V95

USE OF ANCILLARY INFORMATION IN ENSURING A REPRESENTATIVE SAMPLE

VIJAY KUMAR

Dissertation submitted in fulfilment of the
requirements of Post-Graduate Diploma
in Agricultural Statistics of the
Institute of Agricultural
Research Statistics
New Delhi-12

DE 04

INSTITUTE OF AGRICULTURAL RESEARCH STATISTICS
LIBRARY AVENUE, NEW DELHI - 12

1977

CERTIFICATE

This is to certify that the work incorporated in the dissertation entitled ' USE OF ANCILLARY INFORMATION IN ENSURING A REPRESENTATIVE SAMPLE ' by VIJAY KUMAR and submitted for the award of Post-Graduate Diploma in Agricultural Statistics of the Institute of Agricultural Research Statistics (I.C.A.R.) , New Delhi was done under my guidance.

Padam Singh

(PADAM SINGH)
Associate Professor of Statistics
Institute of Agricultural Research Statistics
Library Avenue, New Delhi -12

ACKNOWLEDGEMENTS

I express my profound indebtedness to Dr. Padam Singh, Associate Professor of Statistics, Institute of Agricultural Research Statistics, New Delhi for his encouragement and valuable advice during the course of investigation and for his continuous help in the preparation of this dissertation.

I express my sincere gratitude to Dr. Daroga Singh, Director, Institute of Agricultural Research Statistics, New Delhi for providing necessary facilities to carry out the research work.

My sincere thanks are due to Indian Council of Agricultural Research for providing financial assistance in the form of research fellowship.

I am highly grateful to some of my friends too, for their valuable assistance during the study of the problem.

The help rendered by Sh. Kewal Kher in programming is also gratefully acknowledged.

Finally, I am thankful to Shri Gaya Prashad Pal for typing this dissertation in time.

Vijay Kumar

(VIJAY KUMAR)
I.A.R.S. (I.C.A.R.)
New Delhi - 12

CONTENTS

CHAPTER - 1

INTRODUCTION

The importance of sampling in providing data with greater speed, greater accuracy and lesser cost is well recognised. The element or group of elements on which information can be taken is known as the ' sampling unit '. The group of units specified by the objectives of survey is referred to as ' population '. A population having countable number of units is called ' finite ' and a population having uncountable number of units is called ' infinite '. In sampling, we generally deal with finite populations and derive approximations for infinite populations.

The population to be sampled is called 'Sampled Population' and the population about which information is wanted is called 'Target Population'. Former is more restricted than the latter.

Part of population selected by some procedure is called a ' sample ' and the process of selection is called ' sampling '. A ' representative sample ' is that sample which has characteristics similar to that of the population. Sampling is said to be ' with replacement ' if the unit once selected is replaced back into the population for further selection and ' without replacement ' if the unit once selected is not given the chance of further selection in the sample. The procedure of selecting the sample unit by unit is known as ' sampling scheme '. The set of all possible samples is called ' sample space '. The sample space together with the associated probability measure is called the ' sampling design '. A sampling scheme gives rise to a sampling design. The procedure of estimating population parameter together with the standard error on the basis of sample values is called 'Estimation Procedure'.

Let y be the character defined on the population which takes value y_i for i-th unit of the population. Any function of population values is termed as the ' parameter '. The corresponding function based on sample values is known as ' estimator '. The particular value which the estimator takes is called the ' estimate '. A sampling design together with an estimation procedure is termed as ' sampling strategy '.

A scientific and objective method of selecting the sample units and Collecting information on selected units of the population is called a sample survey or a ' survey ' and collecting information on each unit of the population is called complete enumeration or ' census '. A sample survey helps in deeper scientific investigation of a population with limited trained staff. It also provides the measure of error involved in estimating population parameter such as mean etc. The census is used when cost is no consideration, characters can be easily recorded, non sampling errors are not large and detailed break up for smallest administrative unit is required.

The standard error is generally taken as 'measure of error'.

The character in which we are interested is called the 'study character'. In most of the cases another character highly correlated with the study character is available for all the units of the population. Such a character is called ' auxiliary character ' and the information based on it the ' ancillary information '. Such information is used for obtaining efficient results in estimation of population mean or total. The ancillary information can be made use of either for selection of sampling units from the population as in PPS sampling or for stratification of the units in the population or for ratio and regression methods of estimation.

When the units vary considerably in size, the selection of sampling units with simple random sampling is not advantageous since it does not take into account the importance to be attached to larger units in the population. In such situations it becomes utmost important to select the sampling units with probability proportional to size (PPS). The PPS sampling makes use of the available information on some auxiliary character. This scheme was initially proposed by Hansen and Horwitz (1943). Das (1951), Narain (1951), Horwitz and Thompson (1952), Lahiri (1951), Midzuno (1952), Yates and Grundy (1953), Des Raj (1956), Hartley and Rao (1962) , Rao , Hartley and

Cochran (1962), Hanurav (1967), Durbin (1967), Sampford (1962), Das and Mohanty (1973) and many others also developed various methods for selection of sampling units with varying probabilities.

In stratification, the population is to be divided into strata such that within strata variation is as small as possible. Dalenius (1950) attempted this problem of determining optimum points of stratification on the basis of study variable. The method used by him was iterative in nature and required the knowledge of certain parameters which are functions of boundary points. Dalenius and Gurney (1951) suggested the use of information on auxiliary character for the purpose of stratification. Taga (1967) and Singh (1968) also obtained points of stratification for the character under study based on auxiliary character.

Further, the suitable use of auxiliary character at the estimation stage also results in considerable reduction in variance of the estimate. The ratio and regression methods of estimation make use of auxiliary character for obtaining estimates of population parameters. The ratio method of estimation was first introduced by Cochran (1940) whereas theoretical basis of regression method of estimation was also first discussed by

Cochran (1942). Product method of estimation was first proposed by Murthy (1964). Sukhatme (1944), Quenouille (1956), Robson (1957), Olkin (1958), Goodman and Hartley (1958), Nanjamma, Murthy and Sethi (1959), Singh (1965), [illegible] ([illegible]) and others made further contributions to the theory of ratio, regression and product methods of estimation.

Of all the sampling procedures considered above, none provides any control on sampling error which is an important aspect of sampling. In the procedure under investigation an attempt has been made to use the ancillary information for selecting a representative sample i.e. a sample for which relative error of sample estimate from population parameter is within specified margin. [This defition is somewhat different from the one normally used.] It is of importance to study the probability that sample estimate for study character will also differ from population parameter by specified margin of error. This has been investigated in chapter - II and its behaviour has been studied with sample size, correlation coefficient, coefficient of variation and margins of error. For the procedure suggested in chapter - II, the inclusion probabilities for individual and pairwise units have been worked out in chapter - III to get Horwitz Thompson estimator. Further, the selection procedure has been suitably modified providing non-zero inclusion probabilities for individual and pairwise units.

A sampling technique is defined as introducing control into the selection of n out of N sampling units when it increases the probabilities of selection for preferred samples and thus decreases the probabilities for non preferred samples. The objective of controlled selection is to reduce the variances of estimates for most of sampling techniques at a given cost. Controlled selection in probability sampling was first introduced by Goodman and Kish (1950). Further, Avadhani and Sukhatme (1965) and (1966) proposed controlled simple random sampling and its use in ratio and regression method of estimation.

This concept of controlled selection has been used in the suggested procedure to provide larger probabilities of selection to the representative samples as compared to the remaining samples. The calculation of inclusion probabilities and the estimation procedure have also been discussed for the controlled selection. The emperical investigations have been carried out to study the relative efficiencies of the suggested procedures with some existing procedures.

CHAPTER - II

USE OF ANCILLARY INFORMATION IN ENSURING A REPRESENTATIVE SAMPLE

2.1 Introduction : The main aim of sampling in general consists in selecting a sample and then building an estimator of the population parameter. In practice it is desirable that the estimator so formed should be within some preassigned margin of error from the corresponding parameter. Many times the information on all the units of the population is available for the auxiliary character, highly correlated with the character under study. The use of this ancillary information in selecting sample whose mean is not different from the population mean for the auxiliary character beyond a specified margin of error has been suggested in this chapter. The procedure of selecting such sample has been explained with an example. The probability of estimating the population mean with desired margin of error has also been worked out. The properties of this probability have been examined in respect of sample size, correlation coefficient between study and auxiliary variables and margins of error. Some numerical illustrations have also been given with different values of these parameters.

2.2 The Suggested Procedure : Suppose that the population under study consists of N distinct and identifiable units. Let y_i and x_i

be the values of the study and auxiliary characters for the i-th unit of the population. Further, let x_i be known for all i, ($i = 1,2,\ldots\ldots, N$). Then the suggested procedure consists of the following steps :

Step - I: Select a simple random sample without replacement of size n from the population of size N.

Step - II : Calculate $(\bar{x} - \mu_x)/\mu_x$ for the selected sample and test whether $|(\bar{x} - \mu_x)/\mu_x| \leq \epsilon_x$ for some preassigned ϵ_x, where $\bar{x}$ and μ_x are respectively the sample mean and population mean for the character x.

Step - III : If $|(\bar{x} - \mu_x)/\mu_x| \leq \epsilon_x$, retain the sample, otherwise proceed to step - I for selection of another sample till a sample satisfying $|(\bar{x} - \mu_x)/\mu_x| \leq \epsilon_x$ is obtained.

2.2.1 Illustrative Example : Suppose that we have a population of size 6 with x values given in the table below ,

Sr. No.	1	2	3	4	5	6
x	1	2	2	3	3	4

Suppose a sample of size 4 is desired to be drawn from the 15 possible samples of size 4 listed below. If ϵ_x is .05 then the five samples which can be selected are at Sr. Nos. 5,6,8,9 and 11 .

Sample No.	Sample units	$\bar{x}$	$\lvert(\bar{x}-\mu_x)/\mu_x\rvert$
1	1,2,3,4	2.00	.20
2	1,2,3,5	2.00	.20
3	1,2,3,6	2.25	.10
4	1,2,4,5	2.25	.10
5	1,2,4,6	2.50	.00
6	1,2,5,6	2.50	.00
7	1,3,4,5	2.25	.10
8	1,3,4,6	2.50	.00
9	1,3,5,6	2.50	.00
10	1,4,5,6	2.75	.10
11	2,3,4,5	2.50	.00
12	2,3,4,6	2.75	.10
13	2,3,5,6	2.75	.10
14	2,4,5,6	3.00	.20
15	3,4,5,6	3.00	.20

2.3. Probability of of estimation : Under this procedure it is of interest to find the probability of estimating the population mean for y with the relative margin of error ϵ_y. i.e.

$$P\left[\,\left|(\bar{y}-\mu_y)/\mu_y\right| \leq \epsilon_y \;\middle|\; \left|(\bar{x}-\mu_x)/\mu_x\right| \leq \epsilon_x\right] = p\,(y \mid x)$$

where $\bar{y}$ and μ_y are respectively the sample mean and the population mean for character y. To calculate this probability let us assume that x and y follow a bivariate normal distribution with coefficient of correlation ρ. This joint probability density function of $(\bar{x}, \bar{y})$ is given by :

$$f(\bar{x}, \bar{y}) = \frac{n}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp. \left[\frac{-1}{2(1-\rho^2)} \left\{ \left(\frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}} \right)^2 \right. \right.$$

$$\left. \left. - 2\rho \left(\frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}} \right) \left(\frac{\bar{y}-\mu_y}{\sigma_y/\sqrt{n}} \right) + \left(\frac{\bar{y}-\mu_y}{\sigma_y/\sqrt{n}} \right)^2 \right\} \right]$$

where σ_x and σ_y are the standard deviations for x and y respectively. The required probability is then given by :

$$P_{(y|x)} = \frac{\int_{\bar{y}}\int_{\bar{x}} f(\bar{x}, \bar{y})\, d\bar{x}\, d\bar{y}}{\int_{-\infty}^{\infty}\int_{\bar{x}} f(\bar{x}, \bar{y})\, d\bar{x}\, d\bar{y}}$$

with the limits for $\bar{x}$ and $\bar{y}$ as under,

$\mu_x (1 - \epsilon_x) \leq \bar{x} \leq \mu_x (1 + \epsilon_x)$ and

$\mu_y (1 - \epsilon_y) \leq \bar{y} \leq \mu_y (1 + \epsilon_y)$.

For evaluating $P_{(y|x)}$ we put

$\dfrac{\bar{y}-\mu_y}{\sigma_y/\sqrt{n}} = u$ with $d\bar{y} = \dfrac{\sigma_y}{\sqrt{n}}\, du$ and

$\dfrac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}} = v$ with $d\bar{x} = \dfrac{\sigma_x}{\sqrt{n}}\, dv$. Thus we get ,

$$P_{(y|x)} = \frac{\int_{u}\int_{v} f(u, v)\, dv\, du}{\int_{u}\int_{-\infty}^{\infty} f(u, v)\, dv\, du}$$

where $f(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp.\left[\frac{-1}{2(1-\rho^2)} (u^2 - 2\rho uv + v^2)\right]$

The limits for u and v in the integral will be as under ,

$$-\frac{\epsilon_x \mu_x}{\sigma_x/\sqrt{n}} \leq v \leq \frac{\epsilon_x \mu_x}{\sigma_x/\sqrt{n}}$$

$$-\frac{\epsilon_y \mu_y}{\sigma_y/\sqrt{n}} \leq u \leq \frac{\epsilon_y \mu_y}{\sigma_y/\sqrt{n}}$$

The above probability can be calculated if ρ , ϵ_x , ϵ_y , $\sigma_x / \mu_x = c_x$, $\sigma_y / \mu_y = c_y$ and n are known . It may be mentioned here that most of the parameters such as ρ and c_y are generally unknown. But some guess value of these parameters can always be known from the previous data on the same characteristics or from the sample selected.

2.4 Numerical Illustration : To study the behaviour of above probability with sample size , correlation coefficient , the coefficients of variation and the margin of error , some emperical results have been given below in tables 2.1 to 2.5. To calculate this probability , ' Tables for Statisticians and Biometricians', Part 2 by Pearson. K ,(Tables VIII and IX) have been used . These tables provide the value of $\iint_{u\,v} f(u, v)\, dv\, du$ for positive ρ as well as negative ρ.

Making use of these tables,

$$p_{(y|x)} = \frac{\int_{-h}^{h}\int_{-k}^{k} f(u,v)\, dv\, du}{\int_{-\infty}^{\infty}\int_{-k}^{k} f(u,v)\, dv\, du} \quad \text{with} \quad h = \frac{e_y \mu_y}{\sigma_y/\sqrt{n}} \quad \text{and}$$

$$k = \frac{e_x \mu_x}{\sigma_x/\sqrt{n}}$$ has been computed as described below :

The numerator of $p_{(y|x)}$ is given by

$$2\Big[\int_0^\infty\int_0^\infty f(u,v)\, dv\, du + \int_h^\infty\int_k^\infty f(u,v)\, dv\, du$$
$$- \int_h^\infty\int_0^\infty f(u,v)\, dv\, du - \int_0^\infty\int_k^\infty f(u,v)\, dv\, du\Big]$$

+ 2 (same four integrals with ρ negative).

Similarly the denominater of the required probability $p_{(y|x)}$ is given by $2\Big[\int_0^\infty\int_0^\infty f(u,v)\, dv\, du + \int_\infty^\infty\int_k^\infty f(u,v)\, dv\, du$

$$- \int_\infty^\infty\int_0^\infty f(u,v)\, dv\, du - \int_{-\infty}^\infty\int_k^\infty f(u,v)\, dv\, du\Big]$$

+ 2 (same four integrals with ρ negative).

As tables provide values of bivariate integral for the range of h and k from 0 to 2.6 only, the values for h and k (> 2.6) have been taken at 2.6 in the integration as an approximation. All probabilities have been rounded upto 4 digits for simplicity.

Tables showing the value of the probability for different values of n and ρ

Table 2.1 $\epsilon_x = \epsilon_y = .01$, $c_x = c_y = .10$

n/ρ	.5	.6	.7	.8	.9
16	.3536	.3784	.4145	.4768	.7884
36	.5047	.5330	.5750	.6365	.7986
64	.6313	.6580	.6949	.7464	.8372
100	.7319	.7544	.7832	.8215	.8778

Table 2.2 $\epsilon_x = \epsilon_y = .01$, $c_x = c_y = .08$

n/ρ	.5	.6	.7	.8	.9
16	.6313	.6580	.6949	.7464	.8372
36	.8102	.8185	.8476	.8742	.9116
64	.9135	.9205	.9228	.9401	.9569
100	.9679	.9699	.9721	.9761	.9815

Table 2.3 $\epsilon_x = \epsilon_y = .05$, $c_x = c_y = .10$

n/ρ	.5	.6	.7	.8	.9
16	.9679	.9699	.9721	.9761	.9815
36	1.0000	1.0000	1.0000	1.0000	1.0000
64	1.0000	1.0000	1.0000	1.0000	1.0000
100	1.0000	1.0000	1.0000	1.0000	1.0000

Table : 2.4 $\epsilon_x = .01$, $\epsilon_y = .05$, $c_x = c_y = .10$

n/ρ	.5	.6	.7	.8	.9	1.0
16	.9804	.9886	.9936	.9986	1.0000	1.0000
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
64	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table : 2.5 $\epsilon_x = \epsilon_y = .05$, $c_x = c_y = .05$

n/ρ	.5	.6	.7	.8	.9	1.0
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
64	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

It is clear from the above tables that the probability $p_{(y|x)}$ increases monotonically with increase of n the sample size, ρ the correlation coefficient, ϵ_y margin of error for y and decreases with increase of coefficient of variation and ϵ_x the margin of error for x. Further, it can be seen that for $\epsilon_x = \epsilon_y = .05$ and $\epsilon_x = .01$, $\epsilon_y = .05$ the probability is of the order 1 but for $\epsilon_x = \epsilon_y = .01$ it is of high order only for high correlation or large n. Thus in practical situations with $\epsilon_x = \epsilon_y = .05$ or $\epsilon_x = .01$, $\epsilon_y = .05$ and for known c_y, c_x, n and ρ the suggested procedure can be used satisfactorily as the probability of estimation of the character under study within given margin of error is of high order.

2.5 Summary : In this chapter, a method of utilising ancially information, in ensuring a representative sample, has been suggested. The probability of estimating the mean of the character under study within the specified margin of error has been obtained for the suggested procedure. It has been seen that this probability increases monotically with sample size, the correlation coefficient between study and auxiliary variables and the margin of error for the study variable. It has also been observed that generally the probability of estimating the mean of study character is of high order and thus the suggested procedure can be used satisfactorily.

CHAPTER - III

MODIFIED PROCEDURES FOR UNBIASED ESTIMATION

3.1 Introduction : In chapter - II , the probability of estimating the mean of the character under study with desired margin of error had been worked out for the selection procedure which makes use of auxiliary character. It has been observed that this probability increases with the increase of sample size , correlation coefficient and margin of error for the study variable. However, the sample mean is not unbiased for the population mean under the suggested procedure. It is therefore , important to suggest the estimation procedure which provides unbiased estimate of population mean along with unbiased variance estimator under the selection procedure of chapter - II. This has been attempted in this chapter. For obtaining Horwitz - Thompson's estimates of population mean and its variance the knowledge of π_i's for all i and π_{ij}'s for all $i \neq j$ is required, the calculation of which has also been discussed. It may be mentioned that some π_i's or π_{ij}'s can be zero for certain population which imposes a restriction on the suggested estimation procedure. Therefore, the procedure for the selection of the sample has been modified ensuring non zero π_i's and π_{ij}'s. Further the concept of controlled selection has also been used in modifying the procedure to obtain non zero π_i's and π_{ij}'s. The relative efficiency of the suggested procedures has also been compared with a number of known procedures emperically.

3.2 <u>Estimation Procedure :</u> Let M be the number of samples out of ${}^{N}c_{n}$ possible samples which satisfy the condition $|(\bar{x} - \mu_x)/\mu_x| \leq e_x$ of the procedure suggested in chapter - II. Then the probability of selecting a sample out of these M samples is given by $P_s = 1/M$. Under this scheme it is easy to see that

$$\pi_i = \sum_{s \ni i} P_s = \frac{K_i}{M}$$

and

$$\pi_{ij} = \sum_{s \ni i,j} P_s = \frac{K_{ij}}{M} \quad (i \neq j)$$

where K_i is the number of samples out of M having i-th unit and K_{ij} is the number of samples containing both i-th and j-th units. Knowing the values of the units and their π_i's and π_{ij}'s the Horwitz-Thompson estimate can be used for estimating the population mean. If $\pi_i > 0$ for all i, we have

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

The Yates - Grundy form of variance of above estimator is given by

$$V(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i}^{N} \sum_{i<j}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Also if $\pi_{ij} > 0$ for all $i \neq j$, the Yates-Grundy form of estimate of variance is given by

$$\widehat{V}(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i}^{n} \sum_{i<j}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

If a selection procedure provides larger inclusion probability for the dissimilar units and if π_i's are approximately of the same order, (it has been explained for symmetrical distributions) it is expected that the procedure would result in considerable gain in efficiency as compared to simple random sampling. In the suggested procedure the inclusion probabilities for dissimilar units are expected to be large ~~[illegible]~~ as compared to similar units in respect of auxiliary character which would result in gain in efficiency.

3.2.1 <u>Illustration :</u> For the example considered in chapter - II, with $n = 4$ and $N = 6$ there are in all 5 samples out of 15 samples satisfying $|(\bar{x} - \mu_x)/\mu_x| \leq \epsilon_x$. The π_i's and π_{ij}'s for different units are given by

$$\pi_1 = 4/5,\ \pi_2 = 3/5,\ \pi_3 = 3/5,\ \pi_4 = 3/5,\ \pi_5 = 3/5,\ \pi_6 = 4/5$$

$$\pi_{ij} = 1/5 \begin{array}{c|ccccc|c} i/j & 2 & 3 & 4 & 5 & 6 & \\ \hline & 2 & 2 & 2 & 2 & 4 & 1 \\ & & 1 & 2 & 2 & 2 & 2 \\ & & & 2 & 2 & 2 & 3 \\ & & & & 1 & 2 & 4 \\ & & & & & 2 & 5 \end{array} \quad (i < j)$$

It may be remarked that if the distribution of x is symmetrical π_i's will be more or less equal for all the units. Also the π_{ij} for the pair of units whose mean is closer to the population mean will be more than the other pair whose mean is different from the population mean, in respect of auxiliary character.

3.3 <u>Modified Procedure</u> : For the sampling procedure described in chapter - II, π_i or π_{ij} might be zero for some unit or pair of units which imposes a serious limitation on the use of Horwitz-Thompson estimates. To overcome this difficulty the procedure has been modified for the selection of the sample. The modified selection procedure consists of the following steps :

<u>Step - I :</u> Draw a ~~simple~~ random sample without replacement (SWOR) of size (n - 2) from the population of size N by the selection procedure given in chapter - II.

<u>Step - II :</u> Supplement the above sample by 2 units drawn by SRSWOR from the remaining (N - n + 2) units of the population.

Now if M' is the number of samples of size (n - 2) out of $^{N}c_{(n-2)}$ possible samples which satisfy $|(\bar{x}' - \mu_x)/\mu_x| \leqslant \epsilon_x$, where $\bar{x}'$ is the mean based on (n - 2) units then π_i's and π_{ij}'s will be given by

$$\pi_i = \frac{K_i'}{M'} + \left(1 - \frac{K_i'}{M'}\right) \frac{2}{N - n + 2} \quad \text{and}$$

$$\pi_{ij} = \frac{K_{ij}'}{M'} + \frac{K_i' + K_j' - 2K_{ij}'}{M'} \cdot \frac{2}{N-n+2} + \left(1 - \frac{K_i' + K_j' - K_{ij}'}{M'}\right) \frac{2}{(N-n+1)(N-n+2)}$$

where K_i' is the number of samples out of M' which contain i-th unit and K_{ij}' is the number of samples containing both i-th and j-th units. Using these π_i's and π_{ij}'s which are non zero, estimates of the population mean and its variance can be obtained by the formulae given in section 3.2. In some cases ~~it~~ ~~under~~ this procedure may not hold good for $|(\bar{x}-\mu_x)/\mu_x| < \epsilon_x$.

3.3.1 <u>Illustration</u>: Again for the example considered in chapter-II, the π_i's and π_{ij}'s for $n = 2$ are as under,

$\pi_1 = 1/5$, $\pi_2 = 2/5$, $\pi_3 = 2/5$, $\pi_4 = 2/5$, $\pi_5 = 2/5$, $\pi_6 = 1/5$

$$\pi_{ij} = 1/5 \begin{array}{c|ccccc|c} i/j & 2 & 3 & 4 & 5 & 6 & \\ \hline & 0 & 0 & 0 & 0 & 1 & 1 \\ & & 0 & 1 & 1 & 0 & 2 \\ & & & 1 & 1 & 0 & 3 \\ & & & & 0 & 0 & 4 \\ & & & & & 0 & 5 \end{array} \quad (i < j)$$

Also for the modified selection procedure for $n = 4$, π_i's and π_{ij}'s are given by

$\pi_1 = 6/10$, $\pi_2 = 7/10$, $\pi_3 = 7/10$, $\pi_4 = 7/10$, $\pi_5 = 7/10$, $\pi_6 = 6/10$

$$\pi_{ij} = 1/30 \begin{array}{c|ccccc|c} i/j & 2 & 3 & 4 & 5 & 6 & \\ \hline & 11 & 11 & 11 & 11 & 10 & 1 \\ & & 13 & 14 & 14 & 11 & 2 \\ & & & 14 & 14 & 11 & 3 \\ & & & & 13 & 11 & 4 \\ & & & & & 11 & 5 \end{array} \quad (i < j)$$

3.4 Controlled Selection Procedure : In the selection procedure suggested in chapter - II , a sample is selected if it satisfies $|(\bar{x}-\mu_x)/\mu_x| \leq \epsilon_x$ otherwise it is rejected. This essentially implies that the sample space of SRSWOR is divided into two parts, the first of consisting of preferred samples for which $|(\bar{x}-\mu_x)/\mu_x| \leq \epsilon_x$ and the second consisting of non preferred samples for which $|(\bar{x}-\mu_x)/\mu_x| > \epsilon_x$. The non preferred samples have a zero probability of selection under the procedure suggested in chapter-II which results sometimes in $\pi_i = 0$ or $\pi_{ij} = 0$. The non-preferred samples can be assigned non-zero probability of selection (however small) which will provide non zero inclusion probabilities for all units and pairs of units. The controlled selection procedure for desired control say α consists of the following steps :

Step I : Draw a sample of size n by SRSWOR from the population of size N.

Step II : Test whether $|(\bar{x}-\mu_x)/\mu_x| \leq \epsilon_x$ for given ϵ_x .

Step III : (i) If $|(\bar{x}-\mu_x)/\mu_x| \leq \epsilon_x$, perform a Bernoulli trial with probability of success P_1 , for selecting the sample and (ii) if $|(\bar{x}-\mu_x)/\mu_x| > \epsilon_x$ perform a Bernoulli trial with probability of success P_2 for selecting the sample, otherwise go to Step I. The values of P_1 and P_2 are determined from the equations,

$P_1 = a/M$ and $P_2 = (1-a)/({}^N c_n - M)$. Generally P_2 is assigned a very small value and P_1 is determined by the relation $M P_1 + ({}^N c_n - M) P_2 = 1$.

Under the scheme of controlled selection π_i's and π_{ij}'s can be obtained by the formulae

$$\pi_i = P_1 K_i + P_2 ({}^{N-1}c_{n-1} - K_i) \quad \text{and}$$

$$\pi_{ij} = P_1 K_{ij} + P_2 ({}^{N-2}c_{n-2} - K_{ij})$$

Knowing π_i's and π_{ij}'s for all individual units and pair of units, Horwitz-Thompson estimates of mean and corresponding variance can be obtained by the formulae given in section 3.2.

3.4.1 Illustration : For the example considered in chapter-II for $N = 6$, $n = 4$, and prefixing $P_2 = .001$ we get $P_1 = .198$. Thus π_i's and π_{ij}'s are given by

$\pi_1 = .798$, $\pi_2 = .601$, $\pi_3 = .601$, $\pi_4 = .601$, $\pi_5 = .601$, $\pi_6 = .798$.

$$\pi_{ij} = \begin{array}{c|ccccc|l} i \backslash j & 2 & 3 & 4 & 5 & 6 & \\ \hline 1 & .400 & .400 & .400 & .400 & .794 & \\ 2 & & .203 & .400 & .400 & .400 & (i<j) \\ 3 & & & .400 & .400 & .400 & \\ 4 & & & & .203 & .400 & \\ 5 & & & & & .400 & \end{array}$$

3.5 Efficiency : In an attempt to compare the suggested procedure with some existing procedures five hypothetical populations have been considered. The populations have correlation coefficient ranging from .64 to .95 and are presented in table 3.1.

Table 3.1 Table showing five hypothetical populations with different values of ρ .

Population

Unit No.	ρ = .9863		.8703		.8451		.7441		.6441	
	I		II		III		IV		V	
	X	Y	X	Y	X	Y	X	Y	X	Y
1	1	1	1	1	1	1	1	1	1	1
2	2	1	2	1	2	2	2	1	2	2
3	3	2	2	1	3	2	2	1	3	3
4	4	2	3	2	4	2	3	1	4	3
5	5	3	3	2	5	2	3	2	5	3
6	6	3	4	2	6	3	4	2	6	3

A natural population presented in table 3.2 has also been considered for comparison. It has been taken from ' Annals of Mathematical Statistics , 13, 179-206 ' and is based on estimation of volume of timber stands by strip sampling. x is here in 10 chain units and represents blockwise sum of strip lengths while y is in 1000 ft. board measurement units and represents blockwise sum of timber b volumes.

Table 3.2 Table showing natural population (ρ = .917)

Unit No.	X	Y	Unit No.	X	Y
1	180	8781	6	140	5622
2	142	6756	7	81	3677
3	76	1786	8	91	4368
4	104	2425	9	191	7919
5	109	4655	10	117	3861

For each population the variance for the estimate based on sample size 4 have been worked out for various procedures and results are presented in Table 3.3. The procedures considered for comparison are the following :

1. Horwitz-Thompson estimate in the suggested procedure (ϵ_x = .05)
2. Horwitz-Thompson estimate in modified procedure (ϵ_x = . 05)
3. HT estimate in controlled selection procedure (ϵ_x = .05 , P_z = .001)
4. Simple mean in SRSWOR
5. Usual estimate in PPSWR
6. Regression estimate in SRSWOR

Table 3.3 Table showing the variances of different procedures

Procedure	POPULATION I	II	III	IV	V	Natural
(1)	.0000	.0069	.0000	.0069	.0417	274336.838
(2)	.0556	.0007	.0278	.0008	.0556	1364083.847
(3)	.0007	.0037	.0003	.0074	.0194	329950.150
(4)	.0667	.0250	.0333	.0231	.0583	780200.000
(5)	.0400	.0278	.1130	.0440	.1220	1841318.750
(6)	.0057	.0061	.0095	.0103	.0341	124100.000

Nothing should be concluded with certainty from these results as these are emperical comparisons only. However, it can be seen that H.T. estimate under the suggested procedure of chapter-II provides generally more efficient results than usual estimates of SRSWOR and PPSWR sampling procedures. The suggested procedures are also competitive with regression estimate in many situations.

3.6 Summary : In this chapter, a estimation procedure has been suggested which provides unbiased estimate of population mean along with unbiased variance estimator, under the selection procedure of chapter-II. But under this procedure π_i or π_{ij} may be zero for some i or i ≠ j, therefore the procedure has been modified to provide non zero probabilities of inclusion for all the units and pair of units. Further the concept of controlled selection has been used to provide non zero probabilities of inclusion. All these procedures have been compared with some of the existing procedures emperically and the results have been found to be highly satisfactory.

SUMMARY

It is well known that ancillary information helps in obtaining efficient results in estimation of population parameter as mean etc. A number of procedures based on ancillary information have been developed but none provides control on sampling error which is an very important auspect in sampling. In this dissertation, a selection procedure has been suggested which makes use of ancillary information to ensure the selection of a representative sample i.e., a sample for which relative margin of error of sample estimate from population parameter is within specified margin. Further, the probability that sample estimate for study character will also differ from population parameter by a specified margin of error has been calculated, for different ranges of sample size, correlation coefficient between study and auxillary variables, margins of errors and coefficients of variation. This probability is generally of high order and thus the suggested procedure can be used satisfactorily in most of practical situations.

To obtain Horwitz-Thompson's estimates of mean and variance, method of calculating π_i's and π_{ij}'s for the suggested procedure has also been given. In case $\pi_i = 0$ or $\pi_{ij} = 0$ for some unit or pair of units, the selection procedure and method of computing π_i's and π_{ij}'s for all the units have been modified.

The controlled selection procedure as modification of the suggested procedure has been given and the method of obtaining π_i's and π_{ij}'s based on controlled selection has also been discussed. Some theoretical and natural populations have also been considered to compare the suggested schemes with existing procedures emperically. The suggested schemes have been found to be highly satisfactory in all the situations.

REFERENCES

1. Avadhani, M.S. and Sukhatme, B.V. (1965). " Controlled simple Random Sampling ", JISAS, 17, 34-42.

2. Avadhani, M.S. and Sukhatme, B.V. (1966) . " A note on the ratio and regression methods of estimation in controlled simple random sampling " JISAS, 18, 17-20.

3. Cochran, W.G. (1940) . " The estimation of yield of cereal experiments by sampling from the ratio of grain to total produce, " J. Agri. Sc. 37, 199-212.

4. Cochran, W.G. (1942). " Sampling theory when samling units are of unequal sizes, " JASA 37, 199-212.

5. Dalenius, T. (1950) . " Problems of Optimum strata - I " , Skand, AKF 33, 203-13.

6. Dalenius, T. and Gurney. M. (1951). " Problem of Optimum strata-II " Skand AKF, 34, 133-48.

7. Das, A.C. (1951). " On two phase sampling and sampling with varying probabilities " , Bull. Int. Stat. Inst. 33, 105-42.

8. Das, M.N. and Mohanty, S. (1973) . " On PPS sampling without replacement ensuring selection probabilities exactly proportional to sizes, " AJS, 13.

9. Des Raj (1965). " On a method of using multi-auxiliary information in sample surveys ".

10. Des Raj (1956) . " Some estimators in sampling with varying probabilities without replacement", JASA , 51 i

11. Durbin, J. (1967). " Design of multistage surveys for the estimation of sampling errors," Applied Statistics , 16.

12. Goodman, R. and Kish, L. (1950). " Controlled selection - A technique in probability sampling ", JASA , 45, 350-72.

13. Goodman, L.A. and Hartley, H.O. (1958), " The precision of unbiased ratio type estimators " , JASA 53, 491-508.

14. Hansen, M.N. and Horwitz, W.N. (1943). " On theory of sampling from finite population ", AMS - 14, 333-62.

15. Hanurav, T.V. (1967) . " Optimum utilization of auxiliary information - πPS sampling of two units from a strata " , J. Roy. Stat. Soc, Series, B 29, 374-391.

16. Hartley, H.O. and Rao, J.N.K. (1962). " Sampling with unequal probability without replacement ", Ann. Math. Stat. 33, 350-374.

17. Horwitz, D.G. and Thompson, D.J. (1952). " A generalisation of sampling without replacement from a finite population " JASA , 48, 663-85.

18. Lahiri, D.B. (1951). " A method of sample selection providing unbiased ratio estimates, " BISI , 33.

19. Midzuno, H. (1952). " On sampling system with probability proportional to sums of sizes " , Ann. Inst. Stat. Math., Japan 3, 99-107.

20. Murthy, M.N. (1964). " Product of estimation ", Sankhya 26, (A), 69-74.

21. Nansama, N.S. Murthy, M.N. and Sethi, V.K. (1959). " Some sampling system providing unbiased ratio estimations " Sankhya 21, 299-314.

22. Narain, R.D. (1951). " On sampling without replacement with varying probability " , Jour. Ind. Soc. Ag. Stat. 3, 169-74.

23. Olkin I (1958). " Multivariate ratio estimation for finite population" Biometrika 45, 154-165.

24. Quenouilli, M.N. (1956). " Note on bias in estimation, " Biometrika 43, 353-360.

25. Rao, J.N.K., Hartley H.O. and Cochran W.G. (1962). " A sample procedure of unequal probability sampling without replacement,"JRSS,(B) , 24.

26. Robson, D.S. (1957). " Application of multivariate polykays to the theory of unbiased ratio type estimators ", JASA 52, 511-522.

27. Sampford , M.R. (1962). " Method of cluster sampling with and without replacement for clusters of unequal sizes, Biometrika, 49.

28. Singh, R. (1968). " Some contribution to theory of construction of strata ", unpublished Ph.D. thesis submitted to I.A.R.I., New Delhi.

29. Singh M.P. (1965). " On the estimation of ratio and product of the population parameters " Sankhya, 27 (B), 321-328.

30. Sukhatme, P.V. (1944). " Moments and Product Moments of Moment Statistics for samples of the finite and infinite populations, " Sankhya, 6, 363-82.

31. Taga, Y. (1967). " On optimum strata for the objective of variable based on concomitant variable using prior information " Ann. Stat. Math. 19, 101-30.

32. Yates, F. and Grundy, P.M. (1953). " Selection without replacement from within strata with PPS ", Jour. Roy. Stat. Soc. Series B 15, 253-261.