$; $i = 1,2,...,M(=n!)$ and $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p'_{si}$ are the ordered and unordered estimators of θ respectively, then

(i) $E(\hat{\theta}_u) = E(\hat{\theta}_0)$

(ii) $Var(\hat{\theta}_u) \le Var(\hat{\theta}_0)$

where z_{s_i} is a function of s_i^{th} ordered sample (hence a random variable) and p_{s_i} is the probability of selection of s_i^{th} ordered sample and $p'_{si} = \frac{p_{si}}{p_s}$.

Proof: Total number of ordered samples = $n!\binom{N}{n}$

$$(i)\ E(\hat{\theta}_0) = \sum_{s=1}^{\binom{N}{n}}\sum_{i=1}^{M} z_{si}p_{si}$$

$$E(\hat{\theta}_u) = \sum_{s=1}^{\binom{N}{n}}\left(\sum_{i=1}^{M} z_{si}p'_{si}\right)p_s$$

$$= \sum_{s}\left(\sum_{i} z_{si}\frac{p_{si}}{p_s}\right)p_s$$

$$= \sum_{s}\sum_{i} z_{si}p_{si}$$

$$= E(\hat{\theta}_0)$$

(ii) Since $\hat{\theta}_0 = z_{si}$, so $\hat{\theta}_0^2 = z_{si}^2$ with probability p_{si}, $i = 1,2,...,M$, $s = 1,2,...,\binom{N}{n}$.

Similarly, $\hat{\theta}_u = \sum_{i=1}^{M} z_{si}p'_{si}$, so $\hat{\theta}_u^2 = \left(\sum_{i=1}^{M} z_{si}p'_{si}\right)^2$ with probability p_s

Consider

$$Var(\hat{\theta}_0) = E(\hat{\theta}_0^2) - \left[E(\hat{\theta}_0)\right]^2$$

$$= \sum_s \sum_i z_{si}^2 p_{si} - \left[E(\hat{\theta}_0)\right]^2$$

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

$$= \sum_s \left(\sum_i z_{si} p'_{si}\right)^2 p_s - \left[E(\hat{\theta}_0)\right]^2$$

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i z_{si}^2 p_{si} - \sum_s \left(\sum_i z_{si} p'_{si}\right)^2 p_s$$

$$= \sum_s \sum_i z_{si}^2 p_{si} + \sum_s \left(\sum_i z_{si} p'_{si}\right)^2 p_s$$

$$-2\sum_s \left(\sum_i z_{si} p'_{si}\right)\left(\sum_i z_{si} p'_{si}\right) p_s$$

$$= \sum_s \left[\sum_i z_{si}^2 p_{si} + \left(\sum_i z_{si} p'_{si}\right)^2 \left(\sum_i p_{si}\right) - 2\left(\sum_i z_{si} p'_{si}\right)\left(\sum_i z_{si} p'_{si}\right) p_s\right]$$

$$= \sum_s \left[\sum_i \left\{z_{si}^2 p_{si} + \left(\sum_i z_{si} p'_{si}\right)^2 p_{si} - 2\left(\sum_i z_{si} p'_{si}\right) z_{si} p_{si}\right\}\right]$$

$$= \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p'_{si})^2 p_{si}\right] \geq 0$$

$$\Rightarrow Var(\hat{\theta}_0) - Var(\hat{\theta}_u) \geq 0$$

or $Var(\hat{\theta}_u) \leq Var(\hat{\theta}_0)$

Estimate of $Var(\hat{\theta}_u)$

Since

$$Var(\hat{\theta}_0) - Var(\hat{\theta}_u) = \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p'_{si})^2 p_{si}\right]$$

$$Var(\hat{\theta}_u) = Var(\hat{\theta}_0) - \sum_s \sum_i \left[(z_{si} - \sum_i z_{si} p'_{si})^2 p_{si}\right]$$

$$= \sum_i p'_{si} Var(\hat{\theta}_0) - \sum_i p'_{si} (z_{si} - \sum_i z_{si} p'_{si})^2.$$

Based on this result, now we use the ordered estimators to construct an unordered estimator. It follows from this theorem that the unordered estimator will be more efficient than the corresponding ordered estimators.

Murthy's unordered estimator corresponding to Des Raj's ordered estimator for the sample size 2

Suppose y_i & y_j are the values of units u_i & u_j selected in the First and second draw respectively with varying probability and WOR in a sample of size 2 and let P_1 & P_2 be the

corresponding initial probabilities of selection. So now we have two ordered estimates corresponding to the ordered samples S_1* and S_2* as follows

$$s_1^{\bullet} = (y_i, y_j) \text{ with } (U_i, U_j)$$
$$s_2^{\bullet} = (y_j, y_i) \text{ with } (U_j, U_i)$$

which are given as

$$\bar{z}(s_1^{\bullet}) = \frac{1}{2N}\left[(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_i + \frac{y_i}{p_i} + \frac{y_j(1-p_i)}{p_j}\right]$$

and

$$\bar{z}(s_2^{\bullet}) = \frac{1}{2N}\left[(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right]$$

where the corresponding Des Raj estimator is given by

$$\frac{1}{2N}\left[y_j + \frac{y_j}{p_j} + \frac{y_i(1-p_j)}{p_i}\right].$$

The probabilities corresponding to $\bar{z}(s_1^{\bullet})$ and $\bar{z}(s_2^{\bullet})$ are

$$p(s_1^{\bullet}) = \frac{p_i p_j}{1-p_i}$$

$$p(s_2^{\bullet}) = \frac{p_j p_i}{1-p_j}$$

$$p(s) = p(s_1^{\bullet}) + p(s_2^{\bullet})$$

$$= \frac{p_i p_j (2-p_i-p_j)}{(1-p_i)(1-p_j)}$$

$$p'(s_1^{\bullet}) = \frac{1-p_j}{2-p_i-p_j}$$

$$= \frac{\frac{1}{2N}\left[(1-p_j)\frac{y_i}{p_i}\{(1+p_i)+(1-p_i)\} + (1-p_i)\frac{y_j}{p_j}\{(1-p_j)+(1+p_j)\}\right]}{2-p_i-p_j}$$

$$= \frac{(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}}{N(2-p_i-p_j)}.$$

Murthy's unordered estimate $\bar{z}(u)$ corresponding to Des Raj's ordered estimate is given as

$$\bar{z}(u) = \bar{z}(s_1^{\bullet})p'(s_1) + \bar{z}(s_2^{\bullet})p'(s_2)$$

$$= \frac{\bar{z}(s_1^{\bullet})p(s_1^{\bullet}) + \bar{z}(s_2^{\bullet})p(s_2^{\bullet})}{p(s_1^{\bullet}) + p(s_2^{\bullet})}$$

$$= \frac{\left[\frac{1}{2N}\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}\left(\frac{p_i p_j}{1-p_i}\right)\right] + \left[\frac{1}{2N}\left\{(1+p_j)\frac{y_j}{p_j} + (1-p_j)\frac{y_i}{p_i}\right\}\left\{\frac{p_j p_i}{1-p_j}\right\}\right]}{\frac{p_i p_j}{1-p_i} + \frac{p_j p_i}{1-p_j}}$$

$$= \frac{\frac{1}{2N}\left[\left\{(1+p_i)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}\right\}(1-p_j) + \left\{(1+p_j)\frac{y_j}{p_i} + (1-p_j)\frac{y_i}{p_i}\right\}(1-p_i)\right]}{(1-p_j)+(1-p_i)}$$

$$= \frac{\frac{1}{2N}\left[(1-p_j)\frac{y_i}{p_i}\{(1+p_i)+(1-p_i)\} + (1-p_i)\frac{y_j}{p_j}\{(1-p_j)+(1+p_j)\}\right]}{2-p_i-p_j}$$

$$= \frac{(1-p_j)\frac{y_i}{p_i} + (1-p_i)\frac{y_j}{p_j}}{N(2-p_i-p_j)}.$$

Unbiasedness:

Note that y_i and p_i can take any one of the values out of $Y_1, Y_2, \ldots, Y_N$ and $P_1, P_2, \ldots, P_N$, respectively. Then y_j and p_j can take any one of the remaining values out of $Y_1, Y_2, \ldots, Y_N$ and $P_1, P_2, \ldots, P_N$, respectively, i.e., all the values except the values taken at the first draw. Now

$$E[\bar{z}(u)] = \frac{1}{N}\sum_{i\neq j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i} + (1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_i P_j}{1-P_i} + \frac{P_i P_j}{1-P_j}\right\}}{2-P_i-P_j}$$

$$= \frac{1}{2N}2\sum_{i<j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i} + (1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_i P_j}{1-P_i} + \frac{P_j P_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$= \frac{1}{2N}\sum_{i\neq j}\frac{\left[\left\{(1-P_j)\frac{Y_i}{P_i} + (1-P_i)\frac{Y_j}{P_j}\right\}\right]\left\{\frac{P_i P_j}{1-P_i} + \frac{P_j P_i}{1-P_j}\right\}}{2-P_i-P_j}$$

$$= \frac{1}{2N}\sum_{i\neq j}\left[\left\{(1-P_j)\frac{Y_i}{P_i} + (1-P_i)\frac{Y_j}{P_j}\right\}\left\{\frac{P_i P_j}{(1-P_i)(1-P_j)}\right\}\right]$$

$$= \frac{1}{2N}\sum_{i\neq j}\left[\frac{Y_i P_j}{1-P_i} + \frac{Y_j P_i}{1-P_j}\right]$$

Using result $\sum_{i\neq j=1}^{N} a_i b_j = \sum_{i=1}^{N} a_i\left\{\sum_{j=1}^{N} b_j - b_i\right\}$, we have

$$E[\bar{\Xi}(u)] = \frac{1}{2N}\left[\left\{\sum_{i=1}^{N}\frac{Y_i}{1-P_i}\left(\sum_{j=1}^{N}P_j - P_j\right)\right\} + \left\{\sum_{j=1}^{N}\frac{Y_j}{1-P_j}\left(\sum_{i=1}^{N}P_i - P_j\right)\right\}\right]$$

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

$$= \frac{1}{2N}\left\{\sum_{i=1}^{N}Y_i + \sum_{j=1}^{N}Y_j\right\}$$

$$= \frac{\bar{Y}+\bar{Y}}{2}$$

$$= \bar{Y}.$$

Variance: The variance of $\bar{\Xi}(u)$ can be found as

$$Var[\bar{\Xi}(u)] = \frac{1}{2}\sum_{i\neq j=1}^{N}\frac{(1-P_i-P_j)(1-P_i)(1-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2\frac{P_iP_j(2-P_i-P_j)}{(1-P_i)(1-P_j)}$$

$$= \frac{1}{2}\sum_{i\neq j=1}^{N}\frac{P_iP_j(1-P_i-P_j)}{N^2(2-P_i-P_j)}\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2$$

Using the theorem that $Var(\hat{\theta}_u) \le Var(\hat{\theta}_0)$ we get

$$Var[\bar{\Xi}(u)] \le Var[\bar{\Xi}(s_1^*)]$$

and $Var[\bar{\Xi}(u)] \le Var[\bar{\Xi}(s_2^*)]$

Unbiased estimator of $V[\bar{\Xi}(u)]$

An unbiased estimator of $Var(\bar{\Xi}|u)$ is

$$Var[\bar{\Xi}(u)] = \frac{(1-p_i-p_j)(1-p_i)(1-p_j)}{N^2(2-p_i-p_j)^2}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2.$$

7.5 COMPARISON OF ESTIMATORS:

- Des Raj: Ordered, simpler, useful for small samples, more sensitive to order.
- Murthy: Unordered, averages over permutations, less variance.
- Both are unbiased.
- Murthy's estimator is generally preferred when computational effort is acceptable.

7.6 SUMMARY:

- For sample size two, both estimators serve specific needs.
- Des Raj: Efficient for ordered selection in PPS.
- Murthy: Broader applicability, especially with unordered or complex schemes.

- Choosing the right estimator depends on sampling design**,** order of selection**,** and practical considerations like ease of computation **vs.** variance reduction**.**

7.7 KEY WORDS:

- Probability Proportional to Size (PPS) Sampling
- Estimator
- Unbiased Estimator
- Sample Size Two
- Ordered Estimates
- Unordered Estimates
- Des Raj Estimator
- Murthy Estimator
- With Replacement (WR)
- Without Replacement (WOR)
- Inclusion Probability
- Joint Inclusion Probability
- Sampling Design
- Estimation of Population Total
- Efficiency of Estimator
- Variance of Estimator
- Symmetric Estimator
- Sequential Selection
- Permutation of Units
- Comparison of Estimators

7.8 SELF-ASSESSMENT QUESTIONS:

1. What is the basic idea behind Des Raj's estimator?
2. What is the condition under which Murthy's estimator is preferred over Des Raj's?
3. Why are Des Raj's and Murthy's estimators specifically useful for sample size two?
4. How do these estimators account for the unequal probabilities of selection?
5. Is Des Raj's estimator unbiased? Justify your answer.
6. What is the difference between ordered and unordered estimators in PPS sampling?
7. What does the Murthy estimator attempt to correct compared to Des Raj's estimator?

7.9 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
3. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
4. Singh, D. and Chaudhary, F.S. (1986) *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.
5. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society

Dr. N.Viswam

LESSON- 8

HORVITZ THOMPSON ESTIMATOR

OBJECTIVES:

Upon completion of this lesson, learners will be able to:

- Understand the theory and application of the Horvitz-Thomson estimator in the context of unequal probability sampling.
- Explain Grundy's estimator**,** including its assumptions, derivation, and comparison with the Horvitz-Thomson estimator.
- Describe the Midzuno-Sen sampling scheme**,** its selection procedure, and situations where it is advantageous.
- Compute unbiased estimates of population totals or means using these methods.
- Evaluate the variances of the estimators and understand their relative efficiencies.
- Apply the estimators to real-world survey sampling problems involving unequal probabilities.
- Compare the estimators through theoretical and numerical illustrations.

STRUCTURE:

8.1 INTRODUCTION:

In survey sampling, particularly in unequal probability sampling, estimating population parameters accurately requires specialized techniques. Among the most prominent are the **Horvitz-Thomson estimator**, **Grundy's estimator**, and the **Midzuno-Sen sampling scheme**. These methods aim to provide unbiased or nearly unbiased estimates of population totals or means, even when selection probabilities differ across units.

1. **Horvitz-Thomson Estimator (HT Estimator)**:
   Introduced by Horvitz and Thompson in 1952, this estimator is a design-unbiased estimator for the population total. It is applicable under unequal probability sampling with or without replacement, where the inclusion probabilities are strictly positive. The HT estimator adjusts each sampled unit by the inverse of its inclusion probability, making it robust to unequal selection chances.
2. **Grundy's Estimator**:
   An extension of the Horvitz-Thomson approach, Grundy's estimator modifies the estimation when **second-order inclusion probabilities** are known or available. It aims to improve efficiency, particularly in complex survey designs. It is generally used for **variance estimation** or improving estimator precision when more information about the sampling design is available.
3. **Midzuno-Sen Sampling Scheme**:
   Proposed independently by Midzuno (1951) and Sen (1952), this is a **probability proportional to size (PPS) sampling scheme without replacement**. The scheme ensures that one unit is selected using PPS, and the rest are selected using simple random sampling (SRS) among the remaining units. It maintains unbiasedness while simplifying implementation and variance estimation.

   Together, these methods and estimators provide a strong theoretical foundation for dealing with practical problems in unequal probability sampling, enabling statisticians to make valid inferences from sample data.

8.2 ORCHARD ESTIMATOR:

Daroga Singh Das (1951) &Des Raj(1956) have proposed estimators which are based on the order of units. These estimators do not require calculations. And use of conditional Probabilities without calculating π_i &π_{ij} generally difficult to compute Sampling schemes.

Des-Raj ordered estimator→ depends on conditional Probability. Here, we shall consider the estimator proposed by Des-Raj, first for the case when n=2, and then generalize the result.

Go for Horvitz-Thomson estimator.

8.3 DEFINITION OF HORVITZ-THOMSON ESTIMATOR:

Suppose that y_i is the value of ith unit with π_i the probability of inclusion in the sample. The Horvitz-Thomson estimator for the population total "Y" is defined by $\hat{Y}_{HT} = \sum_{i=1}^{N} \frac{y_i}{\pi_i}$.

Theorem: In PPS Sampling without replacement, $\hat{Y}_{HT}$ is unbiased and its sampling variance is given by $V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j} y_i y_j$.

Let π_i: Probability that the i^{th} unit is included in the Sample.

π_{ij}: Probability that $i^{th}$and j^{th} units are included in the Sample.

Proof: Let $t_i (i = 1,2, \dots, N)$ be a random variable that takes the values '1' If the ith unit is

drawn and 'zero' otherwise. Then t_i follows the Binomial distribution for a sample of size '1' with probability π_i .

Thus , $E(t_i) = \pi_i$ and $V(t_i) = \pi_i(1-\pi_i)$

$$cov(t_i, t_j) = E(t_i t_j) - E(t_i)E(t_j)$$
$$= \pi_{ij} - \pi_i \pi_j$$

Hence, regarding the y_i as fixed and the t_i as the r.v's,

$E(\hat{Y}_{HT}) = E\left(\sum_{i=1}^{N} \frac{t_i y_i}{\pi_i}\right) = \sum_{i=1}^{N} \frac{y_i}{\pi_i} E(t_i) = \sum_{i=1}^{N} \frac{y_i}{\cancel{\pi_t}} \cancel{\pi_t}$=Y.

$\therefore \hat{Y}_{HT}$is unbiased estimate of Y.

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \left(\frac{y_i}{\pi_i}\right)^2 V(t_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} cov(t_i, t_j)$$

$$= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \cancel{\pi_t} (1-\pi_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{y_i}{\pi_i} \cdot \frac{y_j}{\pi_j} (\pi_{ij} - \pi_i \pi_j)$$

$\therefore V(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j$.

Corollary: An unbiased sample estimator of $V(\hat{Y}_{HT})$ is given by

$$\therefore v(\hat{Y}_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Provided that none of the π_{ij} in the population vanishes.

8.4 GRUNDY'S ESTIMATOR:

Grundy's estimator, also known as the **Yates-Grundy estimator**, focuses on **variance estimation** rather than point estimation. It provides an unbiased estimate of the **variance of the Horvitz-Thomson estimator**.

- **Formula** (for variance):

$$V(\hat{Y}_{HT}) = \frac{1}{2} \sum_{i \neq j} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

- **Key Use:** Variance estimation under unequal probability sampling without replacement.
- **Assumption:** Known joint inclusion probabilities π_{ij}.

8.5 YATES ESTIMATOR:

Yates and Grundy form of variance

Since there are exactly n values of α_i which are 1 and $(N-n)$ values which are zero, so

$$\sum_{i=1}^{N}\alpha_i = n.$$

Taking expectation on both sides

$$\sum_{i=1}^{N}E(\alpha_i) = n.$$

Also

$$E\left(\sum_{i=1}^{N}\alpha_i\right)^2 = \sum_{i=1}^{N}E(\alpha_i^2) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_j)$$

$$E(n)^2 = \sum_{i=1}^{N}E(\alpha_i) + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_j) \text{ (using } E(\alpha_i) = E(\alpha_i^2))$$

$$n^2 = n + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_j)$$

$$\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}E(\alpha_i\alpha_j) = n(n-1)$$

Thus $E(\alpha_i\alpha_j) = P(\alpha_i = 1, \alpha_j = 1)$

$$= P(\alpha_i = 1)P(\alpha_j = 1 | \alpha_i = 1)$$

$$= E(\alpha_i)E(\alpha_j | \alpha_i = 1)$$

Therefore

$$\sum_{j(\neq i)=1}^{N}\left[E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\right]$$

$$= \sum_{j(\neq i)=1}^{N}\left[E(\alpha_i)E(\alpha_j | \alpha_i = 1) - E(\alpha_i)E(\alpha_j)\right]$$

$$= E(\alpha_i)\sum_{j(\neq i)=1}^{N}\left[E(\alpha_j | \alpha_i = 1) - E(\alpha_j)\right]$$

$$= E(\alpha_i)\left[(n-1) - (n - E(\alpha_i)\right]$$

$$= -E(\alpha_i)\left[1 - E(\alpha_i)\right]$$

$$= -\pi_i(1-\pi_i) \qquad (1)$$

Similarly

$$\sum_{i(\neq j)=1}^{N}\left[E(\alpha_i\alpha_j) - E(\alpha_i)E(\alpha_j)\right] = -\pi_j(1-\pi_j). \qquad (2)$$

The expression for π_i and π_{ij} can be written for any given sample size.

For example, for $n=2$, assume that at the second draw, the probability of selecting a unit from the units available is proportional to the probability of selecting it at the first draw. Since

$E(\alpha_i) =$ Probability of selecting Y_i in a sample of two

$= P_{i1} + P_{i2}$

where P_{ir} is the probability of selecting Y_i at r^{th} draw $(r=1,2)$. If P_i is the probability of selecting the i^{th} unit at the first draw $(i=1,2,...,N)$ then we had earlier derived that

$$P_{i1} = P_i$$

$$P_{i2} = P\begin{bmatrix} y_i \text{ is not selected} \\ \text{at } 1^{st} \text{ draw} \end{bmatrix} P\begin{bmatrix} y_i \text{ is selected at } 2^{nd} \text{ draw} \mid \\ y_i \text{ is not selected at } 1^{st} \text{ draw} \end{bmatrix}$$

$$= \sum_{j(\neq i)=1}^{N} \frac{P_j P_i}{1-P_j}$$

We had earlier derived the variance of HT estimator as

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_j)z_i z_j\right]$$

Using (1) and (2) in this expression, we get

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2}\left[\sum_{i=1}^{N}\pi_i(1-\pi_i)z_i^2 + \sum_{j=1}^{N}\pi_j(1-\pi_j)z_j^2 - 2\sum_{i\neq j=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})z_i z_j\right]$$

$$= \frac{1}{2n^2}\left[-\sum_{i=1}^{N}\left\{\sum_{j(\neq i)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_i^2\right.$$

$$\left.-\sum_{j=1}^{N}\left\{\sum_{i(\neq j)=1}^{N}E(\alpha_i\alpha_j)-E(\alpha_i)E(\alpha_j)\right\}z_j^2 - 2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{n}\left\{E(\alpha_i)E(\alpha_j)-E(\alpha_i\alpha_j)\right\}z_i z_j\right]$$

$$= \frac{1}{2n^2}\left[\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_i^2 + \sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(-\pi_{ij}+\pi_i\pi_i)z_j^2 + 2\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_{ij}-\pi_i\pi_i)z_i z_j\right]\right.$$

$$= \frac{1}{2n^2}\left[\sum_{i(\neq j)=1}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})(z_i^2+z_j^2-2z_i z_j)\right].$$

Again

$$E(\alpha_i\alpha_j) = \text{Probability of including both } y_i \text{ and } y_j \text{ in a sample of size two}$$

$$= P_{i1}P_{j2|i} + P_{j1}P_{i2|j}$$

$$= P_i\frac{P_j}{1-P_i} + P_j\frac{P_i}{1-P_j}$$

$$= P_iP_j\left[\frac{1}{1-P_i} + \frac{1}{1-P_j}\right] + P_i.$$

Estimate of Variance

The estimate of variance is given by

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2}\sum_{i(\neq j)}^{n}\sum_{j=1}^{n}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}(z_i - z_j)^2.$$

8.6 MIDZUNO-SEN SAMPLING SCHEME:

Under this system of selection of probabilities, the unit in the first draw is selected with unequal probabilities of selection (i.e., pps), and remaining all the units are selected with SRSWOR at all subsequent draws.

Under this system

$$E(\alpha_i) = \pi_i = P\ (\text{unit } i\ (U_i) \text{ is included in the sample})$$

$$= P(U_i \text{ is included in } 1^{st} \text{ draw}) + P(U_i \text{ is included in any other draw})$$

$$= P_i + \begin{pmatrix}\text{Probability that } U_i \text{ is not selected at the first draw and} \\ \text{is selected at any of subsequent } (n-1) \text{ draws}\end{pmatrix}$$

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

$$= \frac{N-n}{N-1}P_i + \frac{n-1}{N-1}.$$

Similarly,

$$E(\alpha_i\alpha_j) = \text{Probability that both the units } U_i \text{ and } U_j \text{ are in the sample}$$

$$= \begin{pmatrix} \text{Probability that } U_i \text{ is selected at the first draw and} \\ U_j \text{ is selected at any of the subsequent draws } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that } U_j \text{ is selected at the first draw and} \\ U_i \text{ is selected at any of the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$+ \begin{pmatrix} \text{Probability that neither } U_i \text{ nor } U_j \text{ is selected at the first draw but} \\ \text{both of them are selected during the subsequent } (n-1) \text{ draws} \end{pmatrix}$$

$$= P_i\frac{n-1}{N-1} + P_j\frac{n-1}{N-1} + (1-P_i-P_j)\frac{(n-1)(n-2)}{(N-1)(N-2)}$$

$$= \frac{(n-1)}{(N-1)}\left[\frac{N-n}{N-2}(P_i+P_j) + \frac{n-2}{N-2}\right]$$

$$\pi_{ij} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(P_i+P_j) + \frac{n-2}{N-2}\right].$$

Similarly,

$$E(\alpha_i\alpha_j\alpha_k) = \pi_{ijk} = \text{Probability of including } U_i, U_j \text{ and } U_k \text{ in the sample}$$

$$= \frac{(n-1)(n-2)}{(N-1)(N-2)}\left[\frac{N-n}{N-3}(P_i+P_j+P_k) + \frac{n-3}{N-3}\right].$$

By an extension of this argument, if $U_i, U_j, ..., U_r$ are the r units in the sample of size $n(r<n)$, the probability of including these r units in the sample is

$$E(\alpha_i\alpha_j...\alpha_r) = \pi_{ij...r} = \frac{(n-1)(n-2)...(n-r+1)}{(N-1)(N-2)...(N-r+1)}\left[\frac{N-n}{N-r}(P_i+P_j+...+P_r) + \frac{n-r}{N-r}\right]$$

Similarly, if $U_1, U_2, ..., U_q$ be the n units, the probability of including these units in the sample is

$$E(\alpha_i\alpha_j...\alpha_q) = \pi_{ij...q} = \frac{(n-1)(n-2)...1}{(N-1)(N-2)...(N-n+1)}(P_i+P_j+...+P_q)$$

$$= \frac{1}{\binom{N-1}{n-1}}(P_i+P_j+...+P_q)$$

which is obtained by substituting $r = n$.

Thus if $P_i's$ are proportional to some measure of size of units in the population then the probability of selecting a specified sample is proportional to the total measure of the size of units included in the sample.

Substituting these $\pi_i, \pi_{ij}, \pi_{ijk}$ etc. in the HT estimator, we can obtain the estimator of population's mean and variance. In particular, an unbiased estimate of variance of HT estimator given by

$$Var(\hat{\bar{Y}}_{HT}) = \frac{1}{2n^2} \sum_{i \neq j=1}^{n} \sum_{j=1}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} (z_i - z_j)^2$$

where

$$\pi_i \pi_j - \pi_{ij} = \frac{N-n}{(N-1)^2} \left[(N-n) P_i P_j + \frac{n-1}{N-2} (1 - P_i - P_j) \right].$$

The main advantage of this method of sampling is that it is possible to compute a set of revised probabilities of selection such that the inclusion probabilities resulting from the revised probabilities are proportional to the initial probabilities of selection. It is desirable to do so since the initial probabilities can be chosen proportional to some measure of size.

8.7 COMPARISON OF ESTIMATORS:

- **HT Estimator** is preferred for general-purpose use due to its simplicity and unbiasedness.
- **Grundy's formula** is essential when precision (variance) needs to be estimated accurately.
- **Midzuno-Sen** offers a practical alternative when unequal probability sampling is required, especially in field surveys.
- **Orchard's Estimator** has limited use today but is conceptually significant in understanding early developments in sampling theory.

8.8 SUMMARY:

- Estimators like Horvitz-Thomson and its variants are crucial for unbiased estimation in unequal probability sampling.
- Each estimator has trade-offs between simplicity, variance, and implementation.
- Midzuno-Sen offers a practical compromise with good properties in many applications.
- The choice of estimator should consider the sampling design, availability of inclusion probabilities, and computational feasibility.

8.9 KEY WORDS:

- Unbiased Estimator
- Inclusion Probability
- Horvitz-Thomson Estimator
- Orchard Estimator
- Grundy's Estimator
- Yates-Grundy Variance
- Midzuno-Sen Scheme
- PPS Sampling
- Joint Inclusion Probability
- Variance Estimation

8.10 SELF-ASSESSMENT QUESTIONS:

1. What is the need for unbiased estimation in survey sampling?
2. Define the concept of an estimator and the importance of its variance.
3. State the formula for the Horvitz-Thomson (HT) estimator.
4. Under what sampling scheme is the HT estimator unbiased?
5. What is the main advantage of using the HT estimator in unequal probability sampling?
6. Define Horvitz – Thompson estimator of the population mean and derive the variance of this estimator.
7. Explain the concept of Yates and Grundy Form of Variance**.**
8. What is the Yates and Grundy form of variance?
9. How does it improve the estimation of variance under unequal probability sampling Write down the Yates-Grundy variance formula and explain the terms involved.
10. Describe the steps involved in the Midzuno-Sen sampling scheme.

8.11 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. S. K. Thompson**,** Title**:** *Sampling* (Wiley Series in Probability and Statistics).
3. P. Mukhopadhyay**,** Title**:** *Theory and Methods of Survey Sampling*
4. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
5. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
6. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society

LESSON- 9
TWO STAGE SAMPLING

OBJECTIVES:

After completing this lesson, learners will be able to:

- **Understand the Concept of Two-Stage Sampling:** Comprehend the rationale behind using two-stage sampling in large-scale surveys.
- **Describe the Structure of Two-Stage Sampling Design:**Define Primary Sampling Units (PSUs) and Secondary Sampling Units (SSUs).
- **Estimate the Population Mean:**Derive an unbiased estimator of the population mean using two-stage sampling with equal SSUs per PSU.
- Interpret the formula in terms of sample means at both stages.
- **Compute the Variance of the Estimator:**Derive the expression for the variance of the population mean estimator under this sampling scheme.
- Understand the contribution of variation at both stages (between PSUs and within PSUs).
- **Estimate the Variance from Sample Data:** Learn the methods for estimating the variance of the sample mean from actual survey data.
- Construct confidence intervals for the population mean using estimated variance.
- **Apply Concepts to Real-world Problems:** Use the discussed methods to analyze data from complex surveys.
- Evaluate the efficiency and practicality of two-stage sampling in field applications.
- **Compare with Other Sampling Methods:** Understand when two-stage sampling is preferable over stratified or simple random sampling.

STRUCTURE:

9.1 INTRODUCTION:

In cluster sampling, all the elements in the selected clusters are surveyed. Moreover, the efficiency in cluster sampling depends on the size of the cluster. As the size increases, the efficiency decreases. It suggests that higher precision can be attained by distributing a given number of elements over a large number of clusters and then by taking a small number of clusters and enumerating all elements within them. This is achieved in sub sampling.

We have, seen that the larger the cluster, the less efficient it will be. It is thus, logical to expect that for a given number of elements, greater precision will be attained by distributing them over a large number of clusters andthen sampling a larger number of elements from each of them or completely enumerating them. We can eliminate this disadvantage through sub-sampling or two stage sampling.

The procedure of first selecting clusters and then choosing specified number of elements from each selected cluster is known as sub-sampling or two-stage sampling. The clusters which form the units of sampling at the first stage are called the first stage units and the elements groups of elementswithin clusters which form the units of sampling at the second stage arecalled sub-units or second stage units. The procedure can be easily generalized to three or more stages and is termed as multi-stage sampling.

For example, in a crop survey

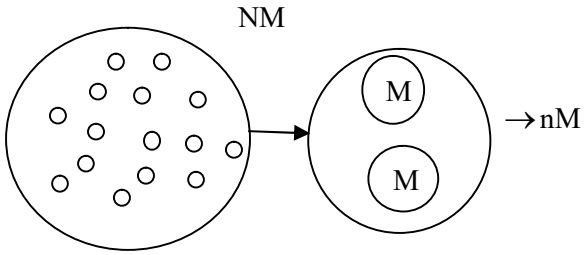
- villages are the first stage units.
- fields within the villages are the second stage units and
- plots within the fields are the third stage units.

In another example, to obtain a sample of fishes from a commercial fishery

- first take a sample of boats.
- then take a sample of fishes from each selected boat.

9.2 TWO STAGE SAMPLING (OR) SUB SAMPLING WITH UNITS OF EQUAL SIZE:

Two stage sampling is also called as sub-sampling $(n\times M \text{ elements})$



Description:

Suppose that each unit in the population can be divided into a number of smaller units or elements. A sample of n-units as being selected. If elements within a selected unit give similar results, it seems uneconomical to measure them all. A common practice is to select and measure a sample of the elements in any chosen unit. This technique is called sub-sampling, since the unit is not measured completely but it is itself a sample. Another name due to Mahalanobis is two-stage sampling, because the sample is taken in two steps.

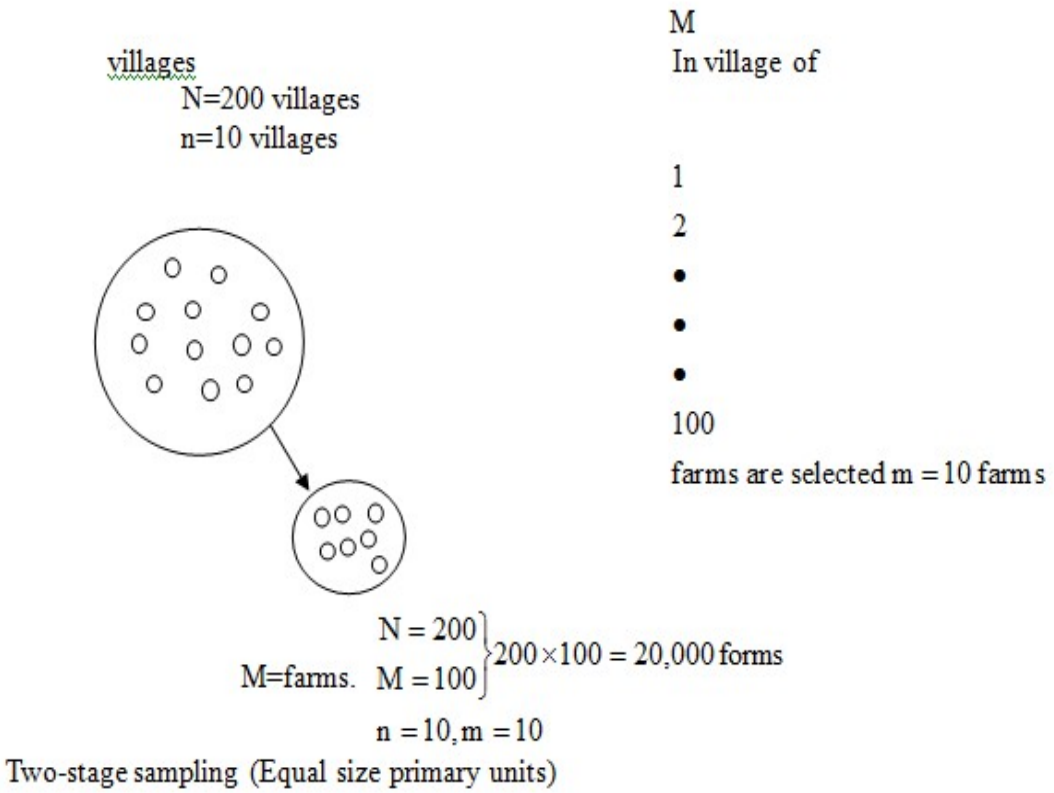
The first is to select a sample of units often called primary units (or first stage units) and the second is to select a sample of elements (or second stage units) from each chosen primary unit.

9.3 APPLICATIONS:

1. Whenever any process involves chemical, physical or biological tests that can be performed on a small amount of material, it is likely to be drawn as a sub sample from a larger amount which is itself a sample.
2. In crop surveys for estimating the yield of a crop in a district, villages may be considered as first stage units and the crop fields of fixed size are the 2nd stage units of sampling.

Note: Two-stage sampling can be expected to be

1) Less efficient than single stage random sampling and more efficient than cluster sampling from the sampling variability point of view.
2) More efficient than single stage random sampling and less efficient than cluster sampling from the cost and operational point of view.
3) The main advantage of this sampling, procedure is that at first stage, the sampling frame of first stage units (f s u) is required can be prepared easily. At the 2nd stage, the sampling frame of the second stage units (s s u) is required only for the selected first stage units.



Two-stage sampling (Equal size primary units)

9.4 ADVANTAGES OF TWO STAGE SAMPLING:

The principle advantage of two stage sampling is that it is more flexible than the one-stage sampling. It reduces to one stage sampling when $m = M$ but unless this is the best choice of m, we have the opportunity of taking some smaller value that appears more efficient. As usual, this choice reduces to a balance between statistical precision and cost. When units of the first stage agree very closely, then consideration of precision suggests a small value of m. On the other hand, it is sometimes as cheap to measure the whole of a unit as to a sample. For example, when the unit is a household and a single respondent can give as accurate data as all the members of the household.

Notations:

y_{ij} = value obtained for j^{th} element in the i^{th} primary unit, $\begin{matrix} i = 1,2,---,N \\ j = 1,2,---,M \end{matrix}$

$Y_i = \sum_{j=1}^{M} y_{ij} = i^{th}$ cluster total.

$Y = \text{Population total} = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$

$\overline{Y}_i = i^{th}$ cluster mean $= \frac{Y_i}{M} = \frac{\sum_{j=1}^{M} y_{ij}}{M}$

$\bar{\bar{Y}} = \text{Population mean for element} = \frac{\sum_{i=1}^{N} \overline{Y}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{NM} = \frac{Y}{NM}$

$\bar{y}_i = \sum_{j=1}^{m} \frac{y_{ij}}{m} = Sample\ mean\ per\ element\ in\ the\ i^{th}\ cluster$

$\bar{\bar{y}} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm} =$ Over all sample mean for element

$\therefore \bar{\bar{y}} = \hat{\bar{\bar{Y}}}$

$\hat{Y} = NM\bar{\bar{Y}} = NM\bar{\bar{y}}$.

$\bar{\bar{y}}$ is an estimate of $\bar{\bar{Y}}$ (population mean for element) and $\hat{Y}$ is an estimate of Y(Population total)

Both the estimates $\hat{Y}$ and $\hat{\bar{\bar{Y}}}$ are of the farm

$$y' = y_1' + y_2' + ---+ y_n' \rightarrow (1)$$

Where y_i' is an estimate made from the sub sample drawn from the i^{th} primary unit.

Let $Y_i' = E\left(y_i' / i\right) \rightarrow (2)$, where the symbol $E(\bullet/i)$ denotes a mean taken over all sub

samples drawn from the i^{th} primary unit.

If these means were know, we could construct the estimate

$$\hat{Y}' = Y_1' + Y_2' + ---+ Y_n' = \sum_{i=1}^{n} Y_i' \rightarrow (3)$$

Let $\prod_i$ denote the probability that the i^{th} primary unit is drawn into the sample.

9.5 CONCEPT OF TWO STAGE SAMPLING POPULATION MEAN:

THEOREM-1: If the primary unit are drawn without replacement and sub-samples are chosen independently in different units, y' is an unbiased estimate of $Y' = \sum_{i=1}^{N} \prod_i Y_i'$ with variance of y' is $V(y') = V(\hat{Y}') + \sum_{i=1}^{N} \prod_i \sigma_{2i}^2$, where $\sigma_{2i}^2 = E\left[\left(y_i' - Y_i'\right)^2 / i\right]$ is the variance of y_i' in repeated sub sampling from the i^{th} primary unit.

Let $v_c(y')$ be a copy of $v(\hat{Y}')$, obtained by replacing Y_i' by y_i' , whenever Y_i' appears.

THEOREM-2: Under the conditions of th-10.1, an unbiased estimate of $V(y')$ is $v(y') = v_c(y') + \sum_{i=1}^{n} \prod_i \hat{\sigma}_{2i}^2$, where $\hat{\sigma}_{2i}^2$ is any unbiased sample estimate of σ_{2i}^2.

THEOREM-3: If the n-units and the m-sub units from each chosen unit are selected by simple random sampling random sampling, $\bar{\bar{y}}$ is an unbiased estimate of $\bar{\bar{Y}}$ with variance of $\bar{\bar{y}}$ is

$$V(\bar{\bar{y}}) = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn}$$

where $S_1^2 = \frac{\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{N-1}$ = variance among primary unit mean. $S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i\right)^2}{N(M-1)}$ = Variance among elements within primary units.

Proof: In the notation of theorem-10.1;

Take $y_i' = \frac{\bar{y}_i}{n}$

Then $\bar{\bar{y}} = y' = y_1' + y_2' + ---+ y_n'$

$$\bar{\bar{y}} = \frac{\bar{y}_1}{n} + \frac{\bar{y}_2}{n} + ---+ \frac{\bar{y}_n}{n} = \frac{\sum_{i=1}^{n}\bar{y}_i}{n} = \hat{\bar{Y}}$$

Now, $Y_i' = \frac{1}{n}\bar{Y}_i$, $\hat{Y}' = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i$, $\Pi_i = \frac{n}{N}$

From theorem-1[we know that y' is an unbiased estimator of Y']

$$E(\bar{\bar{y}}) = E(y') = Y' = \sum_{i=1}^{N} \Pi_i Y_i' = \sum_{i=1}^{N} \frac{n}{N} \cdot \frac{\overline{Y}_i}{n}$$

$$\therefore \quad E(\bar{\bar{y}}) = \frac{1}{N} \sum_{i=1}^{N} \overline{Y}_i = \bar{\bar{Y}}$$

Since $\hat{Y}'$ is the mean of the n-values of $\overline{Y}_i$, using the result of single stage sampling for primary units(i.e., SRSWOR for clusters) we have

$$V(\hat{Y}') = \frac{N-n}{Nn} \sum_{i=1}^{N} \frac{(\overline{Y}_i - \bar{\bar{Y}})^2}{N-1} = \frac{N-n}{N} \cdot \frac{S_1^2}{n} \rightarrow (1)$$

Also, applying the result of SRSWOR for the element of i^{th} cluster [i.e, first stage unit (f s u)] since m-elements are selected out of M from the i^{th} primary unit,

$V(\bar{y}_i) = \frac{M-m}{M} \cdot \frac{S_{2i}^2}{m}$, where $S_{2i}^2 = \frac{\sum_{j=1}^{M} (y_{ij} - \overline{Y}_i)^2}{M-1} =$ variance among elements in the i^{th} cluster.

$$\therefore V(y_i') = V\left(\frac{y_i}{n}\right) = \frac{1}{n^2} V(\bar{y}_i) = \frac{1}{n^2} \cdot \frac{M-m}{M} \frac{S_{2i}^2}{m} = \sigma_{2i}^2 \rightarrow (2)$$

Hence, by variance of $\bar{\bar{y}}$ by using equation (1) & (2) we get

$$V(\bar{\bar{y}}) = V(\hat{Y}') + \sum_{i=1}^{N} \Pi_i \sigma_{2i}^2 \quad (\because \Pi_i = \frac{n}{N})$$

$$= \frac{N-n}{N} \cdot \frac{S_1^2}{n} + \sum_{i=1}^{N} \left(\frac{n}{N}\right) \cdot \frac{1}{n^2} \left(\frac{M-m}{M}\right) \frac{S_{2i}^2}{m}$$

$$\therefore V(\bar{\bar{y}}) = \frac{N-n}{N} \frac{S_1^2}{n} + \left(\frac{M-m}{M}\right) \frac{S_2^2}{mn} \rightarrow (3) \cdot$$

Where $S_2^2 = \sum_{i=1}^{N} \frac{S_{2i}^2}{N}$

i. If $f_1 = \frac{n}{N}$ and $f_2 = \frac{m}{M}$ are sampling fractions in the 1st and 2nd stages, a form of the result that is easier to remember.

$$V(\bar{\bar{y}}) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2 \rightarrow (4) \quad [\because \text{this is for population mean estimate}]$$

$V(\hat{Y}) = N^2 M^2 V(\bar{\bar{y}})$ [$\because \hat{Y} = NM\bar{\bar{Y}} = NM\bar{\bar{y}}$][$\because$ this is for population total estimate]

9.6 ESTIMATION OF VARIANCE:

If the 'n' primary unit means $\overline{Y}_i$ were known as unbiased estimate of the variance of their mean $\hat{Y}'$ would be $v\left(\hat{Y}'\right)=\frac{1-f_1}{n}\cdot\frac{\sum_{i=1}^{n}\left(\overline{Y}_i-\hat{Y}'\right)^2}{n-1}$

The copy of $v\left(\hat{Y}'\right)$ from a two-stage sample is $v_c\left(\bar{y}'\right)=v_c\left(\bar{\bar{y}}\right)=\frac{1-f_1}{n}\cdot\frac{\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n-1}\rightarrow(5)$

$\left[\because f_1=\frac{n}{N}\right]$

For theorem-2: we require also an unbiased estimate of σ_{2i}^2. Since sub samples are chosen by simple random sampling, this is given by $\hat{\sigma}_{2i}^2=\frac{M-m}{M}\frac{s_{2i}^2}{mn^2}=\left(1-f_2\right)\frac{s_{2i}^2}{mn^2}\rightarrow(6)$,

where $s_{2i}^2=\sum_{j=1}^{m}\frac{\left(y_{ij}-\bar{y}_i\right)^2}{m-1}$ $\left[\because f_2=\frac{m}{M}\right]$

THEOREM-4: An unbiased estimate of $V\left(\bar{\bar{y}}\right)$ is $v\left(\bar{\bar{y}}\right)=\frac{1-f_1}{n}s_1^2+\frac{f_1\left(1-f_2\right)}{mn}s_2^2\rightarrow(7)$

where $s_1^2=\frac{\sum_{i=1}^{n}\left(\bar{y}_i-\bar{\bar{y}}\right)^2}{n-1}, s_2^2=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)}{n(m-1)}\rightarrow(8)$

Proof: By theorem-10.2; an unbiased estimate of $V\left(\bar{\bar{y}}\right)$ is $v\left(\bar{\bar{y}}\right)$[from th-10.2]

$$\therefore v\left(\bar{\bar{y}}\right)=v_c\left(\bar{\bar{y}}\right)+\sum_{i=1}^{n}\Pi_i\hat{\sigma}_{2i}^2$$

Using equations (5) & (6) and $\Pi_i=\frac{n}{N}$

This gives $v\left(\bar{\bar{y}}\right)=\frac{\left(1-f_1\right)}{n}s_1^2+\frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{n}{N}\right)\left(\frac{1-f_2}{m}\right)s_{2i}^2$. But $s_2^2=\sum_{i=1}^{n}\frac{s_{2i}^2}{n}$

Hence, $v\left(\bar{\bar{y}}\right)=\frac{1-f_1}{n}s_1^2+\frac{f_1\left(1-f_2\right)}{mn}s_2^2\rightarrow(9)$ $\left[\because f_1=\frac{n}{N}, f_2=\frac{m}{M}\right]$

Equation-(9) is used for estimation of the variance of the estimate of the population total

$$v\left(\hat{Y}\right)=N^2\cdot M^2 v\left(\bar{\bar{y}}\right)$$

Note: If m=M; i.e., f_2=1, formula (9) becomes that appropriate to SRS of the units.

i.e., $v\left(\bar{\bar{y}}\right)=\frac{1-f_1}{n}s_1^2+\frac{f_1(1-1)}{mn}s_2^2$

It becomes that appropriate to SRS of units

$$v\left(\bar{\bar{y}}\right) = \frac{1-f_1}{n} s_1^2 + \frac{f_1(1-1)}{mn} s_2^2 = \frac{1-f_1}{n} s_1^2 + 0 = \frac{1-f_1}{n} s_1^2$$

9.7 SUMMARY AND CONCLUSION:

- Two-stage sampling is an efficient method for large-scale population surveys, especially when complete listings are impractical.
- It provides flexibility, cost-effectiveness, and robustness in estimation.
- The method retains the unbiasedness of estimators while allowing a practical sampling framework.
- Variance estimation is straightforward and enables proper inference with confidence intervals.

9.8 KEY WORDS:

- Two-stage sampling
- Primary Sampling Units (PSUs)
- Secondary Sampling Units (SSUs)
- Equal-size sub-sampling
- Unbiased estimator
- Between-PSU variance
- Within-PSU variance
- Estimation of variance
- Cost-effective sampling

9.9 SELF-ASSESSMENT QUESTIONS:

1. Define two-stage sampling. How does it differ from single-stage sampling?
2. What are the key features of two-stage sampling when secondary units are of equal size?
3. Explain how the population mean is estimated in a two-stage sampling design. Provide the formula.
4. List at least three real-life applications of two-stage sampling. Why is this method preferred in such cases?
5. What are the advantages of using two-stage sampling in large-scale surveys?
6. State and explain the main theorem related to the variance of the two-stage sampling estimator.
7. Derive the expression for the variance of the estimated population mean in a two-stage sampling with equal-sized SSUs.
8. How is the variance of the sample mean estimated from two-stage sampling data? Provide the estimation formula.

9. In two-stage sampling, what do the terms 'between PSU variance' and 'within PSU variance' refer to? Why are both important?
10. What assumptions are necessary for the sample mean to be an unbiased estimator of the population mean in two-stage sampling?

9.10 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. S. K. Thompson**,** Title**:** *Sampling* (Wiley Series in Probability and Statistics).
3. P. Mukhopadhyay**,** Title**:** *Theory and Methods of Survey Sampling*
4. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
5. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
6. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society.
7. Singh, D., and Chaudhary, F.S. (1986) . *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.

Dr. N.Viswam

LESSON -10

DOUBLE SAMPLING (TWO PHASE SAMPLING)

OBJECTIVES:

After completing this lesson, learners will be able to:

- **Understand the Concept of Double Sampling:** Define double sampling (also known as two-phase sampling). Distinguish between single-phase and double-phase sampling methods.
- **Apply Double Sampling in Stratified Populations:** Explain the role of stratification in improving the efficiency of estimators. Illustrate how double sampling can aid in forming effective strata when stratification information is not available initially.
- **Estimate the Population Mean Using Double Sampling:** Derive the estimator for the population mean in double sampling. Compute the variance of the estimated mean under double sampling.
- **Evaluate and Interpret the Variance:** Derive and interpret the expression for the variance of the estimator in two-phase sampling. Compare the variance of double sampling estimators with single-phase estimators to assess efficiency.
- **Understand and Apply Optimum Allocation in Double Sampling:** Learn the concept of optimum allocation for sample sizes in both phases. Derive the condition for optimum allocation that minimizes the variance for a fixed cost or minimizes cost for a fixed variance.
- **Develop Practical Understanding:** Apply double sampling methods to real-life sampling problems. Understand the trade-offs between cost and precision in survey design using two-phase sampling.
- **Use Double Sampling for Cost-Effective Data Collection:** Design a double sampling scheme where preliminary data collection is inexpensive and followed by a more detailed second phase.

STRUCTURE:

10.1 INTRODUCTION:

Double sampling, also known as two-phase sampling, is a statistical technique where data is collected in two stages to improve accuracy and efficiency. In the first stage, a large sample is taken to gather preliminary or inexpensive information. In the second stage, a smaller subset of this sample is selected for more detailed and often more costly measurements. This method is particularly useful in cases where certain measurements are expensive, time-consuming, or difficult to obtain for an entire population.

The main advantage of double sampling is that it allows researchers to make better estimates while minimizing costs. By using a large initial sample to gain general insights and a smaller second sample for precise measurements, it optimizes resource allocation. This technique is widely used in fields such as survey research, quality control, and environmental studies. For example, in quality control, a company may conduct a quick inspection of many products before selecting a smaller subset for rigorous testing. Similarly, in surveys, a broad preliminary study might be conducted to identify key characteristics before conducting in-depth interviews with a smaller group.

10.2 CONCEPT OF DOUBLE SAMPLING:

Two-Phase Sampling is also called Double sampling. In Two-phase sampling the study variables are two. They are X and Y.

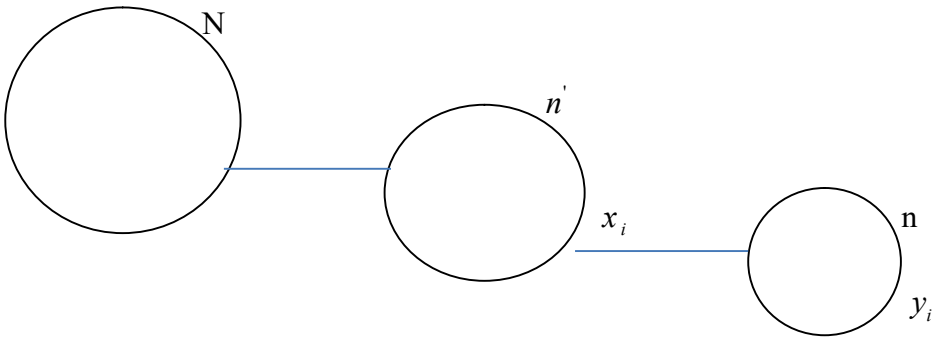
X - auxiliary variables or helping variable

Y - study variable

X - variate helps in estimating better study variate.

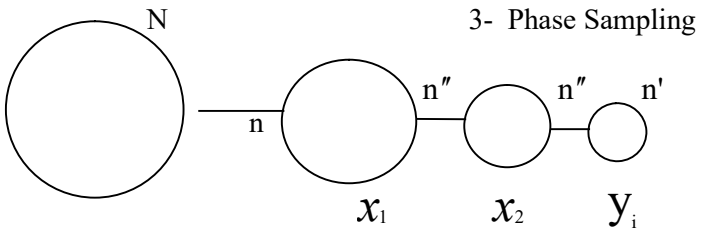
In two- stage sampling we are having only one variate Y.

In each unit we find (x_i, y_i)



Weight is x_i, Area is y_i

Example: y_i - is the household's income; x_i is number of households. Collection of information x_i is less cost, less time. In the collection of information Y_i is more cost and more time will be taken.



The method of selection consists in selecting a sample of units in the first Phase for collecting data on some, suitable and the auxiliary variables and then selecting a sub sample of these units for the main survey by utilizing the auxiliary information obtained in the first phase for arrangement, stratification and selection or for estimation. This procedure is termed as two-phase sampling or double sampling.

As an illustration, the use of double sampling may be given by considering the question of estimating the total consuming expenditure in a town through a sample survey, when only just a list of all households in the town is available, without any other particulars about the households. One procedure is to select a sample of households and collect data on consumer expenditure but such a procedure may require a rather large sample and hence the cost involved may be considerable, if there is a large variation among the households. An alternative procedure in such a case, which is likely to be more economical would be to collect data on some simple characteristics related to consumer expenditure such as household size, means of livelihood, etc. For a sample of households selected in the first phase and to use this information for arrangement, stratification and selection of the 2nd phase sample of household for the collection of data on consumer expenditure. The farmer is uni-phase sampling whereas, the latter method is two-phase sampling. It may be noted that different sampling procedures may be used at the different phases depending on the information available for the sample units.

10.3 DOUBLE SAMPLING FOR STRATIFICATION:

The population is to be stratified into a number of classes according to the values of 'x_i'. The first sample is a SRS of size 'n'. Let $W_h = \frac{N_h}{N}$ = proportion of population falling into stratum h and $w_h = \frac{n'_h}{n'}$ = proportion of 1st sample falling into stratum h, then w_h is an estimate of W_h, The second sample is a stratified random sample of size n(generally $n < n'$) in which y_i is measured : n_h units are drawn from stratum h. The second sample is often a subsample from the 1st sample but it may be drawn independently if this is more convenient.

The cost of the two samples is assumed to be $C = nC_n + n'C'_n$; where C_n is usually large in relation to C'_n.

The problem is to chose n' and the n_h (and consequently n) to minimize the variance of the estimate for a given cost. We must then verify whether the minimum variance is smaller than that can be attained by a SRS in which y_i alone is measured.

The first step is to set up the estimate and determine its variance. The population mean is

$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$ .As an estimate we use $\overline{y}_{st} = \sum_{h=1}^{L} w_h \overline{y}_h$. Whenever a new sample is drawn this implies a fresh drawing of both 1st and 2nd samples. Thus, the w_h and the sample means $\overline{y}_h$ are both random variables.

Q). Define an estimate of population mean $(\overline{Y})$ in double sampling $(\overline{y}_h)$ and show that it is unbiased.

10.4 ESTIMATION OF VARIANCE FOR DOUBLE SAMPLING:

Theorem-1: The estimate $\overline{y}_{st}$ is unbiased estimate of $\overline{Y}$

Proof: Average first over samples in which the w_h are fixed. Since $\overline{y}_h$ is the mean of a SRS from the stratum $E(\overline{y}_h) = \overline{Y}_h$, but when the average is taken over different selections of the first sample $E(w_h) = W_h$. Since the first sample is also a SRS,

Hence, $E(\overline{y}_{st}) = E\left[E\left(\sum_{h=1}^{L} w_h \overline{y}_h / w_h \right) \right]$

$= E\left[\sum_{h=1}^{L} w_h E(\overline{y}_h) \right] = E\left[\sum_{h=1}^{L} w_h \overline{Y}_h \right] = \sum_{h=1}^{L} W_h \overline{Y}_h = \overline{Y}$.

$\therefore \overline{y}_{st}$ is an unbiased estimate of $\overline{Y}$.

Theorem-2: If the first sample is random and of size n', the second sample is(drawn) a random sub-sample of the first, of size $n_h = \nu_h n_h'$, where $0 < \nu_n \leq 1$ and ν_h are fixed then, $V(\overline{y}_{st}) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_{h=1}^{L} \frac{W_h S_h^2}{n'}\left(\frac{1}{\nu_h} - 1\right)$, where s^2 is the population variance.

Proof: Suppose that the y_{hi} were measured on all n_h' first sample units in stratum h, not just on the subsample of n_h. Then, since $w_h = \frac{n_h'}{n'}; \sum_{h=1}^{L} w_h \overline{y}_h' = \overline{y}'$ is the mean of a SRS of size n,

$V(\overline{y}') = \frac{N - n'}{N}\frac{S^2}{n'} \rightarrow (1) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right)$ --------→(1)

But we know that $\overline{y}_{st} = \sum_{h=1}^{L} w_h \overline{y}_h = \sum_{h=1}^{L} w_h \overline{y}_h' + \sum_{h=1}^{L} w_h (\overline{y}_h - \overline{y}_h')$

$V(\overline{y}_{st}) = V(\overline{y}') + V\left[\sum_{h=1}^{L} w_h (\overline{y}_h - \overline{y}_h') \right] \rightarrow (2)$

Hence, for fixed w_h, $V_2\left[\sum_{h=1}^{L} w_h\left(\bar{y}_h - \bar{y}'_h\right)\right] = \sum_{h=1}^{L} w_h^2 S_h^2\left(\frac{1}{n_h} - \frac{1}{n'_h}\right)$

$$\left[\because V\left(\bar{y}_h\right) = S_h^2\left(\frac{1}{n_h} - \frac{1}{n'_h}\right) \Rightarrow \bar{y}_h - \bar{y}'_h \Rightarrow \bar{y}_h > \bar{y}'_h\right]$$

[Where S_h^2 is the variance of finite population consisting of combined domains]

$$= \sum_{h=1}^{L} w_h S_h^2 \frac{n'_h}{n'}\left(\frac{1}{n_h} - \frac{1}{n'_h}\right) = \sum_{h=1}^{L} \frac{w_h S_h^2}{n'}\left(\frac{n'_h - n_h}{n_h n'_h}\right) n'_h$$

$$= \sum_{h=1}^{L} \frac{w_h S_h^2}{n'}\left(\frac{n'_h}{n_h} - 1\right) = \sum_{h=1}^{L} \frac{w_h S_h^2}{n'}\left(\frac{1}{\nu_h} - 1\right) \rightarrow (5)$$

Where $\nu_h = \frac{n_h}{n'_h}$ and since $E\left(V_2\left(\sum_{h=1}^{L} w_h\right)\right) = \sum_{h=1}^{L} W_h$.

Averaging over the distribution of w_h, obtained b repeated selections of the first sample, we have from equations (1),(2), and(5)

$$V\left(\bar{y}_{st}\right) = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \sum_{h=1}^{L} \frac{W_h S_h^2}{n'}\left(\frac{1}{\nu_h} - 1\right) \rightarrow (6)$$

Theorem-3: If the second sample is selected independently of the first sample then(or if the values of n_h do not depend on the w_h) then

$$V\left(\bar{y}_{st}\right) = \sum_{h=1}^{L}\left\{\left[W_h^2 + \frac{g' W_h\left(1 - W_h\right)}{n'}\right]\frac{\left(1 - f_h\right)}{n_h} S_h^2 + \frac{g' \sum_{h=1}^{L} W_h\left(\bar{Y}_h - \bar{Y}\right)}{n'}\right\} \rightarrow (1)$$

Where $g' = \left((N - n')/N - 1\right)$ and $f_h = \frac{n_h}{N_h}; W_h = \frac{N_h}{N}; n_h = \nu_h n'_h$

Proof: Average first over samples in which w_h are fixed over these samples. The mean of $\bar{y}_{st}$ is $\sum_{h=1}^{L} w_h \bar{Y}_h$ so that there is a bias of amount $\sum_{h=1}^{L}\left(w_h - W_h\right)\bar{Y}_h$.

The conditional variance of $\bar{y}_{st}$ is given by (Th.5.3) $V\left(\bar{y}_{st}\right) = \sum_{h=1}^{L} \frac{w_h^2 S_h^2}{n_h}\left(1 - f_h\right)$

$$\left[\bar{y}_{st} = \sum_h W_h \bar{y}_h = \bar{Y} \rightarrow \text{stratified, but } \bar{y}_{st} = \sum_h w_h \bar{y}_h \neq \bar{Y} \rightarrow \text{in double sampling}\right]$$

Hence the mean square error is

$\text{MSE} = \text{Variance} + \left(\text{bias}\right)^2$. Difference is biased.

$$E\left[\left(\bar{y}_{st}-\bar{Y}\right)^2 / w_h\right]=\sum_{h=1}^{L}\frac{w_h^2(1-f_h)S_h^2}{n_h}+\left(\sum_{h=1}^{L}(w_h-W_h)\bar{Y}_h\right)^2$$

Taking expectation on both sides

$$E\left\{E\left[\left(\bar{y}_{st}-\bar{Y}\right)^2 / w_h\right]\right\}=\sum_{h=1}^{L}\frac{(1-f_h)S_h^2}{n_h}E(w_h^2)+E\left(\sum_{h=1}^{L}(w_h-W_h)\bar{Y}_h\right)^2 \rightarrow (2)$$

Now, average when w_h vary

$$V(w_h)=\frac{W_h(1-W_h)(N-n')}{n'(N-1)}$$

Substituting g' in the above equation

$$V(w_h)=\frac{W_h(1-W_h)g'}{n'} \rightarrow (3)$$

Now, $E(w_h^2)=\left[E(w_h)\right]^2+V(w_h)$

$$E(w_h^2)=W_h^2+\frac{g'W_h(1-W_h)}{n'} \rightarrow (4)\ [\because E(w_h)=W_h]$$

Also, $E\left[(w_h-W_h)(w_j-W_j)\right]=\text{cov}(w_h,w_j)=\frac{-g'}{n'}W_hW_j;(h\neq j) \rightarrow (5)$

Since, $\text{cov}(a_i,a_j)=E(a_i,a_j)-E(a_i)-E(a_j)=\frac{-n}{N(N-1)}\left(1-\frac{n}{N}\right)$

$$V\left(\sum_i a_iu_i\right)=\sum a_i^2V(u_i)+2\sum_i\sum_{j>1}\bar{a}_i\bar{a}_j\text{cov}(u_i,u_j)$$

Now, considering $\bar{Y}_h$ as a constant in equation (2), we write

$$E\left[\sum_{h=1}^{L}(w_h-W_h)\bar{Y}_h\right]^2=\sum_{h=1}^{L}\bar{Y}_h^2V(w_h)+2\sum\sum\bar{Y}_h\bar{y}_j\text{cov}(w_h,w_j)$$

$$=\sum_h\bar{Y}_h^2\frac{g'W_h(1-W_h)}{n'}+2\sum\sum\bar{Y}_h\bar{Y}_j\left(\frac{-g'}{n'}W_hW_j\right)$$

$$=\frac{g'}{n'}\left[\sum_h\bar{Y}_h^2W_h(1-W_h)-2\sum\sum W_hW_j\bar{Y}_h\bar{Y}_j\right]$$

$$=\frac{g'}{n'}\left[\sum W_h\bar{Y}_h^2-\sum W_h^2\bar{Y}_h^2-2\sum\sum W_hW_j\bar{Y}_h\bar{Y}_j\right] \qquad E$$

$$\left(\sum_{h=1}^{L}(w_h-W_h)\bar{Y}_h\right)^2=\frac{g}{n'}\left[\sum W_h\bar{Y}_h^2-\left(\sum_{h=1}^{L}W_h\bar{Y}_h\right)^2\right]$$

$$\left(\because\left(\Sigma a_i\right)^2=a_1^2+a_2^2+---+a_i^2+2a_1a_2+2a_1a_3+.....\right.$$

$$\left.=\sum a_i^2+2\sum\sum a_ia_j\right)$$

$$= \frac{g'}{n'}\left[\sum W_h \overline{Y}_h^2 - \overline{Y}^2\right] = \frac{g'}{n'}\left[\sum W_h \overline{Y}_h^2 - \overline{Y}^2 - 2\overline{Y}^2\right]$$

$$= \frac{g'}{n'}\left[\sum W_h \overline{Y}_h^2 + \sum_{h=1}^{L} W_h \overline{Y}^2 - 2\left(\sum W_h \overline{Y}_h\right)\overline{Y}\right]$$

$$= \frac{g'}{n'}\sum_{h=1}^{L} W_h\left(\overline{Y}_h - \overline{Y}\right)^2 \rightarrow (6)$$

Substituting equation (4) and (6) in equation (2) we get the result

$$V\left(\overline{y}_{st}\right) = \sum_{h=1}^{L}\left\{\left[W_h^2 + \frac{g' W_h\left(1 - W_h\right)}{n'}\right]\frac{\left(1 - f_h\right)}{n_h} s_h^2 + \frac{g'}{n'}\sum_{h=1}^{L} W_h\left(\overline{Y}_h - \overline{Y}\right)^2\right] \rightarrow (7)$$

10.5 APPLICATIONS OF DOUBLE SAMPLING (TWO-PHASE SAMPLING):

Double sampling, also known as **two-phase sampling**, is widely used in survey sampling, especially when some auxiliary information is available or easily collectible for a large sample, while the main variable of interest is expensive or difficult to measure. Here are the main applications:

1. Stratification When Strata Are Unknown

- Often used when strata (subgroups within a population) are not known in advance.
- In the first phase, a large sample is drawn to estimate an auxiliary variable, which is used to define strata.
- In the second phase, a smaller subsample is taken from these strata to collect information on the main study variable.

Example: In agricultural surveys, land area (auxiliary variable) can be used to stratify farms before measuring crop yield (study variable).

2. Cost Reduction

- Reduces overall survey cost by collecting inexpensive auxiliary information in the first phase.
- Only a subsample in the second phase is measured for the costly variable.

Example: In health surveys, demographic data (first phase) may be easy to collect, while medical tests (second phase) are expensive.

3. Improvement of Estimator Precision

- By using auxiliary variables correlated with the study variable, double sampling helps in constructing **ratio or regression estimators** with lower variance.

Example: In forestry, tree height (cheap and easy to measure) can be used as an auxiliary variable for estimating timber volume (harder to measure).

4. Non-response Adjustment

- The first phase can help identify and adjust for **non-response bias**.

- In the second phase, additional effort is made to collect data from non-respondents.

Example: A follow-up survey of non-respondents to a mail questionnaire to adjust bias in the original estimate.

5. Environmental and Ecological Studies

- Used where remote sensing (satellite or aerial images) provides broad first-phase data, and ground truthing (field visits) forms the second phase.

Example: Estimating forest cover using satellite imagery (first phase) and field visits to a subsample of locations (second phase).

6. Agricultural and Industrial Surveys

- Used where certain variables like acreage, manpower, or input use can be easily obtained initially.
- Production, efficiency, or income data are then gathered in a subsample.

7. Surveys Involving Sensitive Topics

- In some surveys, people are more willing to provide general data.
- The second phase focuses on gathering sensitive or personal information in a smaller, more trusted subsample.

EXAMPLE: THREE STAGE SAMPLING:

For instance, for conducting a Socio economic survey in a district, where generally household is taken as the ultimate stage unit,(i.e., element) population is district.

In this case element is household a sample of households may be selected in three stages by selecting first a sample of Mandals, then a sample of villages from each selected Mandals after making a list of all the villages in it and finally a sample of households from each selected villages after listing all the households in it. Since, the selection is done in three stages; this procedure is termed as three stage sampling. Here, Mandals are taken as first stages unit (f s u), villages as second stage units(s s u) and households as third stage units (t s u); $M_i = M$ and $m_i = m$.

Where M_i and m_i are different:[for population]

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_i = \frac{N}{n}\sum_{i=1}^{n} \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij}$$

$$V(\hat{Y}) = \frac{N^2}{n}(1-f_2)\overline{M}^2 S_b'^2 + \frac{N}{n}\frac{\sum_{i=1}^{N} M_i^2(1-f_{2i})}{m_i} S_{wi}^2$$

Where $\overline{M} = \sum_{i=1}^{N}\frac{M_i}{N}$, $f_{2i} = \frac{m_i}{M_i}$

$$S_b'^2 = \frac{\sum_{i=1}^{N}\left(\left(\frac{M_i \overline{Y}_i}{\overline{M}}\right) - \overline{\overline{Y}}\right)^2}{N-1}$$ (this becomes S_1^2 when $M_i = M$)

$$S_{wi}^2 = \frac{\sum_{i=1}^{M_i}\left(y_{ij} - \overline{Y}_i\right)^2}{(M_i - 1)}$$

$$S_w^2 = \frac{\sum_{i=1}^{N} S_{wi}^2}{N} \text{ (this becomes } S_2^2 \text{ when } M_i = M)$$

When $M_i = M$ and $m_i = m$:

$$V(\hat{Y}) = N^2 M^2 \left[(1 - f_1)\frac{S_1^2}{n} + \frac{(1 - f_2)}{nm} S_2^2\right] \text{(Population)}$$

When M_i and m_i are different: [for element of the sample units]

$$v(\hat{Y}) = \frac{N^2(1 - f_1)}{n} s_b'^2 + \frac{N}{n}\sum_{i=1}^{n} \frac{M_i^2(1 - f_{2i})}{m_i} s_{wi}^2 \qquad \left[\because f_1 = \frac{n}{N}; f_2 = \frac{m}{M}\right]$$

Where $s_b'^2 = \dfrac{\sum_{i=1}^{n}\left(M_i\overline{y}_i - \sum_{i-1}^{n}\frac{M_i\overline{y}_i}{n}\right)^2}{n-1}$

$$s_{wi}^2 = \sum_{j=1}^{m_i} \frac{\left(y_{ij} - \overline{y}_i\right)^2}{m_i - 1}$$

$$s_w^2 = \sum_{i=1}^{m} \frac{s_{wi}^2}{n}$$

When $M_i = M$ and $m_i = m$:

$$v(\hat{Y}) = N^2 M^2 \left[\frac{(1 - f_1)}{n} s_1^2 + \frac{f_1(1 - f_2)}{nm} s_2^2\right] \text{ (sample).}$$

10.6 OPTIMUM ALLOCATION IN DOUBLE SAMPLING:

We know that,

$$V(\overline{y}_{st}) = \sum_{h=1}^{L}\left\{\left[w_h^2 + \frac{g' w_h(1 - w_h)}{n}\right]\frac{(1 - f_h)}{n_h} s_h^2 + \frac{g'}{n'}\sum_{h=1}^{L} W_h\left(\overline{Y}_h - \overline{Y}\right)^2\right\}$$

In most applications $f_h = \dfrac{n_h}{N_h}$ will be negligible frequently, $\dfrac{n'}{N}$ is also small. So that g' can be replaced by unity.

$$g' = \frac{N - n'}{N - 1} = \frac{1 - \frac{n'}{N}}{1 - \frac{1}{N}} \cong 1$$

Ignoring the term $W_h(1 - W_h)$ and assuming $f_h, \dfrac{n'}{N}$ are negligible.

We obtain n,

$$V\left(\bar{y}_{st}\right)=\sum_{h=1}^{L}\frac{W_h^2S_h^2}{n_h}+\sum_{h=1}^{L}\frac{W_h}{n'}\left(\bar{Y}_h-\bar{Y}\right)^2\rightarrow(1)$$

Substitute $n_h=\frac{n\,W_hS_h}{\sum_{h=1}^{L}W_hS_h}$ in equation (1) we get [In Stratified Random Sampling $V\left(\bar{y}_{st}\right)$ is minimized for a fixed total size of sample 'n' if $n_h=\frac{n\,W_hS_h}{\sum_{h=1}^{L}W_hS_h}$]

$$V_{opt}=\sum_{h=1}^{L}\frac{W_h^2S_h^2}{nW_hS_h}\times\sum_{h=1}^{L}W_hS_h+\sum_{h=1}^{L}\frac{W_h}{n'}\left(\bar{Y}_h-\bar{Y}\right)^2$$

$$V_{opt}=\frac{\left(\Sigma W_hS_h\right)^2}{n}+\frac{\Sigma W_h\left(\bar{Y}_h-\bar{Y}\right)^2}{n'}$$

$$V_{opt}=\frac{V}{n}+\frac{V'}{n'}\rightarrow(2)\ \text{Where}\ V=\left(\Sigma W_hS_h\right)^2\ \text{and}\ V'=\Sigma W_h\left(\bar{Y}_h-\bar{Y}\right)^2.$$

This approximation expression for the variance is now minimized by choice of n and n' for a given cost of

$$C_0=nC+n'C'\rightarrow(3)$$

Write $MSE\left(\hat{\bar{Y}}_{St}\right)=V_{opt}=\frac{V}{n}+\frac{V'}{n'}$ where V and V' contains all the terms containing n and n' respectively.

The cost function is $C_0=nC+n'C'$ where c and C' are the cost per unit for selecting the samples n and n' respectively. Now, we find the optimum sample sizes n and n' for fixed cost C_0. The Lagrangian function is

$$\varphi=\frac{V}{n}+\frac{V'}{n'}+\lambda\left(nC+n'C'-C_0\right)$$

$$\frac{\partial\varphi}{\partial n}=0\Rightarrow\lambda C=\frac{V}{n^2}$$

$$\frac{\partial\varphi}{\partial n'}=0\Rightarrow\lambda C'=\frac{V'}{n'^2}$$

Thus $\lambda C\,n^2=V$ or $n=\sqrt{\frac{V}{\lambda C}}$ or $\sqrt{\lambda}\,nC=\sqrt{VC}$

Similarly $\sqrt{\lambda}\,n'C'=\sqrt{V'C'}$

Thus, $$\sqrt{\lambda}=\frac{\sqrt{VC}+\sqrt{V'C'}}{C_0}$$

And so Optimum $$n=\frac{C_0}{\sqrt{VC}+\sqrt{V'C'}},\sqrt{\frac{V}{C}}=n_{opt},\text{say}$$

$$\text{Optimum } n' = \frac{c_0}{\sqrt{VC} + \sqrt{V'C'}}, \sqrt{\frac{V'}{C'}} = n'_{opt}, \text{say}$$

$$V_{opt}\left(\hat{\bar{Y}}_{St}\right) = \frac{V}{n_{opt}} + \frac{V'}{n'_{opt}}$$

$$= \frac{\left(\sqrt{VC} + \sqrt{V'C'}\right)^2}{C_0}$$

10.7 SUMMARY:

Double sampling enhances sampling efficiency by collecting information in two phases. It supports:

- Stratification when classification variables are unavailable,
- Cost reduction with acceptable precision,
- Improved estimators through the use of auxiliary variables.

Key aspects include designing two-phase samples, computing variances, and determining optimum sample sizes based on cost and variance trade-offs.

10.8 KEY WORDS:

- Double Sampling / Two-phase Sampling
- Auxiliary Variable
- Stratification
- Subsampling
- First-phase Sample / Second-phase Sample
- Variance Estimation
- Optimum Allocation

10.9 SELF-ASSESSMENT QUESTIONS:

1. What is double sampling (two-phase sampling)? Explain the purpose and process of double sampling.
2. What are the advantages of using double sampling for stratification over direct stratification?
3. Derive the expression for the variance of the estimated mean in double sampling.
4. Explain how the estimated variance changes with the sample sizes in the first and second phases.
5. Define optimum allocation in the context of double sampling.
6. Obtain the variance of an estimate for the population mean under double sampling with SRSWR at the first stage and SRSWR at the second stage.
7. Write briefly about Two Phase sampling for stratification.

10.10 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. S. K. Thompson**,** Title**:** *Sampling* (Wiley Series in Probability and Statistics).
3. P. Mukhopadhyay**,** Title**:** *Theory and Methods of Survey Sampling*
4. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
5. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
6. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society.
7. Singh, D., and Chaudhary, F.S. (1986) . *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.
8. Tikkiwal, B.D."Theory of Sample Surveys"

Dr. N.Viswam

LESSON -11
MULTI PHASE SAMPLING

OBJECTIVES:

By the end of this lesson, the learner will be able to:

- Understand the concept and importance of multiphase sampling in survey methodology and its distinction from single-phase sampling.
- Explain the procedure of double (two-phase) sampling**,** a specific case of multiphase sampling.
- Apply double sampling for difference estimation**,** including formulation, assumptions, and interpretation.
- Use double sampling for ratio estimation effectively in appropriate survey contexts.
- Compute unbiased estimates and estimate variances under difference and ratio estimation models in double sampling.
- Compare the efficiency of double sampling methods with conventional single-phase approaches.
- Identify practical situations where multiphase sampling improves cost-efficiency and accuracy.

STRUCTURE:

11.1 INTRODUCTION:

In sample surveys the information on an auxiliary variate x is required many times, either for estimation or for selection or stratification to increase the efficiency of the estimator. When such information is lacking and it is relatively cheaper to obtain information on x, we can consider taking a large preliminary sampling for estimating $\bar{x}$ on distribution of x as the case may be, and only a small sample (sometimes a sub sample) for measuring the y-variate, the character of interest for estimation. This could mean to devoting a part of the resources to this large preliminary sample and, therefore, reduction in sample size for

measuring the study variate. This technique is known as Double sampling (or) Two-phase sampling and was proposed for the first time by Neyman (1938). When the sample for the main survey is selected in three or more phases, the sampling procedure is termed as multiphase sampling.

The difference between Multiphase sampling and multistage sampling procedures is that in multiphase sampling it is necessary to have a complete sampling frame of the units whereas in multistage sampling, a sampling frame of the next stage units is necessary only for the sample units selected at the stage. This design is advantageous when the gain in precision is substantial as compared to the increase in cost due to collection of information on the auxiliary variate for large samples.

A multiphase sampling is a sampling procedure in which it collects some information from the whole unit sample and additional information also is collected, at the same time or by later. Usually the additional information is collected to provide more detail information about the sample. The multiphase sampling is known as “two-phase sampling” where double or more phase sampling procedures can be done in at the same time or by later. The first sampling is to collect “basic information” from a large sample of unit and then followed by the second sampling collects more about “detailed information”.

As instance, there is a situation where a man is responsible to carry out a health survey on participants regarding some basic question about their diet, smoking habits, exercise routines and alcohol consumption. Mean while, another survey is required to collects detailed information of the respondent by as asking them to perform physical tests such as running on a treadmill or having their blood pressure and cholesterol level to be measured by filling out the questionnaires and interviewing participant is relatively economized procedure. Hence, the best approach to conduct this survey by approaching this two-phase sampling. In the first phase, the interviews are performed on an appropriately sized sample. Then a smaller sample is drawn from that sample. The second sample will be continued in the medical test.

11.2. MULTISTAGE SAMPLING:

Multistage sampling divides large populations into stages to make the sampling process more practical. A combination of stratified sampling or cluster sampling and simple random sampling is usually used.

Advantages and Disadvantages:

Multistage sampling is flexible, cost effective and easy to implement. You can use as many stages as you need to reduce the sample to a workable size, with no restrictions on how you divide the groups.

However, as the method has a subjective component, it has problems with external validity. It is also less accurate than simple random sampling.

11.3 DIFFERENCE BETWEEN MULTI-STAGE AND MULTI-PHASE SAMPLING:

Multi-stage	Multi-phase
1. The procedure of two-stage sampling can be generalized to two or more stages. Then it is termed as Multi-stage sampling.	1. When the sample for the main survey is selected to three or more phases. Then the sample procedure is termed as Multi-phase sampling.
2. In Multi-stage sampling the sampling frame of next stage units are necessary only for sample units are selected at the stage.	2. In Multi - phase sampling, it is necessary to have a complete sampling frame of the units.
3. This design is advantages when the gain in precision is substantial.	3. Where as in this design compared to increase in cost due to collection of information on auxiliary variate for large samples.
4. Multi-stage is very useful in practice this is most feasible procedure is being used in large scale surveys.	4. The use of Multi-phase sampling may be given by question of estimating total consumer expenditure in a town through the sample surveys.
5. Multi-stage for instance of conducting a socio-economic survey in district, here selection is done in three stages. Here mandals are taken as first stage units, villages are second stage units and households are ultimate stage units.	5. In Multi-phase sampling for instance the question of estimating the total consumer expenditure in a town through sample survey only list of households are available but any particular one procedure is select some simple characteristics related to consumer expenditure.

11.4 DOUBLE SAMPLING FOR DIFFERENCE ESTIMATOR:

A difference estimator for estimating the population mean $\bar{Y}$ where information on x is not available in advance and it is considered important to use the auxiliary variate to derive more precise estimate, is discussed here. A priliminary random sample WOR, of size 'n' is taken and the information on x is collected.

A sub sample of size 'n' is drawn WOR from the preliminary sample and information on y is measured. The difference estimator of $\bar{y}$ may be defined by

$$\overline{y_{dd}} = \bar{y} + \beta\,(\overline{x'} - \bar{x})$$

Where 'β' is the population

$\bar{y}$, $\bar{x}$ are the sub-sample means for y and x respectively.

$\overline{x'}$ is the preliminary sample mean of x.

Theorem-1:

Show that $\overline{y_{dd}}$ is an unbiased estimator of the population mean , its sampling variance is given by

$$V(\overline{y_{dd}}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)(s_y^2 + \beta^2 s_x^2 - 2p\beta s_x^1 s_y^1)$$

Proof:

Given the first sample, let $\bar{y}$ be the mean value.

$$E\left(\frac{\overline{y_{dd}}}{\bar{X}'}\right) = E\{(\bar{y} + \beta(\bar{X}' - \bar{X}))/\bar{X}'\}$$

$$= \bar{y}'$$

$$E(\bar{y}') = \bar{y}$$
$$E(\overline{y_{dd}}) = \bar{y}$$

This shows that the estimator is unbiased. For the sampling variance in relation th-2 can be written as

$$V(\overline{y_{dd}}) = V_1E_2(\overline{y_{dd}}/\bar{y}') + E_1V_2(\overline{y_{dd}}/\bar{y}')$$

Here

$$V_1E_2(\overline{y_{dd}}/\bar{y}') = (\frac{1}{n'}-\frac{1}{N})s_y^2$$

$$E_1V_2(\overline{y_{dd}}/\bar{y}') = E_1(\frac{1}{n}-\frac{1}{n'})\frac{\sum_{n=1}^{N}(y_1+\beta x_1-\bar{y}'+\beta\bar{X}')}{(n'-1)}$$

$$= (\frac{1}{n}-\frac{1}{n'})\frac{\sum_{n=1}^{N}(y_1-\beta x_1+\bar{y}+\beta\bar{x})^2}{(N-1)}$$

$$= (\frac{1}{n}-\frac{1}{n'})(s_y^2+\beta^2 s_x^2-2p\beta s_x s_y)$$

Combining both results, we prove the theorem

Corollary-1:

An unbiased estimator of the sampling variance can be written as

$$V(\overline{y_{dd}}) = (\frac{1}{n}-\frac{1}{N})s_y^2 + (\frac{1}{n}-\frac{1}{n'})s_d^2$$

Where $s_y^2 = \frac{\sum_{n=1}^{N}(y_1-\bar{y})^2}{(n-1)}$

$$s_d^2 = \frac{\sum_{n=1}^{N}[y_1-\bar{y}-\beta(x_1-\bar{x})]^2}{(n-1)}$$

Corollary-2:

If a direct random sample is taken without using the doubt sampling procedure, the sample size for the same cost obtained by $C = a + nc + nc'$

$n_0 = \frac{C}{c} = n + \frac{a+n'c'}{c}$ and the sampling variance of sample mean will be

$V(\bar{y}_d) = (\frac{1}{n_0}-\frac{1}{N})s_y^2$

Corollary-3:

Taking $\beta = k \cdot S_y/S_x$ the condition that double sampling is more precise than a precise than a direct random sampling will be obtained by

$$2P > \left[K\left(1-\frac{n}{n^1}\right)\left(1+\frac{nc}{n'c'}\right)\right]^{-1}$$

A method using – auxiliary information in the first sample has been discussed by DesRaj (1965) showing how this information may be used for achieving the higher precision by applying the double sampling techniques.

11.5 DOUBLE SAMPLING FOR RATIO ESTIMATOR:

If the population mean $\bar{X}$ is not known, then the double sampling technique is applied. Take a large initial sample of size n' by SRSWOR to estimate the population mean $\bar{X}$ as

$$\hat{\bar{X}} = \bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i .$$

Then a second sample is a subsample of size n selected from the initial sample by SRSWOR. Let $\bar{y}$ and $\bar{x}$ be the means of y and x based on the subsample. Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$.

The ratio estimator under double sampling now becomes

$$\hat{\bar{Y}}_{Rd} = \frac{\bar{y}}{\bar{x}}\bar{x}' .$$

The exact expressions for the bias and mean squared error of $\hat{\bar{Y}}_{Rd}$ are difficult to derive. So we find their approximate expressions using the same approach mentioned while describing the ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}, \quad \varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}, \quad \varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_0) = E(\varepsilon_1) = E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2$$

$$E(\varepsilon_1\varepsilon_2) = \frac{1}{\bar{X}^2} E(\bar{x} - \bar{X})(\bar{x}' - \bar{X})$$

$$= \frac{1}{\bar{X}^2} E_1\left[E_2(\bar{x} - \bar{X})(\bar{x}' - \bar{X}) \mid n'\right]$$

$$= \frac{1}{\bar{X}^2} E_1\left[(\bar{x}' - \bar{X})^2\right]$$

$$= \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}$$

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

$$= E(\varepsilon_2^2).$$

$$
\begin{aligned}
E(\varepsilon_0\varepsilon_2) &= \frac{1}{\bar{X}\bar{Y}} Cov(\bar{y}, \bar{x}') \\
&= \frac{1}{\bar{X}\bar{Y}} Cov\left[E(\bar{y} \mid n'), E(\bar{x}' \mid n')\right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}, \bar{x}') \mid n'\right] \\
&= \frac{1}{\bar{X}\bar{Y}} Cov\left[\bar{Y}, \bar{X}\right] + \frac{1}{\bar{X}\bar{Y}} E\left[Cov(\bar{y}', \bar{x}')\right] \\
&= \frac{1}{\bar{X}\bar{Y}} Cov\left[(\bar{y}', \bar{x}'\right] \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}} \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}} \\
&= \left(\frac{1}{n'} - \frac{1}{N}\right) \rho C_x C_y
\end{aligned}
$$

where $\bar{y}'$ is the sample mean of $y's$ based on the sample size n'.

$$
\begin{aligned}
E(\varepsilon_0\varepsilon_1) &= \frac{1}{\bar{X}\bar{Y}} Cov(\bar{y}, \bar{x}) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}} \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \rho \frac{S_x}{\bar{X}} \frac{S_y}{\bar{Y}} \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \rho C_x C_y
\end{aligned}
$$

$$
\begin{aligned}
E(\varepsilon_0^2) &= \frac{1}{\bar{Y}^2} Var(\bar{y}) \\
&= \frac{1}{\bar{Y}^2} \left[V_1\left\{E_2(\bar{y} \mid n')\right\} + E_1\left\{V_2(\bar{y}_n \mid n')\right\}\right] \\
&= \frac{1}{\bar{Y}^2} \left[V_1(\bar{y}_n') + E_1\left\{\left(\frac{1}{n} - \frac{1}{n'}\right) s_y'^2\right\}\right] \\
&= \frac{1}{\bar{Y}^2} \left[\left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) S_y^2\right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2}{\bar{Y}^2} \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) C_y^2
\end{aligned}
$$

where $s_y'^2$ is the mean sum of squares of y based on an initial sample of size n'.

$$E(\varepsilon_1\varepsilon_2) = \frac{1}{\bar{X}^2} Cov(\bar{x}, \bar{x}')$$

$$= \frac{1}{\bar{X}^2}\left[Cov\{E(\bar{x} \mid n'), E(\bar{x}' \mid n')\} + 0 \right]$$

$$= \frac{1}{\bar{X}^2} Var(\bar{X}')$$

where $Var(\bar{X}')$ is the variance of mean of x based on an initial sample of size n'.

11.6 ESTIMATION ERROR & BIAS:

Estimation error of $\hat{\bar{Y}}_{Rd}$

Write $\hat{\bar{Y}}_{Rd}$ as

$$\hat{\bar{Y}}_{Rd} = \frac{(1+\varepsilon_0)}{(1+\varepsilon_1)}(1+\varepsilon_2)\frac{\bar{Y}}{\bar{X}}\bar{X}$$

$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1+\varepsilon_1)^{-1}$$

$$= \bar{Y}(1+\varepsilon_0)(1+\varepsilon_2)(1-\varepsilon_1+\varepsilon_1^2-\ldots)$$

$$\simeq \bar{Y}(1+\varepsilon_0+\varepsilon_2+\varepsilon_0\varepsilon_2-\varepsilon_1-\varepsilon_o\varepsilon_1-\varepsilon_1\varepsilon_2+\varepsilon_1^2)$$

up to the terms of order two. Other terms of degree higher than two are assumed to be negligible.

Bias of $\bar{Y}_{Rd}$

$$E(\hat{\bar{Y}}_{Rd}) = \bar{Y}\left[1+0+0+E(\varepsilon_0\varepsilon_2)-0-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

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

$$= \bar{Y}\left[E(\varepsilon_0\varepsilon_2)-E(\varepsilon_0\varepsilon_1)-E(\varepsilon_1\varepsilon_2)+E(\varepsilon_1^2)\right]$$

$$= \bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y - \left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2 + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2\right]$$

$$= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)\left(C_x^2 - \rho C_x C_y\right)$$

$$= \bar{Y}\left(\frac{1}{n}-\frac{1}{n'}\right)C_x(C_x - \rho C_y).$$

The bias is negligible if n is large and relative bias vanishes if $C_x^2 = C_{xy}$, i.e., the regression line passes through the origin.

MSE of $\hat{\bar{Y}}_{Rd}$:

$$MSE(\hat{\bar{Y}}_{Rd}) = E(\hat{\bar{Y}}_{Rd} - \bar{Y})^2$$

$$\simeq \bar{Y}^2 E(\varepsilon_0 + \varepsilon_2 - \varepsilon_1)^2 \quad \text{(retaining the terms upto order two)}$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_1\varepsilon_2\right]$$

$$= \bar{Y}^2 E\left[\varepsilon_0^2 + \varepsilon_1^2 + \varepsilon_2^2 + 2\varepsilon_0\varepsilon_2 - 2\varepsilon_0\varepsilon_1 - 2\varepsilon_2^2\right]$$

$$= \bar{Y}^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2 + \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2 - \left(\frac{1}{n'}-\frac{1}{N}\right)C_x^2 + 2\left(\frac{1}{n'}-\frac{1}{N}\right)\rho C_x C_y - 2\left(\frac{1}{n}-\frac{1}{N}\right)\rho C_x C_y\right]$$

$$= \bar{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(C_x^2 + C_y^2 - 2\rho C_x C_y\right) + \bar{Y}^2\left(\frac{1}{n'}-\frac{1}{N}\right)C_x(2\rho C_y - C_x)$$

$$= MSE\text{(ratio estimator)} \quad + \bar{Y}^2\left(\frac{1}{n'}-\frac{1}{N}\right)\left(2\rho C_x C_y - C_x^2\right).$$

The second term is the contribution of the second phase of sampling. This method is preferred over the ratio method if

$$2\rho C_x C_y - C_x^2 < 0$$

or $$\rho < \frac{1}{2}\frac{C_x}{C_y}$$

Choice of n and n'

Write

$$MSE(\hat{\bar{Y}}_{Rd}) = \frac{V}{n} + \frac{V'}{n'}$$

where V and V' contain all the terms containing n and n' resp

The cost function is $C_0 = nC + n'C'$ where C and C' are the co and n' respectively.

Now we find the optimum sample sizes n and n' for fixed cost

$$\varphi = \frac{V}{n} + \frac{V'}{n'} + \lambda(nC + n'C' - C_0)$$

$$\frac{\partial\varphi}{\partial n} = 0 \Rightarrow \lambda C = \frac{V}{n^2}$$

$$\frac{\partial\varphi}{\partial n'} = 0 \Rightarrow \lambda C' = \frac{V'}{n'^2}.$$

Thus $\lambda C n^2 = V$

or $n = \sqrt{\frac{V}{\lambda C}}$

or $\sqrt{\lambda}\, nC = \sqrt{VC}.$

Similarly $\sqrt{\lambda}\, n'C' = \sqrt{V'C'}.$

Thus

$$\sqrt{\lambda} = \frac{\sqrt{VC} + \sqrt{V'C'}}{C_0}$$

and so

$$\text{Optimum } n = \frac{C_0}{\sqrt{VC} + \sqrt{V'C'}} \sqrt{\frac{V}{C}} = n_{opt}, \text{ say}$$

$$\text{Optimum } n' = \frac{C_0}{\sqrt{VC} + \sqrt{V'C'}} \sqrt{\frac{V'}{C'}} = n'_{opt}, \text{ say}$$

$$Var_{opt}(\bar{\hat{Y}}_{Rd}) = \frac{V}{n_{opt}} + \frac{V'}{n'_{opt}}$$

$$= \frac{(\sqrt{VC} + \sqrt{V'C'})^2}{C_0}$$

Comparison with SRS

If X is ignored and all resources are used to estimate $\bar{Y}$ by $\bar{y}$, then required sample size $= \frac{C_0}{C}$.

$$Var(\bar{y}) = \frac{S_y^2}{C_0 / C} = \frac{CS_y^2}{C_0}$$

$$\text{Relative effiiency} = \frac{Var(\bar{y})}{Var_{opt}(\bar{\hat{Y}}_{Rd})} = \frac{CS_y^2}{(\sqrt{VC} + \sqrt{V'C'})^2}$$

11.7 SUMMARY:

- Multiphase sampling allows step-wise data collection, starting with inexpensive information and refining estimates with detailed data from subsamples.
- It is particularly effective when auxiliary data is available **or** complete data is expensive to obtain.
- Double sampling is a special case of two-phase sampling, with difference and ratio estimators enhancing precision.

- Careful selection of auxiliary variables and sample sizes ensures unbiased, efficient estimation.
- Compared to multistage sampling**,** multiphase sampling keeps the same sampling units, focusing on variable complexity rather than population hierarchy.

Multiphase sampling is a powerful strategy in modern surveys, balancing cost**,** accuracy**,** and data richness**.**

11.8 KEYWORDS:

- Multiphase sampling
- Double sampling
- First-phase sample
- Second-phase sample
- Auxiliary variable
- Difference estimator
- Ratio estimator
- Preliminary information
- Improved estimation
- Cost-effective sampling
- Precision improvement
- Sampling phases
- Estimation bias
- Efficiency gain
- Correlation with auxiliary variable
- Estimator of population mean
- Estimator of population total
- Variance reduction
- Sampling error
- Non-response adjustment

11.9 SELF-ASSESSMENT QUESTIONS:

1. Define Multiphase Sampling**.** Explain its purpose and advantages in statistical surveys. Illustrate with suitable examples.
2. Distinguish between Multiphase Sampling and Multistage Sampling**.** Provide at least two examples to highlight the practical differences in application.

3. Explain the procedure of Double Sampling for the Difference Estimator**.** Derive the estimator and discuss its variance. Under what conditions is this method preferred?
4. Describe Double Sampling for the Ratio Estimator**.** Derive the expression for the ratio estimator and its variance under double sampling. When is this method more efficient than simple random sampling?
5. What is Estimation Error and Bias in the context of multiphase sampling? How can they be reduced? Illustrate with appropriate statistical expressions.
6. Discuss in detail the advantages and limitations of using double sampling techniques in large-scale surveys.
7. Compare and contrast the Efficiency of double sampling with that of single-phase sampling using both difference and ratio estimators**.** Support your answer with formulas.
8. Discuss the practical considerations (cost, time, availability of auxiliary variables, etc.) involved in choosing between single-phase, multiphase, and multistage sampling in large-scale surveys.

11.10 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. P. Mukhopadhyay**,** Title**:** *Theory and Methods of Survey Sampling*
3. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
4. Des Raj and Chandhok, P. (1998) *Sampling Theory*, Narosa Publishing House.
5. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society.
6. Singh, D., and Chaudhary, F.S. (1986). *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.

Dr. U. Ramkiran

LESSON -12

DOUBLE SAMPLING FOR REGRESSION ESTIMATOR

OBJECTIVES:

After completing this lesson, learners will be able to:

- **Understand the concept of double sampling for regression estimator**: Define and explain the purpose of regression estimation in survey sampling.
- Describe the methodology and advantages of using double (two-phase) sampling in regression estimation.
- Derive unbiased estimators and estimate variances using auxiliary variables in two-phase sampling.
- **Apply the principles of optimum allocation in sampling**: Understand the need for optimum allocation under cost constraints and fixed population structures.
- Derive formulas for optimum allocation in double sampling (e.g., Neyman allocation).
- Apply optimum allocation techniques to minimize variance or cost in practical survey designs.
- **Understand varying probability sampling techniques**: Explain the concept of Probability Proportional to Size (PPS) and unequal probability sampling.
- Discuss and apply Horvitz–Thompson estimator and its properties.
- Compare different selection methods (e.g., cumulative total method, Lahiri's method) for varying probability sampling.
- **Evaluate estimation efficiency**:
- Compare regression, ratio, and difference estimators under double sampling.
- Analyze the impact of auxiliary variable correlation on the efficiency of regression estimators.
- Use simulations or examples to assess estimator performance.

STRUCTURE:

12.1 INTRODUCTION:

In survey sampling, precision and cost-efficiency are critical. To achieve both, advanced sampling techniques like double sampling**,** optimum allocation**,** and varying probability sampling are used. These techniques improve the reliability of estimators and reduce survey costs, especially when auxiliary information is available or when complete frames are difficult to obtain in the first phase.

Double Sampling for Regression Estimator

Double sampling (or two-phase sampling) involves selecting a large preliminary sample in the first phase to collect auxiliary information, followed by a smaller second-phase sample to collect the main variable of interest. When a linear relationship exists between the study variable YYY and an auxiliary variable XXX, the regression estimator in double sampling improves the efficiency of estimates. This is especially useful when XXX is easy to obtain but YYY is costly or time-consuming to measure.

Optimum Allocation in Double Sampling

Optimum allocation is a strategy used to allocate sample sizes among strata (or phases) to minimize variance for a fixed total cost, or to minimize cost for a fixed precision level. In the context of double sampling, optimum allocation determines the best allocation of sample sizes between the first and second phases by taking into account cost functions, variances, and correlation between variables. It ensures resources are used efficiently to achieve reliable estimates.

Varying Probability Sampling

In many practical situations, selecting units with probabilities proportional to size (PPS) or other known characteristics (e.g., revenue, population size) can enhance efficiency. Varying probability sampling refers to such designs where each unit in the population has a known but unequal probability of selection. This method is especially useful when some units contribute more information than others. Proper estimation techniques must be used to account for the unequal selection probabilities and maintain unbiasedness.

12.2 DOUBLE SAMPLING FOR REGRESSION ESTIMATOR:

When the population mean of the auxiliary variable $\bar{X}$ is not known, then double sampling is used as follows:

- A large sample of size n' is taken from of the population by SRSWOR from which the population mean $\bar{X}$ is estimated as $\bar{x}'$, i.e. $\hat{\bar{X}} = \bar{x}'$.
- Then a subsample of size n is chosen from the larger sample and both the variables x and y are measured from it by taking $\bar{x}'$ in place of $\bar{X}$ and treat it as if it is known.

Then $E(\bar{x}') = \bar{X}$, $E(\bar{x}) = \bar{X}$, $E(\bar{y}) = \bar{Y}$. The regression estimate of $\bar{Y}$ in this case is given by

$$\hat{\bar{Y}}_{regd} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$$

where $\hat{\beta} = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is an estimator of $\beta = \frac{S_{xy}}{S_x^2}$ based on the sample of size n.

It is difficult to find the exact properties like bias and mean squared error of $\hat{\bar{Y}}_{regd}$, so we derive the approximate expressions.

Let

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

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

$$\varepsilon_2 = \frac{\bar{x}' - \bar{X}}{\bar{X}} \Rightarrow \bar{x}' = (1+\varepsilon_2)\bar{X}$$

$$\varepsilon_3 = \frac{s_{xy} - S_{xy}}{S_{xy}} \Rightarrow s_{xy} = (1+\varepsilon_3)S_{xy}$$

$$\varepsilon_4 = \frac{s_x^2 - S_x^2}{S_x^2} \Rightarrow s_x^2 = (1+\varepsilon_4)S_x^2$$

$$E(\varepsilon_1) = 0, E(\varepsilon_2) = 0, E(\varepsilon_3) = 0, E(\varepsilon_4) = 0$$

Define

$$\mu_{21} = E\left[(\bar{x} - \bar{X})^2(y - \bar{Y})\right]$$

$$\mu_{30} = E\left[\bar{x} - \bar{X}\right]^3$$

Estimation error:

Then

$$\hat{\bar{Y}}_{regd} = \bar{y} + \hat{\beta}(\bar{x}' - \bar{x})$$

$$= \bar{y} + \frac{S_{xy}(1+\varepsilon_3)}{S_x^2(1+\varepsilon_4)}(\varepsilon_2 - \varepsilon_1)\bar{X}$$

$$= \bar{y} + \bar{X}\frac{S_{xy}}{S_x^2}(1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1+\varepsilon_4)^{-1}$$

$$= \bar{y} + \bar{X}\beta(1+\varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - \ldots)$$

Retaining the powers of $\varepsilon's$ up to order two assuming $|\varepsilon_3|<1$, (using the same concept as detailed in the case of ratio method of estimation)

$$\hat{\bar{Y}}_{regd} \cong \bar{y} + \bar{X}\beta(\varepsilon_2 + \varepsilon_2\varepsilon_3 - \varepsilon_2\varepsilon_4 - \varepsilon_1 - \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4).$$

12.3 BIAS & MEAN SQUARE ERROR:

Bias:

The bias of $\hat{\bar{Y}}_{regd}$ upto the second order of approximation is

$$E(\hat{\bar{Y}}_{regd}) = \bar{Y} + \bar{X}\beta\left[E(\varepsilon_2\varepsilon_3) - E(\varepsilon_2\varepsilon_4) - E(\varepsilon_1\varepsilon_3) + E(\varepsilon_1\varepsilon_4)\right]$$

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

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)\right]$$

$$-\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x}' - \bar{X})(s_x^2 - S_x^2)}{\bar{X}S_x^2}\right)$$

$$-\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_{xy} - S_{xy})}{\bar{X}S_{xy}}\right)$$

$$+\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N}\sum\left(\frac{(\bar{x} - \bar{X})(s_x^2 - S_x^2)}{\bar{X}S_x^2}\right)$$

$$= \bar{X}\beta\left[\left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} - \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2} - \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{21}}{\bar{X}S_{xy}} + \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\mu_{30}}{\bar{X}S_x^2}\right]$$

$$= -\beta\left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\mu_{21}}{S_{xy}} - \frac{\mu_{30}}{S_x^2}\right).$$

Mean squared error:

$$MSE(\hat{\bar{Y}}_{regd}) = E(\bar{Y}_{regd} - \bar{Y})^2$$

$$= \left[\bar{y} + \hat{\beta}(\bar{X}' - \bar{x}) - \bar{Y}\right]^2$$

$$= E\left[(\bar{y} - \bar{Y}) + \bar{X}\beta(1 + \varepsilon_3)(\varepsilon_2 - \varepsilon_1)(1 - \varepsilon_4 + \varepsilon_4^2 - \ldots)\right]^2$$

Retaining the powers of $\varepsilon's$ up to order two, the mean squared error up to the second order of approximation is

$$
\begin{aligned}
MSE(\hat{\bar{Y}}_{regd}) &= E\left[(\bar{y}-\bar{Y})+\bar{X}\beta(\varepsilon_2+\varepsilon_2\varepsilon_3-\varepsilon_2\varepsilon_4-\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4)\right]^2 \\
&\simeq E(\bar{y}-\bar{Y})^2+\bar{X}^2\beta^2E(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2)+2\bar{X}\beta E[(\bar{y}-\bar{Y})(\varepsilon_1-\varepsilon_2)] \\
&= E(\bar{y}-\bar{Y})^2+\bar{X}^2\beta^2E(\varepsilon_1^2+\varepsilon_2^2-2\varepsilon_1\varepsilon_2)+2\bar{X}\bar{Y}\beta E[\varepsilon_0(\varepsilon_1-\varepsilon_2)] \\
&= Var(\bar{y})+\bar{X}^2\beta^2\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}+\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}-2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_x^2}{\bar{X}^2}\right] \\
&\quad -2\beta\bar{X}\bar{Y}\left[\left(\frac{1}{n'}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}\bar{Y}}-\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_{xy}}{\bar{X}\bar{Y}}\right] \\
&= Var(\bar{y})+\beta^2\left(\frac{1}{n}-\frac{1}{n'}\right)S_x^2-2\beta\left(\frac{1}{n}-\frac{1}{n'}\right)S_{xy} \\
&= Var(\bar{y})+\left(\frac{1}{n}-\frac{1}{n'}\right)(\beta^2S_x^2-2\beta S_{xy}) \\
&= Var(\bar{y})+\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}^2}{S_x^4}S_x^2-2\frac{S_{xy}}{S_x^2}S_{xy}\right) \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\left(\frac{S_{xy}}{S_x}\right)^2 \\
&= \left(\frac{1}{n}-\frac{1}{N}\right)S_y^2-\left(\frac{1}{n}-\frac{1}{n'}\right)\rho^2S_y^2 \quad (\text{using } S_{xy}=\rho S_xS_y) \\
&\approx \frac{(1-\rho^2)S_y^2}{n}+\frac{\rho^2S_y^2}{n'}. \quad \text{(Ignoring the finite population correction)}
\end{aligned}
$$

Clearly, $\hat{\bar{Y}}_{regd}$ is more efficient than the sample mean SRS, i.e. when no auxiliary variable is used.

Now we address the issue of whether the reduction in variability is worth the extra expenditure required to observe the auxiliary variable.

Let the total cost of the survey is

$$C_0 = C_1n + C_2n'$$

where C_1 and C_2 are the costs per unit observing the study variable y and auxiliary variable x, respectively.

Now minimize the $MSE\,(\hat{\bar{Y}}_{regd})$ for fixed cost C_0 using the Lagrangian function with Lagrangian multiplier λ as

$$
\begin{aligned}
\varphi &= \frac{S_y^2(1-\rho^2)}{n}+\frac{\rho^2S_y^2}{n'}+\lambda(C_1n+C_2n'-C_0) \\
\frac{\partial\varphi}{\partial n} &= 0 \Rightarrow -\frac{1}{n^2}S_y^2(1-\rho^2)+\lambda C_1 = 0 \\
\frac{\partial\varphi}{\partial n'} &= 0 \Rightarrow -\frac{1}{n'^2}S_y^2\rho^2+\lambda C_2 = 0
\end{aligned}
$$

Thus $n=\sqrt{\frac{S_y^2(1-\rho^2)}{\lambda C_1}}$

and $n'=\frac{\rho S_y}{\sqrt{\lambda C_2}}$.

Substituting these values in the cost function, we have

$$C_0 = C_1 n + C_2 n'$$
$$= C_1\sqrt{\frac{S_y^2(1-\rho^2)}{C_1\lambda}} + C_2\sqrt{\frac{\rho^2 S_y^2}{\lambda C_2}}$$
$$\text{or } C_0\sqrt{\lambda} = \sqrt{C_1 S_y^2(1-\rho^2)} + \sqrt{C_2\rho^2 S_y^2}$$
$$\text{or } \lambda = \frac{1}{C_0^2}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]^2.$$

Thus the optimum values of n and n' are

$$n_{opt}^{'} = \frac{\rho S_y C_0}{\sqrt{C_2}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]}$$
$$n_{opt} = \frac{C_0 S_y\sqrt{1-\rho^2}}{\sqrt{C_1}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]}.$$

The optimum mean squared error of $\hat{\bar{Y}}_{regd}$ is obtained by substituting $n = n_{opt}$ and $n' = n_{opt}^{'}$ as

$$MSE(\hat{\bar{Y}}_{regd})_{opt} = \frac{S_y^2(1-\rho^2)\left[\sqrt{C_1}\left(\sqrt{C_1 S_y^2(1-\rho^2)} + \rho S_y\sqrt{C_2}\right)\right]}{C_0\sqrt{S_y^2(1-\rho^2)}}$$
$$+\frac{S_y^2\rho^2\sqrt{C_2}\left[S_y\left(\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right)\right]}{\rho S_y C_0}$$
$$= \frac{1}{C_0}\left[S_y\sqrt{C_1(1-\rho^2)} + \rho S_y\sqrt{C_2}\right]^2$$
$$= \frac{S_y^2}{C_0}\left[\sqrt{C_1(1-\rho^2)} + \rho\sqrt{C_2}\right]^2$$

The optimum variance of $\bar{y}$ under SRS for SRS where no auxiliary information is used is

$$Var(\bar{y}_{SRS})_{opt} = \frac{C_1 S_y^2}{C_0}$$

which is obtained by substituting $\rho = 0, C_2 = 0$ in $MSE(\hat{\bar{Y}}_{SRS})_{opt}$. The relative efficiency is

$$RE = \frac{Var(\bar{y}_{SRS})_{opt}}{MSE(\hat{\bar{Y}}_{regd})_{opt}} = \frac{C_1 S_y^2}{S_y^2\left[\sqrt{C_1(1-\rho^2)} + \rho\sqrt{C_2}\right]^2}$$
$$= \frac{1}{\left[\sqrt{1-\rho^2} + \rho\sqrt{\frac{C_2}{C_1}}\right]^2}$$
$$\le 1.$$

Thus the double sampling in regression estimator will lead to gain in precision if

$$\frac{C_1}{C_2} > \frac{\rho^2}{\left[1-\sqrt{1-\rho^2}\right]^2}.$$

12.4 OPTIMUM ALLOCATION VARYING PROBABILITY SAMPLING:

Suppose it is desired to select the sample with probability proportional to an auxiliary variable x but information on x is not available. Then, in this situation, the double sampling can be used. An initial sample of size n' is selected with SRSWOR from a population of size N, and information on x is collected for this sample. Then a second sample of size n is selected with replacement and with probability proportional to x from the initial sample of size n'. Let $\bar{x}'$ denote the mean of x for the initial sample of size n', Let $\bar{x}$ and $\bar{y}$ denote means respectively of x and y for the second sample of size n. Then we have the following theorem.

Theorem:

(1) An unbiased estimator of the population mean $\bar{Y}$ is given as

$$\hat{\bar{Y}} = \frac{x'_{tot}}{n'n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right),$$

where x'_{tot} denotes the total for x in the first sample.

(2) $Var(\hat{\bar{Y}}) = \left(\frac{1}{n'}-\frac{1}{N}\right)S_y^2 + \frac{(n'-1)}{N(N-1)nn'}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{X_{tot}}{X_{tot}}} - Y_{tot}\right)^2$, where X_{tot} and Y_{tot} denote the totals of x and y respectively in the population.

(3) An unbiased estimator of the variance of $\hat{\bar{Y}}$ is given by

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'}-\frac{1}{N}\right)\frac{1}{n(n'-1)} + \left[x'_{tot}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x'^2_{tot}(A-B)}{n'(n-1)}\right] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x'_{tot}y_i}{n'x_i} - \hat{\bar{Y}}\right)$$

where $A = \left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2$ and $B = \sum_{i=1}^{n}\frac{y_i^2}{x_i^2}$

Proof. Before deriving the results, we first mention the following result proved in varying probability scheme sampling.

Result: In sampling with varying probability scheme for drawing a sample of size n from a population of size N and with replacement.

(i) $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ is an unbiased estimator of the population mean $\bar{y}$ where $z_i = \frac{y_i}{Np_i}$, p_i being the

probability of selection of i^{th} unit. Note that y_i and p_i can take any one of the N values $Y_1, Y_2, ..., Y_N$ with initial probabilities $P_1, P_2, ..., P_N$, respectively.

(ii) $Var(\bar{z}) = \frac{1}{nN^2}\left[\sum_{i=1}^{N}\frac{Y_i^2}{P_i} - N^2\bar{Y}^2\right] = \frac{1}{nN^2}\sum_{i=1}^{N}P_i\left(\frac{Y_i}{P_i} - \bar{Y}\right)^2$..

(iii) An unbiased estimator of the variance of $\bar{z}$ is

$$Var(\bar{z}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{Np_i} - \bar{z}\right)^2..$$

Let E_2 denote the expectation of $\hat{\bar{Y}}$, when the first sample is fixed. The second is selected with probability proportional to x, hence using the result (i) with $P_i = \frac{x_i}{x'_{tot}}$, we find that

$$E_2\left(\frac{\hat{\bar{Y}}}{n'}\right) = E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{n'\frac{x_i}{x'_{tot}}}\right]$$
$$= E_2\left[\frac{x'_{tot}}{nn'}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)\right]$$
$$= \bar{y}'$$

where $\bar{y}'$ is the mean of y for the first sample. Hence

$$E(\hat{\bar{Y}}) = E_1\left[E_2\left(\hat{\bar{Y}} \mid n'\right)\right]$$
$$= E_1(\bar{y}_{n'})$$
$$= \hat{\bar{Y}},$$

which proves the part (1) of the theorem. Further,

$$Var(\hat{\bar{Y}}) = V_1E_2\left(\hat{\bar{Y}} \mid n'\right) + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= V_1(\bar{y}') + E_1V_2\left(\hat{\bar{Y}} \mid n'\right)$$
$$= \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + E_1V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Now, using the result (ii), we get

$$V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\sum_{i=1}^{n'}\frac{x_i}{x'_{tot}}\left(\frac{y_i}{\frac{x_i}{x'_{tot}}} - y'_{tot}\right)^2$$
$$= \frac{1}{nn'^2}\sum_{i=1}^{n'}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

and hence

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{i<j}^{n'}x_ix_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2,$$

using the probability of a specified pair of units being selected in the sample is $\frac{n'(n'-1)}{N(N-1)}$. So we can express

$$E_1 V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'^2}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_{tot}}}{\frac{X_{tot}}{X_{tot}}} - Y_{tot}\right)^2.$$

Substituting this in $V_2\left(\hat{\bar{Y}} \mid n'\right)$, we get

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 + \frac{(n'-1)}{nn'N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{\frac{X_{tot}}{X_{tot}}} - Y_{tot}\right)^2.$$

This proves the second part (2) of the theorem.

We now consider the estimation of $Var(\hat{\bar{Y}})$. Given the first sample, we obtain

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

where $p_i = \frac{x_i}{x'_{tot}}$. Also, given the first sample,

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}\right)^2\right] = V_2(\hat{\bar{Y}}) = E_2(\hat{\bar{Y}}^2) - \bar{y}'^2.$$

Hence

$$E_2\left[\hat{\bar{Y}}^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{n' p_i} - \hat{\bar{Y}}^2\right)^2\right] = \bar{y}'^2.$$

Substituting $\hat{\bar{Y}} = \frac{x'_{tot}}{n' n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)$ and $p_i = \frac{x_i}{x'_{tot}}$ the expression becomes

$$E_2\left[\frac{x'^2}{nn'^2(n-1)}\left\{\left(\sum_{i=1}^{n}\frac{y_i}{x_i}\right)^2 - \left(\sum_{i=1}^{n}\frac{y_i^2}{x_i^2}\right)\right\}\right] = \bar{y}'^2$$

Using

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i}\right] = \sum_{i=1}^{n'} y_i^2,$$

we get

$$E_2\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \frac{x_{tot}^{'}}{x_i} - \frac{x_{tot}^{'2}}{nn'(n-1)}(A-B)\right] = \sum_{i=1}^{n'} y_i^2 - n'\bar{y}^{'2}$$

where $A = \left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right)^2$, and $B = \sum_{i=1}^{n} \frac{y_i^2}{x_i^2}$ which further simplifies to

$$E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = s_y^{'2},$$

where $s_y^{'2}$ is the mean sum of squares of y for the first sample. Thus, we obtain

$$E_1E_2\left[\frac{1}{n(n'-1)}\left\{x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right\}\right] = E_1(s_y^{'2}) = S_y^2 \qquad (1)$$

which gives an unbiased estimator of S_y^2. Next, since we have

$$E_1V_2\left(\hat{\bar{Y}} \mid n'\right) = \frac{1}{nn'}\frac{(n'-1)}{N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{X_{tot}} - Y_{tot}\right)^2,$$

and from this result we obtain

$$E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i x_{tot}^{'}}{n' x_i} - \hat{\bar{Y}}\right)^2\right] = V_2\left(\hat{\bar{Y}} \mid n'\right).$$

Thus

$$E_1E_2\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n' x_i} - \hat{\bar{Y}}\right)^2\right] = \frac{(n'-1)}{nn'N(n-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{X_{tot}} - Y_{tot}\right)^2 \qquad (2)$$

when gives an unbiased estimator of

$$\frac{(n'-1)}{nn'N(N-1)}\sum_{i=1}^{N}\frac{x_i}{X_{tot}}\left(\frac{\frac{y_i}{x_i}}{X_{tot}} - Y_{tot}\right)^2.$$

Using (1) and (2) an unbiased estimator of the variance of $\hat{\bar{Y}}$ is obtained as

$$Var(\hat{\bar{Y}}) = \left(\frac{1}{n'} - \frac{1}{N}\right)\frac{1}{n(n'-1)}\left[x_{tot}^{'}\sum_{i=1}^{n}\frac{y_i^2}{x_i} - \frac{x_{tot}^{'2}(A-B)}{n'(n-1)}\right] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{x_{tot}^{'}y_i}{n' x_i} - \hat{\bar{Y}}\right)^2$$

12.5 SUMMARY:

- Double Sampling with Regression Estimators leverages auxiliary variables to increase estimation efficiency.
- Bias and MSE are vital for understanding estimator quality.

- Varying Probability Sampling ensures that information-rich units are adequately represented in the sample.
- Double sampling is a cost-effective method when auxiliary information is inexpensive.
- Error analysis (bias and MSE) guides estimator reliability.
- Optimum allocation balances cost and efficiency.
- Varying probability sampling introduces flexibility and efficiency in heterogeneous populations.

These tools together empower researchers to design more precise, cost-effective, and representative surveys.

12.6 KEY WORDS:

- Double Sampling
- Regression Estimator
- Auxiliary Variable
- Estimation Error
- Mean Square Error
- Optimum Allocation
- Varying Probability Sampling
- Probability Proportional to Size (PPS)

12.7 SELF-ASSESSMENT QUESTIONS:

1. Explain the basic idea behind two-phase (double) sampling with an example.
2. What are the steps involved in implementing double sampling for a regression estimator?
3. Derive the formula for the double sampling regression estimator.
4. Why is an auxiliary variable used in regression estimation?
5. How does the use of double sampling improve the regression estimator?
6. What is the formula for the mean square error (MSE) of a double sampling regression estimator?
7. What is meant by optimum allocation in the context of double sampling?
8. What is varying probability sampling and how is it used in double sampling?
9. Describe probability proportional to size (PPS) sampling improve estimator efficie

12.8 SUGGESTED READINGS:

1. Cochran, W.G. (1977) *Sampling Techniques* (3rd Edition), Wiley.
2. P. Mukhopadhyay**,** Title**:** *Theory and Methods of Survey Sampling*

3. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984) *Sampling Theory of Surveys with Applications*, Indian Society of Agricultural Statistics.
4. Murthy, M.N. *Title: Sampling Theory and Methods*, *Publisher:* Statistical Publishing Society.
5. Singh, D., and Chaudhary, F.S. (1986) . *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.

Dr. U. Ramkiran

LESSON -13

NON-SAMPLING & NON-RESPONSE ERRORS

OBJECTIVES:

After completing this lesson, learners will be able to:

- To understand the concept of non-sampling errors: Differentiate between sampling and non-sampling errors
- Comprehend the impact of non-sampling errors on survey accuracy
- To identify various sources and types of non-sampling errors: Explore sources such as measurement, processing, coverage, and response errors.
- Classify types like interviewer bias, respondent bias, and data handling errors
- To study non-response errors and their implications: Understand types of non-response: item and unit non-response
- Learn the effects of non-response on data quality and estimation bias
- To learn techniques for adjustment of non-response
- Study weighting adjustments, imputation methods, and follow-up surveys
- Understand the trade-offs in applying each technique
- To apply the Hansen and Hurwitz Technique for non-response adjustment
- Learn the procedure and assumptions behind the Hansen-Hurwitz method
- Calculate unbiased estimates using sub sampling of non-respondents
- To understand and apply Deming's model of total survey error
- Explore Deming's framework integrating both sampling and non-sampling errors
- Use this model to design more accurate and efficient surveys.

STRUCTURE:

13.1 INTRODUCTION:

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

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

In general, the sampling errors decrease as the sample size increases, whereas non-sampling error increases as the sample size increases. In some situations, the non-sampling errors may be large and deserve greater attention than the sampling error.

In any survey, it is assumed that the value of the characteristic to be measured has been defined precisely for every population unit. Such a value exists and is unique. This is called the true value of the characteristic for the population value. In practical applications, data collected on the selected units are called survey values and they differ from the true values. Such difference between the true and observed values is termed as the observational error or response error. Such an error arises mainly from the lack of precision in measurement techniques and variability in the performance of the investigators.

13.2 SOURCES OF NON-SAMPLING ERRORS:

Non sampling errors can occur at every stage of planning and execution of survey or census. It occurs at the planning stage, fieldwork stage as well as at tabulation and computation stage. The main sources of the non-sampling errors are

- ♣ lack of proper specification of the domain of study and scope of the investigation,
- ♣ incomplete coverage of the population or sample,
- ♣ faulty definition,
- ♣ defective methods of data collection and
- ♣ tabulation errors.

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

- The data specification may be inadequate and inconsistent with the objectives of the survey or census.
- Due to the imprecise definition of the boundaries of area units, incomplete or wrong identification of units, faulty methods of enumeration etc., the data may be duplicated or may be omitted.

- The methods of interview and observation collection may be inaccurate or inappropriate.
- The questionnaire, definitions and instructions may be ambiguous.
- The investigators may be inexperienced or not trained properly.
- The recall errors may pose difficulty in reporting the true data.
- The scrutiny of data is not adequate.
- The coding, tabulation etc. of the data may be erroneous.
- There can be errors in presenting and printing the tabulated results, graphs etc.
- In a sample survey, the non-sampling errors arise due to defective frames and faulty selection of sampling units.

These sources are not exhaustive but surely indicate the possible source of errors.

Non-sampling errors may be broadly classified into three categories.

13.3 TYPES OF NON-SAMPLING ERRORS:

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

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

c) **Tabulation errors**: These errors occur at tabulation stage due to various reasons, e.g., inadequate scrutiny of data, errors in processing the data, errors in publishing the tabulated results, graphs etc.

Ascertainment errors may be further sub-divided into

(i) **Coverage errors** owing to over-enumeration or under-enumeration of the population or the sample, resulting from duplication or omission of units and from the non-response.

(ii) **Content errors** relating to the wrong entries due to the errors on the part of investigators and respondents.

Same division can be made in the case of tabulation error also. There is a possibility of missing data or repetition of data at tabulation stage which gives rise to coverage errors and also of errors in coding, calculations etc. which gives rise to content errors.

Treatment of non-sampling errors:

Some conceptual background is needed for the mathematical treatment of non-sampling errors.

Total error: Difference between the sample survey estimate and the parametric true value being estimated is termed as total error.

Sampling error:

If complete accuracy can be ensured in the procedures such as determination, identification and observation of sample units and the tabulation of collected data, then the **total error** would consist only of the error due to sampling, termed as sampling error.

The measure of sampling error is mean squared error (*MSE*). The *MSE* is the difference between the estimator and the true value and has two components:
- square of sampling bias.
- sampling variance.

If the results are also subjected to non-sampling errors, then the total error would have both sampling and non-sampling error.

13.4 NON-RESPONSE ERRORS:

The non-response error may occur due to refusal by respondents to give information or the sampling units may be inaccessible. This error arises because the set of units getting excluded may have characteristic so different from the set of units actually surveyed as to make the results biased. This error is termed as non response error since it arises from the exclusion of some of the anticipated units in the sample or population. One way of dealing with the problem of non-response is to make all the efforts to collect information from a sub-sample of the units not responding in the first attempt.

Measurement and control of errors:

Some suitable methods and adequate procedures for control can be adopted before initiating the main census or sample survey. Some separate programmes for estimating the different types of non-sampling errors are also required. Some such procedures are as follows:

1. Consistency checks:

Certain items in the questionnaires can be added, which may serve as a check on the quality of the collected data. To locate the doubtful observations, the data can be arranged in increasing order of some basic variable. Then they can be plotted against each sample unit. Such graph is expected to follow a certain pattern and any deviation from this pattern would help in spotting the discrepant values.

2. Sample check:

An independent duplicate census or sample survey can be conducted on a comparatively smaller group by trained and experienced staff. If the sample is properly designed and if the checking operation is efficiently carried out, then it is possible to detect the presence of non-sampling errors and to get an idea of their magnitude. Such a procedure is termed as the **method of sample check.**

3. Post-census and post-survey checks:

It is a type of sample check in which a sample (or subsample) is selected of the units covered in the census (or survey) and re-enumerate or re-survey it by using better trained and more experienced survey staff than those involved in the main investigation. This procedure is called as post-survey check or post census.

The effectiveness of such check surveys can be increased by
- re-enumerating or re-surveying immediately after the main census to avoid recall error
- taking steps to minimize the conditioning effect that the main survey may have on the work of the check-survey.

4. External record check:

Take a sample of relevant units from a different source, if available, and to check whether all the units have been enumerated in the main investigation and whether there are discrepancies between the values when matched. The list, from which the check-sample is drawn for this purpose, need not be a complete one.

5. Quality control techniques:

The use of tools of statistical quality control like control chart and acceptance sampling techniques can be used in assessing the quality of data and in improving the reliability of final results in large scale surveys and census.

6. Study or recall error:

Response errors arise due to various factors like the attitude of respondents towards the survey, method of interview, skill of the investigators and recall errors. Recall error depends on the length of the reporting period and on the interval between the reporting period and data of survey. One way of studying recall error is to collect and analyze data related to more than one reporting period in a sample (or sub-sample) of units covered in the census or survey.

7. Interpenetrating sub-samples:

The use of interpenetrating sub-sample technique helps in providing an appraisal of the quality of information as the interpenetrating sub-samples can be used to secure information on non-sampling errors such as differences arising from differential interviewer bias, different methods of eliciting information etc. After the sub-samples have been surveyed by different groups of investigators and processed by different team of workers at the tabulation stage, a comparison of the final estimates based on the subsamples provides a broad check on the quality of the survey results.

13.5 TECHNIQUES FOR ADJUSTMENT OF NON-RESPONSE:

Non-response errors occur when some selected participants in a survey or study don't respond, leading to potential bias in the results. To mitigate these errors, techniques like weighting adjustments, imputation, and response propensity weighting are employed.

Types of Non-response:

- **Unit Non-response:** Occurs when an entire survey or a unit (e.g., a household) doesn't respond.
- **Item Non-response:** Happens when respondents don't answer specific questions within a survey Techniques for Adjustment:

1. **Weighting Adjustments:**
   - **Post-stratification:** Dividing the population into subgroups (strata) and adjusting weights to match known population totals within each stratum.
   - **Raking:** Iteratively adjusting weights to match multiple control totals (e.g., age, gender, income).
   - **Response Propensity Weighting:** Assigning weights based on the probability of a unit responding, often estimated using logistic regression.

2. **Imputation:**
   - **Mean Imputation:** Replacing missing values with the average response from those who did answer.
   - **Regression Imputation:** Using a regression model to predict missing values based on other variables.
   - **Multiple Imputation:** Creating several plausible imputed datasets to account for uncertainty in the imputed values.

3. **Other Techniques:**
   - **Calibration Weighting:** Adjusting weights to match external benchmarks, like census data.
   - **Response Rate Maximization:** Employing strategies to encourage participation and reduce non-response during the data collection phase.

Key Considerations:

- **Pattern of Missingness:**

Understanding why non-response occurs (e.g., refusals, not-at-homes) is crucial for choosing appropriate adjustment methods.

- **Assumptions:**

Weighting adjustments often rely on assumptions about the similarity of respondents and non-respondents.

- **Data Integrity:**

Imputation methods should be carefully considered to minimize bias and maintain data integrity.

By using these techniques, researchers can reduce the impact of non-response errors and produce more reliable survey results.

13.6 HANSEN AND HURWITZ TECHNIQUE:

Let a sample s of size n be selected by simple random sampling without replacement (SRSWOR) method. Suppose n_1 units responded and remaining $n_2(=n-n_1)$ did not respond at the first attempt. The set of responded and nonresponded units will be denoted, respectively, by s_1 and s_2. A simple random subsample s_2^* of size $m = vn_2$ (assuming integer) is selected from s_2 where v is a known fraction. Responses from all the units of s_2^* are obtained by using more intensive method, which is obviously expensive. Hansen and Hurwitz (1946) assumed that the population U under study of size N can be divided into two strata according to the nature of respondents. The first stratum H_1 consisting of N_1 (unknown) units that respond at the first attempt and the remaining set of units $N_2(=N-N_1)$ that do not respond at the first attempt comprise the stratum H_2. It is assumed that the members who belong to H_1 always respond at the first attempt while members of H_2 never respond at the first attempt, but they will always respond if a more persuasive method is employed.

Let $\bar{y}(s_1)$ and $\bar{y}(s_2^*)$ be the sample means of the variable under study y for the samples s_1 and s_2^*, respectively. Let

$$\bar{y}_w = w_1\bar{y}(s_1) + w_2\bar{y}(s_2^*) \tag{15.4.1}$$

with $w_1 = n_1/n$ and $w_2 = n_2/n$.

Let $\bar{Y}, S_y^2, S_{1y}^2$, and S_{2y}^2 denote the population mean, population variance, population variances of response, and nonresponse strata, respectively.

Substituting $n' = n$, $\gamma_1 = 1$, $\gamma_2 = v$, $n_2 = m$ in Theorem 10.3.2, we obtain the following result.

Theorem 15.4.1

(i) $\bar{y}_w$ is unbiased for $\bar{Y}$

(ii) The variance of $\bar{y}_w$ is

$$V(\bar{y}_w) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{W_2}{n}\left(\frac{1}{v} - 1\right)S_{2y}^2$$

(iii) An unbiased estimator of $V(\bar{y}_w)$ is

$$\begin{aligned}\hat{V}(\bar{y}_w) = {} & \frac{(N-n)(n_1-1)}{N(n-1)}w_1\frac{\hat{S}_{1y}^2}{n_1} \\ & + \frac{(N-1)(n_2-1)-(n-1)(m-1)}{N(n-1)}w_2\frac{\hat{S}_{2y}^2}{m} \\ & + \frac{N-n}{N(n-1)}\left[w_1\{\bar{y}(s_1)-\bar{y}_w\}^2 + w_2\{\bar{y}(s_2^*)-\bar{y}_w\}^2\right]\end{aligned}$$

where $W_2 = N_2/N$, $\hat{S}_{1y}^2$ and $\hat{S}_{2y}^2$ denote the sample variance of response and nonresponse stratum, respectively.

Remark 15.4.1

The bias of $\overline{y}(s_1)$ (sample mean of the response stratum) for estimating $\overline{Y}$ is given by

$$\begin{aligned} B[\overline{y}(s_1)] &= E[\overline{y}(s_1)] - \overline{Y} \\ &= \overline{Y}_1 - \overline{Y} \\ &= W_2(\overline{Y}_1 - \overline{Y}_2) \end{aligned} \qquad (15.4.2)$$

where $\overline{Y}_1$ and $\overline{Y}_2$ denote the population mean of response and nonresponse stratum, respectively. The bias in Eq. (15.4.2) is negligible if at least one of the quantities W_2 and $(\overline{Y}_1 - \overline{Y}_2)$ is negligible. From Theorem 15.4.1, we note that estimator $\overline{y}_w$ is less efficient than the sample mean based on all the n observed, and the loss of efficiency is negligible if W_2 or S_{2y}^2 is small.

13.7 DEMING'S MODEL:

William Edwards Deming (October 14, 1900 – December 20, 1993) was an American business theorist, composer, economist, industrial engineer, management consultant, statistician, and writer. Educated initially as an electrical engineer and later specializing in mathematical physics, he helped develop the sampling techniques still used by the United States Census Bureau and the Bureau of Labor Statistics. He is also known as the father of the quality movement and was hugely influential in post-WWII Japan, credited with revolutionizing Japan's industry and making it one of the most dominant economies in the world. He is best known for his theories of management.

The Deming System of Profound Knowledge:

The prevailing style of management must undergo transformation. A system cannot understand itself. The transformation requires a view from outside. The aim of this chapter is to provide an outside view—a lens—that I call a system of profound knowledge. It provides a map of theory by which to understand the organizations that we work in.

The first step is transformation of the individual. This transformation is discontinuous. It comes from understanding of the system of profound knowledge. The individual, transformed, will perceive new meaning to his life, to events, to numbers, to interactions between people.

Once the individual understands the system of profound knowledge, he will apply its principles in every kind of relationship with other people. He will have a basis for judgment of his own decisions and for transformation of the organizations that he belongs to.

Deming advocated that all managers need to have what he called a System of Profound Knowledge, consisting of **four parts:**

1. *Appreciation of a system*: understanding the overall processes involving suppliers, producers, and customers (or recipients) of goods and services (explained below);
2. *Knowledge of variation*: the range and causes of variation in quality, and use of statistical sampling in measurements;

3. *Theory of knowledge*: the concepts explaining knowledge and the limits of what can be known.
4. *Knowledge of psychology*: concepts of human nature.

The System of Profound Knowledge is the basis for application of Deming's famous 14 Points for Management, described below.

W. E. Deming's 14 Key Principles:

1. **Constancy of purpose:** Create constancy of purpose for continual improvement of products and service to society, allocating resources to provide for long range needs rather than only short term profitability, with a plan to become competitive, to stay in business, and to provide jobs.
2. **The new philosophy:** Adopt the new philosophy. We are in a new economic age, created in Japan. We can no longer live with commonly accepted levels of delays, mistakes, defective materials and defective workmanship. Transformation of Western management style is necessary to halt the continued decline of business and industry.
3. **Cease dependence on mass inspection:** Eliminate the need for mass inspection as the way of life to achieve quality by building quality into the product in the first place. Require statistical evidence of built in quality in both manufacturing and purchasing functions.
4. **End lowest tender contracts:** End the practice of awarding business solely on the basis of price tag. Instead require meaningful measures of quality along with price. Reduce the number of suppliers for the same item by eliminating those that do not qualify with statistical and other evidence of quality. The aim is to minimize total cost, not merely initial cost, by minimizing variation. This may be achieved by moving toward a single supplier for any one item, on a long term relationship of loyalty and trust. Purchasing managers have a new job, and must learn it.
5. **Improve every process:** Improve constantly and forever every process for planning, production, and service. Search continually for problems in order to improve every activity in the company, to improve quality and productivity, and thus to constantly decrease costs. Institute innovation and constant improvement of product, service, and process. It is management's job to work continually on the system (design, incoming materials, maintenance, improvement of machines, supervision, training, retraining).
6. **Institute training on the job:** Institute modern methods of training on the job for all, including management, to make better use of every employee. New skills are required to keep up with changes in materials, methods, product and service design, machinery, techniques, and service.
7. **Institute leadership:** The aim of supervision should be to help people and machines and gadgets to do a better job. Supervision of management is in need of overhaul, as well as supervision of production workers.
8. **Drive out fear:** Encourage effective two-way communication and other means to drive out fear throughout the organization so that everybody may work effectively and more productively for the company.

9. **Break down barriers:** Break down barriers between departments and staff areas. People in different areas, such as Leasing, Maintenance, Administration, must work in teams to tackle problems that may be encountered with products or service.
10. **Eliminate exhortations:** Eliminate the use of slogans, posters and exhortations for the work force, demanding Zero Defects and new levels of productivity, without providing methods. Such exhortations only create adversarial relationships; the bulk of the causes of low quality and low productivity belong to the system, and thus lie beyond the power of the work force.
11. **Eliminate arbitrary numerical targets:** Eliminate work standards that prescribe quotas for the work force and numerical goals for people in management. Substitute aids and helpful leadership in order to achieve continual improvement of quality and productivity.
12. **Permit pride of workmanship:** Remove the barriers that rob hourly workers, and people in management, of their right to pride of workmanship. This implies, among other things, abolition of the annual merit rating (appraisal of performance) and of Management by Objective. Again, the responsibility of managers, supervisors, foremen must be changed from sheer numbers to quality.
13. **Encourage education:** Institute a vigorous program of education, and encourage self improvement for everyone. What an organization needs is not just good people; it needs people that are improving with education. Advances in competitive position will have their roots in knowledge.
14. **Top management commitment and action:** Clearly define top management's permanent commitment to ever improving quality and productivity, and their obligation to implement all of these principles. Indeed, it is not enough that top management commit themselves for life to quality and productivity. They must know what it is that they are committed to-that is, what they must do. Create a structure in top management that will push every day on the pre ceding 13 Points, and take action in order to accomplish the transformation. Support is not enough: action is required!

13.8 SUMMARY:

This unit focused on **non-sampling errors**, which are errors not related to the process of selecting a sample but rather to other factors that can affect the accuracy and reliability of survey results.

- **Introduction** introduced the concept of non-sampling errors and highlighted their significance in survey methodology.
- **Sources of Non-sampling Errors** included human errors, data processing mistakes, faulty questionnaire design, and interviewer bias.
- **Types of Non-sampling Errors** were classified into four categories: **coverage errors**, **response errors**, **non-response errors**, and **processing errors**.
- **Non-response Errors** were emphasized as a major source of bias, occurring when selected units do not participate or provide incomplete information.

- **Techniques for Adjustment of Non-response** described various methods such as weighting adjustments, imputation techniques, and follow-up strategies to minimize bias.
- **Hansen and Hurwitz Technique** presented a practical approach where a sub-sample of non-respondents is re-contacted to estimate and adjust for bias.
- **Deming's Model** provided a theoretical framework showing how total survey error can be decomposed into identifiable components, aiding better design and error control.

In conclusion, while sampling errors can be reduced by increasing the sample size or improving sampling techniques, **non-sampling errors require careful planning, monitoring, and correction**. Understanding their sources and applying suitable adjustment techniques is essential for enhancing the **validity and reliability of survey estimates**.

13.9 KEY WORDS:

- Non-sampling Error
- Response Bias
- Non-response Error
- Measurement Error
- Hansen and Hurwitz Technique
- Deming's Model
- Substitution
- Imputation
- Weighting Adjustment
- Total Survey Error

13.10 SELF-ASSESSMENT QUESTIONS:

1. What are non-sampling errors? How do they differ from sampling errors?
2. Enumerate and briefly explain the different sources of non-sampling errors.
3. Classify non-sampling errors and describe each type with suitable examples.
4. What is a non-response error? How can it affect the reliability of survey results?
5. List and explain the common techniques used to adjust for non-response in survey sampling.
6. Describe the Hansen and Hurwitz technique for handling non-response. How does it work in practice?
7. Explain Deming's model in the context of non-sampling errors. What are its key features?
8. Discuss the impact of interviewer bias and response bias on survey results. How can they be minimized?
9. How does recall bias differ from reporting error in surveys? Give an example of each.
10. Why is it important to address non-sampling errors in large-scale surveys?

13.11 SUGGESTED READINGS:

1. **Cochran, W.G. (1977),** *Sampling Techniques* (3rd Edition), Wiley.
2. **P. Mukhopadhyay**, Title: *Theory and Methods of Survey Sampling,* PHI Learning Pvt. Ltd., New Delhi.
3. **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984).** *Sampling Theory of Surveys with Applications,* Publisher: Iowa State University Press.
4. Singh, D., and Chaudhary, F.S. **(1986).** *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.
5. **Deming, W. Edwards (1960).** *Sample Design in Business Research*, Publisher: Wiley
6. **Hansen, M.H., Hurwitz, W.N., & Madow, W.G. (1953).** *Sample Survey Methods and Theory* **(Vol. I & II)**, Publisher: Wiley.
7. **Kish, Leslie (1965).** *Survey Sampling*, Publisher: Wiley.
8. **Groves, Robert M. et al. (2009).** *Survey Methodology* **(2nd Edition).**

Dr. U. Ramkiran

LESSON -14

RATIO METHOD OF ESTIMATOR

OBJECTIVES:

After studying this lesson, the learner will be able to:

Understand the Concept of Ratio Estimator

- Grasp the motivation behind using auxiliary information in estimation.
- Learn the definition and formulation of the ratio estimator for population mean/total.

Evaluate Bias and Mean Square Error (MSE)

- Derive and interpret the expressions for bias and MSE of the ratio estimator.
- Understand conditions under which the ratio estimator performs better than the simple mean per unit estimator.

Estimate the Variance of Ratio Estimator

- Learn techniques to estimate the variance of a ratio estimator from sample data.
- Understand the importance of variance estimation in practical inference.

Construct Confidence Intervals

- Develop confidence intervals for population parameters using the ratio estimator.
- Understand the assumptions and limitations involved in such constructions.

Compare with Mean Per Unit Estimator

- Analyze the efficiency of the ratio estimator relative to the mean per unit (unbiased) estimator.
- Understand scenarios where the ratio estimator is more suitable.

Apply Ratio Estimator in Stratified Sampling

- Extend the concept of ratio estimation to stratified random sampling designs.
- Learn how stratification affects bias, variance, and efficiency of the ratio estimator.

Apply the Ratio Estimator in Practical Survey Situations

- Use real or simulated survey data to apply ratio estimation techniques.

STRUCTURE:

14.1 INTRODUCTION:

In survey sampling, improving the precision of estimators is a key goal, especially when auxiliary information is available. The **ratio estimator** is one such method that utilizes the known relationship between the study variable and an auxiliary variable to enhance estimation efficiency.

Unlike the simple mean per unit estimator, the ratio estimator takes advantage of the correlation between the variable of interest (say, income) and an auxiliary variable (say, expenditure) to provide more accurate population estimates, particularly when a strong linear relationship exists through the origin.

This method is especially useful when:

- The auxiliary variable is positively correlated with the study variable.
- The population mean or total of the auxiliary variable is known.
- The coefficient of variation of the auxiliary variable is smaller than that of the study variable.

The ratio estimator can be extended to complex designs like **stratified random sampling**, offering further improvements in precision when applied appropriately.

In this chapter, we will explore the concept of ratio estimation, derive expressions for its bias and mean square error (MSE), understand variance estimation and confidence intervals, and compare its performance with the mean per unit estimator. Applications in **stratified sampling** will also be discussed to highlight its practical significance in survey design.

14.2 CONCEPT OF RATIO ESTIMATOR:

Study variable - y_i, Auxiliary variable - x_i

Each unit having a pair of units (x_i, y_i), $Y, \overline{Y}, \hat{Y}, \hat{\overline{Y}}$

We want to estimate $\frac{\overline{Y}}{\overline{X}} = \frac{Y/N}{X/N} = R$ and $\hat{R}$ is called "the ratio of two estimates".

14.3 NOTATIONS AND DEFINITION:

y_i : The value of the characteristic under study for the i^{th} unit of the population (study variable)

x_i : The value of the auxiliary variable on the same unit. x_i is correlated with y_i. [where x_i is the in the two-phase or double sampling possession of advance information about an

auxiliary variatex_i].

Y: The total of y_i values in the population.

y : The total of y_i values in the sample.

X: The total of x_i values in the population.

x: The total of x_i values in the sample.

$R = \frac{Y}{X} = \frac{\overline{Y}}{\overline{X}}$ = the ratio of the population totals (or) population means of variates y_i and x_i.

ρ = correlation coefficient between y_i and x_i in the population.

Suppose it is desired estimate Y or$\bar{Y}$ or R by drawing a SRS of n-units (x_i, y_i), $(i = 1,2, - - -, n)$ from the population. Let us assume that, based on n-pairs of observations $\bar{y}$ and $\bar{x}$ are the sample means of y_i and x_i respectively and the population total X or population mean $\bar{X}$ is known. The ratio estimators of the ratio R, total Y and the mean $\bar{Y}$ may be defined by

$$\hat{R} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}} \rightarrow (1)$$

$$\hat{Y}_R = \frac{\bar{y}}{\bar{x}} X = \hat{R}X \rightarrow (2)$$

$$\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X} = \hat{R}\bar{X} \rightarrow (3)$$

14.3.1 EXAMPLES FOR THE USE OF RATIO ESTIMATES:

1) If x_i is the value of y_i at some previous time, the ratio method uses the sample to estimate the relative change $\frac{Y}{X}$ that has occurred since that time. The estimated relative change $\frac{y}{x}$ is multiplied by the known population total X on the previous occasion to provide an estimate of the current (present) population total.
2) x_i may be the total acreage(large portion of the parks/forest) of a form and y_i be the no. of acres sown to some crop. The ratio estimate will be successful in this case all framers devote about the same percentage of their total average of this crop.
3) The ratio of corn acres to wheat acres, the ratio of expenditures on labour to total expenditures are the examples if the problem is to estimate a ratio.

14.3.2 THEOREM -1: If variates $y_i and x_i$ are measured on each unit of a SRS of size n, assumed large, the variances of $\hat{R}, \hat{Y}_R and \hat{\bar{Y}}_R$ are obtained approximately as

$$V(\hat{R}) \approx \frac{1-f}{n\bar{X}^2}\left[\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1}\right] \rightarrow (1)$$

$$V(\hat{Y}_R) \approx N^2\frac{1-f}{n}\left[\sum_{i=1}^{N}\frac{(y_i - R\,x_i)^2}{N-1}\right] \rightarrow (2)$$

$$V\left(\bar{y}_R\right) \approx \frac{1-f}{n}\left[\sum_{i=1}^{N}\frac{\left(y_i - R x_i\right)^2}{N-1}\right] \qquad \rightarrow (3)$$

Where $R = \frac{\bar{Y}}{\bar{X}}, f = \frac{n}{N}, \hat{R} = \frac{\bar{y}}{\bar{x}}, \hat{Y}_R = \hat{R}X, \bar{\hat{Y}}_R = \hat{R}\bar{X}, \hat{Y}_R = N\bar{x}\hat{R}$

Proof: Let us consider, $\hat{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}}$

If n is large, $\bar{x}$ should not differ greatly from $\bar{X} \therefore \hat{R} - R \approx \frac{\bar{y} - R\bar{x}}{\bar{X}}$ $\rightarrow (4)$

Now average overall simple random samples of size n.

$$E\left(\hat{R} - R\right) = \frac{E\left(\bar{y} - R\bar{x}\right)}{\bar{X}} = \frac{E\left(\bar{y}\right) - R\,E\left(\bar{x}\right)}{\bar{X}} = \frac{\bar{Y} - R\bar{X}}{\bar{X}} = 0, \text{ since } R = \frac{\bar{Y}}{\bar{X}}$$

This shows that to the order of approximation used here $\hat{R}$ is an unbiased estimate of R.
Now from equation (4),

$$E\left(\hat{R} - R\right)^2 = V\left(\hat{R}\right) = \frac{1}{\bar{X}^2}E\left(\bar{y} - R\bar{x}\right)^2$$

These quantity $\bar{y} - R\bar{x}$ is the sample mean of the variate $\because R = \frac{\bar{Y}}{\bar{X}}, \left[\because \bar{Y} = R\bar{X}\right]$

$d_i = y_i - Rx_i$, whose population mean $\bar{D} = \bar{Y} - R\bar{X} = 0 = R\bar{X} - R\bar{X} = 0$.

Hence we can find variance of $\hat{R}$ for the variance of the mean of a SRS to the variate d_i and dividing by $\bar{X}^2$. This gives

$$V\left(\hat{R}\right) = \frac{1}{\bar{X}^2}E\left(\bar{d} - \bar{D}\right)^2 = \frac{1}{\bar{X}^2}V\left(\bar{d}\right) = \frac{1}{\bar{X}^2}\cdot\frac{1-f}{n}\cdot S_d^2 \left[\because V\left(\bar{y}\right) = \frac{1-f}{n}S^2\right]$$

$$= \frac{1-f}{n\bar{X}^2}\sum_{i=1}^{N}\frac{\left(d_i - \bar{D}\right)^2}{N-1} = \frac{1-f}{n\bar{X}^2}\sum_{i=1}^{N}\frac{\left(y_i - Rx_i\right)^2}{N-1} \qquad \rightarrow (5)$$

Since $\hat{Y}_R = \hat{R}.X$

$$V\left(\hat{Y}_R\right) = V\left(\hat{R}.X\right) = X^2 V\left(\hat{R}\right) = N^2\bar{X}^2 V\left(\hat{R}\right)$$

$$V\left(\hat{Y}_R\right) = N^2.\bar{X}^2\frac{1-f}{n\bar{X}^2}\sum_{i=1}^{N}\frac{\left(y_i - Rx_i\right)^2}{N-1} \qquad \text{[From (5)]}$$

$$= N^2\frac{1-f}{n}\sum_{i=1}^{N}\frac{\left(y_i - Rx_i\right)^2}{N-1}$$

Since $\bar{\hat{Y}}_R = \hat{R}.X$

$$V\left(\bar{\hat{Y}}_R\right) = V\left(\hat{R} - \bar{X}\right) = \bar{X}^2 V\left(\hat{R}\right) \qquad \text{[from 5]}$$

$$V\left(\bar{\hat{Y}}_R\right) = \bar{X}^2\frac{1-f}{n\bar{X}^2}\sum_{i=1}^{N}\frac{\left(y_i - Rx_i\right)^2}{N-1}$$

$$V\left(\bar{y}_R\right)=\frac{1-f}{n}\sum_{i=1}^{N}\left[\frac{\left(y_i-Rx_i\right)^2}{N-1}\right]$$

Corollary-1: There are various alternative forms of the above results.

Since $\bar{Y}=R\,\bar{X}$, we may write

$$V\left(\bar{y}_R\right)=\frac{N^2(1-f)}{n(N-1)}\sum_{i=1}^{N}\left[\left(y_i-\bar{Y}\right)-R\left(x_i-\bar{X}\right)\right]^2$$

$$=\frac{N^2(1-f)}{n(N-1)}\left[\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2+R^2\sum_i\left(x_i-\bar{X}\right)^2-2R\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)\right]$$

The correlation coefficient ρ between y_i and x_i in the finite population is defined by the equation

$$\rho=\frac{E\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{\sqrt{E\left(y_i-\bar{Y}\right)^2.E\left(x_i-\bar{X}\right)^2}}\quad,\quad\therefore\rho=\frac{\frac{\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{N}}{\sqrt{\sum_i\frac{\left(y_i-\bar{Y}\right)^2}{N}.\sum_i\frac{\left(x_i-\bar{X}\right)^2}{N}}}$$

$$\rho=\frac{\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{\sqrt{\sum_i\left(y_i-\bar{Y}\right)}\sqrt{\sum_i\left(x_i-\bar{X}\right)}}$$

$$=\frac{\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{\sqrt{(N-1)S_y^2}\sqrt{(N-1)S_x^2}}$$

$$=\frac{\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)}{(N-1)S_yS_x}$$

$$\left[\because\ S_y^2=\frac{\sum_i\left(y_i-\bar{Y}\right)}{N-1}\Rightarrow(N-1)S_y^2=\sum_i\left(y_i-\bar{Y}\right)^2\right.$$

$$\left.S_x^2=\frac{\sum_i\left(x_i-\bar{X}\right)}{N-1}\Rightarrow(N-1)S_x^2=\sum_i\left(x_i-\bar{X}\right)^2\right]$$

This leads to the result in $V\left(\bar{y}_R\right)$ is

$$V\left(\bar{y}_R\right)=\frac{N^2(1-f)}{n(N-1)}\left[\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2+R^2\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2-2R\sum_i\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)\right]$$

$$V\left(\bar{y}_R\right)=\frac{N^2(1-f)}{n}\left[S_y^2+R^2S_x^2-2R\rho S_yS_x\right]$$

$$=\frac{N^2(1-f)}{n}\bar{Y}^2\left(\frac{S_y^2}{\bar{Y}^2}+\frac{R^2S_x^2}{\bar{Y}^2}-\frac{2R\rho S_yS_x}{\bar{Y}^2}\right)\left(\because\ \text{Divide and multiply by }\frac{\bar{X}^2}{\bar{Y}^2}\right)$$

$$=\frac{(1-f)}{n}N^2\bar{Y}^2\left(\frac{S_y^2}{\bar{Y}^2}+\frac{S_x^2}{\bar{Y}^2}-\frac{2S_yS_x}{\bar{Y}\bar{X}}\right)\left(\because\ R=\frac{\bar{Y}}{\bar{X}}\Rightarrow R^2=\frac{\bar{Y}^2}{\bar{X}^2}\ R^2\bar{X}^2=\bar{Y}^2\ \frac{2RS}{R^2X^2}=R\bar{X}\ \bar{X}\right)$$

Where $S_{yx}=\rho S_yS_x$ is the covariance between y_i & x_i.

This relation may also be written as

$$V\left(\bar{y}_R\right)=\frac{1-f}{n}Y^2\left(c_{yy}+c_{xx}-2c_{yx}\right)\quad----------(6)$$

Where c_{yy},c_{xx} are the squares of the coefficient of variance (cv) of y_i & x_i respectively and

c_{yx} is the relative co-variance.

Corollary-2:- since $\bar{y}_R$, $\bar{y}_R$, $\bar{R}$ differ only by known multipliers, the coefficient of variation is same for all the three estimates

$$(cv)^2 = \frac{V(\bar{y}R)}{Y^2} = \frac{1-f}{n}\left(c_{yy} + c_{xx} - 2c_{yx}\right) \text{ ---}(7)$$

The quantity $(cv)^2$ has been called relative variance. Then

$$(cv)^2 = \frac{V(\bar{y}_R)}{\bar{Y}^2} = \frac{\frac{V(\bar{Y}_R)}{N^2}}{\bar{Y}^2} = \frac{\frac{1-f}{n}Y^2\left(c_{yy} + c_{xx} - 2c_{yx}\right)\Big/N^2}{\bar{Y}^2} \qquad \left[\because Equ'n-(6)\right]$$

$$= \frac{\frac{1-f}{n}\bar{Y}^2\left(c_{yy} + c_{xx} - 2c_{yx}\right)}{\bar{Y}^2} = \frac{1-f}{n}\left(c_{yy} + c_{xx} - 2c_{yx}\right) \text{---} \rightarrow (8)\left[\because \frac{Y^2}{N^2} = \bar{Y}^2\right]$$

$$\text{Again} \quad (cv)^2 = \frac{V(\bar{R})}{R^2} = \frac{V(\bar{y}_R)\Big/X^2}{R^2} = \frac{\left(\frac{1-f}{n}\right)Y^2\left(c_{yy} + c_{xx} - 2c_{yx}\right)\Big/X^2}{R^2}\left[\because \bar{R} = \frac{\bar{y}_R}{X}\right] \quad \left[\because from\ equ'n(6)\right]$$

$$= \frac{\left(\frac{1-f}{n}\right)\frac{Y^2}{N^2}\left(c_{yy} + c_{xx} - 2c_{yx}\right)\Big/\left(\frac{X^2}{N^2}\right)}{R^2}$$

$$= \left(\frac{1-f}{n}\right)\frac{\left(c_{yy} + c_{xx} - 2c_{yx}\right)}{R^2}\,\frac{\left(Y^2/N^2\right)}{\left(X^2/N^2\right)}$$

$$= \left(\frac{1-f}{n}\right)\frac{\left(c_{yy} + c_{xx} - 2c_{yx}\right)}{R^2}\left(R^2\right) \quad \left[\because R = \frac{\bar{Y}}{\bar{X}}\bar{Y}^2 = R^2\bar{X}^2\bar{X}^2 = R^2\bar{Y}^2\right]$$

$$(cv)^2 = \frac{1-f}{n}\left(c_{yy} + c_{xx} - 2c_{yx}\right)$$

14.4 BIAS OF THE RATIO ESTIMATE:

Result-1:- obtain the bias of the ratio estimate and its relative bias

Proof:- we know that

$$\bar{R} - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} = \frac{\bar{y} - R\bar{x}}{\bar{X} + \left(\bar{x} - \bar{X}\right)} = \frac{\bar{y} - R\bar{x}}{\bar{X}\left[1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right]} = \frac{\bar{y} - R\bar{x}}{\bar{X}}\left[1 + \frac{\bar{x} - \bar{X}}{\bar{X}}\right]^{-1}$$

Expanding by a Taylor's series, we get

$$\bar{R} - R \square \frac{\bar{y} - R\bar{x}}{\bar{x}}\left(1 - \frac{\bar{x} - \bar{X}}{X} + ----\right)$$

Ignoring the terms of the second and higher orders, we have

$E(\hat{R}-R)=\frac{1}{\bar{X}}\left[E\left(\bar{y}-R\bar{x}\right)-\frac{1}{\bar{X}}E\left\{\left(\bar{y}-R\bar{x}\right)\left(\bar{x}-\bar{X}\right)\right\}\right]$,since $E(\bar{y}-R\bar{x})=\bar{Y}-R\bar{X}=0$,since $R=\frac{\bar{Y}}{\bar{X}}$

$E(\hat{R}-R)=-\frac{1}{\bar{X}^2}E\{(\bar{y}-R\bar{x})(\bar{x}-\bar{X})\}$ $=-\frac{1}{\bar{X}^2}\{E[\bar{y}(\bar{y}-\bar{x})]-R.E[\bar{x}(\bar{x}-\bar{X})\}$

$=-\frac{1}{\bar{X}^2}\{E[(\bar{y}-\bar{Y})(\bar{y}-\bar{x})]-R.E[(\bar{x}-\bar{X})^2\}$ $=-\frac{1}{\bar{X}^2}\left[\frac{(1-f)}{n}\rho S_y S_x-R\frac{(1-f)}{n}{S_x}^2\right]$

$B(\hat{R})$ $=-\frac{1-f}{n\bar{X}^2}\left[R{S_x}^2-\rho S_y S_x\right]$

This is the bias of the ratio estimate

$\frac{B(\hat{R})}{R}=relativeBias=\frac{bias}{R}=-\frac{1-f}{n\bar{X}^2\frac{\bar{Y}}{\bar{X}}}\left[R{S_x}^2-\rho S_y S_x\right]$

$\frac{B(\hat{R})}{R}=\frac{1-f}{n}\left(C_{xx}-C_{yx}\right)$ ----- (1) $\left[\because C_{xx}=\frac{{S_x}^2}{\bar{x}}\right]$

Result -2:-An upper bound to the ratio of the bias to the standard error is given by

$$\frac{B\left(\hat{R}\right)}{\sigma_{\hat{R}}}\le\frac{\frac{\sigma_x}{\bar{X}}}{}=cv\ of\ \bar{x}$$

Proof:- consider the covariance in SRS's of size n,of the quantities $\hat{R}$ and $\bar{x}$ .We have

$$\operatorname{cov}\left(\hat{R},\bar{x}\right)=E\left(\frac{\bar{y}}{\bar{x}}\bar{x}\right)-E\left(\hat{R}\right)E\left(\bar{x}\right)=\bar{Y}-\bar{X}E\left(\hat{R}\right)$$

$$hence\ \ E\left(\hat{R}\right)=\frac{\bar{Y}}{\bar{X}}-\frac{1}{\bar{X}}\operatorname{cov}\left(\hat{R},\bar{x}\right)=R-\frac{1}{\bar{X}}\operatorname{cov}\left(\hat{R},\bar{x}\right)$$

$$the\ bias\ in\ \hat{R}\ is\ B\left(\hat{R}\right)=E\left(\hat{R}\right)-R=\frac{-\operatorname{cov}\left(\hat{R},\bar{x}\right)}{\bar{X}}$$

$$\left|B\left(\hat{R}\right)\right|=\frac{\left|\rho_{\hat{R},\bar{x}}\ \sigma_{\hat{R}}\sigma_{\bar{x}}\right|}{\bar{X}}\le\frac{\sigma_{\hat{R}}\sigma_{\bar{x}}}{\bar{X}}\ ,\quad \text{since}\left|\rho_{\hat{R},\bar{x}}\right|\le 1.$$

$$Hence\ \frac{\left|B\left(\hat{R}\right)\right|}{\hat{\sigma}_{\hat{R}}}\le\frac{\frac{\sigma_x}{\bar{X}}}{}=cv\ \text{of}\ \bar{x}.$$

Theorem-6.3:- In large sample, with simple random sampling, the ratio estimate $\hat{Y}_R$ has a smaller variance then the estimate $\hat{Y}=N\bar{y}$ obtained by simple random sampling if

$$\rho>\frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right)=\frac{cv\ of\ x_i}{2\left(cv\ of\ \mathrm{y}_i\right)}=\frac{C_x}{2C_y}.$$

Proof:- For $\hat{Y}$ we have

$$V\left(\hat{Y}\right)=N^2\left(\frac{1-f}{n}\right)S_y^2\quad ----------(1)$$

For the ratio estimate we have

$$V\left(\bar{y}_R\right) = N^2\left(\frac{1-f}{n}\right)\left(S_y^2 + R^2S_x^2 - 2R\rho S_yS_x\right) \quad ----------(2)$$

Hence the ratio estimate has the smaller variance if

$$\therefore \text{var}\left(\bar{y}_R\right) < \text{var}\left(\bar{y}\right)$$

$$\therefore S_y^2 + R^2S_x^2 - 2R\rho S_yS_x < S_y^2$$

$$R^2S_x^2 < 2R\rho S_yS_x$$

$$\text{RS}_x < 2\rho S_y \Rightarrow \frac{RS_x}{2S_y} < \rho \qquad \left[\because R = \frac{\bar{Y}}{\bar{X}}\right]$$

$$\Rightarrow \rho > \frac{RS_x}{2S_y} = \frac{1}{2}\frac{\bar{Y}S_x}{\bar{X}S_y} = \frac{1}{2}\left(\frac{S_x}{\bar{X}}\right)\Big/\left(\frac{S_y}{\bar{Y}}\right)$$

$$\rho > \frac{1}{2}\frac{\left(cv \text{ of } x_i\right)}{\left(cv \text{ of } y_i\right)} = \frac{C_x}{2C_y}$$

Note:-

This theorem shows that the ratio estimate may be either more or less precise than a SRS estimate. The issue depends on the size of the correlation coefficient between y_i & x_i and on the CV's of these two variables. The variability of the auxiliary variate x_i is an important factor, if its cv is more than twice that of y_i, the ratio estimate is always less precise, since ∫ cannot exceed one . when x_i is the value of y_i at some previous time, the two CV's may be about equal. In this event the ratio estimate is superior if ρ >0.5.

14.4.1 BEST LINEAR UNBIASED ESTIMATE(BLUE):-

Consider all estimates that are linear functions of the sample values y_i, i.e., that are of the form $l_1y_1 + l_2y_2 + ---------+ l_ny_n$. Where the l's do not depend on the y_i, although they may be functions of x_i.

The choice of theses restricted to those that give unbiased estimates of $\bar{Y}$. The estimate that has be smallest variance is called Best Linear Unbiased Estimate (BLUE).

Theorem-6.4:-With simple random sampling from an infinite population, the ratio estimate of $\bar{Y}$ is a (BLUE) if two conditions are satisfied.

1. The relation between y_i and x_i is a straight line through the origin.
2. The variance of y_i about this line is proportional to x_i.

Proof:- The mathematical model is $y_i = Bx_i + e_i$

where e_i are independent of the x_i. In arrays in which x_i is fixed e_i has mean zero and variance λx_i

Hence $\bar{Y} = B\bar{X}$

It was shown by gauss that the BLUE of $B\overline{X}$ is $b\overline{X}$ where b is the least squares estimate of B. The least squares estimate is

$$b = \frac{\sum_i W_i y_i x_i}{\sum_i W_i x_i^2}, \text{ where } W_i = \frac{1}{\sigma^2_{e_i}} = \frac{1}{\lambda x_i} \text{ this gives,}$$

$$b = \frac{\sum_i \frac{1}{\lambda x_i} y_i x_i}{\sum_i \frac{1}{\lambda x_i} x_i^2} = \frac{\sum_{i=1}^{n} y_i/n}{\sum_{i=1}^{n} x_i/n} = \frac{\bar{y}}{\bar{x}} = R$$

Consequently the optimum estimate of $\overline{Y}$ is the ratio estimate

$$\left(\frac{\bar{y}}{\bar{x}}\right)\overline{X} = (R)\overline{X} = \overline{Y}_R$$

14.5 BIAS AND MEAN SQUARE ERROR OF RATIO ESTIMATOR:

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

$$E(\hat{\overline{Y}}_R) = \frac{1}{\binom{N}{n}} \sum_{i=1}^{\binom{N}{n}} \frac{\bar{y}_i}{\bar{x}_i} \overline{X}$$

$\neq \overline{Y}$ (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

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

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

Since SRSWOR is being followed, so

$E(\varepsilon_0) = 0$

$E(\varepsilon_1) = 0$

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

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

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

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

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

Similarly,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

If we assume that only terms of ε_0 and ε_1 involving powers more than two are negligibly small (which is more realistic than assuming that powers more than one are negligibly small), then the estimation error of $\hat{\bar{Y}}_R$ can be approximated as

$\hat{\bar{Y}}_R$ can be approximated as

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

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

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

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

upto the second order of approximation. The bias generally decreases as the sample size grows large.

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

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

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

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

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

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

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

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

Now, to find the mean squared error, consider

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

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

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

$$=\frac{\bar{Y}^2 f}{n}\left[C_X^2+C_Y^2-2\rho C_X C_Y\right]$$

up to the second-order of approximation.

14.6 ESTIMATION OF VARIANCE, CONFIDENCE INTERVAL AND COMPARISON WITH MEAN PER UNIT ESTIMATOR:

Estimation of Variance:

We have $V\left(\hat{R}\right)=\frac{1-f}{n\bar{X}^2}S_u^2$ .....(1)

Where

$S_u^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(U_i-\bar{U}\right)^2$, $U_i=Y_i-Rx_i$, i=1,2,.....,N

Hence a natural estimator of $V\left(\hat{R}\right)$ is

$$v_1\left(\hat{R}\right)=\frac{1-f}{n\bar{X}^2}S_u^2$$

Where

$$S_u^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}-\hat{R}\left(x_i-\bar{x}\right)\right)^2$$

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

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2+\hat{R}^2\sum_{i=1}^{n}x_i^2-2\hat{R}\sum_{i=1}^{n}x_i y_i\right]$$.....(2)

Since, in estimating $R,\bar{X}$ is not often known, an alternative estimator is

$$v_2\left(\hat{R}\right)=\frac{1-f}{n\bar{X}^2}S_u^2$$.....(3)

Two variance estimators of $\hat{y}_R$ are, therefore, approximately

$$v_1\left(\hat{y}_R\right)=X^2 v_1\left(\hat{R}\right)$$.....(4)

$$v_2\left(\hat{y}_R\right)=X^2 v_2\left(\hat{R}\right)$$

All these estimators $v_1\left(\hat{R}\right), v_2\left(\hat{R}\right), v_1\left(\hat{y}_R\right), v_2\left(\hat{y}_R\right)$ are biased.

Confidence Interval:

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

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

and

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

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

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

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

Comparison with Mean per Unit Estimator:

Under SRSWOR,

$$V(\bar{y}) = \frac{(1-f)S_y^2}{n}$$

$$V\left(\hat{\bar{y}}_R\right) = \frac{(1-f)\left[S_y^2 + R^2 S_x^2 - 2RS_{xy}\right]}{n}$$

Hence, $\hat{\bar{y}}_R$ will have a smaller variance than $\bar{y}$ in SRSWOR, if

$$R < 2\rho\frac{S_y}{S_x}$$

i.e.,

$$\rho > \frac{cv(x)}{2cv(y)} \quad .$$

14.7 RATIO ESTIMATOR IN STRATIFIED RANDOM SAMPLING:

When the population is stratified and units are drawn by simple random sampling method from each stratum. There are two ways of obtaining a ratio estimate of the population total Y.

1. Separate ratio estimate
2. Combined ratio estimate

1. **Separate ratio estimate:-**

If y_h, x_h are the sample totals in the h^{th} stratum and X_h is the stratum total of the x_{hi}, the separate ratio estimate $\hat{Y}_{RS}$ (s for separate) is defined as

$$\hat{Y}_{RS} = \sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{\bar{y}_h}{\bar{x}_h} X_h \qquad ---(1)$$

This estimate requires knowledge of the separate totals X_h.

2. **Combined ratio estimate**:-

From the sample data we compute

$$\hat{Y}_{st} = \sum_h N_h y_h, \hat{X}_{st} = N\bar{x}_h$$

These are standard estimates of the population totals Y and X respectively made from a stratified sample. Then the combined ratio estimate $\hat{Y}_{RC}$ (c for combined) is defined as

$$\hat{Y}_{RC} = \frac{\hat{Y}_{st}}{\hat{X}_{st}} X = \frac{N\bar{y}_{st}}{N\bar{x}_{st}} X = \frac{\bar{y}_{st}}{\bar{x}_{st}} x \qquad ---(2)$$

Where $\bar{y}_{st} = \frac{\hat{Y}_{st}}{N}, \bar{x}_{st} = \frac{\hat{X}_{st}}{N}$ are the estimated population means from a stratified sample. The estimate $\hat{Y}_{RC}$ does not require a knowledge of the X_h but only of X.

Theorem :-1

If the sample sizes n_h are large in all strata,

$$V\left(\hat{Y}_{RS}\right) = \sum_{h=1}^{L} \frac{N_h^2 (1 - f_h)}{n_h} \left[S_{yh}^2 + R_h^2 + S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh} \right] \qquad ---(3)$$

Where $R_h = \frac{Y_h}{X_h}$ is the two ratio in stratum h, and ρ_h is the correlation coefficient between y_{hi} and x_{hi} in h^{th} stratum.

Proof:- Write

$$\hat{Y}_{Rh} = \frac{y_h}{x_h} X_h$$

$$\hat{Y}_{Rs} = \sum_h \hat{Y}_{Rh}$$

$$Then \quad \hat{Y}_{Rs} - Y = \sum \left(\hat{Y}_{Rh} - Y_h \right) \qquad \left[\because Y = \sum_{h=1}^{L} Y_h \right]$$

$$Hence \quad V\left(\hat{Y}_{Rs}\right) = E\left(\hat{Y}_{Rs} - Y\right)^2$$

$$= \sum_h E\left(\hat{Y}_{Rh} - Y_h\right)^2 + 2\sum_h \sum_{j>h} E\left(\hat{Y}_{Rh} - Y_h\right)\left(\hat{Y}_{Rj} - Y_j\right)$$

$$= \sum V\left(\hat{Y}_{Rh}\right) + 22\sum_h \sum_{j>h} E\left(\hat{Y}_{Rh} - Y_h\right)\left(\hat{Y}_{Rj} - Y_j\right) \qquad ----------(4)$$

since $\hat{Y}_{Rh}$ is the ratio estimate made from a SRS with in stratum h, using the result of $V\left(\hat{Y}_R\right)$ [from corollary(1)] for the approximate variance of $V\left(\hat{Y}_{Rh}\right)$,we have

$$V\left(\hat{Y}_{Rh}\right) = \frac{N_h^2 (1 - f_h)}{n_h} \left[S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h \rho_h S_{yh} S_{xh} \right] \qquad ----------(5)$$

The cross product terms vanish because the sampling is independent in the different

strata and to the order of approximation used in the variance formula $\bar{Y}_{Rh}$ is an estimate of $\bar{Y}h$. Substitute eq^n (5) in eq^n (4), we get

$$V\left(\bar{Y}_{Rs}\right)=\sum_h \frac{N_h^2\left(1-f_h\right)}{n_h}\left[S_{yh}^2+R_h^2S_{xh}^2-2R_h\rho_hS_{yh}S_{xh}\right] .$$

Theorem 2: If the total sample size n is large,

$$v(\bar{Y}_{RC})=\sum_h N_h^2(1-f_h)/n_h[S_{yh}^2+R^2S_{xh}^2-2R\rho_hS_{yh}S_{xh}] \quad ----------(6)$$

Proof: We know that; $\bar{Y}_{RC}=\overline{y_{st}}/\ \bar{x}_{st}(X)=\overline{y_{st}}/\overline{x_{st}}(N\overline{X})$

$R=\overline{Y}/\overline{X};\ Y=N\overline{Y}$

$(\bar{Y}_{RC}-Y)=N\overline{X}/\overline{x_{st}}[\overline{y_{st}}-R\overline{x_{st}}]=N[\overline{y_{st}}-R\overline{x_{st}}]$ ----------(7)

since n is large $\overline{x_{st}}\ \square\ \overline{X}$

Now consider the variate $u_{hi}=y_{hi}-R\,x_{hi}$

The right hand side of eq^n (7) is $N\overline{u_{st}}$,where $\overline{u_{st}}$ the weighted mean of the variate u_{hi} in a stratified sample. Further ,the population mean $\overline{U}$ of u_{hi} is

$\overline{U}=\overline{Y}-R\overline{X}=\overline{Y}-(\overline{Y}/\overline{X})\overline{X}=0$

$E\left(\bar{Y}_{Rh}-Y\right)^2=V\left(\bar{Y}_{RC}\right)=N^2E\left(\bar{u}_{st}-\overline{U}\right)^2$

Applying the result of th-5.3[of] to we set

$$V\left(\bar{Y}_{RC}\right)=N^2V\left(\bar{u}_{st}\right)=N^2\left[\frac{1}{N^2}\sum_h N_h\left(N_h-n_h\right)\frac{S_{uh}^2}{n_h}\right]$$

$$\therefore V\left(\bar{Y}_{RC}\right)=\frac{\sum_h N_h\left(N_h-n_h\right)}{n_h}S_{uh}^2 \qquad \rightarrow(8)$$

Where $S_{uh}^2=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left(u_{hi}-\bar{u}_h\right)^2$

$$=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left[\left(y_{hi}-Rx_{hi}\right)-\left(\overline{Y}_h-R\overline{X}_h\right)\right]^2$$

$$=\frac{1}{N_h-1}\sum_{i=1}^{N_h}\left[\left(y_{hi}-\overline{Y}_h\right)-\left(x_{hi}-\overline{X}_h\right)\right]^2$$

$$S_{uh}^2=\sum_i\frac{\left(y_{hi}-\overline{Y}_h\right)^2}{N_h-1}+R^2\sum_i\frac{\left(x_{hi}-\overline{X}_h\right)^2}{N_h-1}-2R\sum_i\frac{(y_{hi}-\overline{Y}_h)(x_{hi}-\overline{X}_h)}{N_h-1}$$

$$=S_{yh}^2+R^2S_{xh}^2-2R\rho_hS_{yh}S_{xh} \qquad \rightarrow(9)$$

Where $\rho_h=\frac{\sum_i\left(y_{hi}-\overline{Y}_h\right)\left(x_{hi}-\overline{X}_h\right)}{\left(N_h-1\right)S_{yh}S_{xh}}$

Substituting equation (9) in equation (8) we get

$$V\left(\bar{Y}_{RC}\right)=\sum_h\frac{N_h^2}{N_h}\cdot\frac{\left(N_h-n_h\right)}{n_h}S_{uh}^2$$

$$V\left(\bar{Y}_{RC}\right)=\sum_{h}\frac{N_h^2\left(1-f_h\right)}{n_h}\left[S_{yh}^2+R^2S_{xh}^2-2R\rho_h\,S_{yh}.S_{xh}\right]$$

14.8 SUMMARY:

- **Ratio Estimator** is a widely used method in survey sampling that improves the estimation of population parameters by incorporating **auxiliary information**.
- It is particularly useful when there is a **strong positive correlation** between the study variable y and the auxiliary variable x.
- The basic form of the ratio estimator for the population mean is: $\hat{Y}_R=\bar{y}.\frac{X}{\bar{x}}$

  Where $\bar{y}$ and $\bar{x}$ are sample means of y and x, and X is the known population mean of $\bar{x}$
- **Notations and definitions** are introduced for precise understanding, and the estimator is analyzed both theoretically and through practical examples.
- **Bias and Mean Square Error (MSE)** are derived to understand the estimator's accuracy. Although the ratio estimator is biased, the bias is generally small when the sample size is large.
- The **variance and confidence intervals** are estimated to quantify uncertainty, and the estimator is compared with the **mean per unit estimator**. Ratio estimator is more efficient when the correlation ρ between y and x is high.
- The **stratified ratio estimator** is an extension applied within strata to further increase the efficiency of estimates by reducing within-stratum variability.
- The ratio estimator is a powerful technique for improving survey accuracy, especially when reliable auxiliary data is available.
- While it introduces a small bias, it **reduces variance**, often resulting in a **lower MSE** compared to unbiased estimators like the mean per unit estimator.
- Its utility is enhanced in **stratified random sampling**, making it suitable for complex survey designs.
- Overall, the ratio estimator is a **practical and efficient** tool in survey sampling when used under appropriate conditions.

14.9 KEY WORDS:

- **Ratio Estimator**
- **Auxiliary Variable**
- **Bias**
- **Mean Square Error (MSE)**
- **Efficiency**
- **Confidence Interval**
- **Stratified Sampling**
- **Sample Mean**

- **Population Mean**
- **Correlation Coefficient**
- **Estimator Variance**

14.10 SELF-ASSESSMENT QUESTIONS:

1. Define a ratio estimator and state its formula.
2. Under what condition is the ratio estimator preferred over the mean per unit estimator?
3. Derive the approximate bias of the ratio estimator.
4. Explain how variance of the ratio estimator is estimated.
5. How is ratio estimation extended to stratified random sampling?
6. Write about Ratio Estimator and derive its variance
7. Derive Mean Square Error for Ratio Estimator.
8. Explain ratio estimation. Obtain the variances of ratio estimates in stratified sampling.

14.11 SUGGESTED READINGS:

1. **Cochran, W.G. (1977),** *Sampling Techniques* (3rd Edition), Wiley.
2. **P. Mukhopadhyay**, Title: *Theory and Methods of Survey Sampling,* PHI Learning Pvt. Ltd., New Delhi.
3. **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984).** *Sampling Theory of Surveys with Applications,* Publisher: Iowa State University Press.
4. Singh, D., and Chaudhary, F.S. **(1986).** *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.

Dr. U. Ramkiran

LESSON -15

DIFFERENCE ESTIMATOR

OBJECTIVES:

By the end of this module, learners will be able to:

- Understand the concept and application of the Difference Estimator and the Regression Estimator**.**
- Derive and interpret the bias**,** mean square error (MSE)**,** and variance for these estimators.
- Apply the estimators in simple random sampling and stratified random sampling frameworks.
- Compare efficiency of estimators such as Mean per Unit**,** Ratio**,** Difference, and Regression estimators.
- Develop and analyze confidence intervals using auxiliary information in survey sampling.

STRUCTURE:

15.1 INTRODUCTION:

A "difference estimator" and a "regression estimator" are both statistical techniques used to improve the accuracy of population parameter estimates by leveraging information from an auxiliary variable that is highly correlated with the variable of interest, but while a difference estimator simply calculates the difference between the study variable and a weighted version of the auxiliary variable, a regression estimator utilizes a linear regression model to establish a more precise relationship between the two variables, resulting in potentially more efficient estimates, particularly when the correlation between the variables is strong.

Key points to remember:

- **Difference estimator:**
  - Uses a simple weighted difference between the study variable and the auxiliary variable, where the weight is a constant value determined based on the known population means of both variables.
  - Best suited for situations where the relationship between the study and auxiliary variables is relatively linear and the correlation is moderate.
- **Regression estimator:**
  - Leverages a regression line to estimate the relationship between the study variable and auxiliary variable, allowing for a more nuanced adjustment based on the observed data.
  - Generally considered more efficient than a difference estimator, especially when the correlation between variables is strong and the relationship is not strictly linear.

When to use each:

a **difference estimator** is used When the relationship between the study variable and the auxiliary variable is relatively simple and a linear approximation is sufficient.

a **regression estimator** is used When there is a strong, potentially non-linear relationship between the study variable and the auxiliary variable, and you want to leverage the full information from the regression model.

15.2 CONCEPT OF DIFFERENCE ESTIMATOR:

The ratio estimator is most suitable when the relation between y and x is a straight line passing through the origin. i.e., when the regression equation of y on x is $y = kx$. In case x, y are related such that for unit increase in value of x, y increases by an amount k, where k is a constant, we may assume that

$\bar{Y} - \bar{y} = k\left(\bar{X} - \bar{x}\right)$

i.e., if $\bar{x}$ is below $\bar{X}$ by an amount $\left(\bar{X} - \bar{x}\right)$, then $\bar{y}$ is expected to be below $\bar{Y}$ by an amount $k\left(\bar{X} - \bar{x}\right)$. Hence, we get an estimator of $\bar{Y}$ as

$\hat{\bar{y}}_D = \bar{y} + k\left(\bar{X} - \bar{x}\right)$ ……(1)

$\hat{\bar{y}}_D$ is the difference estimator of $\bar{Y}$,first considered by Hansen, Hurwitz and Madow (1953). In general, for an arbitrary sampling design, the difference estimator of $\bar{Y}$ is $\hat{\bar{Y}} + k\left(\bar{X} - \hat{\bar{X}}\right)$, where $\hat{\bar{Y}}$, $\hat{\bar{X}}$ are respectively estimators of $\bar{Y}$, $\bar{X}$ under this sampling procedure.

15.3 THEOREM-1:

In SRSWOR, (N, n), $\hat{\bar{y}}_D$ is an unbiased estimator of $\bar{Y}$ with

$$V\left(\hat{\bar{y}}_D\right) = \frac{1-f}{n}\left(S_y^2 + k^2 S_x^2 - 2kS_{xy}\right) \text{……(2)}$$

and an unbiased variance estimator

$$v\left(\hat{\bar{y}}_D\right)=\frac{1-f}{n}\left(S_y^2+k^2S_x^2-2ks_{xy}\right)\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Proof: Since k is a constant, unbaisedness follows readily. Writing $U_i=Y_i-kx_i$, $\bar{u}=\sum_{i=1}^{n}u_i/n$

$$V\left(\hat{\bar{y}}_D\right)=V\left(\bar{u}\right)=\frac{\left(1-f\right)S_u^2}{n}$$

Where

$$S_u^2=\sum_{i=1}^{n}\frac{\left(u_i-u\right)^2}{N-1}\frac{1}{N-1}\sum_{I=1}^{N}\left(Y_i-\bar{Y}-k\left(x_i-\bar{X}\right)\right)^2$$

Clearly,

$$V\left(\hat{\bar{y}}_D\right)=\frac{\left(1-f\right)S_u^2}{n}$$

Where

$$S_u^2=\sum_{i=1}^{n}\frac{\left(u_i-u\right)^2}{n-1}\frac{1}{n-1}\sum_{I=1}^{n}\left(y_i-\bar{y}-k\left(x_i-\bar{x}\right)\right)^2$$

Corollary-1: $\hat{\bar{y}}_D$ is superior to $\bar{y}$ (in the smaller variance sense) if $k\left(k-2B\right)<0$ or $0<\mathrm{k}<2\mathrm{B}$, where $\mathrm{B}=\rho\mathrm{S_y}/\mathrm{S}_x$, the finite population regression coefficient.

Corollary-2: For $k=R=Y/X$, the variance of $\hat{\bar{y}}_D$ is the same as the asymptotic expression for $V\left(\hat{\bar{y}}_R\right)$.

Corollary-3: For $k=1$,

$$V\left(\hat{\bar{y}}_D\right)=V\left(\bar{y}\right)\left\{1+\frac{S_x^2}{S_y^2}-2\rho\frac{S_x}{S_y}\right\}\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Hence, in this case, $\bar{y}_D$ has smaller variance than $\bar{y}$ if $\rho>S_x/2S_y$. Comparing (3) with the approximate expression for $V\left(\hat{\bar{y}}_R\right)$, (assuming R > 0), $V\left(\hat{\bar{y}}_D\right)<V\left(\hat{\bar{y}}_R\right)\ for k=1, if$

$$\rho>\left(<\right)\frac{\left(R+1\right)S_x}{2S_y}\ \text{ for } 0<\mathrm{R}<1\ \left(\mathrm{R}>1\right)$$

The optimum values of k, which minimize $V\left(\hat{\bar{y}}_D\right)$ are obtained by differentiating $V\left(\hat{\bar{y}}_D\right)$ in eq$^{\mathrm{n}}$ (2) with respect to k and equating it to zero. The gives the optimum k as

$$k^*=\rho\frac{S_y}{S_x}=B\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

At k^*,the second derivative of $V\left(\hat{\bar{y}}_D\right)$ is positive and hence $V\left(\hat{\bar{y}}_D\right)$ attains its minimum value which is

$$\frac{1-f}{n}\left(S_y^2-B^2S_x^2\right)=\frac{\left(1-f\right)S_y^2\left(1-\rho^2\right)}{n}\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

In general B will be unknown. However, if a good guesses value b_0 of B is available. Say, from a past survey, this may be used for k. Cochran has shown that if the proportional increase in $V\left(\hat{\bar{y}}_D\right)$

over eq^n (6) is to be less than α , we must have,

$$\left|b_0/B-1\right|<\sqrt{\alpha\left(1-\rho^2\right)/\rho^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

Hence, if ρ is very high, α small, b_0/B must be close to unity; however, if ρ is only moderate, it may depart substantially from unity. As an example, for $\alpha=0.1,\rho=0.4, 0.2754<b_0/B<1.7246.$ Thus for moderate ρ ,the choice of b_0 is somewhat robust to the moderate departures from B.
In case x is the value of y measured at a previous time, B is close to unity and k should be set of equal to unity.

15.4 EXAMPLE:

- Imagine estimating the average income of households in a city. If it's difficult to collect income data for every household, but data on the average house value is readily available, you could use the average house value as the auxiliary variable in a difference estimator to improve the accuracy of your income estimate.

Let's say we are estimating the average weight y of a group of students. We also know their average height x and the population mean height $\overline{X}$.

From the sample:

- $\overline{y}=62kg$
- $\overline{x}=165cm$
- $\overline{X}=167cm$

Using the difference estimator:

$$\hat{Y}_D=\overline{y}\left(\overline{X}-\overline{x}\right)$$
$$\hat{Y}_D=62\left(167-165\right)$$
$$\hat{Y}_D=64\text{kg}$$

So, the adjusted estimate of the population mean weight is **64 kg**.

15.5 DIFFERENCE ESTIMASTOR IN STRATIFIED RANDOM SAMPLING:

Consider a stratified random sampling with N_h, n_h as population size and sample size respectively for the h^{th} stratum. The separate difference estimator (like the separate ratio estimator), assuming the strata regression coefficients $B_h\left(h=1,2,.....,L\right)$ to be known, is

$$\hat{\overline{y}}_{DS}=\sum_{h=1}^{L}W_h\left[\overline{y}_h+B_h\left(\overline{X}_h-\overline{x}_h\right)\right]\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

with

$$V\left(\hat{\bar{y}}_{DS}\right)=\sum_{h=1}^{L}\frac{W_h^2\left(1-f_n\right)\left(S_{yh}^2-B_h^2S_{xh}^2\right)}{n_h}$$

$$V\left(\hat{\bar{y}}_{DS}\right)=\sum_{h=1}^{L}\frac{W_h^2\left(1-f_n\right)S_{yh}^2\left(1-\rho_h^2\right)}{n_h} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Where

$$W_h=\frac{N_h}{N},\ \mathrm{f}_h=\frac{n_h}{n},\ \rho_h=\frac{S_{xyh}}{S_{xh}S_{yh}}$$

A combined difference estimator (like a combined ratio estimator) is

$$\hat{\bar{y}}_{DC}=\bar{y}_{st}+k\left(\bar{X}-\bar{x}_{st}\right)$$

Where

$\bar{y}_{st}=\sum W_h\bar{y}_h,\ \bar{x}_{st}=\sum W_h\bar{x}_h$ and k is a suitable constant. To determine the optimum value of k, one has to minimize

$$V\left(\hat{\bar{y}}_{DC}\right)=V\left(\bar{y}_{st}\right)+k^2V\left(\bar{x}_{st}\right)-2Cov\left(\bar{y}_{st},\bar{x}_{st}\right)$$

with respect to k**.** It is seen that the optimum value of k is

$$k^*=\frac{Cov\left(\bar{x}_{st},\bar{y}_{st}\right)}{V\left(\bar{y}_{st}\right)}=\frac{\sum W_h^2\mathrm{S}_{xyh}\left(1-f_h\right)/n_h}{\sum W_h^2\mathrm{S}_{xh}\left(1-f_h\right)/n_h}$$

$$k^*=\frac{\sum\lambda_hB_h}{\sum\lambda_h}=B^* \text{ (say)} \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Key Points:

- **Choice of Auxiliary Variable:**

The accuracy of the difference estimator depends heavily on the selection of a highly correlated auxiliary variable (x) with the study variable (y).

- **Advantages:**
  - More efficient than simple random sampling when there is significant variation within strata but less variation between strata.
  - Can be particularly useful when information on the auxiliary variable is readily available.

- **Limitations:**
  - Requires prior knowledge of the population structure to properly define strata.
  - If the auxiliary variable is poorly correlated with the study variable, the estimator may not be accurate.

Study variable - y_i, Auxiliary variable - x_i

Each unit having a pair of units (x_i, y_i), $Y, \bar{Y}, \hat{Y}, \hat{\bar{Y}}$

We want to estimate $\frac{\bar{Y}}{\bar{X}}=\frac{Y/N}{X/N}=R$ and $\hat{R}$ is called "the ratio of two estimates".

Problem Statement:

A population of households is divided into 2 strata based on income levels:

- **Stratum 1:** Low-income households
- **Stratum 2:** High-income households

We are interested in estimating the **average electrici ↓ onsumption (Y)** using the **number of electrical appliances (X)** as an auxiliary variable.

From each stratum, a **simple random sample without replacement** is drawn. The following data are available:

Stratum	N_h	n_h	$\bar{y}_h$	$\bar{x}_h$	$\bar{X}_h$
1	500	50	120 kWh	4	5
2	300	30	200 kWh	6	6.5

Here:

- N_h: Population size in stratum h
- n_h: Sample size in stratum h
- $\bar{y}_h$: Sample mean of study variable Y in stratum h
- $\bar{x}_h$: Sample mean of auxiliary variable X in stratum h
- $\bar{X}_h$: Population mean of auxiliary variable X in stratum h (known)

Step 1: Compute the Difference Estimator for each stratum

The **difference estimator** for each stratum is:

$$\hat{Y}_{d,h} = \bar{y}_h + (\bar{X}_h - \bar{x}_h)$$

- For **Stratum 1**:

$$\hat{Y}_{d,1} = 120 + (5 - 4) = 121$$

- For **Stratum 2**:

$$\hat{Y}_{d,2} = 200 + (6.5 - 6) = 200.5$$

Step 2: Calculate the Weighted Stratified Estimate

The overall stratified difference estimator is:

$$\hat{Y}_d = \sum_{h=1}^{L} \frac{N_h}{N} \hat{Y}_{d,h}$$

Where total population size $N = N_1 + N_2 = 500 + 300 = 800$

$$\hat{Y}_d = \left(\frac{500}{800} \cdot 121\right) + \left(\frac{300}{800} \cdot 200.5\right)$$

$$\hat{Y}_d = (0.625 \cdot 121) + (0.375 \cdot 200.5)$$

$$\hat{Y}_d = 75.625 + 75.1875 = \boxed{150.8125 \text{ kWh}}$$

Conclusion:

The stratified difference estimator for average electricity consumption in the population is approximately 150.81 kWh.

15.6 SUMMARY:

- The difference estimator uses auxiliary information to improve the estimation of the population mean.
- It is most effective when the auxiliary variable xxx is positively correlated with the study variable y.
- The estimator is unbiased under SRSWOR and often more efficient than the sample mean if the variability in $y - x$ is small.
- The method is also extended to stratified sampling, further enhancing precision.
- The choice between ratio and difference estimator depends on the correlation structure and variance of y and x.
- When properly used, it offers substantial gains over the sample mean estimator.

15.7 KEY WORDS:

- **Difference Estimator**
- **Auxiliary Variable**
- **Sample Mean**
- **Population Mean**
- **Unbiased Estimator**
- **Stratified Sampling**

15.8 SELF-ASSESSMENT QUESTIONS:

1. What is the basic idea behind the difference estimator?
2. Under what conditions is the difference estimator more efficient than the sample mean?
3. Write the formula for the difference estimator and define each term.
4. State and explain the theorem regarding the bias of the difference estimator.
5. Derive the variance formula of the difference estimator under SRSWOR.
6. What is the role of the auxiliary variable in constructing a difference estimator?
7. How is the difference estimator extended to stratified sampling?
8. Compare the efficiency of the sample mean and the difference estimator.
9. What are the assumptions required for the difference estimator to perform well?
10. Give a numerical example of how to compute a difference estimator.

15.9 SUGGESTED READINGS:

1. **Cochran, W.G. (1977),** *Sampling Techniques* (3rd Edition), Wiley.
2. **P. Mukhopadhyay**, Title: *Theory and Methods of Survey Sampling,* PHI Learning Pvt. Ltd., New Delhi.
3. **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984).** *Sampling Theory of Surveys with Applications,* Publisher: Iowa State University Press.

Dr. U. Ramkiran

LESSON -16

REGRESSION METHOD OF ESTIMATOR

OBJECTIVES:

After completing this unit, learners will be able to:

- Understand the Concept of Regression Estimator
  - Grasp the logic behind using auxiliary variables to improve the estimation of population parameters.
  - Explain the theoretical foundation and practical application of the regression estimator.
- Identify Key Notations and Formulas
  - Familiarize with important terms such as regression coefficient, sample mean, population mean, covariance, and variance.
  - Use standard notation for expressing regression-based estimates.
- Evaluate the Bias and Mean Square Error (MSE)
  - Derive expressions for the bias and mean square error of the regression estimator.
  - Understand under what conditions the regression estimator becomes unbiased or achieves minimum MSE.
- Construct Variance Estimates and Confidence Intervals
  - Learn how to estimate the variance of regression estimators.
  - Construct confidence intervals for population parameters using the regression approach.
- Compare Regression Estimator with Other Estimators
  - With Mean per Unit Estimator**:** Understand when regression estimators are more efficient and under what assumptions.
- With Ratio Estimator**:** Recognize the difference in applicability depending on the relationship between variables.
- Apply Regression Estimator in Stratified Sampling
  - Adapt the regression estimator for stratified random sampling scenarios.
  - Compute combined regression estimates using stratum-wise regression adjustments.
- Analyze Efficiency and Practical Use Cases
  - Evaluate efficiency gains through numerical examples.
  - Identify practical conditions for using regression estimators in survey sampling.

STRUCTURE:

16.1 INTRODUCTION:

The ratio method of estimation uses the auxiliary information, which is correlated with the study variable to improve the precision, which results in improved estimators when the regression of *Y* on *X* is linear and passes through the origin. When the regression of on *X* is linear, the line doesn't need to always pass through the origin. Under such conditions, it is more appropriate to use the regression type estimator to estimate the population means.

In the ratio method, the conventional estimator sample mean $\bar{y}$ was improved by multiplying it by a factor $\frac{\bar{X}}{\bar{x}}$ where $\bar{x}$ is an unbiased estimator of the population mean $\bar{X}$ which is chosen as the population mean of the auxiliary variable. Now, we consider another idea based on the differences.

Consider an estimator of $\bar{x}$ of $\bar{X}$ for which $E\left(\bar{x} - \bar{X}\right) = 0$

Description: Like the ratio estimate, the linear regression estimate is designed to increase precision by the use of an auxiliary variate x_i which is correlated with y_i. When the relation between y_i and x_i which is correlated with examined, it may be found that although the relation is approximately linear. The line does not go through the origin. This suggest an estimate based on the linear regression of y_i and x_i rather than on the ratio of the two variables we suppose that y_i and x_i are each obtained for every unit in the sample and that the population mean $\bar{X}$ of the x_i is known. The linear regression estimate of $\bar{y}$, the population mean of the y_i is

$$\bar{y}_{lr} = \bar{y} + b\left(\bar{X} - \bar{x}\right) = \bar{y} - b\left(\bar{x} - \bar{X}\right)$$

where the subscript "$_{lr}$" denotes linear regression and ''b" is an estimate of the change in y when x is increased by unity. The rational of this estimate is left, since $\bar{x}$ is below the average by an amount $b\left(\bar{X} - \bar{x}\right)$ be of the regression of y_i y_i on x_i.

For an estimate of the population total Y, we take $\hat{Y}_{lr} = N\bar{y}_{lr}$.

Examples:

Applications (or) Situations:

(1) Watson (1937) used a regression of leaf area on leaf weight to estimate the average area of the leaves on a plant. The procedure was to weight all the leaves on the plant. For a small sample of leaves, the area and the weight of each leaf were then adjusted by means of the regression on leaves weight. The point of the application is ofcourse that the weight of a leaf can be found quickly but determination of it's area is more time consuming. This example illustrate a general situation in which regression estimates are helpful.

(2) A rat expert might make a quick eye estimate of the no. of rats in each block in a city are and then determine by trapping the actual number of rats in each of a SRS of the blocks.

(3) An eye estimate of the volume of timber was made on each of a population of $\frac{1}{10}$ acre plots and the actual timber volume was measured for a sample of plots.

16.2 CONCEPT OF REGRESSION ESTIMATOR:

The regression estimator makes use of the linear relationship between two variables:

$$Y = \alpha + \beta X + \epsilon$$

In sampling, we use the sample regression line to estimate the population mean or total of Y based on the known population mean $\bar{X}$ of the auxiliary variable.

The regression estimator of the population mean $\bar{Y}$ is:

$$\hat{Y}_{reg} = \bar{y} + b(\bar{X} - \bar{x})$$

Where:

- $\bar{y}$: Sample mean of Y
- $\bar{x}$: Sample mean of X
- $\bar{X}$: Population mean of X (known)
- b: Sample regression coefficient (slope)

16.3 NOTATIONS AND DEFINITION:

Let:

- Y_i: Value of the study variable for the ith unit
- X_i: Value of the auxiliary variable for the ith unit
- N: Population size
- n: Sample size
- $\bar{Y}$, $\bar{X}$: Population means of Y and X
- $\bar{y}$, $\bar{x}$: Sample means of Y and X
- S_{xy}: Sample covariance
- S_x^2: Sample variance of X
- $b = \frac{S_{xy}}{S_x^2}$: Estimated regression coefficient

Regression Estimation with Preassigned (b):

Although in most applications , b is estimated from the results of the sample, it is some time reasonable to choose the value of b in advance. In repeated surveys, previous calculations may have shown that the sample values of b remain fairly constant.

Since the sampling theory of regression estimates when b is pre assigned in both simple and information, this case is considered first.

16.3.1 THEOREMS:

Statement: In SRSing , in which b_0 is a pre-assigned constant, the linear regression estimate. $\bar{y}_{lr} = \bar{y} + b_0(\bar{X} - \bar{x})$ is unbiased with variance

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\sum_{i=1}^{N}\frac{\left[(y_i-\bar{Y})-b_0(X_i-\bar{X})\right]^2}{N-1}$$

$$= \frac{1-f}{n}(S_y^2 - 2b_0S_{yx} + b_0^2S_x^2)$$

Proof : since b_0 is a constant in repeated sampling

$$E(\bar{y}_{lr}) = E(\bar{y}) - b_0E(\bar{x} - \bar{X})$$

$$\therefore E(\bar{y}_{lr}) = \bar{Y} - 0 = \bar{Y}$$

$\therefore \bar{y}_{lr}$ is an unbiased estimate of $\bar{Y}$.

Further, $\bar{y}_{lr}$ is the sample mean of the quantities $u_i = y_i - b_0(x_i - \bar{X})$

Whose population mean is $\bar{Y}$.

When $V(\bar{y}_{lr}) = \frac{1-f}{n}\sum_{i=1}^{N}\frac{\left[(y_i-\bar{Y})-b_0(X_i-\bar{X})\right]^2}{N-1}$

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left[\sum_{i=1}^{N}\frac{(y_i-\bar{Y})^2}{N-1} + b_0^2\frac{\sum_{i=1}^{N}(x_i-\bar{X})^2}{N-1} - 2b_0\sum_{i=1}^{N}\frac{(y_i-\bar{Y})(x_i-\bar{X})}{N-1}\right]$$

$$V(\bar{y}_{lr}) = \frac{1-f}{n}\left[S_y^2 + b_0^2S_x^2 - 2b_0\rho S_yS_x\right]$$

$$= \frac{1-f}{n}\left[S_y^2 + b_0^2S_x^2 - 2b_0S_{yx}\right], \text{where } \rho = \frac{\sum_i(y_i-\bar{Y})(x_i-\bar{X})}{(N-1)S_yS_x}$$

Corollary: an unbiased estimate of $V(\bar{y}_{lr})$ is $v(\bar{y}_{lr}) = \frac{1-f}{n}\sum_{i=1}^{N}\frac{\left[(y_i-\bar{Y})-b_0(x_i-\bar{X})\right]^2}{n-1}$

Theorem 7.2 :- The value of b_0 which minimize $V(\bar{y}_{lr})$ is

$$b_0 = B = \frac{S_{yx}}{S_x^2} = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})}{\sum_{i=1}^{N}(x_i-\bar{X})^2} = \text{Cov(Y.X)/Var(X)}$$

Which may be called the linear regression coefficient of y and x in the finite population.

The resulting minimum variance is

$$V_{min}\left(\bar{y}_{lr}\right)=\frac{1-f}{n}S_y^2\left(1-\rho\right)^2$$

Proof : we know that $V\left(\bar{y}_{lr}\right)=\frac{1-f}{n}\left(S_y^2+b_0^2S_x^2-2b_0S_{yx}\right)$

Put $b_0=B+d=\frac{S_{yx}}{S_x^2}+d$

This gives, $V\left(\bar{y}_{lr}\right)=\frac{1-f}{n}\left(S_y^2+S_x^2\left(\frac{S_{yx}^2}{S_x^4}+2d\frac{S_{yx}}{S_x^2}+d^2\right)-2S_{yx}\left(\frac{S_{yx}}{S_x^2}+d\right)\right)$

$$=\frac{1-f}{n}\left(S_y^2+\frac{S_{yx}^2}{S_x^2}+2d\,S_{yx}+d^2S_x^2-2\frac{S_{yx}^2}{S_x^2}-2d\,S_{yx}\right)$$

$$=\frac{1-f}{n}\left[\left(S_y^2-\frac{S_{yx}^2}{S_x^2}\right)+d^2S_x^2\right]$$

Clearly this is minimized when d=0 (as other terms depend on data), since $\rho^2=\frac{S_{yx}^2}{S_x^2S_y^2}$

$$\therefore V_{min}\left(\bar{y}_{lr}\right)=\frac{1-f}{n}\left(S_y^2-\rho^2S_y^2\right)$$

$$V_{min}\left(\bar{y}_{lr}\right)=\frac{1-f}{n}S_y^2\left(1-\rho^2\right)$$

Result: Show that in samples of size n the quantity $\left(b-B\right)$ is of order $\frac{1}{\sqrt{n}}$ where b is the least squares estimate of B

Proof: Define the variate e_i, by the relation

$$e_i=\left(y_i-\bar{Y}\right)-B\left(x_i-\bar{X}\right) \qquad \rightarrow(1)$$

It follows that,

$$\sum_{i=1}^{N}e_i\left(x_i-\bar{X}\right)=\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)\left(x_i-\bar{X}\right)-B\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2$$

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

$$\sum_{i=1}^{N}e_i\left(x_i-\bar{X}\right)=0 \qquad \rightarrow(2)$$

$$b=\frac{\sum_{i=1}^{n}y_i\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{\sum_{i=1}^{n}\left[\bar{Y}+B\left(x_i-\bar{X}\right)+e_i\right]\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \qquad [\because \text{using equation(1)}]$$

$$= \frac{\sum_{i=1}^{n} \bar{Y}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + B\frac{\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n} e_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

R.H.S of I -term is

i.e., $\sum_{i=1}^{n} \bar{Y}(x_i - \bar{x}) = \bar{Y}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Since sum of the deviations from its mean is zero.

R.H.S of II -term is

i.e., $B\frac{\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = B$

When n is large $\bar{X}$ and $\bar{x}$ may not differ much to simplify this

$$\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\therefore b = B + \frac{\sum_{i=1}^{n} e_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \rightarrow (3)$$

But $\frac{\sum_{i=1}^{n} e_i(x_i - \bar{x})}{(n-1)}$ is an unbiased estimate of $\frac{\sum_{i=1}^{n} e_i(x_i - \bar{X})}{(N-1)}$ =0. by using(2)].

In repeated samples of size n the sample co-variance is. $\frac{\sum_{i=1}^{n} e_i(x_i - \bar{x})}{(n-1)}$ is therefore distributed about a zero mean. The standard error of a sample covariance is known to be of order $\frac{1}{\sqrt{n}}$. This in samples of size n, $\frac{\sum_{i=1}^{n} e_i(x_i - \bar{X})}{(n-1)}$ will be of order $\frac{1}{\sqrt{n}}$.

But the quantity , $\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{(n-1)} = s_x^2$ is of order unity in samples of size n.

Hence from equation (3) $(b-B)$ is of order $\frac{1}{\sqrt{n}}$.

Theorem -7.3: If b is the least squares estimate of B and $\bar{y}_{lr} = \bar{y} + b(X_i - \bar{x}) \rightarrow (1)$ then in SRS of size n,

$V(\bar{y}_{lr}) = \frac{1-f}{n} S_y^2(1 - \rho^2)$ is provided that n is large enough so that terms of order $\frac{1}{\sqrt{n}}$ is negligible.

Proof: Define the variate e_i as

$$e_i = y_i - \bar{Y} - B\left(x_i - \bar{x}\right) \quad \rightarrow (2)$$

Averaging this over the units in the sample, we have

$$\bar{e} = \bar{y} - \bar{Y} - B\left(\bar{x} - \bar{X}\right)$$

$$\bar{y} = \bar{Y} + B\left(\bar{x} - \bar{X}\right) + \bar{e}$$

Substitute this value of $\bar{y}$ in equation (1)

$$\bar{y}_{lr} = \bar{Y} + B\left(\bar{x} - \bar{X}\right) + \bar{e} + b\left(\bar{X} - \bar{x}\right)$$

$$\bar{y}_{lr} = \bar{Y} + \left(b - B\right)\left(\bar{X} - \bar{x}\right) + \bar{e}$$

From equation (2) , it is clear that she population mean e_i is zero. Hence $\bar{e}$ is of order $\frac{1}{\sqrt{n}}$. But we know that $\left(b - B\right)$ is order of $\frac{1}{\sqrt{n}}$. Since $\left(\bar{x} - \bar{X}\right)$ is also of order $\frac{1}{\sqrt{n}}$, their product $\left(b - B\right)\left(\bar{x} - \bar{X}\right)$ is order $\frac{1}{n}$. Consequently their product can be ignored relative to $\bar{e}$ is in terms of order $\frac{1}{\sqrt{n}}$ are negligible.

This gives $\bar{y}_{lr} = \bar{Y} + \bar{e} \Rightarrow \bar{y}_{lr} - \bar{Y} = \bar{e}$

Since $E\left(\bar{e}\right) = 0$ then $E\left(\bar{e}^2\right)$ is the variance of the mean of the quantities e_i in a SRS .

Hence $E\left(\bar{y}_{lr} - \bar{Y}\right)^2 = E\left(\bar{e}^2\right)$

$$V\left(\bar{y}_{lr}\right) = \frac{1-f}{n} S_e^2 = \frac{1-f}{n} \frac{\sum_{i=1}^{N} e_i^2}{N-1} \quad \rightarrow (3)$$

Now $\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[\left(y_i - \bar{Y}\right) - B\left(x_i - \bar{X}\right)\right]^2$ [$\because$ from equation (2)]

$$= \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 - 2B \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right) + B^2 \sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2$$

$$= \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 - 2B^2 \sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2 + B^2 \sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2 \left[\because \left(y_i - \bar{Y}\right) = B\left(x_i - \bar{X}\right)\right]$$

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 - B^2 \sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2 \quad \rightarrow (4)$$

$$= \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 - \rho^2 \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2$$

Where $\rho = \frac{\sum_i \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{\sqrt{\sum \left(x_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2}}$

$$\rho = \frac{B\sqrt{\sum \left(x_i - \bar{X}\right)^2}}{\sqrt{\sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2}}$$

$$\Rightarrow \rho^2 \Sigma\left(y_i - \bar{Y}\right)^2 = B^2 \Sigma\left(x_i - \bar{X}\right)^2$$

$$\Sigma e_i^2 = \sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2\left(1-\rho^2\right) = (N-1)S_y^2\left(1-\rho^2\right) \quad \rightarrow (5)$$

Substituting (5) in (3) we have

$$V\left(\bar{y}_{lr}\right) = \frac{(1-f)}{n} S_y^2\left(1-\rho^2\right)$$

16.4 BIAS OF THE REGRESSION ESTIMATE:

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

Now, assuming that the random sample (x_i, y_i), $i = 1, 2, ..., n$ is drawn by SRSWOR,

$$\begin{aligned} E(\hat{\bar{Y}}_{reg}) &= E(\bar{y}) + \beta_0\left[\bar{X} - E(\bar{x})\right] \\ &= \bar{Y} + \beta_0\left[\bar{X} - \bar{X}\right] \\ &= \bar{Y} \end{aligned}$$

Thus $\hat{\bar{Y}}_{reg}$ is an unbiased estimator of $\bar{Y}$ when β is known.

Variance of $\hat{\bar{Y}}_{reg}$

$$\begin{aligned} Var(\hat{\bar{Y}}_{reg}) &= E\left[\hat{\bar{Y}}_{reg} - E(\hat{\bar{Y}}_{reg})\right]^2 \\ &= E\left[\bar{y} + \beta_0(\bar{X} - \bar{x}) - \bar{Y}\right]^2 \\ &= E\left[(\bar{y} - \bar{Y}) - \beta_0(\bar{x} - \bar{X})\right]^2 \\ &= E\left[(\bar{y} - \bar{Y})^2 + \beta_0^2(\bar{x} - \bar{X})^2 - 2\beta_0 E(\bar{x} - \bar{X})(\bar{y} - \bar{Y})\right] \\ &= Var(\bar{y}) + \beta_0^2 Var(\bar{x}) - 2\beta_0 Cov(\bar{x}, \bar{y}) \\ &= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 S_{XY}\right] \\ &= \frac{f}{n}\left[S_Y^2 + \beta_0^2 S_X^2 - 2\beta_0 \rho S_X S_Y\right] \end{aligned}$$

where

$$f = \frac{N-n}{N}$$

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

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

ρ : Correlation coefficient between X and Y.

Comparing $Var(\hat{\bar{Y}}_{reg})$ with $Var(\bar{y})$, we note that

$$Var(\hat{\bar{Y}}_{reg}) < Var(\bar{y})$$

if $\beta_0^2 S_X^2 - 2\beta_0 S_{XY} < 0$

or $\beta_0 S_X^2\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) < 0$

which is possible when

either $\beta_0 < 0$ and $\left(\beta_0 - \frac{2S_{XY}}{S_X^2}\right) > 0 \Rightarrow \frac{2S_{XY}}{S_X^2} < \beta_0 < 0$.

Optimal value of β

Choose β such that $Var(\hat{\bar{Y}}_{reg})$ is minimum.

So

$$\frac{\partial Var(\hat{\bar{Y}}_{reg})}{\partial \beta} = \frac{\partial}{\partial \beta}\left[S_Y^2 + \beta^2 S_X^2 - 2\beta\rho S_X S_Y\right] = 0$$

$$\Rightarrow \beta = \rho\frac{S_Y}{S_X} = \frac{S_{XY}}{S_X^2}.$$

The minimum value of the variance of $\hat{\bar{Y}}_{reg}$ with optimum value of $\beta_{opt} = \frac{\rho S_Y}{S_X}$ is

$$Var_{min}(\hat{\bar{Y}}_{reg}) = \frac{f}{n}\left[S_Y^2 + \rho^2\frac{S_Y^2}{S_X^2}S_X^2 - 2\rho\frac{S_Y}{S_X}\rho S_X S_Y\right]$$

$$= \frac{f}{n}S_Y^2(1-\rho^2).$$

Since $-1 \le \rho \le 1$, so

$$Var(\hat{\bar{Y}}_{reg}) \le Var_{SRS}(\bar{y})$$

which always holds true. So, the regression estimator is always better than the sample mean under SRSWOR.

Estimate of variance

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

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

Note that the variance of $\hat{\bar{Y}}_{reg}$ increases as the difference between β_0 and β_{opt} increases.

Regression estimates when β is computed from the sample

Suppose a random sample of size n on paired observations on (x_i, y_i), $i = 1,2,..,n$ is drawn by SRSWOR. When β is unknown, it is estimated as

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}$$

and then the regression estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{\beta}(\bar{X}-\bar{x}).$$

It isn't easy to find the exact expressions of $E(\bar{Y}_{reg})$ and $Var(\hat{\bar{Y}}_{reg})$. So we approximate them using the same methodology as in the case of the ratio method of estimation.

Let

$$\varepsilon_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}} \Rightarrow \bar{y} = \bar{Y}(1+\varepsilon_0)$$
$$\varepsilon_1 = \frac{\bar{x}-\bar{X}}{\bar{X}} \Rightarrow \bar{x} = \bar{X}(1+\varepsilon_1)$$
$$\varepsilon_2 = \frac{s_{xy}-S_{XY}}{S_{XY}} \Rightarrow s_{xy} = S_{XY}(1+\varepsilon_2)$$
$$\varepsilon_3 = \frac{s_x^2-S_X^2}{S_X^2} \Rightarrow s_x^2 = S_X^2(1+\varepsilon_3)$$

Then

$$E(\varepsilon_0) = 0, \quad E(\varepsilon_1) = 0,$$
$$E(\varepsilon_2) = 0, \quad E(\varepsilon_3) = 0,$$
$$E(\varepsilon_0^2) = \frac{f}{n}C_Y^2,$$
$$E(\varepsilon_1^2) = \frac{f}{n}C_X^2,$$
$$E(\varepsilon_0\varepsilon_1) = \frac{f}{n}\rho C_X C_Y$$

and

$$\hat{\bar{Y}}_{reg} = \bar{y} + \frac{S_{xy}}{S_x^2}(\bar{X} - \bar{x})$$

$$= \bar{Y}(1+\varepsilon_0) + \frac{S_{XY}(1+\varepsilon_2)}{S_x^2(1+\varepsilon_3)}(-\varepsilon_1 \bar{X}).$$

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

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

where $\beta = \frac{S_{XY}}{S_X^2}$ is the population regression coefficient.

Assuming $|\varepsilon_3| < 1$,

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

Retaining the terms up to the second power of $\varepsilon's$ and ignoring other terms, we have

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

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

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

Now, the bias of $\hat{\bar{Y}}_{reg}$ up to the second order of approximation is

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

$$= -\frac{\beta\bar{X}f}{n}\left[\frac{\mu_{21}}{\bar{X}S_{XY}} - \frac{\mu_{30}}{\bar{X}S_X^2}\right]$$

where $f = \frac{N-n}{N}$ and $(r, s)^{th}$ cross-product moment is given by

$$\mu_{rs} = E\left[(x-\bar{X})^r (y-\bar{Y})^s\right]$$

So that

$$\mu_{21} = E\left[(x-\bar{X})^2(y-\bar{Y})\right]$$

$$\mu_{30} = E\left[(x-\bar{X})^3\right].$$

Thus

$$E(\hat{\bar{Y}}_{reg}) = -\frac{\beta f}{n}\left[\frac{\mu_{21}}{S_{XY}} - \frac{\mu_{30}}{S_X^2}\right].$$

Also,

$$E(\hat{\bar{Y}}_{reg}) = E(\bar{y}) + E[\hat{\beta}(\bar{X} - \bar{x})]$$

$$= \bar{Y} + \bar{X}E(\hat{\beta}) - E(\hat{\beta}\bar{x})$$

$$= \bar{Y} + E(\bar{x})E(\hat{\beta}) - E(\hat{\beta}\bar{x})$$

$$= \bar{Y} - Cov(\hat{\beta}, \bar{x})$$

$$Bias(\hat{\bar{Y}}_{reg}) = E(\hat{\bar{Y}}_{reg}) - \bar{Y} = -Cov(\hat{\beta}, \bar{x})$$

16.5 BIAS AND MEAN SQUARE ERROR:

MSE of $\hat{\bar{Y}}_{reg}$

To obtain the MSE of $\hat{\bar{Y}}_{reg}$, consider

$$E(\hat{\bar{Y}}_{reg}-\bar{Y})^2 \approx E\left[\varepsilon_0\bar{Y}-\beta\bar{X}(\varepsilon_1-\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_2)\right]^2$$

Retaining the terms of $\varepsilon's$ up to the second power second and ignoring others, we have

$$\begin{aligned}E(\hat{\bar{Y}}_{reg}-\bar{Y})^2 &\approx E\left[\varepsilon_0^2\bar{Y}^2+\beta^2\bar{X}^2\varepsilon_1^2-2\beta\bar{X}\bar{Y}\varepsilon_0\varepsilon_1\right]\\ &=\bar{Y}^2E(\varepsilon_0^2)+\beta^2\bar{X}^2E(\varepsilon_1^2)-2\beta\bar{X}\bar{Y}E(\varepsilon_0\varepsilon_1)\\ &=\frac{f}{n}\left[\bar{Y}^2\frac{S_Y^2}{\bar{Y}^2}+\beta^2\bar{X}^2\frac{S_X^2}{\bar{X}^2}-2\beta\bar{X}\bar{Y}\rho\frac{S_XS_Y}{\bar{X}\bar{Y}}\right]\end{aligned}$$

$$\begin{aligned}MSE(\hat{\bar{Y}}_{reg}) &= E(\hat{\bar{Y}}_{reg}-\bar{Y})^2\\ &=\frac{f}{n}(S_Y^2+\beta^2S_X^2-2\beta\rho S_XS_Y)\end{aligned}$$

Since $\beta=\frac{S_{XY}}{S_X^2}=\rho\frac{S_Y}{S_X}$,

so substituting it in MSE$(\hat{\bar{Y}}_{reg})$, we get

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

So, up to the second order of approximation, the regression estimator is better than the conventional sample mean estimator under SRSWOR. This is because the regression estimator uses some extra information. Moreover, such additional information requires some extra cost. This shows a false superiority in some sense. So, the regression estimators and SRS estimates can be combined if the cost aspect is also considered.

16.6 ESTIMATION OF VARIANCE, CONFIDENCE INTERVAL AND COMPARISON WITH MEAN PER UNITESTIMATOR:

For the comparisons the sample size n must be large enough so that approximation formulas for the variances of the ratio and regression estimates are valid. The three comparable variances for the estimated population mean $\overline{Y}$ are as follows

$$V\left(\bar{y}_{lr}\right)=\frac{N-n}{Nn}S_y^2\left(1-\rho^2\right)$$

$$V\left(\bar{Y}_R\right)=\frac{N-n}{Nn}\left[S_y^2+R^2S_x^2-2R\rho S_yS_x\right]$$

$$V\left(\bar{y}\right)=\frac{N-n}{Nn}S_y^2$$

(1) It is apparent that the variance of the regression estimate is smaller than that of the SRS estimate unless $\rho=0$, in which case the two variances are equal.

(2) The variance of the regression estimate is less than that of the ratio estimate if

$$-\rho^2S_y^2<R^2S_x^2-2R\rho S_yS_x$$

This is equivalent to the inequality

$$\left(\rho S_y - R S_x\right)^2 > 0 \ \left(\rho = \frac{S_{yx}}{S_x S_y} \Rightarrow \rho S_y = \frac{S_{yx}}{S_x}, B = \frac{S_{yx}}{S_x^2} \Rightarrow \rho S_y = B S_x\right)$$

$$\Rightarrow \left(B S_x - R S_x\right)^2 > 0 \qquad \Rightarrow S_x^2 \left(B - R\right)^2 > 0$$

Thus the regression estimate is more precise than the ratio estimate unless B=R. This occurs when the relation between y_i and x_i is a straight line through the origin.

16.7 REGRESSION ESTIMATOR IN STRATIFIED SAMPLING:

As with the ratio estimate, two types of regression estimates can be made in stratified random sampling, in the first estimate $\bar{y}_{lrs}$ (s- for separate). A separate regression estimate is computed for each stratum mean.

i.e., $\bar{y}_{lrh} = \bar{y}_h + b_h\left(\bar{X}_h - \bar{x}_h\right) \quad \rightarrow (1)$

Then $\bar{y}_{lrs} = \sum_{h=1}^{L} W_h \bar{y}_{lrh} \quad \rightarrow (2) \quad W_h = \frac{N_h}{N}$

This estimate is appropriate when (it is thought that the true) regression coefficients B_h vary from stratum to stratum. The second regression estimate, $\bar{y}_{lrc}$ (c for combined), is appropriate when the B_h are presumed to the same in all strata. To compute $\bar{y}_{lrc}$, we first find $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, $\bar{x}_{st} = \sum_h W_h \bar{x}_h$ then

$$\bar{y}_{lrc} = \bar{y}_{st} + b\left(\bar{X} - \bar{x}_{st}\right) \quad \rightarrow (3)$$

When the concept is applied to each stratum, $\bar{y}_{lrh}$ is an unbiased estimate of $\bar{y}_h$, so that $\bar{y}_{lrs}$ is an unbiased estimate of $\bar{Y}$. Further, since sampling is independent in different strata, if follows from theorem-7.1, that

$$V\left(\bar{y}_{lrs}\right) = \sum_h W_h^2 \frac{\left(1 - f_h\right)}{n_h}\left(S_{yh}^2 - 2 b_h S_{yxh} + b_h^2 S_{xh}^2\right) - \rightarrow (4)$$

Theorem-7.2 shows that, $V\left(\bar{y}_{lrs}\right)$ is minimized when $b_h = B_h$, the true regression coefficient in stratum h. The minimum value of the variance may be written

$$V_{min}\left(\bar{y}_{lrs}\right) = \sum_{h=1}^{L} W_h^2 \frac{\left(1 - f_h\right)}{n_h}\left(S_{yh}^2 - \frac{S_{yxh}^2}{S_{xh}^2}\right)$$

Turing to the combined regression estimate with pre assigned b, $\bar{y}_{lrc}$ is also an unbiased estimate of $\bar{Y}$. Since $\bar{y}_{lrc}$ is the usual estimate from a stratified sample for the variate $y_{hi} + b\left(\bar{X} - x_{hi}\right)$. We may apply Th-5.3 to this variate, giving the result

$$V\left(\bar{y}_{lrc}\right) = \sum_h W_h^2 \frac{\left(1 - f_h\right)}{n_h}\left(S_{yh}^2 - 2 b S_{yxh} + b^2 S_{xh}^2\right) \quad \rightarrow (5)$$

From equation (4) & (5), we observe that the difference is b_h in equation (4) and b in equation (5).

Q.2) Show that regression estimator is optimum?

Ans: Consider the survey population as a random sample from a super population. The population vector Y is therefore a realization of a random vector $y = (y_1, y_2, ..., y_N)$. It is assumed that the joint distribution of y is such that y_i's are independent with

$$\left.\begin{aligned} \in(y_i/x_i) &= \alpha+\beta x_i \\ V(y_i/x_i) &= \sigma^2 \end{aligned}\right\} \quad \to (1)$$

$\in$,V denoting respectively expectation and variance operators w.r.t super population models.

Hence, given the data, a model-unbiased predictor of $y = \sum_{i=1}^{N} y_i$ is $e_s(y) = \sum_{i\in s} y_i + \hat{U}_s$

Where $\in(\hat{U}_s) = \in\left(\sum_{i\in\bar{s}} y_i\right)\perp = (N-n)\alpha+\beta\sum_{i\in\bar{s}} x_i \quad \to (2)$

For all 's' with $P(s) > 0$. Hence BLUP(Best Linear Unbiased Prediction) of y for a given 's' is

$$e_s^*(y) = \sum_{i\in s} y_i + \hat{U}_s^* \quad \to (2A)$$

Where $V(\hat{U}_s^*) \le V(\hat{U}_s^{'}) \forall$ linear predictors $\hat{U}_s^{'}$ satisfying (2).

By the Gauss-Markoff theorem, BLUP of α, β are α^*, β^* respectively and hence.

$$\hat{U}_s^* = (N-1)\hat{\alpha}^* + \hat{\beta}^* \sum_{i\in\bar{s}} x_i \quad \to (2B)$$

Where $\hat{\alpha}^* = \bar{y}_s - \hat{\beta}^* \bar{x}, \ \bar{y}_s = \sum_{i\in s} \frac{y_i}{n} \quad \to (3)$

$$\hat{\beta}^* = \frac{\sum_{i\in s}(y_i - \bar{y}_s)(x_i - \bar{x}_s)}{\sum_{i\in s}(x_i - \bar{x}_s)^2} \text{ and hence}$$

$$e_s^*(y) = \sum_{i\in s} y_i + \left(\bar{y}_s - \hat{\beta}^* \bar{x}\right)(N-n) + \hat{\beta}^* \sum_{\bar{s}} x_i \qquad (\because \text{ from 2A,2B and(3)})$$

$$= n\bar{y}_s + N\bar{y}_s - N\hat{\beta}^*\bar{x} - n\bar{y}_s + n\hat{\beta}^*\bar{x} + \hat{\beta}^* \sum_{\bar{s}} x_i$$

$$= N\left[\bar{y}_s + \hat{\beta}^*(\bar{X} - \bar{x})\right] \quad \to (4) \qquad (\because \text{ from (3) and } \sum_i \frac{x_i}{N} = \bar{X} \Rightarrow n\bar{y}_s = \sum y_i)$$

Again, for all s, the variance of e_s^*,

$$V\left[e_s^*(y)\right] = \in\left[e_s^*(y)-y\right]^2 = V\left[N\left(\bar{y}_s + \hat{\beta}^*\left(\bar{X}-\bar{x}\right)\right)\right]$$

$$= N^2 V\left(\bar{y}_s\right) + N^2 V\left(\hat{\beta}^*\left(\bar{X}-\bar{x}\right)\right) = N^2\left(\frac{1-f}{n}\right)\sigma^2 + N^2\left(\bar{X}-\bar{x}\right)^2 V\left(\hat{\beta}^*\right)$$

(since$\left(\left(\bar{X}-\bar{x}\right)\right.$ is constant, $V\left(\bar{X}-\bar{x}\right) = \left(\bar{X}-\bar{x}\right)^2, V\left(\hat{\beta}^*\right) = \frac{\sigma^2}{\sum_{i\in s}(x_i-\bar{x})^2}$)

$$= N^2\left[\left(\frac{1-f}{n}\right)\sigma^2 + \sigma^2\left(\frac{\left(\bar{X}-\bar{x}\right)^2}{\sum_{i\in s}(x_i-\bar{x})^2}\right.\right] = N^2\sigma^2\left[\frac{1-f}{n} + \frac{(\bar{X}-\bar{x})^2}{\sum_{i\in s}(x_i-\bar{x})^2}\right] \quad \rightarrow (5)$$

(Since from(4) $y=y_1, y_2, \ldots y_N$)

Equation (4) states that for a given s, $e_s^*(y)$ is the BLUP of y and equation (5) states that the best fixed sample size(n) to use $e_s^*(y)$ is to choose a sample $s = s_b$ (say) for which $\bar{x}_{sb} = \bar{X}$. An optimum sampling design is, therefore, a, purposive sampling design. The minimum value of model-variance of $e_s^*(y)$ is then $N^2\sigma^2\frac{(1-f)}{n}$. Clearly, such sampling designs are difficult to realize in practice. Samples satisfying the property $\bar{X} = \bar{x}$ are called balanced samples.

16.8 SUMMARY:

In this chapter, we explored the **regression estimator** as a method of improving the estimation of population parameters using auxiliary information. The key idea is to exploit the linear relationship between the study variable Y and an auxiliary variable X to enhance precision.

- The **regression estimator** is given by:

  $\hat{Y}_{reg} = \bar{y} + b\left(\bar{X}-\bar{x}\right)$

  where 'b' is the regression coefficient estimated from the sample.
- This estimator is particularly effective when there is a **strong linear correlation** between Y and X.
- It is **approximately unbiased**, and its **mean square error (MSE)** is often **lower** than that of the mean per unit estimator.
- The **variance** of the regression estimator can be estimated, and it is used to construct **confidence intervals** for population parameters.
- In comparison to:
  - **Mean per unit estimator**, the regression estimator is more efficient when auxiliary information is available and properly used.
  - **Ratio estimator**, the regression estimator is better when the line relating Y and X **does not pass through the origin**.
- In **stratified sampling**, the regression estimator is applied within each stratum, leading to further gains in precision by reducing within-stratum variance.

Conclusion:

The regression estimator is a powerful and versatile tool in survey sampling. By incorporating auxiliary information, it provides more reliable and precise estimates than basic methods. Its application in both simple random and stratified sampling frameworks makes it an essential technique for statisticians and researchers involved in practical data collection and estimation.

16.9 KEY WORDS:

- Regression Estimator
- Auxiliary Variable
- Bias
- Mean Square Error
- Confidence Interval
- Stratified Sampling
- Regression Coefficient
- Sample Mean
- Population Mean
- Efficiency

16.10 SELF-ASSESSMENT QUESTIONS:

1. What is the basic principle behind the regression estimator?
2. How is the regression estimator different from the ratio estimator?
3. Derive the expression for the bias of the regression estimator.
4. When is the regression estimator more efficient than the mean per unit estimator?
5. How is the regression estimator applied in stratified sampling?
6. Explain Regression Estimator in Stratified Sampling
7. Explain the comparison of Regression Estimator with Mean per uit Estimator.

16.11 SUGGESTED READINGS:

1. **Cochran, W.G. (1977),** *Sampling Techniques* (3rd Edition), Wiley.
2. **P. Mukhopadhyay**, Title: *Theory and Methods of Survey Sampling,* PHI Learning Pvt. Ltd., New Delhi.
3. **Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., & Asok, C. (1984).** *Sampling Theory of Surveys with Applications,* Publisher: Iowa State University Press.
4. Singh, D., and Chaudhary, F.S. **(1986).** *Theory and Analysis of Sample Survey Designs*, Wiley Eastern Ltd.
5. **Murthy, M.N.** (1977). *Sampling Theory and Methods*. Statistical Publishing Society.

Dr. U. Ramkiran