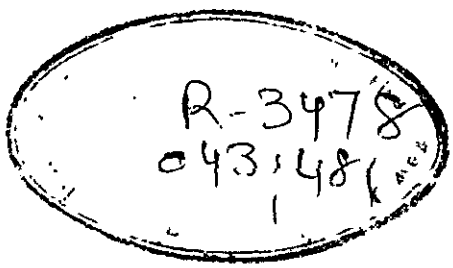
SOME INVESTIGATIONS

ON

THE ROLE OF SAMPLING FOR PREDICTION

KALI CHARAN GAUTAM



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

INSTITUTE OF AGRICULTURAL RESEARCH STATISTICS
(I.C.A.R.)
LIBRARY AVENUE, NEW DELHI-12

1975

CERTIFICATE

This is to certify that the work incorporated in the dissertation entitled " SOME INVESTIGATION ON THE ROLE OF SAMPLING FOR PREDICTION " by K.C. Gautam and submitted for the award of Post-Graduate Diploma in Agricultural Statistics of the Institute of Agricultural Research Statistics, New Delhi was done under my guidance.

Padam Singh

(PADAM SINGH)
Statistician
cum
Associate Professor
Institute of Agricultural Research Statistics
(I.C.A.R.)
Library Avenue, New Delhi - 12.

ACKNOWLEDGEMENTS

I have great pleasure in expressing my deep sense of gratitude to Shri Padam Singh, Statistician-cum-Associate Professor, I.A.R.S., New Delhi, for his guidance, keen interest and constant encouragement throughout the course of this investigation and preparation of thesis.

I am grateful to Dr. D. Singh, Director, I.A.R.S. for providing adequate facilities and encouragement for the research work done in the thesis.

I am grately indebted to Shri M. P. Jha, Senior Statistician and Shri A.K. Shrivastava, Junior Statistician of the Institute for their valuable suggestions. My thanks are also due to Shri D.K. Aggrawal, Statistical Investigator of the Institute for his Computer Programming help.

The work was supported by a Senior Research Fellowship of the Institute of Agricultural Research Statistics. The assistance is gratefully acknowledged.

At the last but not the least, I thanks Shri Prem Kumar Grover, Electronic Typist for presenting the material in this thesis in a beautiful format.

I.A.R.S., NEW DELHI-12.
6 OCT, 1975.

Kali Charan Gautam
(KALI CHARAN GAUTAM)

CONTENTS

CHAPTER - I

INTRODUCTION

1.1 Preliminaries

Reliable statistical information (data) is required for framing important policies such as export, import, price, distribution, storage, planning etc. and considerable cost is involved in the collection of such data. Complete enumeration is impracticable if not impossible. The role of sample surveys in the collection of reliable statistical data needs no emphasise. For clarity and reference some preliminaries and necessary review are discussed in what follows.

A 'population' is an aggregate of individuals (sampling units) and it is said to be 'finite' if it contains countable number of units. A 'sample' is a finite ordered sequence of (not necessarily distinct) units and is a sub set of the population. A 'sampling procedure' is the method of selection of sample from the population. Any function of the values of the units in the population such as population mean or population variance is termed as 'parameter' whereas the functions of the values of the units in the sample is known as statistic. The aim of sampling theory in general is to estimate some parameter of the population with associated standard error from the sample.

A 'functional relation' is a relationship among the parameters of the distribution of different variables and a

'regression relation' expresses the expected value of one variable in terms of the observed values of other variables. The functional relation is based on the variation in the true values whereas the regression relation is based on the variation in both the true values and random errors. If the independent variables are free from error then the functional relation and regression relation coincide each other. In practical situations actually the functional relation is required but only the regression function is obtained because this is based on observed values only.

1.2 A Brief Review

Regression relation problem was first considered by Adcock (1877), who defined the line of best fit as the one for which the sum of squares of the normal deviates of the observed points from the line becomes a minimum. Two main objections were raised against the method. First there is no justification for minimizing the sum of squares of normal deviates in preference to the deviation in some other direction. Second, the straight line obtained by this method is not invariant under the transformation of coordinate system. This point had been emphasized by Roos (1928). It is a common feature of Roos general formula and all other methods prepared in recent years that the fitting of straight line can not be determined without any prior assumptions (independence of observations) regarding the standard deviations of errors in

variables x and y. That is to say that the standard deviations of the errors in variables are involved in the formula of the fitted straight line and no method is available by which these standard deviations can be estimated by means of observed values of x and y. Another method of determining the best fitted straight line was given by Allen (1939) where in he emphasised that the line can be determined only if the values of any two quantities, standard deviations of errors in independent and dependent variables, and the correlation between the errors in two variables are given a-priori.

For the case involving only two variables, Abraham Wald (1940) had given a method, commonly known as "method of grouping" which does not depend on apriori. knowledge of the standard deviations of the errors in independent as well as dependent variables and the correlation between them (errors in the variables). Wald considered the problem under the conditions that —

i) The random variables, errors in x_i's denoted by ϵ_i $(i = 1, \ldots, N)$ each have the same distribution and are uncorrelated.

ii) The random variables, errors in y_i's denoted by η_i $(i = 1, \ldots, N)$ each have the same distribution and are uncorrelated.

iii) Random variables ϵ_i and η_i are uncorrelated.

iv) A single linear relation holds between the true values x and y i.e. $y_i = \alpha + \beta x_i$ $(i = 1, \ldots, N)$.

Wald concluded that :

(a) the fitted straight line can be determined without making a-prior assumptions regarding the standard deviations of errors

(b) standard deviations of errors can also be estimated by means of observed values of x and y and

(c) the precision of estimates increases with the increase in number of observations.

Under this method the pairs are arranged firstly according to the increasing magnitude of independent variable and the total number of observations are then devided into two groups. The estimates of β and α are given by

$$\hat{\beta} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \beta \bar{x} ,$$

where $(\bar{y}_i , \bar{x}_i)$ are means of corresponding variables of i-th group $(i = 1,2)$ and $\bar{y} = \sum_{i=1}^{N} y_i / N$, $\bar{x} = \sum_{i=1}^{N} x_i / N$.

Bartlett (1942) presented a modification of Wald's method which generally results in greater accuracy. In the modified procedure suggested by him, the total number of

observations, arranged according to the magnitude (increasing) of independent variable are devided into three non-over lapping groups of as nearly equal as possible sizes. The join of the mean coordinates $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_3, \bar{y}_3)$ for the two extreme groups is used to determine the slope of best fitted line. Instead of devided the observations into two or three groups of equal size as required Wald's and Bartlett's procedure, Mandansky (1959) suggested that the observations may be first arranged by magnitude of independent variable x. The first p_1 fraction of observations from group G_1 and last p_2 fraction of observations from G_3 and remaining $(1 - p_1 - p_2).N$ observations, where N is the total number of pairs of observations, constitute group G_2. The estimates of β, is then given by $\hat{\beta} = (\bar{y}_3 - \bar{y}_1)/(\bar{x}_3 - \bar{x}_1)$ where $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_3, \bar{y}_3)$ are the mean of groups G_1 and G_3 respectively. The procedure given by Wald and Bartlett are particular cases of this when $p_1 = p_2 = \frac{1}{2}$ and $p_1 = p_2 = \frac{1}{3}$ respectively.

Generalisation of the method to the case of two independent variables was discussed by Hooper and Tholl (1957). Since the regression equation in case of two independent variables involve three unknowns, three centres of gravity are sufficient for determination of regression equation. The total number of observations are, devided into three non-over lapping groups of sizes n_1, n_2, n_3 where $\sum_{i=1}^{3} n_i = n$.

Using the minimization of generalised variance as a criterian, optimum grouping of observations presented practical difficulties as it involved the joint distribution of the independant variables.

Gibson and Jowett (1957 a), (1957 b) extended the Bartlett's method to multiple regression with two independent variables and presented an easy method of estimating the standard errors of regression coefficients. The optimum allocation of points to the three groups have been derived, when the distribution of independent variables is normal, rectangular bell shaped, U. shaped, J shaped and skew. There are many others who have also contributed to the theory of regression analysis.

1.3 Orientation of the Problem

Almost all the procedures known so far assume that the population under consideration is infinite which is not so because actually population consists of always some finite (may be very large) number of elements. In practice population exists in the form of certain groups called strata or can be considered as devided into strata. For finite populations there are number of sampling procedures available in the literature suiting the administrative and economic constraints. The existing procedures cannot be successfully used in finite population situations and the results may be

many times misleading.

Most of the survey statisticians have concentrated their efforts for estimating population mean or population total. The relative performances of various sampling procedures for the estimation of mean or total have been very well discussed in literature. However, no attempts appear to have been made to investigate the problem of optimum sampling design for prediction. In the present investigation the problem of determining appropriate sampling procedures for prediction on the basis of regression equations has been attempted and the procedure has been explained with numerical illustrations.

In Chapter - II the general problem of regression analysis from finite population has been considered. The expressions for the variances of prediction for various sampling procedures have been worked out for different situations.

In Chapter - III the method of determining an optimum sampling design for prediction has been discussed. Few selection procedures have been compared for a wide range of population.

The application of the suggested procedures has been explained in Chapter - IV with the help of data collected under the Pre-harvest Forecasting scheme on Jute for

predicting the yield rate on the basis of biometrical characters. A solution to various statistical problems involved in the prediction has been indicated.

In the end a summary incorporating the main findings and conclusions has been included in the thesis.

CHAPTER - II

SAMPLING AND PREDICTION

2.1 Introduction

Most of the work in sample surveys has concentrated around estimation of population mean or population total. The relative efficiencies of various sampling procedures for estimating population mean or total have been very well discussed in all standard text books. Determination of functional relationship in two or more than two variables is a field in which the role of sampling is yet to be investigated. In this chapter the problem of determining regression relationship in two variables on the basis of a sample has been discussed. The expressions for sampling variance have been worked out for various situations. It has also been attempted to compare some of the equal probability sampling schemes on the basis of variance of prediction.

2.2 Regression Analysis and Sampling

Consider a finite population consisting of N units and let y and x be two variables taking values y_j and x_j for the j-th unit of the population, $j = 1, \ldots, N$. Let a functional relationship of the type $y = \alpha + \beta x$ is desired to be determined for the population. For this, let a sample of size 'n' be drawn and let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be the pairs of values of x and y respectively on 'n' units of the population. The procedure of fitting the regression equation consist in estimating

α and β such that the sum of squares of the deviations from the line of regression is minimum. Following least square technique, the estimates of α and β on the basis of sample are given by

$$\hat{\beta} = b = \frac{\sum_{j=1}^{n} (x_j - \bar{x}_n)(y_j - \bar{y}_n)}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \qquad \ldots \quad (2.1)$$

and

$$\hat{\alpha} = a = \bar{y}_n - \hat{\beta}\,\bar{x}_n \qquad \ldots \quad (2.2)$$

where $\bar{x}_n$ and $\bar{y}_n$ are the sample means of characters x and y respectively.

Having obtained the estimates of α and β by least square technique the next aspect of regression analysis is to predict the value of y corresponding to some value of x. The value of x may be any one of the following :

(a) Some given value which is decided subjectively.

(b) The value of a unit selected from the units of the sample.

(c) The value of a unit selected from the units not in the sample.

(d) The value of unit chosen from the whole population.

Lot of work relating to the situation (a) such as response to given level of fertilizer, milk yield for a given intake, output for given input etc. has been well discussed in literature. There are other types of situations such as prediction of yield corresponding to plant population or plant height, prediction of supply

corresponding to price level etc. where the plant population, plant height and the price level themselves are random variables. We shall be concerning mainly with the unbiasedness of the prediction and the error associated with the last three situations (b), (c) and (d) mentioned above as these are of interest to us. However, for clearity and completeness the case corresponding to situation (a) has also been included here.

2.2.1 Case (a)

This pertains to the situation where the unit for which the value of y has to be predicted is selected subjectively. It is also assumed that x_j's are fixed constants and not the random variables. The assumptions in the prediction model

$$y_j = \alpha + \beta x_j + \epsilon_j, \quad (j = 1, \ldots, N) \qquad \ldots \quad (2.3)$$

are

(1) ϵ_j is a random variable with mean zero and variance σ_e^2 (unknown) ... (2.4)

(2) ϵ_j and ϵ_i are uncorrelated for $j \neq i$ ... (2.5)

(3) ϵ_j is normally distributed random variable with mean zero and variance σ_e^2 ... (2.6)

It is known that $\hat{\beta} = b = \dfrac{\sum_{j=1}^{n} (x_j - \bar{x}_n)(y_j - \bar{y}_n)}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2}$ ... (2.7)

$$b = \frac{\sum_{j=1}^{n} (x_j - \bar{x}_n) \cdot y_j}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \quad \ldots \quad (2.8)$$

since the terms removed from the numerator is

$$= \sum_{j=1}^{n} (x_j - \bar{x}_n) \cdot \bar{y}_n \quad \ldots \quad (2.9)$$

$$= \bar{y}_n \sum_{j=1}^{n} (x_j - \bar{x}_n) \quad \ldots \quad (2.10)$$

$$= 0$$

Now the variance of the function

$$F = a_1 y_1 + a_2 y_2 + \ldots + a_n y_n \quad \ldots \quad (2.11)$$

is $$V(F) = a_1^2 V(y_1) + a_2^2 V(y_2) + \ldots + a_n^2 V(y_n) \ldots \quad (2.12)$$

since y_j's are pair wise uncorrelated and a_j's are constants

$$\text{or } V(F) = (a_1^2 + a_2^2 + \ldots + a_n^2)\, \sigma_e^2 \quad \ldots \quad (2.13)$$

$$\text{or } V(F) = \left(\sum_{j=1}^{n} a_j^2\right) \sigma_e^2 \quad \ldots \quad (2.14)$$

In the expression for b, $a_j = \dfrac{x_j - \bar{x}_n}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2}$ ... (2.15)

Hence the variance of regression coefficient is given by

$$V(b) = \frac{\sigma_e^2}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \quad \ldots \quad (2.16)$$

Further the regression equation between y and x is known to be

$$y = \bar{y}_n + b(x - \bar{x}_n) \qquad \ldots \quad (2.17)$$

Now if a_j and c_j are constants and

$$a = a_1 y_1 + a_2 y_2 + \ldots + a_n y_n \qquad \ldots \quad (2.18)$$

$$c = c_1 y_1 + c_2 y_2 + \ldots + c_n y_n \qquad \ldots \quad (2.19)$$

Then as y_j and y_i are uncorrelated for $j \neq i$, we have

$$\mathrm{Cov}(a, c) = (a_1 c_1 + a_2 c_2 + \ldots + a_n c_n)\sigma_e^2 \qquad \ldots \quad (2.20)$$

It follows that setting $a = \bar{y}$ implies $a_j = \frac{1}{n}$, and setting $c = b$, implies

$$c_j = \frac{(x_j - \bar{x}_n)}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \qquad \ldots \quad (2.21)$$

So that

$$\mathrm{Cov}(\bar{y}_n, b) = 0 \qquad \ldots \quad (2.22)$$

that is, $\bar{y}_n$ and b are uncorrelated random variables. Thus the variance of the predicted mean value of y, y_k at a specific value x_k, of x is

$$V(y_k) = V(\bar{y}_n) + (x_k - \bar{x}_n)^2 V(b) \qquad \ldots \quad (2.23)$$

$$= \frac{\sigma_e^2}{n} + \frac{(x_k - \bar{x}_n)^2 \sigma_e^2}{\sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \qquad \ldots \quad (2.24)$$

This is minimum when $x_k = \bar{x}_n$ and increases as

we have x_k away from $\bar{x}_n$ in either direction.

2.2.3 Case (b)

This pertains to situation (b) when x is the value of unit selected from the units in the sample by simple random sampling.

We have regression equation

$$y_j = \bar{y}_n + b(x_j - \bar{x}_n) \qquad \ldots \quad (2.25)$$

The predicted value of y corresponding to x_j, $j = 1, 2, \ldots, n$ is given by

$$\hat{y}_j = \bar{y}_n + b(x_j - \bar{x}_n) \qquad \ldots \quad (2.26)$$

The difference between the true value and the predicted value corresponding to j-th unit is

$$y_j - \hat{y}_j = y_j - \bar{y}_n - b(x_j - \bar{x}_n) \qquad \ldots \quad (2.27)$$

The expectation of above expression has to be taken twice first for the given sample and then for all possible samples. Thus, with usual notations

$$E(y_j - \hat{y}_j) = E_1 E_2 [y_j - \bar{y}_n - b(x_j - \bar{x}_n)] \qquad \ldots \quad (2.28)$$

$$= E_1 [\bar{y}_n - \bar{y}_n - b(\bar{x}_n - \bar{x}_n)] = 0 \qquad \ldots \quad (2.29)$$

Since $E(x_j) = \bar{x}_n$, $E(y_j) = \bar{y}_n$ because of selection and hence the prediction is unbiased. $\ldots \quad (2.30)$

Now the variance of prediction with usual notations, is given by

$$E(y_j - \hat{y}_j)^2 = E_1 V_2 (y_j - \hat{y}_j) + V_1 E_2 (y_j - \hat{y}_j) \quad \ldots (2.31)$$

$$= E_1 V_2 \left[y_j - \bar{y}_n - b(x_j - \bar{x}_n) \right] \quad \ldots (2.32)$$

$$= E_1 \left[s_y^2 (1 - r^2)(1 - \frac{1}{n}) \right] \quad \ldots (2.33)$$

$$= E_1 \left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right] \frac{n-1}{n} \quad \ldots (2.34)$$

$$= \frac{n-1}{n} \sum_{s \in S} p_s \left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right] \quad \ldots (2.35)$$

where p_s is the probability of selecting the sample.

An Illustration : Consider the example of 5 units with x and y values as given below:

S.No.	x	y
1	1	3
2	2	6
3	3	10
4	4	11
5	5	15

... (2.36)

Also assume that the sampling design is as follows :

Samples	Probability of selection	Samples	Probability of selection
s_1 : (1,2,3)	0.20	s_6 : (1,4,5)	0.10
s_2 : (1,2,4)	0.10	s_7 : (2,3,4)	0.05
s_3 : (1,2,5)	0.10	s_8 : (2,3,5)	0.05
s_4 : (1,3,4)	0.10	s_9 : (2,4,5)	0.05
s_5 : (1,3,5)	0.10	s_{10} : (3,4,5)	0.15

TABLE - 1

S.No.	p_n	$\sum (y_i - \bar{y}_n)^2$	$\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n)$	$\sum (x_i - \bar{x}_n)^2$	$\left[\sum (y_i - \bar{y}_n)^2 - \frac{\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum (x_i - \bar{x}_n)^2}\right]^2$	$\left[\sum (y_i - \bar{y}_n)^2 - \frac{\sum (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum (x_i - \bar{x}_n)^2}\right]^2 \frac{p_n}{n}$
1.	.20	24.67	7.00	2.00	.17	.011
2.	.10	32.67	12.34	4.67	.07	.002
3.	.10	78.00	26.00	8.68	.12	.004
4.	.10	38.00	13.00	4.67	1.82	.060
5.	.10	73.66	24.00	8.00	1.66	.055
6.	.10	74.77	25.34	8.66	0.62	.020
7.	.05	14.00	5.00	2.00	1.50	.025
8.	.05	40.67	13.67	4.67	0.66	.011
9.	.05	40.67	13.68	4.67	0.60	.010
10.	.15	14.00	5.00	2.00	1.50	.075

The necessary computation for the variance of prediction are given in the above table and the variance of prediction for given illustration on page 15 is 0.273.

For the simple situation, when the original sample has been drawn by simple random sampling without replacement, the variance of prediction is given by

$$= \frac{n-1}{n} E\left[s_y^2 - \frac{s_{xy}^2}{s_x^2} \right] = \frac{n-1}{n}\left[S_y^2 - E \frac{s_{xy}^2}{s_x^2} \right] \quad \ldots(2.38)$$

Let $m_{ij} = \frac{1}{(n-1)} \Sigma (y_i - \bar{y})^i (x_i - \bar{x})^j$ — by defination ... (2.39)

Now $E \frac{s_{xy}^2}{s_x^2} = E \frac{m_{11}^2}{m_{02}} = E \frac{(\mu_{11} + \epsilon_{11})^2}{(\mu_{02} + \epsilon_{02})}$ ... (2.40)

$$= E \frac{\mu_{11}^2}{\mu_{02}} \left[1 + \frac{\epsilon_{11}^2}{\mu_{11}^2} + \frac{2\epsilon_{11}}{\mu_{11}} \right] \left[1 + \frac{\epsilon_{02}}{\mu_{02}} \right]^{-1} \quad \ldots(2.41)$$

$$= \frac{\mu_{11}^2}{\mu_{02}} E\left[1 + \frac{\epsilon_{11}^2}{\mu_{11}^2} - \frac{2\epsilon_{11}\epsilon_{02}}{\mu_{11}\mu_{02}} + \frac{\epsilon_{02}^2}{\mu_{02}^2} \right] \quad \ldots(2.42)$$

Upto second order approximation

$$= \frac{\mu_{11}^2}{\mu_{02}} \left[1 + \frac{V(m_{11})}{\mu_{11}^2} - \frac{2Cov(m_{11}, m_{02})}{\mu_{11}\mu_{02}} + \frac{V(m_{02})}{\mu_{02}^2} \right] \ldots(2.43)$$

$$= \left[\frac{\mu_{11}^2}{\mu_{02}} + \frac{V(m_{11})}{\mu_{02}} - \frac{2Cov(m_{11}, m_{02})}{\mu_{02}^2} \mu_{11} + \frac{\mu_{11}^2}{\mu_{02}^3} V(m_{02}) \right] \quad \ldots(2.44)$$

We know that

$$V(m_{11}) = \frac{1}{n} \left[\mu_{22} - \mu_{11}^2 \right] \quad \ldots(2.45)$$

$$\text{Cov}(m_{11}, m_{02}) = \frac{1}{n}\left[\mu_{13} - \mu_{11}\mu_{02}\right] \qquad \ldots (2.46)$$

$$V(m_{02}) = \frac{1}{n}\left[\mu_{04} - \mu_{02}^2\right] \qquad \ldots (2.47)$$

Putting (2.45), (2.46) and (2.47) in (2.44), we get

$$E\left(\frac{s_{xy}^2}{s_x^2}\right) = \left[\frac{\mu_{11}^2}{\mu_{02}} + \frac{1}{n\mu_{02}}(\mu_{22} - \mu_{11}^2) - \frac{2\mu_{11}}{n\mu_{02}^2}(\mu_{13} - \mu_{11}\mu_{02}) + \frac{\mu_{11}^2}{n\mu_{02}^3}(\mu_{04} - \mu_{02}^2)\right] \qquad \ldots (2.48)$$

$$= \frac{\mu_{11}^2}{\mu_{02}} + \frac{1}{n\mu_{02}}\left[(\mu_{22} - \mu_{11}^2) - \frac{2\mu_{11}}{\mu_{02}}(\mu_{13} - \mu_{11}\mu_{02}) + \frac{\mu_{11}^2}{\mu_{02}^2}(\mu_{04} - \mu_{02}^2)\right] \qquad \ldots (2.49)$$

$$= \frac{\mu_{11}^2}{\mu_{02}} + \frac{1}{n\mu_{02}}\left[\mu_{22} - \frac{2\mu_{11}\mu_{13}}{\mu_{02}} + \frac{\mu_{11}^2\mu_{04}}{\mu_{02}^2}\right] \qquad \ldots (2.50)$$

Hence the variance of prediction is given by

$$V_p = \frac{n-1}{n}\left[\mu_{20} - \left\{\frac{\mu_{11}^2}{\mu_{02}} + \frac{1}{n\mu_{02}}\left(\mu_{22} - \frac{2\mu_{11}\mu_{13}}{\mu_{02}} + \frac{\mu_{11}^2\mu_{04}}{\mu_{02}^2}\right)\right\}\right] \qquad \ldots (2.51)$$

$$= \frac{n-1}{n}\left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} - \frac{1}{n\mu_{02}}\left(\mu_{22} - \frac{2\mu_{11}\mu_{13}}{\mu_{02}} - \frac{\mu_{11}^2\mu_{04}}{\mu_{02}^2}\right)\right] \qquad \ldots (2.52)$$

which for the first degree of approximation takes the form

$$V_p = \left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} \right] \qquad \ldots \quad (2.53)$$

2.2.3 Case (c)

This pertain to Section (c), when x is the value of a unit selected from the remaining $(N-n)$ units of the population. We have regression equation

$$y_j = \bar{y}_n + b(x_j - \bar{x}_n) \qquad \ldots \quad (2.54)$$

The predicted value of y corresponding to x_j, $j = n+1, \ldots, N$ is given by $\hat{y}_j = \bar{y}_n + b(x_j - \bar{x}_n)$ ... (2.55)

Now the difference between the true value and the predicted value of y of j-th unit is given by

$$y_j - \hat{y}_j = y_j - \bar{y}_n - b(x_j - \bar{x}_n) \qquad \ldots \quad (2.56)$$

Also, since j-th unit has been selected from the remaining $(N-n)$ units we have

$$E_2(x_j) = \bar{X}_{N-n} = \frac{N\bar{X}_N - n\bar{x}_n}{N-n} = \frac{N}{N-n}\bar{X}_N - \frac{n}{N-n}\bar{x}_n \qquad \ldots \quad (2.57)$$

and

$$E_2(y_j) = \bar{Y}_{N-n} = \frac{N}{N-n}\bar{Y}_N - \frac{n}{N-n}\bar{y}_n \qquad \ldots \quad (2.58)$$

Thus,

$$E(y_j - \hat{y}_j) = E_1\left[\frac{N}{N-n}\bar{Y}_N - \frac{n}{N-n}\bar{y}_n - \bar{y}_n - b\left(\frac{N}{N-n}\bar{X}_N - \frac{n}{N-n}\bar{x}_n - \bar{x}_n\right)\right] \qquad \ldots \quad (2.59)$$

$$= - \frac{N}{N-n} E\left[\bar{y}_n - \bar{Y}_N - b(\bar{x}_n - \bar{X}_N) \right] \quad \ldots \quad (2.60)$$

$$= 0$$

and hence the prediction is unbiased.

The variance of prediction in this situation is given by

$$V_p = E_1 V_2\left[y_j - \bar{y}_n - b(x_j - \bar{x}_n) \right] + V_1\left[\bar{y}_n - \bar{Y}_N - b(\bar{x}_n - \bar{X}_N) \right] \left(\frac{N}{N-n}\right)^2 \quad \ldots \quad (2.61)$$

$$= E_1 s_y^2 (1-\rho^2) \frac{(n-1)}{n} + \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 (1-\rho^2) \left(\frac{N}{N-n}\right)^2 \quad \ldots \quad (2.62)$$

$$= \frac{n-1}{n}\left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} - \frac{1}{n\mu_{02}}\left(\mu_{22} - \frac{2\mu_{11}\mu_{13}}{\mu_{02}} + \frac{\mu_{11}^2 \mu_{04}}{\mu_{02}^2}\right)\right] + \left(\frac{1}{n} - \frac{1}{N}\right)\mu_{20}(1-\rho^2)\left(\frac{N}{N-n}\right)^2 \quad \ldots \quad (2.63)$$

which for the first degree of approximation simplifies to

$$V_p = \left(\frac{N}{N-n}\right) \frac{1}{n}(1-\rho^2) S_y^2 + \left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} \right]$$

2.2.4. **Case (d)**

Pertaining to Section (d): when x is the value of a unit selected from the whole population, the predicted value of y corresponding to x_j, $j = 1, \ldots, N$ is given by

$$\hat{y}_j = \bar{y}_n + b(x_j - \bar{x}_n) \quad \ldots \quad (2.64)$$

Now, since the unit under section has drawn from the whole population we have

$$E_2(x_j) = \bar{X}_N \qquad \ldots \quad (2.65)$$

$$E_2(y_j) = \bar{Y}_N \qquad \ldots \quad (2.66)$$

Thus, the difference between the true value and the predicted value is given by

$$y_j - \hat{y}_j = y_j - \bar{y}_n - b(x_j - \bar{x}_n) \qquad \ldots \quad (2.67)$$

and

$$E(y_j - \hat{y}_j) = E_1\left[\bar{Y}_N - \bar{y}_n - b(\bar{X}_N - \bar{x}_n)\right] \ldots \quad (2.68)$$

$$= 0$$

and hence the prediction is unbiased.

The variance of prediction is given by

$$V_p = E_1 V_2\left[y_j - \bar{y}_n - b(x_j - \bar{x}_n)\right] + V_1\left[\bar{Y}_N - \bar{y}_n - b(\bar{X}_N - \bar{x}_n)\right] \qquad \ldots \quad (2.69)$$

$$= E(1-\rho^2)s_y^2\,\frac{n-1}{n} + \left(\frac{1}{n} - \frac{1}{N}\right)s_y^2(1-\rho^2) \qquad \ldots \quad (2.70)$$

which is given by

$$V_p = \left(\frac{1}{n} - \frac{1}{N}\right)(1-\rho^2)\mu_{20} + \frac{n-1}{n}\left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} - \frac{1}{n\mu_{02}}\left(\mu_{22} - \frac{2\mu_{11}\mu_{13}}{\mu_{02}} + \frac{\mu_{11}^2\mu_{04}}{\mu_{02}^2}\right)\right] \qquad \ldots \quad (2.71)$$

which for the first degree of approximation is

$$\approx \left(\frac{1}{n} - \frac{1}{N}\right)\cdot(1-\rho^2)\,\mu_{20} + \left(\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}}\right)$$

From (2.52), (2.69) and (2.71) it can be seen that the variance of prediction is in the order given by

$$V_p(\text{Case b}) \leq V_p(\text{Case d}) \leq V_p(\text{Case c}).$$

2.3 Cluster Sampling

So far we have worked out the expression for the variance of prediction for the situation when the original sample is drawn by simple random sampling. There are other types of sampling procedures also which are some what more suitable from cost and administrative convenience aspects. In this section the variance formula has been worked out for the very simple situation where the original sample is drawn by cluster sampling and the unit for which the value of y is to be predicted is selected from the units within the sample by simple random sampling.

Let

N = total number of clusters in the population,

M = total number of units in each cluster,

n = number of clusters in the sample.

It is known that (with usual notations)

$$(nM-1)s^2 = \sum_{i=1}^{n}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{nM})^2 = n(M-1)\bar{s}_w^2 + M(n-1)s_b^2 \quad \ldots \quad (2.72)$$

$$E(s_{by}^2) = S_{by}^2,\ E(s_{bx}^2) = S_{bx}^2 \text{ and } E(s_{bxy}) = S_{bxy} \quad \ldots \quad (2.73)$$

and

$$E(\bar{s}_{wx}^2) = \bar{S}_{wx}^2,\ E(\bar{s}_{wy}^2) = \bar{S}_{wy}^2 \text{ and } E(\bar{s}_{wxy}) = \bar{S}_{wxy} \quad \ldots \quad (2.74)$$

Let

$$s_{by}^2 = S_{by}^2 + \epsilon_{by}$$

$$s_{bx}^2 = S_{bx}^2 + \epsilon_{bx}$$

$$s_{bxy} = S_{bxy} + \epsilon_{bxy}$$

with $E(\epsilon_{by}) = E(\epsilon_{bx}) = E(\epsilon_{bxy}) = 0$

and

$$E(\epsilon_{by}^2) = V(s_{by}^2),\ E(\epsilon_{bx}^2) = V(s_{bx}^2),\ E(\epsilon_{bxy}^2) = V(s_{bxy}) \quad \ldots \quad (2.75)$$

and similar expressions for within components such as s_{wx}^2 etc.

Now variance of prediction is given by

$$V_p = \frac{n-1}{n}\left[E(s_y^2) - E\left(\frac{s_{xy}^2}{s_x^2}\right)\right] \quad \ldots \quad (2.76)$$

Considering

$$E(s_y^2) = E\left[\frac{n(M-1)}{nM-1}\bar{s}_{wy}^2 + \frac{M(n-1)}{nM-1}s_{by}^2\right]$$

$$= \left[\frac{n(M-1)}{nM-1} \bar{S}^2_{wy} + \frac{M(n-1)}{nM-1} S^2_{by} \right] \quad \ldots \quad (2.77)$$

and

$$E\left(\frac{\sigma^2_{xy}}{\bar{\sigma}^2_x}\right) = E \frac{\left[\frac{n(M-1)}{nM-1}\left(1 + \frac{\epsilon_{wxy}}{\bar{S}_{wxy}}\right)\bar{S}_{wxy} + \frac{M(n-1)}{nM-1}\left(1 + \frac{\epsilon_{bxy}}{S_{bxy}}\right) S_{bxy} \right]^2}{\frac{n(M-1)}{nM-1}\left(1 + \frac{\epsilon_{wx}}{\bar{S}^2_{wx}}\right)\bar{S}^2_{wx} + \frac{M(n-1)}{nM-1}\left(1 + \frac{\epsilon_{bx}}{S^2_{bx}}\right) S^2_{bx}}$$

$$\ldots \quad (2.78)$$

Putting $\frac{n(M-1)}{nM-1} = A$ and $\frac{M(n-1)}{nM-1} = B$ above expression simplifies to

$$= \frac{1}{(A\bar{S}^2_{wx} + BS^2_{bx})} E \left[A^2 \bar{S}^2_{wxy}\left(1 + \frac{2\epsilon_{wxy}}{\bar{S}^2_{wxy}} + \frac{\epsilon^2_{wxy}}{\bar{S}^2_{wxy}}\right) + B^2 S^2_{bxy}\left(1 + \frac{2\epsilon_{bxy}}{S_{bxy}} + \frac{\epsilon^2_{bxy}}{S^2_{bxy}}\right) + 2AB S_{bxy} \bar{S}_{wxy}\left(1 + \frac{\epsilon_{bxy}}{S_{bxy}} + \frac{\epsilon_{wxy}}{\bar{S}_{wxy}}\right) \right] \times$$

$$\left[1 - \frac{A\bar{\epsilon}_{wx} + B\epsilon_{bx}}{A\bar{S}^2_{wx} + BS^2_{bx}} + \frac{A^2\bar{\epsilon}^2_{wx} + B^2\epsilon^2_{bx}}{(\bar{S}^2_{wx}A + BS^2_{bx})^2} \right] \quad \ldots \quad (2.79)$$

$$= \frac{1}{(A\bar{S}^2_{wx} + BS^2_{bx})} \left[(A\bar{S}_{wxy} + BS_{bxy})^2 \left[1 + \frac{A^2 V(\bar{S}^2_{wx}) + B^2 V(s^2_{bx})}{(A\bar{S}^2_{wx} + BS^2_{bx})^2} \right] + \right.$$

$$A^2 V(\bar{s}_{wxy}) + B^2 V(s_{bxy}) + \frac{2}{(A\bar{S}^2_{wx} + BS^2_{bx})} (A\bar{S}_{wxy} + BS_{bxy})$$

$$\left. \left[A^2 \mathrm{Cov}(\bar{s}^2_{wx}, \bar{s}_{wxy}) + B^2 \mathrm{Cov}(s^2_{bx}, s_{bxy}) \right] \right] \quad (2.80)$$

It is further known that

$$(NM-1)S^2 = M(N-1)S_b^2 + N(M-1)\bar{S}_w^2 \quad \ldots \quad (2.81)$$

or

$$S^2 = \frac{1-\frac{1}{N}}{1-\frac{1}{NM}} S_b^2 + \frac{1-\frac{1}{M}}{1-\frac{1}{NM}} \bar{S}_w^2 \quad \ldots \quad (2.82)$$

when N and M are large

$$S^2 \simeq S_b^2 + \bar{S}_w^2 \quad \ldots \quad (2.83)$$

Also since $A = \frac{1-\frac{1}{M}}{1-\frac{1}{nM}}$ and $B = \frac{1-\frac{1}{n}}{1-\frac{1}{nM}}$ $\ldots$ (2.84)

and when n and M are so large $A \longrightarrow 1$ and $B \longrightarrow 1$

$\ldots$ (2.85)

Thus

$$V_p = \frac{n-1}{n}\left[S_y^2 - \frac{1}{S_x^2}\left[S_{xy}^2 \left\{ 1 + \frac{V(\bar{s}_{wx}^2) + V(s_{bx}^2)}{S_x^4} \right\} + V(\bar{s}_{wxy}) + V(s_{bxy}) - \frac{2}{S_x^2} S_{xy} \left\{ Cov(\bar{s}_{wx}^2, \bar{s}_{wxy}) + Cov(s_{bx}^2, s_{bxy}) \right\} \right] \right]$$

We know that under normality assumptions

$$V(\bar{s}_{wx}^2) = V(m_{\bar{w}02}) = \frac{1}{n}\left[\mu_{\bar{w}04} - \mu_{\bar{w}02}^2 \right] \quad \ldots \quad (2.86)$$

$$V(s_{bx}^2) = V(m_{b02}) = \frac{1}{n}\left[\mu_{b04} - \mu_{b02}^2 \right] \quad \ldots \quad (2.87)$$

$$V(s_{bxy}) = V(m_{b11}) = \frac{1}{n}(\mu_{b22} - \mu_{b11}^2) \quad \ldots \quad (2.89)$$

$$V(\bar{s}_{wxy}) = V(m_{\bar{w}11}) = \frac{1}{n}(\mu_{\bar{w}22} - \mu_{\bar{w}11}^2) \quad \ldots \quad (2.90)$$

$$Cov(\bar{s}_{wx}^2, \bar{s}_{wxy}) = Cov(m_{\bar{w}02}, m_{\bar{w}11}) = \frac{1}{n}(\mu_{\bar{w}13} - \mu_{\bar{w}11}\mu_{\bar{w}02}) \quad \ldots \quad (2.91)$$

$$Cov(s_{bx}^2, s_{bxy}) = Cov(m_{b02}, m_{b11}) = \frac{1}{n}(\mu_{b13} - \mu_{b11}\mu_{b02}) \quad \ldots \quad (2.92)$$

Substituting these in the expression (2.76) we get

$$V_p = \frac{n-1}{n}\left[\mu_{20} - \frac{1}{\mu_{02}}\left[\mu_{11}^2\left\{1 + \frac{1}{n\mu_{02}^2}(\mu_{\bar{w}04} - \mu_{\bar{w}02}^2 + \mu_{b04} - \mu_{b02}^2)\right\}\right.\right.$$
$$+ \frac{1}{n}(\mu_{\bar{w}22} - \mu_{\bar{w}11}^2 + \mu_{b22} - \mu_{b11}^2) - \frac{2\mu_{11}}{n\mu_{02}}(\mu_{\bar{w}13} - \mu_{\bar{w}11}\mu_{\bar{w}02}$$
$$\left.\left. + \mu_{b13} - \mu_{b11}\mu_{b02})\right]\right] \quad \ldots \quad (2.93)$$

Above expression can also be written as

$$= \frac{n-1}{n}\left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} - \frac{1}{n\mu_{02}}\left[(\mu_{\bar{w}22} + \mu_{b22} - \mu_{b11}^2 - \mu_{\bar{w}11}^2) - \right.\right.$$
$$\frac{2\mu_{11}}{\mu_{02}}(\mu_{b13} + \mu_{\bar{w}13} - \mu_{\bar{w}11}\mu_{\bar{w}02} - \mu_{b11}\mu_{b02}) +$$
$$\left.\left.\frac{\mu_{11}^2}{\mu_{02}^2}(\mu_{b04} + \mu_{\bar{w}04} - \mu_{b02}^2 - \mu_{\bar{w}02}^2)\right]\right] \quad \ldots \quad (2.94)$$

which for the first degree of approximation reduces to

$$V_p = \left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} \right] \qquad \ldots \quad (2.95)$$

Thus, the variance of prediction in cluster sampling with equal cluster sizes is approximately equal to the variance of prediction for simple random sampling. Similar results may be expected in the situations where clusters are not of exactly equal size but of approximately equal size and the unit for which the value of y is to be predicted is drawn from the whole population or from the units not in the sample selected for building the prediction equation.

2.4. Two Stage Sampling

Multistage sampling has been found very useful in practice and this procedure is being currently used in most of the surveys. Many times multistage sampling may be the only feasible procedure when a satisfactory sampling frame of ultimately observable units is not readily available and the cost of obtaining such a frame is considerable. In this section the expression for variance of prediction has been worked out for the very simple situation where the sample is drawn by two stage sampling and the unit for which y value has to be predicted is further selected at random from the units in the sample. Let

N = total number of psu's in the population,

M = total number of ssu's in each psu of the population,

n = number of psu's in the sample,

m = number of ssu's from each of the selected psu in the sample.

It is known that (with usual notations)

$$(nm-1)s^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\bar{y}_{nm})^2 = m(n-1)s_b^2 + n(m-1)\bar{s}_w^2 \quad \ldots \quad (2.96)$$

and

$$E(s_b^2) = S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 , \; E(\bar{s}_w^2) = \bar{S}_w^2 \quad \ldots \quad (2.97)$$

Thus

$$E(s^2) = \frac{1-\frac{1}{n}}{1-\frac{1}{nm}}\left[S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2\right] + \frac{1-\frac{1}{m}}{1-\frac{1}{nm}}(\bar{S}_w^2) \quad \ldots \quad (2.98)$$

$$= \frac{1-\frac{1}{n}}{1-\frac{1}{nm}}S_b^2 + \frac{1-\frac{1}{m}+\left(\frac{1}{m}-\frac{1}{M}\right)\left(1+\frac{1}{n}\right)}{1-\frac{1}{nm}}\bar{S}_w^2 \quad \ldots \quad (2.99)$$

$$= \frac{1-\frac{1}{n}}{1-\frac{1}{nm}}S_b^2 + \frac{1-\frac{1}{M}-\frac{1}{nm}+\frac{1}{nM}}{1-\frac{1}{nm}}\bar{S}_w^2 \quad \ldots \quad (2.100)$$

$$\simeq S_b^2 + \bar{S}_w^2 , \text{ when } n, M \text{ and } N \text{ are large}$$

$$= S^2 \quad \ldots \quad (2.101)$$

$$\text{or } E(s_y^2) = S_y^2 \simeq \mu_{20} \qquad \ldots \quad (2.102)$$

Now putting

$$C = \frac{1 - \frac{1}{n}}{1 - \frac{1}{nm}} \quad , \quad D = \frac{1 - \frac{1}{M} - \frac{1}{nm} + \frac{1}{nM}}{1 - \frac{1}{nm}} \ldots \quad (2.103)$$

and

$$s_{xy} = C\, S_{bxy} + D\, \bar{S}_{wxy} = C\,(S_{bxy} + \epsilon_{bxy}) + D(\bar{S}_{wxy} + \epsilon_{wxy})$$
$$\ldots \quad (2.104)$$

$$s_x^2 = C\, S_{bx}^2 + D\, \bar{S}_{wx}^2 = C\,(S_{bx}^2 + \epsilon_{bx}) + D(\bar{S}_{wx}^2 + \bar{\epsilon}_{wx})$$
$$\ldots \quad (2.105)$$

etc. with $E(\epsilon_{bxy}) = 0$, $E(\epsilon_{bxy}^2) = V(\hat{S}_{bxy}) \ldots$ (2.106)

$$E(\bar{\epsilon}_{wxy}) = 0 \, , \; E(\bar{\epsilon}_{wxy}^2) = V(\hat{\bar{S}}_{wxy}) \text{ etc.} \ldots (2.107)$$

We get

$$E\left(\frac{s_{xy}^2}{s_x^2}\right) = \frac{1}{(CS_{bx}^2 + D\bar{S}_{wx}^2)} E\left[(CS_{bxy} + D\bar{S}_{wxy})^2 \left\{ 1 + \frac{C^2 V(\hat{S}_{bx}^2) + D^2 V(\hat{\bar{S}}_{wx}^2)}{(CS_{bx}^2 + D\bar{S}_{wx}^2)^2} \right\} \right.$$

$$\left. + C^2 V(\hat{S}_{bxy}) + D^2 V(\hat{\bar{S}}_{wxy}) - \frac{2}{(CS_{bx}^2 + D\bar{S}_{wx}^2)} \left\{ (CS_{bxy} + D\bar{S}_{wxy}) \cdot \right. \right.$$

$$\left. \left. \left\{ C^2 Cov(\hat{S}_{bx}^2, \hat{S}_{bxy}) + D^2 Cov(\hat{\bar{S}}_{wx}^2, \hat{\bar{S}}_{wxy}) \right\} \right\} \right] \ldots (2.108)$$

It may be mentioned here that s_b^2 and $\bar{S}_w^2$ are independent since s_b^2 is a function of $\bar{y}_i$'s and $\bar{S}_w^2$ is a function

of s_i^2 which are known to be independent under normality assumptions. Further the results from (2.85) to (2.92) holds true approximately for two stage sampling also therefore simplifying the above expression and noting that $C \longrightarrow 1$ and $D \longrightarrow 1$ when n, M, N are large, we get

$V_p = \left[\mu_{20} - \frac{\mu_{11}^2}{\mu_{02}} \right]$, upto the first degree of approximation.

As stratified sampling is a particular case of two stage sampling it is expected that when the strata are of approximately equal size and approximately equal number of units are drawn from each stratum, the variance of prediction may be approximately same as that of corresponding simple random sampling situation.

2.5 Summary

In this chapter the problem of prediction from finite populations has been considered for various situations. It has been observed that there is a component in variance of prediction which depends upon the sampling procedure used for the selection of sample. It has been further observed that this component is variance of prediction has approximately same magnitude in the three situations viz. simple random sampling, cluster sampling and two stage sampling. The results point out the possibility of analysing the data as simple random sample

even if procedures like stratified multistage sampling design were used for sampling purpose keeping in view the cost of aspect and administrative convenience.

CHAPTER - III

OPTIMUM SAMPLING DESIGN FOR PREDICTION

3.1 Introduction

It has been observed in Chapter - II that the performance of simple random sampling, cluster sampling, two stage sampling and stratified sampling is approximately the same in the sense of variance of prediction and as such there is nothing to choose between these sampling procedures as all these are equal probability sampling procedures. It is worth while to investigate the possibility of a sampling design for which the variance of prediction is the least. In this chapter this possibility has been examined and it is attempted to work out the optimum probabilities of selection for every unit in the population. Towards the end some empirical results are presented to compare the efficiencies of three sampling procedures for prediction.

3.2 Optimum Sampling Design

For simplicity of expressions in determining the optimum probabilities of selection we consider the simple case of sampling with replacement. Let p_i denotes the probability of selection for the i-th unit, ($i = 1, 2, \ldots, N$) and $\sum_{i=1}^{N} p_i = 1$. As the sampling is with replacement the probability of selection for a particular unit remains the same in all the draws. We

have, from Chapter - II , variance of prediction is given by

$$V_p = E\frac{1}{n}\left[\sum_{i=1}^{n}(y_i-\bar{y}_n)^2-\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right] \quad \ldots (3.1)$$

It is also known that

$$E\left[\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x}_n)^2}\right] = \frac{\left[E\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\right]^2}{E\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]} \cdot$$

$$\left[1+\frac{E\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)\cdot(y_i-\bar{y}_n)\right]^2}{\left[E\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\right]^2}+\frac{E\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]^2}{\left[E\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]^2}\right.$$

$$\left.-\frac{2E\left[\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\cdot\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]}{\left[E\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)\cdot E\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\right]}\right] \quad (3.2)$$

Further

$$\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n) = \frac{1}{n}\left[(n-1)\sum_{i=1}^{n}x_iy_i-\sum_{i\neq j}^{n}x_iy_j\right] \quad \ldots (3.3)$$

$$\sum_{i=1}^{n}(x_i-\bar{x}_n)^2 = \frac{1}{n}\left[(n-1)\sum_{i=1}^{n}x_i^2-\sum_{i\neq j}^{n}x_ix_j\right] \quad \ldots (3.4)$$

and

$$\sum_{i=1}^{n}(y_i-\bar{y}_n)^2 = \frac{1}{n}\left[(n-1)\sum_{i=1}^{n}y_i^2-\sum_{i\neq j}^{n}y_iy_j\right] \quad \ldots (3.5)$$

Since the sampling is with replacement the expectations of

(3.3), (3.4), and (3.5) are respectively given by

$$E\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)=(n-1)\left[\sum_{i=1}^{N}X_iY_ip_i-\sum_{i=1}^{N}X_ip_i\sum_{i=1}^{N}Y_ip_i\right] \quad \ldots (3.6)$$

$$E\sum_{i=1}^{n}(x_i-\bar{x}_n)^2=(n-1)\left[\sum_{i=1}^{N}X_i^2p_i-\left(\sum_{i=1}^{N}X_ip_i\right)^2\right] \quad \ldots (3.7)$$

and

$$E\sum_{i=1}^{n}(y_i-\bar{y}_n)^2=(n-1)\left[\sum_{i=1}^{N}Y_i^2p_i-\left(\sum_{i=1}^{N}Y_ip_i\right)^2\right] \quad \ldots (3.8)$$

$$\text{Also}\left[\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}\right]^2=\left(\sum_{i=1}^{n}x_iy_i\right)^2+\frac{1}{n^2}\left(\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right)^2-\frac{2}{n}\left[\sum_{i=1}^{n}x_iy_i\cdot\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right] \quad \ldots (3.9)$$

$$=\sum_{i=1}^{n}x_i^2y_i^2+\sum_{i\neq j}^{n}x_iy_ix_jy_j+\frac{1}{n^2}\left[\sum_{i=1}^{n}x_i^2y_i^2+\sum_{i\neq j}^{n}x_i^2y_j^2+\sum_{i\neq j\neq k}^{n}x_i^2y_jy_k\right.$$

$$+2\sum_{j\neq k}^{n}x_j^2y_jy_k+\sum_{i\neq j\neq k}^{n}y_i^2x_jx_k+2\sum_{j\neq k}^{n}y_j^2x_jx_k+\sum_{i\neq j\neq k\neq\ell}^{n}x_ix_jy_ky_\ell$$

$$\left.+4\sum_{i\neq j\neq k}^{n}x_ix_jy_iy_k+\sum_{i\neq j}^{n}x_iy_ix_jy_j\right]-\frac{2}{n}\left[\sum_{i=1}^{n}x_i^2y_i^2+\sum_{i\neq j}^{n}x_iy_ix_jy_j\right.$$

$$\left.+\sum_{i\neq j}^{n}x_i^2y_iy_j+\sum_{i\neq j}^{n}y_i^2x_ix_j+\sum_{i\neq j\neq k}x_iy_ix_jy_k\right] \quad \ldots (3.10)$$

Expected value of (3.10) is given by

$$E\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right]^2 = n \sum_{i=1}^{N} X_i^2 Y_i^2 p_i + n(n-1)\left(\sum_{i=1}^{N} X_i Y_i p_i\right)^2$$

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

$$\left(\sum_{i=1}^{N} Y_i p_i\right)^2 + 2(n-1)\sum_{j=1}^{N} X_j^2 Y_j p_j \sum_{k=1}^{N} Y_k p_k + (n-1)(n-2)\sum_{i=1}^{N} Y_i^2 p_i \cdot$$

$$\left(\sum_{i=1}^{N} X_i p_i\right)^2 + 2(n-1)\sum_{j=1}^{N} Y_j^2 X_j p_j \sum_{k=1}^{N} X_k p_k + (n-1)(n-2)(n-3) \cdot$$

$$\left(\sum_{i=1}^{N} X_i p_i\right)^2 \left(\sum_{j=1}^{N} Y_j p_j\right)^2 + 4(n-1)(n-2)\sum_{i=1}^{N} X_i Y_i p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k$$

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

$$+ (n-1)\sum_{i=1}^{N} X_i^2 Y_i p_i \sum_{j=1}^{N} Y_j p_j + (n-1)\sum_{i=1}^{N} Y_i^2 X_i p_i \sum_{j=1}^{N} X_j p_j$$

$$\left. + (n-1)(n-2)\sum_{i=1}^{N} X_i Y_i p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k\right] \quad \ldots (3.11)$$

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

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

$$- 2(n-1)\left[\sum_{j=1}^{N} X_j^2 Y_j p_j \sum_{k=1}^{N} Y_k p_k + \sum_{j=1}^{N} Y_j^2 X_j p_j \sum_{k=1}^{N} X_k p_k\right]$$

$$\left. + (n-2)(n-3)\left(\sum_{i=1}^{N} X_i p_i\right)^2 \left(\sum_{j=1}^{N} Y_j p_j\right)^2 + 2(n-2)\sum_{i=1}^{N} X_i Y_i p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k\right]$$

$$\ldots (3.12)$$

Similarly

$$E\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]^2 = \frac{n-1}{n}\left[(n-1)\sum_{i=1}^{N} X_i^4 p_i + \{1 + (n-1)^2\}\left(\sum_{i=1}^{N} X_i^2 p_i\right)^2\right.$$

$$+ 2(n-1)(n-2)\sum_{i=1}^{N} X_i^2 p_i \left(\sum_{i=1}^{N} X_i p_i\right)^2 - 4(n-1)\sum_{i=1}^{N} X_i^3 p_i \sum_{i=1}^{N} X_i p_i$$

$$\left. + (n-2)(n-3)\left(\sum_{i=1}^{N} X_i p_i\right)^4\right] \qquad \ldots (3.13)$$

Now

$$\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n) = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \qquad \ldots (3.14)$$

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

$$+ \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^3 \sum_{i=1}^{n} y_i \qquad \ldots (3.15)$$

$$= \sum_{i=1}^{n} x_i^3 y_i + \sum_{i \neq j}^{n} x_i^2 x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i + \sum_{i \neq j}^{n} x_i y_j\right)\sum_{i=1}^{n} x_i^2$$

$$- \frac{1}{n}\sum_{i=1}^{n} x_i y_i \left(\sum_{i=1}^{n} x_i^2 + \sum_{i \neq j}^{n} x_i x_j\right) + \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i^3 + 3\sum_{i \neq j}^{n} x_i^2 x_j + \right.$$

$$\left.\sum_{i \neq j \neq k}^{n} x_i x_j x_k\right)\left(\sum_{i=1}^{n} y_i\right) \qquad \ldots (3.16)$$

$$= \left[\sum_{i=1}^{n} x_i^3 y_i + \sum_{i \neq j}^{n} x_i^2 x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^3 y_i + \sum_{i \neq j}^{n} x_i^2 x_j y_j + \sum_{i \neq j}^{n} x_i^3 y_j\right.\right.$$

$$\left.\left. + \sum_{i \neq j}^{n} x_i^2 y_i x_j + \sum_{i \neq j \neq k}^{n} x_i^2 x_j y_k\right) -$$

contd...

$$\frac{1}{n}\Big(\sum_{i=1}^{n} x_i^3 y_i + \sum_{i\neq j}^{n} x_i^2 y_j x_j + 2\sum_{i\neq j}^{n} x_i^2 y_i x_j + \sum_{i\neq j\neq k}^{n} x_i y_i x_j x_k\Big)$$

$$+ \frac{1}{n^2}\Big(\sum_{i=1}^{n} x_i^3 y_i + \sum_{i\neq j}^{n} x_i^3 y_j + 3\sum_{i\neq j}^{n} x_i^2 y_i x_j + 3\sum_{i\neq j}^{n} x_i^2 x_j y_j$$

$$+ 3\sum_{i\neq j\neq k}^{n} x_i^2 x_j y_k + 3\sum_{i\neq j\neq k}^{n} x_i y_i x_j x_k + \sum_{i\neq j\neq k\neq \ell}^{n} x_i x_j x_k y_\ell\Big)\Big] \qquad \ldots (3.17)$$

$$= \frac{1}{n^2}\Big[(n-1)^2\sum_{i=1}^{n} x_i^3 y_i + \Big\{(n-1)^2 + 2\Big\}\sum_{i\neq j}^{n} x_i^2 x_j y_j - (n-1)\sum_{i\neq j}^{n} x_i^3 y_j$$

$$- 3(n-1)\sum_{i\neq j}^{n} x_i^2 y_i x_j - (n-3)\Big(\sum_{i\neq j\neq k}^{n} x_i^2 x_j y_k + \sum_{i\neq j\neq k}^{n} x_i y_i x_j x_k\Big)$$

$$+ \sum_{i\neq j\neq k\neq \ell}^{n} x_i x_j x_k y_\ell\Big] \qquad \ldots (3.18)$$

Expected value of (3.18) is given by

$$E\Big\{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\Big\}$$

$$= \frac{n-1}{n}\Big[(n-1)\sum_{i=1}^{N} X_i^3 Y_i p_i + \Big\{(n-1)^2 + 2\Big\}\sum_{i=1}^{N} X_i^2 p_i \sum_{j=1}^{N} X_j Y_j p_j$$

$$+ -(n-1)\Big\{\sum_{i=1}^{N} X_i^3 p_i \sum_{j=1}^{N} Y_j p_j + 2\sum_{i=1}^{N} X_i^2 Y_i p_i \sum_{j=1}^{N} X_j p_j\Big\} - (n-2)(n-3)\cdot$$

$$\Big\{\sum_{i=1}^{N} X_i^2 p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k + \sum_{i=1}^{N} X_i Y_i p_i \Big(\sum_{i=1}^{N} X_i p_i\Big)^2$$

$$+ \Big(\sum_{i=1}^{N} X_i p_i\Big)^2 \sum_{i=1}^{N} Y_i p_i\Big\}\Big] \qquad \ldots (3.19)$$

Substituting corresponding expected values in (3.1) we get

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

$$\left[(n-1)\sum_{i=1}^{N} X_i^2 Y_i^2 p_i + (n-1)^2\left(\sum_{i=1}^{N} X_i Y_i p_i\right)^2 + \sum_{i=1}^{N} X_i^2 p_i \sum_{i=1}^{N} Y_i^2 p_i + (n-2)\left[\sum_{i=1}^{N} X_i^2 p_i \cdot\right.\right.$$

$$\left.\left(\sum_{i=1}^{N} Y_i p_i\right)^2 + \sum_{i=1}^{N} Y_i^2 p_i \left(\sum_{i=1}^{N} X_i p_i\right)^2\right] - 2(n-1)\left[\sum_{j=1}^{N} X_j^2 Y_j p_j \sum_{k=1}^{N} Y_k p_k +\right.$$

$$\left.\sum_{j=1}^{N} Y_j^2 X_j p_j \sum_{k=1}^{N} X_k p_k\right] + (n-2)(n-3)\left(\sum_{i=1}^{N} X_i p_i\right)^2 \left(\sum_{j=1}^{N} Y_j p_j\right)^2 + 2(n-2)^2 \cdot$$

$$\left.\sum_{i=1}^{N} X_i Y_i p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k\right] \cdot \left[\sum_{i=1}^{N} X_i Y_i p_i - \sum_{i=1}^{N} X_i p_i \sum_{i=1}^{N} Y_i p_i\right]^{-2}$$

$$+ \frac{1}{n(n-1)}\left[(n-1)\sum_{i=1}^{N} X_i^4 p_i + \left[1 + (n-1)^2\right]\left(\sum_{i=1}^{N} X_i^2 p_i\right)^2 + 2(n-1)(n-2)\sum_{i=1}^{N} X_i^2 p_i \left(\sum_{i=1}^{N} X_i p_i\right)^2\right.$$

$$\left.- 4(n-1)\sum_{i=1}^{N} X_i^3 p_i \sum_{i=1}^{N} X_i p_i + (n-2)(n-3)\left(\sum_{i=1}^{N} X_i p_i\right)^4\right] \cdot \left[\sum_{i=1}^{N} X_i^2 p_i - \left(\sum_{i=1}^{N} X_i p_i\right)^2\right]^{-2}$$

$$- \frac{2}{n(n-1)}\left[(n-1)\sum_{i=1}^{N} X_i^3 Y_i p_i + \left[(n-1)^2 + 2\right]\sum_{i=1}^{N} X_i^2 p_i \sum_{j=1}^{N} X_j Y_j p_j - (n-1)\left[\sum_{i=1}^{N} X_i^3 p_i \cdot\right.\right.$$

$$\left.\left.\sum_{j=1}^{N} Y_j p_j + 3\sum_{i=1}^{N} X_i^2 Y_i p_i \sum_{j=1}^{N} X_j p_j\right] + (n-2)(n-3)\left[\sum_{i=1}^{N} X_i^2 p_i \sum_{j=1}^{N} X_j p_j \sum_{k=1}^{N} Y_k p_k\right.\right.$$

$$\left.\left.+ \sum_{i=1}^{N} X_i Y_i p_i \left(\sum_{i=1}^{N} X_i p_i\right)^2 + \left(\sum_{i=1}^{N} X_i p_i\right)^3 \sum_{i=1}^{N} Y_i p_i\right]\right] \cdot \left[\left[\sum_{i=1}^{N} X_i Y_i p_i - \sum_{i=1}^{N} X_i p_i \sum_{i=1}^{N} Y_i p_i\right] \cdot\right.$$

$$\left.\left.\left[\sum_{i=1}^{N} X_i^2 p_i - \left(\sum_{i=1}^{N} X_i p_i\right)^2\right]^{-1}\right]\right] \qquad \ldots (320)$$

For obtaining a solution of p_i's for expression (3.20), variance to have minimum value subject to $\sum_{i=1}^{N} p_i = 1$, we have to differentiate partially (3.20) with respect to p_i and equate to zero. The ultimate expression after differentiation and equating to zero is unmanagable and thus nothing can be concluded. To simplify the calculations further we assume for large sample size [Cochran (1963)]

$$\bar{y}_n \simeq \bar{Y}_N \quad \text{and} \quad \bar{x}_n \simeq \bar{X}_N \qquad \ldots (3.21)$$

Throughout we replace $x_i - \bar{x}_n \simeq x_i - \bar{X}_N$ by x_i and

$$y_i - \bar{y}_n \simeq y_i - \bar{Y}_N \text{ by } y_i$$

Now

$$E \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{N} X_i^2 p_i \qquad \ldots (3.22)$$

$$E \sum_{i=1}^{n} y_i^2 = n \sum_{i=1}^{N} Y_i^2 p_i \qquad \ldots (3.23)$$

and

$$E \sum_{i=1}^{n} x_i y_i = n \sum_{i=1}^{N} X_i Y_i p_i \qquad \ldots (3.24)$$

Also

$$E \left(\sum_{i=1}^{n} x_i^2 \right)^2 = n \left[\sum_{i=1}^{N} X_i^4 p_i + (n-1) \left(\sum_{i=1}^{N} X_i^2 p_i \right)^2 \right] \qquad \ldots (3.25)$$

$$E \left(\sum_{i=1}^{n} y_i^2 \right)^2 = n \left[\sum_{i=1}^{N} Y_i^4 p_i + (n-1) \left(\sum_{i=1}^{N} Y_i^2 p_i \right)^2 \right] \qquad \ldots (3.26)$$

$$E \left(\sum_{i=1}^{n} x_i y_i \right)^2 = n \left[\sum_{i=1}^{N} X_i^2 Y_i^2 p_i + (n-1) \left(\sum_{i=1}^{N} X_i Y_i p_i \right)^2 \right] \qquad \ldots (3.27)$$

$$E\left[\sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i^2\right] = n\left[\sum_{i=1}^{N} x_i^3 y_i p_i + (n-1)\left(\sum_{j=1}^{N} x_j^2 p_j\right)\left(\sum_{i=1}^{N} x_i y_i p_i\right)\right] \quad \ldots (3.28)$$

Hence the variance of prediction reduces to

$$V_p = \left[\sum_{i=1}^{N} y_i^2 p_i - \frac{\left(\sum_{i=1}^{N} x_i y_i p_i\right)^2}{\left(\sum_{i=1}^{N} x_i^2 p_i\right)}\right] + \frac{\left[\sum_{i=1}^{N} x_i^2 y_i^2 p_i + (n-1)\left(\sum_{i=1}^{N} x_i y_i p_i\right)^2\right]}{n\left(\sum_{i=1}^{N} x_i y_i p_i\right)^2}$$

$$+ \frac{\left[\sum_{i=1}^{N} x_i^6 p_i + (n-1)\left(\sum_{i=1}^{N} x_i^2 p_i\right)^2\right]}{n\left(\sum_{i=1}^{N} x_i^2 p_i\right)^2} - 2\left[\frac{\sum_{i=1}^{N} x_i^3 y_i p_i + (n-1)\left(\sum_{i=1}^{N} x_i^2 p_i\right)\sum_{i=1}^{N} x_i y_i p_i}{n\left(\sum_{i=1}^{N} x_i y_i p_i\right)\left(\sum_{i=1}^{N} x_i^2 p_i\right)}\right] \quad \ldots (3.29)$$

If (3.29) is partially differentiated with respect to p_i subject to $\sum_{i=1}^{N} p_i = 1$, the ultimate expression does not simplify to a meaningful form.

However, in order to have some idea we consider the following selection procedures for comparison:

(i) Procedure - A : All units have equal chance of selection i.e.

$$p_i = \frac{1}{N} \quad \ldots (3.30)$$

(ii) Procedure - B : Extreme units have more chance of selection than units in the centre i.e.

$$p_i \propto \left(X_i - \bar{X}_N\right)^2 \quad \ldots (3.31)$$

and

(iii) Procedure - C : Units around the mean have more chance of selection than extreme units i.e.

$$p_i \propto (X_i - \bar{X}_N)^{-2} \qquad \ldots (3.32)$$

For comparison purpose few populations satisfying the model given below were generated

$$Y_i = \beta x_i + \epsilon_{ij} \quad (i = 1,2,\ldots,4 ; j = 1,2,\ldots,5)$$

$$E(\epsilon_{ij} / x_i) = 0$$

$$E(\epsilon_{ij}^2 / x_i) = A x_i^{\beta} .$$

TABLE - 3.1

Table showing the variances of various sampling procedures under considerations for different populations

POPULATION - 1
N = 80, n = 30, A = 100, and β = 0.4

	β = 0.0	0.5	1.0	1.5	2.0
Procedure - A	335.44	2035.23	8368.64	32913.29	128839.27
Procedure - B	772.21	4100.93	16900.01	67021.30	264971.36
Procedure - C	1175.08	6210.90	25547.43	100474.17	393322.27

POPULATION - 2
N = 40, n = 20, A = 100 and β = 0.4

Procedure-A	273.33	1309.99	5140.74	19396.08	72802.96
Procedure-B	553.92	2661.13	10473.87	39712.49	150135.02
Procedure-C	899.60	4029.94	15824.24	59725.07	224205.39

contd...

table - 3.1 (contd.)

POPULATION - 3
N = 30, n = 10, A = 100 and β = 0.4

	θ				
	0.0	0.5	1.0	1.5	2.0
Procedure-A	139.37	610.24	2279.92	8227.72	29510.85
Procedure-B	289.60	1271.01	4759.06	17233.69	62118.52
Procedure-C	440.82	1934.25	7234.83	26132.96	93807.05

POPULATON - 4
N = 12, n = 5, A = 100 and β = 0.4

Procedure-A	441.67	1956.05	6494.10	19514.20	6[illegible]628.19
Procedure-B	957.17	3370.20	11944.69	42604.44	152908.57
Procedure-C	1469.75	5182.05	18356.28	65319.42	233469.33

It can be seen from the above results that out of the three procedures considered here the performance of equal probability sampling scheme is the best throughout. The results presented here are for sampling with replacement and it is expected that for sampling without replacement results may also be of similar type. Thus, for prediction purposes equal probability sampling procedures such as simple random sampling or stratified multistage random sampling can be safely used.

3.8 Summary

In this chapter the possibility of an optimum sampling

design for prediction has been examined. It has been observed that the performance of equal probability sampling procedure is highly satisfactory as compared to the other procedures from the variance of prediction point of view. Thus for prediction purposes the use of sampling schemes like simple random sampling or stratified multistage sampling can be safely recommended.

CHAPTER - IV

AN APPLICATION TO PRE-HARVEST FORECASTING

4.1. Introduction :

Reliable estimate of total crop production well before the harvest is of considerable importance in price and export-import policies. The method of forecasting of crop production being used at present is based on eye estimates and hence is subjective. It needs to be investigated whether forecasting of production would be made more reliable and objective by taking into consideration biometrical observations on the plant during the various stages of its growth. Of the two component of crop production, acreage and yield, acreage presents a comparatively simple problem. Ordinarily, acreage is much more stable than yield per acre. The acreage of a crop is usually influenced by the ratio of the price of that crop to those of competitive crops which prevailed during the previous year. In addition wheather also some times affects the change in acreage.

With a view to bring about improvement in forecasting of Jute crop through the use of measurement of biometrical characters the Institute of Agricultural Research Statistics in Collaboration with a Department of Agriculture, Bihar took a pilot study on Jute crop. In this chapter we discuss the various statistical problems with the help of data collected during 1971-72 to 1973-74. in the Purnea district of Bihar.

4.2 Problems Relating to Crop Fore-casting

The problem of crop forecasting involves number of important problems some of which are being discussed in the following sections:

(i) Choice of sampling design and organisation of survey work:

As indicated in Chapter- II that all procedures of equal probability sampling are approximately equally efficient for prediction purposes, a two stage stratified random sampling design was therefore used for selection of plots mainly for administrative convenience and equality in distribution of work. Four community development blocks covered under the special package programme were taken for the study which constituted the population under study. There were six to seventeen V. L. W. Circles in a block which formed the strata. In each circle five fields growing the Jute crop were selected at random and from each selected field two plots each of area 2 x 2 meter were further selected at random for recording detailed biometrical observations.

In each selected plot total number of plants were counted at each occasion of recording biometrical observations. For measuring the plant height and basal diameter during the crop growth period five plants four corner and one central were

taken in each plot. The first observation was recorded after about a month of sowing of the crop and subsequent observation were recorded periodically at an interval of one month upto and including the time of harvest. The fibre weight of Jute was taken after retting the crop.

(2) Choice of biometrical characters :

The main objective of study is to predict the yield rate on the basis of biometrical characters and the approach to be followed is regression technique of yield on biometrical character influencing the yield after adjusting the data for differences between block, circles and fields. Correlation coefficients between the yield y and (i) number of plants x_1 , (ii) height of the plant x_2 and (iii) basal diameter have been worked out for each recording for different years and are presented in the table given below.

TABLE - 4.1

Year	Period after sowing in months	Total Correlation Coefficients Between Yield and		
		Number of plants	Height of plants	Diameter of plants
1971-72	1.	0.4845**	0.36**	0.41**
	2.	0.4805**	0.62**	0.81**
	3.	0.4759**	0.65**	0.89**
	4.	0.5191**	0.29**	0.14
1972-73	1.	0.3520**	0.13	0.06
	2.	0.3338**	0.18*	- 0.09
	3.	0.3540**	0.10	- 0.11
	4.	0.4920**	0.12	- 0.09

contd...

table - 6.1 (contd.)

Year	Period after sowing in months	Total Correlation Coefficients Between Yield and Number of plants	Height of plants	Diameter of plants
1973-74	1.	0.6398^{oo}	0.19^{o}	0.07
	2.	0.5524^{oo}	0.15	- 0.03
	3.	0.5695^{oo}	0.11	0.01
	4.	0.5673^{oo}	0.06	- 0.01

^{o} and ^{oo} denote significance at 5 per cent and 1 per cent level of significance respectively.

The correlation between fibre weight and number of plants were found significant consistantly in all the cases. The significance of correlation coefficient of fibre weight with height of the plants and diameter of the plants was observed only for the year 1971-72. Secondly, there is a doubt of measurement errors with the measurement on height and basal diameter of the plants. Thus, number of plants can be chosen as predicting variable for yield rate of Jute. Also, as in Chapter-II the prediction problem with one variable has only been discussed therefore, we presents the various statistical problems in pre-harvest forecasting the yield rate with number of plants only.

(3) Choice of prediction model and determination of period of forecasting :

For determination of prediction function following four models were tried for each period of recording using the method of least squares.

Model - 1 : $y = a_1 + b_1 x$

Model - 2 : $\log y = a_2 + b_2 \log x$

Model - 3 : $\sqrt{y} = a_3 + b_3 \sqrt{x}$

Model - 4 : $\frac{1}{y} = a_4 + \frac{b_4}{x}$

where y and x denote yield per plot and number of plants per unit square meter respectively. The results are presented in the following tables.

TABLE - 4.2

Table showing the regression coefficient, correlation coefficient and variance of prediction

YEAR 1971-72

Model	Period	a	b	r	Variance of prediction
I	1.	- 13.695	0.1579	0.4846	590.80
	2.	- 12.645	0.1605	0.4885	591 .13
	3.	- 11.009	0.1549	0.4759	598.28
	4.	- 11.792	0.1730	0.5191	559.04
II	1.	- 1.4567	0.7176	0.5844	647.76
	2.	- 1. 371	0.7092	0.5965	637.97
	3.	- 1. 289	0.6993	0.5911	641.95
	4.	- 1.079	0.6764	0.6110	417.09
III	1.	- 1.048	0.3363	0.5341	867.35
	2.	- 0.968	0.3364	0.5418	864.40
	3.	- 0.841	0.3343	0.5331	868.67
	4.	- 0.848	0.3510	0.5680	856.16
IV	1.	0.115	4.0435	0.6593	9551.84
	2.	0.114	3.7974	0.6946	9318.16
	3.	0.115	3.6261	0.6705	9470.49
	4.	0.105	3.0153	0.6643	939.96

contd...

table - 4.2 (contd.) YEAR 1972 - 73

Model	Period	a	b	r	Variance of prediction
I	1.	-9.457	0.1372	0.3528	1011.49
	2.	-7.919	0.1402	0.3338	1037.20
	3.	-8.223	0.1494	0.3540	1029.29
	4.	-8.419	0.1532	0.4328	456.58
II	1.	- 0.935	0.6327	0.4595	1170.48
	2.	- 0.542	0.5672	0.3939	1237.16
	3.	- 0.758	0.6183	0.4461	1156.74
	4.	- 0.498	0.5762	0.4284	725.59
III	1.	- 0.476	0.3016	0.4172	979.16
	2.	- 0.173	0.2917	0.3828	1010.56
	3.	- 0.315	0.3112	0.4150	971.60
	4.	-0. 195	0.3073	0.4533	512.60
IV	1.	0.116	2.4058	0.4457	3450.24
	2.	0.119	1.6050	0.2725	3854.74
	3.	0.115	2.0591	0.3856	3509.74
	4.	0.107	1.8929	0.2801	2846.44

YEAR 1973 - 74

I	1.	- 11.673	0.1727	0.5398	459.46
	2.	-10.421	0.1779	0.5924	441.28
	3.	-10.736	0.1879	0.5695	429.75
	4.	- 9.789	0.1900	0.5673	439.09
II	1.	- 0.995	0.6664	0.6432	404.29
	2.	- 0.904	0.6639	0.6573	417.00
	3.	- 0.9387	0.6780	0.6534	409.89
	4.	- 0.695	0.6426	0.6062	381.04
III	1.	- 0.709	0.3427	0.5981	412.24
	2.	- 0.568	0.3479	0.6103	400.76
	3.	- 0.617	0.3572	0.6148	391.88
	4.	- 0.408	0.3522	0.5920	353.84
IV	1.	0.108	2.3680	0.6400	1004.76
	2.	0.102	2.2688	0.6792	925.89
	3.	0.106	2.3327	0.6742	946.00
	4.	0.094	2.0179	0.6115	603.28

It seems from the Table - 4.2 that no appreciable difference is observed in the values of correlation coefficient by transforming the data under model 2, 3 and 4. It is also noted that the explainable amount of variation in yield is upto 50 per cent. Further the third and fourth periods of observation give relatively large value of correlation coefficients for almost all the years and all the models.

For the choice between models based on different transformations, the techniques of minimum residual sum square is not strictly applicable, since the dependent variable is in different scales. For that we have to compare the prediction variances associated with each model for predicting yield rate on the basis of number of plants observed on randomly selected plots. For this the approximate formulae for variance of y in terms of variance of $f(y)$ is given by

$$\text{Var}(y) = \frac{V[f(y)]}{\left(\frac{\partial f}{\partial y}\right)^2_{y=\bar{Y}}}$$ (Kendall, Stuart. (1969)) can be used.

The formulae for variances of prediction in the present situation for the various models considered here are given by

$$V(y) = S_y^2(1-\rho^2) \qquad \text{for model - I}$$

$$V(y) = S^2_{\log y}\left(1-\rho^2_{\log y}\right)\bar{Y}^2 \qquad \text{for model - II}$$

$$V(y) = S^2_{\sqrt{y}} \left(1 - \rho^2_{\sqrt{y}}\right) 4\bar{Y} \qquad \text{for model - III}$$

$$V(y) = S^2_{\frac{1}{y}} \left(1 - \rho^2_{\frac{1}{y}}\right) \bar{Y}^4 \qquad \text{for model - IV}$$

The variances of prediction associated with each model are also presented in the last column of Table - 4.2. It can be seen that the variance of prediction is relatively less for the third model followed by the first model. However, the difference of the third model with the first model is not appreciable. Further the last period of recording gives the least variance of prediction for all the years uniformly. Thus for pre-harvest forecasting for yield of Jute the optimum time of forecast is about four month after sowing and prediction model to be used should be simple linear in the original scale or in the square root scale. Considering the simplicity of calculations etc. simple linear regression model between dry fibre weight and number of plants recorded after about four month of sowing should be utilized for forecasting the yield rate of Jute.

(4) Stability of regression coefficients :

Having decided about the suitable prediction model and optimum time of forecasting the problem remains is about the stability of regression coefficient. By taking number of

plants as the predicting variable, there is observed a uniform consistency in the average number of plants and the variability in the number of plants. It can be seen from the Table - 4.2 that the regression coefficients are highly stable for all the three years. It is also revealed that the regression coefficients are not only stable from year to year but also from period to period also. As the coefficients are stable they can be pooled and the combined regression equation to be used for prediction is given by

$$y = -17.1447 + 0.1723\,x$$

(5) Estimation of variance of prediction:

The formula for variance of prediction involves knowledge of S_y^2 and ρ^2. For pre-harvest forecasting purposes where the problem itself is the prediction of y the yield rate on the basis of x alone, it is not possible to estimate unbiasedly S_y^2 and also correlation coefficients between x and y. On the basis of x alone. Thus in real practice we will be having information on x only and not on y. To over come this difficulty some results are being presented in the Table - 4.3.

TABLE - 4.3

Values of S_y^2 / S_x^2 for the years 1971 - 72 to 1973-74

	YEAR		
Period	1971-72	1972-73	1973-74
1.	0.1057	0.1519	0.1022
2.	0.1080	0.1764	0.1040
3.	0.1151	0.1781	0.1033
4.	0.1111	0.1255	0.1132

It can be seen from the above table that the ratio of the variances of yield to the number of plants is constant at the fourth period of recording which is the time proposed for pre-harvest forecasting. Thus on the basis of the variance of number of plants the value of S_y^2 can be estimated making use of the constant of the ratios of the variances. The another problem regarding the estimation of correlation coefficients can be best tackled by pooling the correlation coefficients with inverse of variances. In the present situation the value of pooled correlation coefficient was 0.5089.

4.3. Summary

In this chapter the various statistical problems arising in the forecasting of yield rate have been discussed. It has been shown that number of plants can be taken as a predicting variable

for forecasting the yield rate from all aspects such as variance of prediction, less chance of measurement errors, simplicity in counting etc. Further, it has been observed that a forecast after four months of sowing had relatively smaller variance of prediction and simple linear regression model between yield and number of plants is satisfactory. The regression coefficient were observed to be stable for all the years. It has also been observed that the ratio of variances of yield rate and the number of plants is highly stable which will allow in estimation of variance of y with the help of variance of x . Final model after pooling regression coefficients has been suggested and a method of estimating variance of prediction has also been suggested at the end.

SUMMARY

Regression relations are of very much use in predicting the value of one variable on the basis of the given values of other variables. Recent literature in sampling theory has gathered around the estimation of mean and total only. The relative efficiencies of various sampling procedures have been very well discussed in relation to estimation of population mean or total. No effort, however, seems to have been made to investigate the possibility of an optimum sampling design which can be used for prediction purposes. In this investigation the problem of determining the regression relation on the basis of a sample from finite population has been considered. The expressions for the variance of prediction have been worked out for the situations when the sample is drawn by simple random sampling, cluster sampling, two stage sampling and stratified sampling under different situations. It has been observed that the performance of all the procedures mentioned above is approximately the same in respect of the prediction variance. The results thus point out the possibility of analysing the sample as if it is drawn by simple random sampling even when the procedures like multi-stage stratified sampling etc. are adopted from the administrative and technical conveniences.

The problem of determining an optimum sampling design for prediction has also been considered in this investigation and it has been observed that the solution in general is complex. However,

to have some idea few empirical results are discussed to compare three sampling procedures under wide range of population. It has been observed that the performance of equal probability sampling is highly satisfactory in comparison to other sampling procedures considered. The results thus indicate that simple random sampling can safely be used for determining a linear regression relationship to be used for prediction.

In the end the application of the suggested procedure has been explained to the data collected under pre-harvest forecasting scheme during 1970-71 to 1973-74 of Jute crop in Bihar State. The solutions of various statistical problems involved in pre-harvest forecasting have also been indicated.

REFERENCES

1. Adcock, R.J. (1878). " A Problem on least square". The Analyst 5.

2. Allen, R.G.D. (1939). " Assumptions of linear regression". Econometrica 6.

3. Bartlett, M.S. (1949). " Fitting a straight line when both the variables are subject to error" Biometrics 5.

4. Cochran, W.G. (1963). " Sampling Techniques", New York, John Wiley and Sons.

5. Draper, N.R. and Smith, H. (1966). " Applied regression analysis " New York, John Wiley and Sons.

6. Gibson, W.M. and Jowett, G.H. (1957). " Three group" regression analysis 'I' simple regression. Applied Statistics 6 : 2.

7. Gibson, W.M. and Jowett, G.H. (1957). " Three group " regressions analysis "II" simple regression. Applied Statistics" 6 : 3.

8. Hooper, J.H. and Theil, H. (1958). "Extension of Wald's method of fitting of straight line to the case of two independent variables". Review of International Statistical Institutes.

9. Johnston, J. (1963). " Econometric methods" McGraw Hill Book Company.

10. Madansky, Albert (1959). " Fitting of straight lines when both the variables are subject to error" Journal of American Statistical Association, 54.

11. Mahajan, V.K. (1971). On the Methodology of Pre-harvest Forecasting of Jute crop. (Unpublished M.Sc. thesis, I.A.R.S.)

12. Murthy, M.N. (1967). " Sampling Theory and Methods". Statistical Publishing Society, Calcutta - 35, India.

13. M.G. Kendall and A. Stuart (1969). " The advanced theory of statistics". Charles Griffin and Company Limited, 42, Durby Land, London, W.C. 2.

14. Nair, K.R. and K.S. Banerjee (1942). " Note on fitting of straight line if both the variables are subject to error". Sankhya 6.

15. Opponheim and Roos(1928)" Bulletin of the American Mathematical Society" Vol. 34.

16. Ramesh, B.S. (1974). " Some Methodological Investigations on Pre-harvest Forecasting of Jute Crop based on Biometrical Observations".(Unpublished Diploma thesis, I.A.R.S.)

17. Sanderson, Fred H.(1954)." Method of Crop Fore-casting". Cambridge Massachusetts: Harvard University Press.

18. Singh, H.P., Singh, Padam and Jha, M.P. (1974). " Consolidated report of the pilot study on Pre-harvest forecasting of yield of Jute crop in Purnea (Bihar).

19. Sukhatme, P.V. and Sukhatme, B.V. " Theory of Sampling with Applications" Asia Publishing House, New Delhi.

20. Wald, A. (1940). " Fitting of straight lines if both the variables subject to error". Annals of Mathematical Statistics".

21. Williams, E.J. (1959). " Regression Analysis". New York, John Wiley and Sons.