USE OF DOUBLE SAMPLING
IN REPEATED SURVEYS

By

R. D. Singh.

Dissertation submitted in fulfillment of
the requirements for the award of Dip-
loma in Agricultural & Animal Hus-
bandry Statistics of the Ins-
titute of Agricultural Re-
search Statistics.

New Delhi

1962

ACKNOWLEDGEMENT

I have pleasure in expressing my deepest sense of gratitude to Shri D. Singh, Senior Research Statistician, Institute of Agricultural Research Statistics (I.C.A.R.), New Delhi for his valuable guidance, constant encouragement and constructive criticism during the course of investigation and writing this thesis.

My thanks are also due to the Institute of Agricultural Research Statistics for the facilities provided to me during the course of investigation.

B D Singh
7/9/62
(B.D. Singh)

CONTENTS

INTRODUCTION

For a population which changes with time a single survey on a particular occasion, furnishes information, which is valid for that occasion only, and it does not give any information on the nature of or rate of changes which occur in the population. But many a time the interest of the sampler does not lie only in estimating the value of the character for the most recent occasion but his interest goes beyond, such as estimating the change in the value of the character from one occasion to the next, estimating the average value of the character over all occasions in a given period of time etc. In such cases periodical resurvey must be made on the same population. Once we have decided to study the population on successive occasions, several alternatives are open before us. We can survey the same fixed sample on all occasions, or take independent samples on each occasion, or take a sub-sample of the previous occasion, or supplement the sub-sample with another independent sample taken afresh. The relative advantages of the various types of procedure depend on the relation between the variability of the units and the variability of changes in these units as well as on the relative importance of information on the population means and on the changes in these means.

In case of complex designs, such as multistage and double sampling the alternatives become much more. For example when a multistage design is repeated on several occasions, one may like to retain the first-stage sampling units from occasion to occasion but select each time a fresh sample of second stage units from the selected first stage units or retain only part of the first stage units along with thier samples of second stage units etc.

Suppose that we are free to alter or retain the composition of the sample, and that the total sample size is to be same on all occasions. If we wish to maximise the precision, the following statements can be made about replacement policy:

(i) For estimating change, it is best to retain the same sample throughout all occasions.

(ii) For estimating the average over all ocasions it is best to draw a new sample on each occasion.

(iii) For current estimates equal precision is obtained by keeping the same sample or by changing it on every occasion. Replacement of part of the sample on each occasion may be better than these alternatives.

As Yates has remarked there are two further points which must be borne in mind in connection with sampling on successive occasions. Firstly repeated resurvey of the same units may be inexpedient since resistance to the provision of the necessary information may be engendered and secondly repeated resurvey may result in modifications of these units relative to the rest of the population.

The first attempt to study the theory of sampling on successive occasions with partial replacement of units on each occasion was made by Jessen (1942). However he confined himself to only two occasions. He built two independent estimates for the mean on second occasion, one on the basis of the units common to both occasions and another based on the units selected afresh on the second occasion. The former was a double sampling regression estimate for the mean on second occasion, the information on the first occasion serving as the preliminary large sample for ancillary variate and the latter was a simple average of the units confined to the second occasion only. These two estimates were weighted with reciprocal of their variances to get an estimate with minimum variance. Jessen also gave expression for the optimum proportion of units to be retained on the second occasion.

Yates (1949) was more liberal in his approach. He contended that the most straight forward procedure for estimating the values of the population mean on two successive occasions was to treat each occasion seperately, following whatever method of estimation was appropriate to the sample obtained on that occasion, regardless of values obtained on the other occasions. Such estimates he termed as overall estimates. He considered two important cases, the one when the sample on the second occasion was confined to a sub-sample of the original sample and the other when the sub-sample retained from the first occasion was supplemented with a fresh sample on the second occasion.

Yates also considered the general case of successive sampling for h occasions. Under the limitations that
(i) a given fraction of units is replaced on each ocasion,
(ii) the variability on the different occasions and the correlation r between successive occasions are constant, and
(iii) the correlation between occasions two apart is r^2, that between three apart is r^3 etc.
he obtained the relation

$$\bar{Y}_h = (1 - \phi_h)\left\{ \bar{y}'_h + r\left(\bar{Y}_{h-1} - \left[\bar{y}'_{h-1}\right] \right) \right\} + \phi_h \, \bar{y}''_h$$

where $\bar{Y}_h$ is the most accurate estimate which can be obtained for occasion h, taking into account the result of sampling up to and including this occasion h, and $\bar{Y}_{h-1}$ is a similar estimate for the previous occasion, taking into account the result up to and including occasion h-1 and where suffices indicate the occasion, single dashes units common to occasion h and h-1, the mean on earlier occasion being distinguished by square brackets and the double dashes units occuring on occasion h only. The value of ϕ_h depends on r, the fraction μ replaced on each occasion and h. With increasing h, ϕ_h rapidly tends to a limiting value which depends only on r and μ

Patterson (1950) approached the problem of successive sampling in a different and slightly more general way. He built an estimate as a linear function of a set of variates and developed a set of conditions for that estimate to be the most efficient. Using these conditions, he determined an efficient estimate of the mean on the hth occasion which is same as that given by Yates. He also gave a recurrence relation between ϕ_h and ϕ_{h-1} as

$$(1 - \phi_h)(1 - \phi_{h-1}) - (\alpha + \beta)(1 - \phi_h) + \alpha\beta = 0$$

where α, β are the roots of the quadratic equation obtained by putting $\phi_h = \phi_{h-1} = \phi$. He also found that with increasing h, $1-\phi_h$ tends to numerically smallest root of the quadratic

$$\phi^2 r^2 \lambda + \phi(1-r^2) - \mu(1-r^2) = 0 .$$

Thus he obtained the limiting value of ϕ as

$$\phi = \frac{-(1-r^2) + \sqrt{(1-r^2)\{1-r^2(1-4\mu\lambda)\}}}{2\lambda r^2}, \text{ where}$$

$$\lambda = 1 - \mu.$$

Patterson also gave efficient estimate of the difference between the mean on occasion h and that on occasion h-1. He also considered the case when sample size varies from occasion to occasion.

Tikkiwal (1953) was still more general in his approach. He allowed the correlation between units taken on two successive occasions to vary but assumed that correlation between units two or more than two occasion apart were equal to the product of correlations between units on all pairs of the consecutive occasions, formed by these. If all correlations were assumed to be equal on all occasions, he proved that with limiting Ø, the limiting value of replacement to be effected on different occasions is 50% from above, i.e. under the conditions imposed the replacement fraction is always greater than ½.

D. Singh (1959) investigated the problem of replacement, when the design is multistage. This aspect of the problem, apart from statistical consideration, had practical advantages, since inactual field, frequently the design is multistage and when the character under observation changes with season, it becomes necessary that survey should be repeated over the seasons. He gave expressions for the estimate of mean on second occasion and its variance, when a two-stage design is repeated on two occasions with partial replacement of first stage units only.

Khaturia (1959) extended the case of two-stage sampling repeated on two occasions with partial replacement of first-stage units, to h occasions. He also considered the case of sampling on two occasions with replacement among second stage units also. He investigated the problem of optimum allocation for a given cost function.

In the present study an attempt has been made to find the best unbiased linear estimate for the mean on second occasion and its variance when a double sampling is repeated on two occasions with partial replacement of units on second occasion. Under certain limitations approximate solutions for optimum allocation has also been determined. The double sampling for stratification has been considered both with respect to single as well two-stages. In particular, the case of only two strata of which one contains only zero-elements, has been considered in greater detail.

The results obtained may be applicable to many a surveys that may be conducted in due course. At present one such survey for estimating acreage under coco-nut, number of coco-nut trees and total yield, is being conducted in Assam. This survey is considered in chapter VI.

Lastly the case when sampling with varying probabilities of selection is repeated on two occasions, with partial replacement of units on second occasion has also been considered and is given in the last chapter.

:CHAPTER II:

Double Sampling for Regression Estimates on two occasions.

2.1 Introduction. A number of sampling techniques depend upon the possession of advance information about an auxiliary variate x_1. Ratio and Regression estimates require a knowledge of the population mean $\bar{X}$. When information about $\bar{X}$ is lacking, it is sometimes relatively cheap to take a large preliminary sample in which x_1 alone is measured. The purpose of this sample is to furnish a good estimate of $\bar{X}$ In a survey on a single occasion, it may pay to devote a part of the resources to this preliminary sample, although this means that the size of the sample in the main survey on y_1 must be decreased. This technique is known as double-sampling or two-phase sampling, and is profitable only if the gain in precision from ratio or regression estimates more than offsets the loss in precision due to reduction in the size of the main sample. If in a particular survey it is found that the double sampling for Regression estimate results in more precision on any single occasion, it may be reasonable to assume that if the same basic design is repeated on two or more occasions with partial replacement of units, it will furnish a better estimate for the mean on last occasion than the corresponding estimate furnished by simple random sampling repeated on same number of occasions. In the present chapter, we propose to obtain the best unbiased linear estimate for the mean on second occasion and its variance when a double sampling is repeated on two occasions with partial replacement of units on second occasion.

Suppose that a population has N sampling units and from each of the units of the population two variates, x_i and y_i can be measured. We are intrested in estimating the mean for y_i on second occasion. On the first occasion a sample of size kn, $(k > 1)$ is selected out of N units and only variate x_i is measured on them. A sub-sample of size n is selected from these kn units and variate y_i is also measured on them. A further sub-sample of size np $(p < 1)$ is selected from these n units of y-sample and is retained for second occasion as a part of both x-sample and y-sample. An independent sample of size (k-p)n is selected from N units and supplemented to np units retained from first occasion to complete the preliminary large sample of size kn on second occasion for measuring variate x_i only. From this preliminary large sample a further sub-sample of size n(1-p) is independently chosen and supplemented two the np units retained from the first occasion to complete the y-sample of size n for the second occasion. For sake of simplicity it is assumed that sampling is done with replacement at each stage.

In this way we obtain four sets of means as follows:-

(i) $\bar{\bar{y}}'_{np}, \bar{\bar{y}}'_{nq}$

(ii) $\bar{\bar{y}}_{np}, \bar{\bar{y}}_{nq}$

(iii) $\bar{\bar{x}}'_{np}, \bar{\bar{x}}'_{nq}, \bar{\bar{x}}'_{(k-1)n}$

(iv) $\bar{\bar{x}}_{np}, \bar{\bar{x}}_{nq}, \bar{\bar{x}}_{(k-1)n}$

Where the dash denotes second occasion and suffices denote the sample sizes on which the respective means are based.

The expected values of the means in these four sets are respectively $\bar{\bar{Y}}'$, $\bar{\bar{Y}}$, $\bar{\bar{X}}'$ and $\bar{\bar{X}}$. We want to estimate $\bar{\bar{Y}}'$. To this end in view we form a linear combination of these ten means as

$$a_1\bar{\bar{y}}'_{np} + a_2\bar{\bar{y}}'_{nq} + a_3\bar{\bar{y}}_{np} + a_4\bar{\bar{y}}_{nq} + a_5\bar{\bar{x}}'_{np} + a_6\bar{\bar{x}}'_{nq} + a_7\bar{\bar{x}}'_{n(k-1)}$$

$$+ a_8\bar{\bar{x}}_{np} + a_9\bar{\bar{x}}_{nq} + a_{10}\bar{\bar{x}}_{n(k-1)}$$

If this linear combination is to be an unbiased estimate for $\bar{\bar{Y}}'$ we must have

$$a_1+a_2 = 1$$

$$a_3+a_4 = 0$$

$$a_5+a_6+a_7 = 0$$

$$a_8+a_9+a_{10} = 0$$

Utilising these conditions, the linear combination can be can be written as

$$l_1 \bar{\bar{y}}'_{np} + l_2 \bar{\bar{y}}'_{nq} + l_3 (\bar{\bar{y}}_{np} - \bar{\bar{y}}_{nq}) + l_4 (\bar{\bar{x}}_{np} - \bar{\bar{x}}_{n(k-1)}) +$$

$$l_5 (\bar{\bar{x}}_{nq} - \bar{\bar{x}}_{n(k-1)}) + l_6 (\bar{\bar{x}}'_{np} - \bar{\bar{x}}'_{n(k-1)}) + l_7 (\bar{\bar{x}}'_{nq} - \bar{\bar{x}}'_{n(k-1)}) \qquad (2.1)$$

where $l_1 + l_2 = 1$ $\qquad (2.2)$

The variance for this linear estimation is

$$\begin{bmatrix} l_1, l_2, l_3, l_4, l_5, l_6, l_7 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} & V_{13} & V_{14} & V_{15} & V_{16} & V_{17} \\ V_{21} & V_{22} & V_{23} & V_{24} & V_{25} & V_{26} & V_{27} \\ V_{31} & V_{32} & V_{33} & V_{34} & V_{35} & V_{36} & V_{37} \\ V_{41} & V_{42} & V_{43} & V_{44} & V_{45} & V_{46} & V_{47} \\ V_{51} & V_{52} & V_{53} & V_{54} & V_{55} & V_{56} & V_{57} \\ V_{61} & V_{62} & V_{63} & V_{64} & V_{65} & V_{66} & V_{67} \\ V_{71} & V_{72} & V_{73} & V_{74} & V_{75} & V_{76} & V_{77} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \\ l_7 \end{bmatrix} \qquad (2.3)$$

where $V_{ij} = V_{ji}$ = covariance between the two expressions in (2. whose coefficients are l_i and l_j respectively.

The above expression can be written in the simple form

$L \, A L'$

where L is the row vector $(l_1, l_2, l_3, l_4, l_5, l_6, l_7)$ and L' is the transpose of L and A is the variance-covariance matrix (Vij)

For obtaining the best linear unbiased estimate for $\bar{Y}'$, the population mean for the variate y_i on the second occasion this variance should be minimised with respect to l_i's and subject to the condition $l_1+l_2 = 1$

Consider the expression

$$\phi = L A L' - 2\lambda(L E' - 1) \qquad (2.4)$$

where E is the row-matrix (1,1,0,0,0,0,0), and λ is the undetermined multiplier.

For minimum variance, we have

$$\frac{\phi}{L} = 2AL' - 2\lambda E' = 0$$

$$\text{i.e. } AL' = \lambda E' \qquad (2.5)$$

$$\text{Now let } P = \frac{L}{\lambda} \qquad (2.6)$$

then $AP' = E'$ where $P = (p_1,p_2,p_3,p_4,p_5,p_6,p_7)$

and $p_i = \frac{l_i}{\lambda}$, $i = 1,2,.....,7$.

$$\text{Hen-e or } P' = A^{-1} E' \qquad (2.7)$$

Hence $L' = \lambda A^{-1} E'$

But $l_1 + l_2 = 1$, and therefore $p_1 + p_2 = \frac{1}{\lambda}$

$$\text{or } \lambda = \frac{1}{p_1 + p_2} \qquad (2.8)$$

Thus the best linear unbiased estimate for $\bar{Y}'$ is the linear combination (2.1) where $l_1, l_2, ..., l_7$ are given by

$$L' = \frac{1}{p_1 + p_2} A^{-1} E'$$

p_1 and p_2 being the first two elemnt of row vetor P which itself is given by

$$p' = A^{-1} E'$$

The minimum variance is

$$L\,A L' = \lambda L\,E' = \lambda^2 E\,A^{-1} E'$$

$$= \frac{1}{(p_1+p_2)^2}\,E\,A^{-1} E' \qquad (2.9)$$

Now it remains only to obtain the expressions for V_{ij}'s.

For this we may assume that

(i) $s_x^2 = s_{x'}^2$ (2.10)

(ii) $s_y^2 = s_{y'}^2$ (2.11)

(iii) $\rho_{xy} = \rho_{xy'} = \rho_{x'y} = \rho_{x'y'} = \rho$ (2.12)

(iv) Sampling is done with replacement at each stage on both occasions.

where $s_x^2 = (1/N)\sum_{i=1}^{N}(x_i-\bar{X})^2$, $S_{x'}^2 = \sum_{i=1}^{N}(x_i' - \bar{X}')^2/N$

$S_y^2 = \sum_{i=1}^{N}(y_i-\bar{Y})^2/N$, $S_{y'}^2 = \sum_{i=1}^{N}(y_i' - \bar{Y}')^2/N$ (2.13)

and $\rho_{xy} = \dfrac{S_{xy}}{S_x\;S_y}$, $\rho_{x'y} = \dfrac{S_{x'y}}{S_{x'}\;S_y}$, $\rho_{xy'} = \dfrac{S_{xy'}}{S_x\;S_{y'}}$,

$$\rho_{x'y'} = \frac{S_{x'y'}}{S_{x'}\;S_{y'}}$$

where $S_{xy} = \sum_{i=1}^{N}(x_i-\bar{X})\;(y_i-\bar{Y})/N$ (2.14)

$S_{x'y} = \sum_{i=1}^{N}(x_i'-\bar{X}')\;(y_i-\bar{Y})/N$

$$s_{xy'} = \sum_{i=1}^{N} (x_i - \bar{X})\,(y_i' - \bar{Y}')/N$$

and $$s_{x'y'} = \sum_{i=1}^{N} (x_i' - \bar{X}')\,(y_i' - \bar{Y}')/N$$

Thenthe variance covariance matrix is given by (2.16)

where $$\rho_x = \frac{s_{xx'}}{s_x\ s_{x'}}$$

and $$\rho_y = \frac{s_{yy'}}{s_y\ s_{y'}}$$, $s_{xx'}$ and $s_{yy'}$ being given by

$$s_{xx'} = \sum_{i=1}^{N} (x_i - \bar{X})\,(x_i' - \bar{X}')/N \qquad (2.15)$$

$$s_{yy'} = \sum_{i=1}^{N} (y_i - \bar{Y})\,(y_i' - \bar{Y}')/N$$

$$\begin{bmatrix} \frac{1}{np}s_y^2 & 0 & \frac{\rho_y s_y^2}{np} & \frac{\rho s_x s_y}{np} & 0 & \frac{\rho s_x s_y}{np} & 0 \\ 0 & \frac{s_y^2}{nq} & 0 & 0 & 0 & 0 & \frac{\rho s_x s_y}{nq} \\ \frac{\rho_y s_y^2}{np} & 0 & \frac{s_y^2}{npq} & \frac{\rho s_x s_y}{np} & -\frac{\rho s_x s_y}{nq} & \frac{\rho s_x s_y}{np} & -\frac{\rho s_x s_y}{nq} \\ \frac{\rho s_x s_y}{np} & 0 & \frac{\rho s_x s_y}{np} & \left(\frac{1}{p}+\frac{1}{(k-1)}\right)\frac{s_x^2}{n} & \frac{s_x^2}{n(k-1)} & \left(\frac{1}{p}+\frac{1}{k-1}\right)\frac{\rho_x s_x^2}{n} & \frac{\rho_x s_x^2}{(k-1)n} \\ 0 & 0 & -\frac{\rho s_x s_y}{nq} & \frac{s_x^2}{n(k-1)} & \left(\frac{1}{q}+\frac{1}{k-1}\right)\frac{s_x^2}{n} & \frac{\rho_x s_x^2}{n(k-1)} & \left(\frac{1}{q}+\frac{1}{k-1}\right)\frac{\rho_x s_x^2}{n} \\ \frac{\rho s_x s_y}{np} & 0 & \frac{\rho s_x s_y}{np} & \left(\frac{1}{p}+\frac{1}{k-1}\right)\frac{\rho_x s_x^2}{n} & \frac{\rho_x s_x^2}{n(k-1)} & \left(\frac{1}{p}+\frac{1}{(k-1)}\right)\frac{s_x^2}{n} & \frac{s_x^2}{(k-1)n} \\ 0 & \rho\frac{s_x s_y}{nq} & \frac{\rho s_x s_y}{nq} & \frac{\rho_x s_x^2}{(k-1)n} & \left(\frac{1}{q}+\frac{1}{k-1}\right)\frac{\rho_x s_x^2}{n} & \frac{s_x^2}{(k-1)n} & \left(\frac{1}{q}+\frac{1}{k-1}\right)\frac{s_x^2}{n} \end{bmatrix} = ((v_{ii'})) \tag{2.16}$$

CHAPTER III

Double sampling for stratification on two occasions.

3.1 Introduction. In applying the technique of stratified random sampling, a question which invariably arises is this. How to divide the population which is going to be sampled into various strata. Usually the stratification is done with the help of the frequency distribution of an auxiliary variate x_1. When this information about the frequency distribution of the auxiliary variable is lacking, it may be useful to take a preliminary large sample in which x_1 alone is measured with a view to obtain a fairly reliable and accurate estimate of the frequency distribution of x_1. From this preliminary sample smaller sub-samples can be independently taken in different strata and utilizing the estimates of strata sizes provided by the preliminary sample, we can build an unbiased estimate for the population mean. This technique which is known as double sampling for stratification is useful only if it results in gain in efficiency over simple random sampling, subject to same amount of total cost in both the cases.

In the present chapter we will consider the theory of double sampling for stratification for both single-stage and two-stage sampling repeated on two occasions with partial replacement of units (first-stage units in case of two-stage random sampling) on the second occasion.

3.2 A particular case of double sampling. First we shall consider a very restrictive and particular case of the double sampling for stratification. We assume that the population is divided into two strata, one consists of units on which y_i the variate under consideration assumes values other than zero, while the other consists of only such units for which $y_i = 0$.

Let there be a univariate finite population P with y_i as the variate. Let there be N units in the population. A certain number of Units are known to have $y_i = 0$. The remaining units have values $y_i \neq 0$. The number of zero units is not known. Let this number be N_1 and put $N-N_1 = N_2$. We want to estimate the population mean

$$\bar{\bar{Y}} = \sum_{i=1}^{N} y_i/N = \left(\sum_{i=1}^{N_1} y_i + N_2 . 0\right)/N$$
$$= N_1 \bar{\bar{Y}}_1/N = p_1 \bar{\bar{Y}}_1 \qquad (3.1)$$

where $\bar{\bar{Y}}_1$ is the mean in the non-zero stratum and $p_1 = \frac{N_1}{N}$

To estimate this an estimator of either $\bar{\bar{Y}}$ are $\bar{Y}_1$ will be sufficient. Now two sampling procedures are open to us.

(1) Take a simple random sample of size n from the whole population P consisting of N units.

The estimate for $\bar{\bar{Y}}$ is $\frac{1}{n} \sum^{n} y_i = \bar{\bar{y}}_n$

and its variance is $V(\bar{\bar{y}}_n) = \frac{S^2}{n}$

where $S^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{\bar{Y}})^2$, the sampling being done <u>with replacement.</u>

As is evident some of these y_i's will be zero and their inclusion in the sample tends to increase the sampling error of the estimate.

If with varying expenditures of efforts these units having zero values are found and listed, so that they need not be sampled then

$$S_1^2 = \frac{S^2}{P_1} - \frac{1-P_1}{P_1^2}\bar{Y}^2 \qquad (3.2)$$

where S_1^2 is the variance when all zero units are excluded; If the population is estimated from a simple random sample of size n, it can be shown that with the exclusion of zero units the fractional reduction in the variance of the estimate is

$$\frac{(1-P_1)\,(V^2+1)}{V^2} \qquad (3.3)$$

where $V^2 = \frac{S^2}{\bar{Y}^2}$ is the coefficient of variation in the original population." (Cochran's sampling Tec-hniques page 80)

(ii) However if we cannot find and list the zero-units we can resort to the method of double sampling. We select a preliminary large sample of size n' from the original population P. Suppose that out of n', n_1 units have values different from zero and $n_2 = n'-n_1$ the value zero, for the variate y_i. From n_1 non-zero units, a further sub-sample of size $R(n_1)$ is selected where $R(n_1)$ is a random variable defined as follows.

$R(n_1) = n$ if $n_1 \geqslant n$

$R(n_1) = n_1$ if $n_1 < n$

(3.4)

where n is a fixed positive integer less than n'; n_1 varies from 0 to n' and $R(n_1)$ from 0 to n.

The estimate

$\frac{n_1}{n'}\bar{y}_{R(n_1)}$ gives an unbiased estimate for mean $\bar{Y}$ of population P, where

$$\bar{y}_{R(n_1)} = \left(\sum_{i=1}^{R(n_1)} y_i \right) / R(n_1) \qquad (3.5)$$

and has the conditional variance

$$V\left[\frac{n_1}{n'}\bar{y}_{R(n_1)} \Big/ R(n_1)\right] = \frac{p_1(1-p_1)}{n'}\bar{Y}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{n'}\right]\frac{s_1^2}{R(n_1)} \qquad (3.6)$$

It is worth noting that if $(1-p_1)$ N units have same constant value c for their y-variate instead of zero, then the above formula for conditional variance takes the form

$$\frac{s_{y_1}^2}{R(n_1)}\left[p_1^2 + \frac{p_1(1-p_1)}{n'}\right] + \frac{p_1(1-p_1)}{n'}(\bar{Y}_1 - c)^2. \qquad (3.7)$$

Now $$V\left(\frac{n_1}{n'}\bar{y}_{R(n_1)}\right) = V\left[E\left\{\frac{n_1}{n'}\bar{y}_{R(n_1)} \Big/ R(n_1)\right\}\right] + E\left[V\left\{\frac{n_1}{n'}\bar{y}_{R(n_1)} \Big/ R(n_1)\right\}\right]$$

$$= V(\bar{Y}) + E\left\{V\left[\frac{n_1}{n'}\bar{y}_{R(n_1)} \Big/ R(n_1)\right]\right\}$$

$$= \frac{p_1(1-p_1)}{n'}\bar{Y}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{n'}\right] s_1^2 \, E\frac{1}{R(n_1)} \qquad (3.8)$$

since $V(\bar{Y}) = 0$

Now let

$R(n_1) = n - \epsilon$ where ϵ can take any value from 0 to n.

Then

$$E\frac{1}{R(n_1)} = E\frac{1}{n-\epsilon}$$

$$= \frac{1}{n} E\left(1 - \frac{\epsilon}{n}\right)^{-1}$$

$$= \frac{1}{n} E\left(1 + \frac{\epsilon}{n} + \frac{\epsilon^2}{n^2} + \frac{\epsilon^3}{n^3} + \ldots\ldots\ldots\ldots\right)$$

The expansion being justified, since $\frac{\epsilon}{n} < 1$.

$$= \frac{1}{n}\left[1 + \frac{n - E\,R(n_1)}{n} + \frac{V\,R(n_1)}{n^2}\right] \qquad (3.9)$$

to a second degree of approximation.

Again,

$$E\,R(n_1) = \sum_{r=0}^{n'} r \times \text{Prob}\left[R(n_1) = r\right]$$

$$= \sum_{r=0}^{n-1} r\binom{n'}{r} p_1^r (1-p_1)^{n'-r} + \sum_{r=n}^{n'} n\binom{n'}{r} p_1^r (1-p_1)^{n'-r} \qquad (3.10)$$

and

$$E\,R^2(n_1) = \sum_{r=0}^{n-1} r^2\binom{n'}{r} p_1^r (1-p_1)^{n'-r} + n^2\sum_{r=n}^{n'}\binom{n'}{r} p_1^r (1-p_1)^{n'-r} \qquad (3.11)$$

while

$$V\,R(n_1) = E\,R^2(n_1) - E^2\,R(n_1)$$

If n' is very large compared to n then $R(n_1) = n$ with a probability which is very close to 1. As n' increases this probability converges to 1, and $n - E\,R(n_1)$ and $V\,R(n_1)$ both tend to zero. In any case $n - E\,R(n_1)$ and $V\,R(n_1)$ will be normally so small that $\frac{n - E\,R(n_1)}{n^2}$ and $\frac{V\,R(n_1)}{n^3}$ can be regarded as negligible as compared to $\frac{1}{n}$ which is the lower bound of $E\,\frac{1}{R(n_1)}$.

Thus
$$V\left[\frac{n_1}{n'}\,\bar{\bar{y}}_{R(n_1)}\right] = \frac{p_1(1-p_1)}{n'}\,\bar{\bar{Y}}_1^2 + \left[p_1^2 + \frac{p_1^2(1-p_1)}{n'}\right] \times$$
$$\frac{s_1^2}{n}\left[1 + \frac{n - E\,R(n_1)}{n} + \frac{V\,R(n_1)}{n^2} + \text{terms of order } \frac{1}{n^3}\right]$$
$$\doteq \frac{p_1(1-p_1)}{n'}\,\bar{\bar{Y}}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{n'}\right]\frac{s_1^2}{n} \qquad (3.12)$$

It is worth mentioning here that in most of the text-books on sample surveys the formula (3.12) is given as an exact expression for the variance of double sampling estimate. The procedure described in these books is to select a sub-sample of size n out of n_1 units falling into first stratum i.e. non-zero stratum, in the preliminary sample n' for estimating the strata sizes. However, since n is bounded above by the random variable n_1, $(0 \leq n_1 \leq n')$, it can be never treated as a constant fixed in advance. Even if we choose n as a very small positive integer, there ~~always~~ may exist ~~plenty~~ Some of preliminary samples which contain units from first stratum whose total number is less than n. Obviously we can not select n units from these samples. The method adopted here by defining $R(n_1)$ as in (3.4) seems to be the only alternative. To avoid unnecessary repitions, we will hence-forth use the phrase " A sub-sample of size n is taken from these n_1 units; but" it should be interpreted as " A sub-sample size $R(n_1)$ is taken from these n_1 units, where $R(n_1) = n$ if $n_1 \geq n$ and $R(n_1) = n_1$ if $n_1 < n$. "

3.3 Optimum allocation in the above double sampling. Since $\frac{p_1(1-p_1)}{n'}$ is very small in comparison with p_1^2, the variance formula (3.12) can be further approximated by

$$\frac{p_1^2 \; s_1^2}{n} + \frac{p_1(1-p_1)\; \bar{Y}_1^2}{n'} \qquad (3.13)$$

If c_1 is the cost per unit of measuring the variate y_i and c_2 the cost per unit for determing whether $y_i = 0$ or $y_i \neq 0$, then a suitable cost function for this particular case of double sampling is

$$C = c_1 n + c_2 n' \qquad (3.14)$$

where C is the total cost.

Minimising the variance in (3.13) with respect to n' and n for the cost function (3.14), it is easily seen that

$$\frac{n}{s_1\sqrt{p_1\, c_2}} = \frac{n'}{\bar{Y}_1\{(1-p_1)\ c_1\}^{1/2}}$$

This relation and (3.14) give the value for n' and n which correspond to the optimum allocation. The expression for the minimum variance is

$$V_{opt} = \frac{p_1}{C}\left(s_1\sqrt{p_1\, c_1} + \bar{Y}_1\sqrt{(1-p_1)\, c_2}\,\right)^2 \qquad (3.15)$$

The variance when only a simple random sample is ~~taken~~ is

$$\frac{S^2}{n}$$

The cost function in this case is

$$C = c_1\, n_o + c_2\, n \qquad (3.16)$$

where n_o is the number of non-zero units in the sample of size n, while c_1 and c_2 are as before. Taking the expectation of this cost function we have

$$E\ C = (p_1c_1+c_2)\ n.$$

The optimum variance in this case is evidently $\frac{S^2}{C}\,(p_1c_1+c_2)$

Hence the double sampling will be of any advantage if

$$\frac{S^2}{C}(p_1c_1+c_2) > \frac{p_1}{C}\left(s_1\sqrt{p_1\, c_1} + \bar{Y}_1\sqrt{(1-p_1)\, c_2}\right)^2$$

but $S^2 = p_1\, s_1^2 + p_1\,(1-p_1)\,\bar{Y}_1^2$

when this relation is used, the above inequality reduces to

$$\bar{Y}_1^2\left[\frac{s_1}{\bar{Y}_1} - \sqrt{\frac{p_1(1-p_1)\ c_2}{c_1}}\right]^2 > 0 \qquad (3.17)$$

which is always true, unless $\sqrt{\frac{p_1\,(1-p_1)\ c_2}{c_1}} = \frac{s_1}{\bar{Y}_1}$

the coefficient of variation in the non-zero stratum, in which case the simple random sampling and the double sampling provide equally precise estimates.

<u>3.4 Double sampling on two occasions.</u> Suppose that the sampling scheme on the first occasion is the same as described in (ii) of section 3.2. On the second occasion, retain np units from the original sample of n units and supplement it with nq, (q = 1 - p) independent units selected from the n_1' non - zero units of the preliminary large sample n' taken afresh for the second occasion. For the sake of simplicity we assume that the sampling is done with replacement on both the occasions.

Let
$$\bar{y}_{np} = \frac{1}{np}\sum^{np} y_i$$

= mean in the first stratum for first occasion based on np units which are common to both the occasions.

$$\bar{y}_{nq} = \frac{1}{nq}\sum^{nq} y_i$$

= mean in the first stratum for first occasion based on nq units which are in the sample for the first occasion only.

$$\bar{y}'_{np} = \frac{1}{np}\sum^{np} y_i'$$

= mean in the first stratum for second occasions based on np units which are common to both the occasions.

$$\bar{y}'_{nq} = \frac{1}{nq}\sum^{nq} y_i'$$

= mean in the first stratum for second occasion based on nq units which are independently selected on second occasion.

Now

$\frac{n_1 + n_1'}{2n'} \bar{y}_{np}$ and $\frac{n_1 + n_1'}{2n'} \bar{y}_{nq}$ are unbiased estimates of $\bar{Y}$ the population mean for first occasion, while,

$\frac{n_1 + n_1'}{2n'} \bar{y}'_{np}$ and $\frac{n_1 + n_1'}{2n_1'} \bar{y}'_{nq}$ are unbiased estimates of $\bar{Y}'$ the population mean on second occasion.

We wish to estimate $\bar{Y}'$ by a linear combination of these four estimates;

$$\frac{n_1 + n_1'}{2n'} \left[a\,\bar{y}_{np} + b\,\bar{y}_{nq} + c\,\bar{y}'_{np} + d\,\bar{y}'_{nq} \right] \qquad (3.18)$$

where n_1' is the number of non-zero elements in the preliminary large sample of size n' for second occasion, and $q = 1-p$. The expected value of this linear estimate is

$$(a+b)\,\bar{Y} + (c+d)\,\bar{Y}'$$

If this linear estimate is to be an unbiased estimate for $\bar{Y}'$, we must have

$$a + b = 0 \quad \text{and} \quad c + d = 1$$

Thus, the linear estimate becomes

$$\frac{n_1 + n_1'}{2n'} \left[a\,(\bar{y}_{np} - \bar{y}_{nq}) + c\,\bar{y}'_{np} + (1-c)\,\bar{y}'_{nq} \right] \qquad (3.19)$$

and the variance of this estimate is

$$a^2 (V_{11} + V_{22}) + c^2 V_{33} + (1-c)^2 V_{44} + 2ac\, V_{13}$$

where

$$V_{11} = Var\left(\frac{n_1 + n_1'}{2n'}\ \bar{y}_{np}\right)$$

$$V_{22} = Var\left(\frac{n_1 + n_1'}{2n'}\ \bar{y}_{nq}\right)$$

$$V_{33} = Var\left(\frac{n_1 + n_1'}{2n'}\ \bar{y}'_{np}\right)$$

$$V_{44} = Var\left(\frac{n_1 + n_1'}{2n'}\ \bar{y}'_{nq}\right)$$

and $$V_{13} = Cov\left(\frac{n_1 + n_1'}{2n'}\ \bar{y}_{np}\ ,\ \frac{n_1 + n_1'}{2n'}\ \bar{y}'_{np}\right)$$

Now we want to chose the constants a and c in such a way that variance of this linear unbiased estimate becomes minimum. Differenciating the expression for the variance of this estimate with respect to a and c and equating them to zero, we have

$$2a\,(V_{11} + V_{22}) + 2c\,V_{13} = 0$$

$$2c\,V_{13} + 2(1-c)\,(-1)\,V_{44} + 2a\,V_{13} = 0$$

These two equations give

$$a = -\frac{V_{13}\,V_{44}}{(V_{11} + V_{22})\,(V_{33} + V_{44}) - V_{13}^2}$$

and $$c = \frac{V_{44}\,(V_{11} + V_{22})}{(V_{11} + V_{22})\,(V_{33} + V_{44}) - V_{13}^2}$$

(3.21)

Thus the estimate for the population mean on the second occasion is

$$\frac{(n_1 + n_1')\ /\ 2n'}{(V_{11} + V_{22})\,(V_{33} + V_{44}) - V_{13}^2}\left[\ V_{13}\,V_{44}(\bar{y}_{nq} - \bar{y}_{np}) + V_{44}\,(V_{11} + V_{22})\,\bar{y}'_{np}\left\{V_{33}\,(V_{11} + V_{22}) - V_{13}^2\right\}\bar{y}'_{nq}\right]$$

(3.22)

and the variance for this expression is

$$\frac{V_{44}\left[V_{33}\,(V_{11}+V_{22})-V_{13}^{2}\right]}{(V_{11}+V_{22})\,(V_{33}+V_{44})-V_{13}^{2}} \qquad (3.23)$$

where

$$V_{11} = \frac{p_1(1-p_1)}{2n'}\,\bar{Y}_1^{2} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] s_1^{2}/np$$

$$V_{22} = \frac{p_1(1-p_1)}{2n'}\,\bar{Y}_1^{2} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] s_1^{2}/nq$$

$$V_{33} = \frac{p_1(1-p_1)}{2n'}\,\bar{Y}_1'^{2} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] s_1^{2}/\,np$$

$$V_{44} = \frac{p_1(1-p_1)}{2n'}\,\bar{Y}_1'^{2} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] s_1'^{2}/nq$$

$$V_{13} = \frac{p_1(1-p_1)}{2n'}\,\rho\,\bar{X}_1\,\bar{Y}_1' + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right]\rho\, s_1\, s_1'/np$$

(3.24)

Now if we assume that

$$s_1/\bar{Y}_1 = s_1'/\bar{Y}_1' = C_{v1} \qquad (3.25)$$

i.e. coefficent of variation in the non-zero stratum is same on both occasion, then

$$\frac{V_{11}}{\bar{Y}_1^{2}} = \frac{V_{33}}{\bar{Y}_1'^{2}} = \frac{p_1(1-p_1)}{2n'} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] C_{v1}^{2}/\,np = A$$

$$\frac{V_{22}}{\bar{Y}_1^{2}} = \frac{V_{44}}{\bar{Y}_1'^{2}} = \frac{p_1(1-p_1)}{2n'} + \left[p_1^{2} + \frac{p_1(1-p_1)}{2n'}\right] C_{v1}^{2}/\,nq = B$$

$$\text{and } \frac{V_{12}}{\bar{\bar{Y}}_1 \, Y'_1} = \frac{p_1(1-p_1)}{2n'} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right] \rho C_{v1}^2/np = C$$

(3.26)

$$\text{Then } V_{opt} = \bar{\bar{Y}}_1^2 \; B \quad \frac{A(A+B) - C^2}{(A+B)^2 - C^2}$$

(3.27)

<u>3.5. Generalisation to k strata.</u> Now let us generalise the particular type of double sampling repeated on two occasions, considered above to the general case, of double sampling for stratification, repeated on two occasions. Let there be k strata with strata sizes $N_1, N_2, \ldots\ldots\ldots\ldots, N_k$, where

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

The population mean on the first occasion is

$$\bar{\bar{Y}} = \sum_{i=1}^{K} p_i \bar{\bar{Y}}_i \text{ where } p_i = \frac{N_i}{N}$$

and $\bar{\bar{Y}}_i = \dfrac{1}{N_i} \displaystyle\sum_{j=1}^{N_i} y_{ij}$ is the mean in the i^{th} stratum.

As an estimate of this we can use

$$\bar{\bar{y}}_{st} = \sum_{i=1}^{K} \hat{p}_i \; \bar{\bar{y}}_{i(n_i)} \quad \text{where } \hat{p}_i = \frac{n_i' + n_i''}{2n'}$$

n' being the size of the preliminary large sample and n_i' and n_i'' the number of units of this sample falling in the i^{th} stratum, on first and second occasion respectively, and n_i the size of the subsample for that stratum.

Then

$$V(\bar{\bar{y}}_{st}) = \sum_{i=1}^{K} \left[p_i^2 + \frac{p_i(1-p_i)}{2n'} \right] \frac{s_i^2}{n_i} + \frac{1}{2n'} \sum_{i=1}^{K} p_i(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$$

(3.28)

On the second occasion we retain a portion $t_i n_i$ in the i^{th} stratum from first occasion and supplement it with a fresh sample of size $(1 - t_i)\, n_i$ taken independently on second occasion. Then we have got the following four estimates.

$$(\bar{y}_{st})_1 = \sum_{i=1}^{k} \frac{n_i' + n_i''}{2n'}\, \bar{y}_{t_i n_i}$$

$$(\bar{y}_{st})_2 = \sum_{i=1}^{k} \frac{n_i' + n_i''}{2n'}\, \bar{y}_{(1-t_i)\, n_i}$$

$$(\bar{y}_{st})_3 = \sum_{i=1}^{k} \frac{n_i' + n_i''}{2n'}\, \bar{y}'_{t_i n_i} \qquad (3.29)$$

$$(\bar{y}_{st})_4 = \sum_{i=1}^{k} \frac{n_i' + n_i''}{2n'}\, \bar{y}'_{(1-t_i)\, n_i}$$

To obtain the estimate for the mean on second occasion, we form a linear combination of these four estimates.

$$a(\bar{y}_{st})_1 + b\,(\bar{y}_{st})_2 + c\,(\bar{y}'_{st})_1 + d\,(\bar{y}'_{st})_2$$

If this is to be an unbiased estimate of the population mean on the second occasion, i.e.

$$\bar{Y}' = \sum_{i=1}^{k} p_i\, \bar{Y}_i \text{ , then we must have}$$

$a + b = 0$, and $c + d = 1$

Hence the estimate becomes

$$a\left[(\bar{y}_{st})_1 - (\bar{y}_{st})_2\right] + c\,(\bar{y}'_{st})_1 + (1 - c)\,(\bar{y}'_{st})_2$$

$$(3.30)$$

and its variance is

$$a^2(V_{11} + V_{12}) + c^2 V_{33} + (1 - a)^2 V_{44} + 2ac\, V_{13}$$

(3.31)

where

$$V_{11} = \sum_{i=1}^{K} \left[\left\{ p_i^2 + \frac{p_i(1-p_i)}{2n'} \right\} \frac{s_i^2}{t_i n_i} + \frac{p_i(\dot{\bar{Y}}_i - \dot{\bar{Y}})^2}{2n'} \right]$$

$$V_{22} = \sum_{i=1}^{K} \left[\left\{ p_i^2 + \frac{p_i(1-p_i)}{2n'} \right\} \frac{s_i^2}{(1-t_i)n_i} + \frac{p_i(\dot{\bar{Y}}_i - \dot{\bar{Y}})^2}{2n'} \right]$$

$$V_{33} = \sum_{i=1}^{K} \left[\left\{ p_i^2 + \frac{p_i(1-p_i)}{2n'} \right\} \frac{s_i'^2}{t_i n_i} + \frac{p_i(\dot{\bar{Y}}_i' - \dot{\bar{Y}}')^2}{2n'} \right]$$

$$V_{44} = \sum_{i=1}^{K} \left[\left\{ p_i^2 + \frac{p_i(1-p_i)}{2n'} \right\} \frac{s_i'^2}{(1-t_i)\, n_i} + \frac{p_i(\dot{\bar{Y}}_i' - \dot{\bar{Y}}')^2}{2n'} \right]$$

and

$$V_{13} = \sum_{i=1}^{K} \left[\left[p_i^2 + \frac{p_i^2(1-p_i)}{2n'} \right] \frac{\rho_i\, s_i\, s_i'}{t_i n_i} + p_i \frac{(\dot{\bar{Y}}_i - \dot{\bar{Y}})(\dot{\bar{Y}}_i' - \dot{\bar{Y}}')}{2n'} \right]$$

(3.32)

where s_i^2 is the variance in the i th stratum on first occasion, $s_i'^2$ is the variance in the i th stratum on the second occasion, and ρ_i is the correlation coefficient between observations on first and second occasion of the units in i th stratum.

Minimising the variance (3.31) we have

$$a = \frac{-V_{13} V_{44}}{(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2}$$

and $$c = \frac{V_{44}(V_{11} + V_{22})}{(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2}$$

(3.33)

substituting these in the expressions for the linear unbiased estimate, we get the best linear unbased estimate. The minimum variance is given, as before by

$$\frac{V_{44}\left[V_{33}(V_{11}+V_{22}) - V_{13}^{2}\right]}{(V_{11}+V_{22})(V_{33}+V_{44}) - V_{13}^{2}} \qquad (3.34)$$

where

V_{ij}s are given in (3.32)

CHAPTER IV

Two-Phase Multistage Sampling On Two Occasions.

(4.1) Introduction. In the previous chapter we studied a particular case of double sampling repeated on two occasions, with partial replacement of units on second occasion. The double sampling considered, was applied to a population consisting of single stage units only. But frequently the population is divided into multistage units and its stratification according to the first stage units is considered essential. If the strata sizes are estimated from a preliminary large sample from which a sub - sample of first stage unit is selected in each stratum and from each first stage units of this sub - sample, a number of second stage units is selected and from each of these second stage units a number of third - stage units and so on, then this sampling technique may be called Two - Phase multistage sampling. In the present chapter we propose to obtain the expressions for the unbiased estimate for the population mean and its variance, when this type of two phase multistage sampling is repeated on two occasions with partial replacement of first - stage units on second occasion. The study will be restricted to two stage sampling only. It will be further assumed that all the first stage units can be divided in two strata;

(i) the one consisting of those first - stage units which contain at least one non - zero second stage - units, and

(ii) the other consisting of those first - stage units whose all second stage - units are zero - units. (j^{th} second stage unit in i^{th} first stage unit is a zero unit, if y_{ij}, the value of the variate on that unit is zero).
We shall further assume that replacement on second occasion is done only in first - stage units, that is, all the second - stage units sampled on the first occasion from the first stage units common to both the occasions are retained on second occasion also.

4.2 Two - phase two - stage sampling on two occasions: equal first stage units.

Let

N = number of first stage units in the population

N_1 = number of first stage units in the first stratum,

M = number of second stage units in each of the first stage units,

n' = size of preliminary large sample for estimating the proportion $p_1 = \frac{N_1}{N}$,

n_1 = number of first stage units belonging to the first stratum in the sample n', on the first occasion,

n_1' = number of first stage units belonging to the first stratum in the sample n', on the second occasion,

n = size of the sub - sample from n_1 units belonging to first stratum on first occasion,

n_2 = number of first stage units common to both occasions,

$nq = n(1-p)$ = number of first stage units from n_1' units belonging to the first stratum on the second occasion,

m = number of second stage units from each of the first stage units,

$\bar{\bar{y}}_1 = \frac{1}{nmp} \sum_{i}^{np}\sum_{j}^{m} y_{ij}$ = mean per second stage units in non-zero stratum for the first occasion for nmp units which are common to both occasion,

$\bar{\bar{y}}_2 = \frac{1}{nmq} \sum_{i}^{nq}\sum_{j}^{m} y_{ij}$ = mean per second stage units in non-zero stratum on the second occasion for the nmq units which are in the sample for first occasion only,

$\bar{\bar{y}}'_1 = \frac{1}{nmp} \sum_{i}^{np}\sum_{j}^{m} y_{ij}$ = mean per second stage units in non-zero stratum on the second occasion for the nmp units which are common to both occasions,

$\bar{\bar{y}}'_2 = \frac{1}{nmq} \sum_{i}^{nq}\sum_{j}^{m} y_{ij}$ = mean per second stage units in non-zero stratum on the second occasion for the nmq units which are independently selected for second occasion only.

Now

$$\frac{n_1 + n_1'}{2n'}\,\bar{\bar{y}}_1 \quad \text{and} \quad \frac{n_1 + n_1'}{2n'}\,\bar{\bar{y}}_2 \quad \text{are unbiased estimates}$$

for $\bar{\bar{Y}}$, the population mean on the first occasion,

while $\frac{n_1 + n_1'}{2n'}\,\bar{\bar{y}}'_1$ and $\frac{n_1 + n_1'}{2n'}\,\bar{\bar{y}}'_2$ are unbiased estimates for $\bar{\bar{Y}}'$, the population mean on the second occasion.

It is assumed here that all the first stage units belonging to the first and second stratum respectively on first occasion belong to the same stratum on second occasion also. We wish to estimate $\bar{\bar{Y}}'$ by a linear estimate of the form

$$\bar{\bar{y}}'_1 = (a\bar{\bar{y}}_1 + b\,\bar{\bar{y}}_2 + c\,\bar{\bar{y}}'_1 + d\,\bar{\bar{y}}'_2)\ \frac{n_1 + n'_1}{2n'} \tag{4.1}$$

But

$$E(\bar{\bar{y}}'_1) = (a + b)\,\bar{\bar{X}} + (c + d)\,\bar{\bar{Y}}'$$

If $\bar{\bar{y}}'_1$ is to be an unbiased estimate for $\bar{\bar{Y}}'$ we must have

$$a + b = 0 \qquad \text{and} \qquad c + d = 1. \tag{4.2}$$

Hence

$$\bar{\bar{y}}'_1 = \left[a(\bar{\bar{y}}_1 - \bar{\bar{y}}_2) + c\,\bar{\bar{y}}'_1 + (1-c)\,\bar{\bar{y}}'_2 \right] \frac{n_1 + n'_1}{2n'} \tag{4.3}$$

and

$$V(\bar{\bar{y}}'_E) = a^2 (V_{11} + V_{22}) + c^2 V_{33} + (1-c)^2 V_{44} + 2ac\,V_{13} \tag{4.4}$$

where

$$V_{11} = Var \left(\frac{n_1 + n'_1}{2n'}\,\bar{\bar{y}}_1 \right)$$

$$= \frac{p_1 (1-p_1)}{2n'}\,\bar{\bar{Y}}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \left(\frac{s_{1b}^2}{np} + \frac{\bar{S}_{1w}^2}{nmp} \right)$$

$$V_{22} = Var\,\frac{n_1 + n'_1}{2n'}\,\bar{\bar{y}}_2$$

$$= \frac{p_1(1-p_1)}{2n'}\,\bar{\bar{Y}}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \left(\frac{s_{2b}^2}{nq} + \frac{\bar{S}_{1w}^2}{nmq} \right)$$

$$V_{33} = \text{Var}\left(\frac{n_1 + n_1'}{2n'}\;\bar{\bar{y}}_1'\right)$$

$$= \frac{p_1(1-p_1)}{2n'}\;\bar{\bar{Y}}_1'^{2} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right]\left(\frac{S_{1b}'^{2}}{np} + \frac{\bar{S}_{1w}'^{2}}{nmp}\right)$$

$$V_{44} = \text{Var}\left(\frac{n_1 + n_1'}{2n'}\;\bar{\bar{y}}_2'\right)$$

$$= \frac{p_1(1-p_1)}{2n'}\;\bar{\bar{Y}}_1'^{2} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right]\left(\frac{S_{1b}'^{2}}{nq} + \frac{\bar{S}_{1w}'^{2}}{nmq}\right)$$

and

$$V_{13} = \text{Cov}\left[\frac{n_1 + n_1'}{2n'}\,\bar{\bar{y}}_1,\quad \frac{n_1+n_1'}{2n'}\;\bar{\bar{y}}_1'\right]$$

$$= \frac{p_1(1-p_1)}{2n'}\,\bar{\bar{Y}}_1\,\bar{\bar{Y}}_1' + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right] \times$$

$$\left[\frac{\rho_{1b}\;S_{1b}\;S_{1b}'}{np} + \frac{\rho_{1w}\;\bar{S}_{1w}\;\bar{S}_{1w}'}{nmp}\right] \qquad (4.6)$$

where

$$S_{1b}^2 = \sum_{i=1}^{N_1} \frac{(\bar{\bar{y}}_i - \bar{\bar{Y}}_1)^2}{N_1}$$

= Variance between first stage units on first occasion in first stratum.

$$S_{1b}'^{2} = \sum_{i=1}^{N_1} \frac{(\bar{\bar{y}}_i' - \bar{\bar{Y}}_1')^2}{N_1}$$

= Variance between first stage units on second occasion in the first stratum.

and $s_{1w}^2 = \frac{1}{N_1} \sum_{i=1}^{N_1} s_i^2$ and $s'^2_{2w} = \frac{1}{N_1} \sum_{i=1}^{N_1} s'^2_i$

where $s_i^2 = \frac{1}{M} \sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2$ for $i = 1, 2, \ldots, N_1$

= Variance within i th first stage unit on first occasion

and $s'^2_i = \frac{1}{M} \sum_{j=1}^{M} (y'_{ij} - \bar{y}'_i)^2$ for $i = 1, 2, \ldots N_1$

= Variance withing i th first stage unit on second occasion.

and

$$\rho_{2b} = \frac{\sum_{i=1}^{N_1} (\bar{y}_i - \bar{Y}_1)(\bar{y}'_i - \bar{Y}'_1)}{\sqrt{\sum_{i=1}^{N_1} (\bar{y}_i - \bar{Y}_1)^2 \sum_{i=1}^{N_1} (\bar{y}'_i - \bar{Y}'_1)^2}}$$

and

$$\rho_{2w} = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{M} (y_{ij} - \bar{y}_i)(y'_{ij} - \bar{y}'_i)}{\sqrt{\sum_{i=1}^{N_1}\sum_{j=1}^{M} (y_{ij} - \bar{y}_i)^2 \sum_{i=1}^{N_1}\sum_{j=1}^{M} (y'_{ij} - \bar{y}'_i)^2}} \tag{4.6}$$

The estimate $\bar{y}'_1$ is best possible linear unbiased estimate if the constants a and c are chosen in such way that $V(\bar{y}'_1)$ is minimum. Differentiating the expression on right hand side in (4.4) with respect to a and c and equating the resulting expressions to zero, and solving for a and c we have

$$a = - \frac{V_{13} V_{44}}{(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2} \tag{4.7}$$

and $$c = \frac{V_{44}(V_{11} + V_{22})}{(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2}$$

Thus the best linear unbiased estimate for $\bar{\bar{Y}}'$ is

$$\frac{n_1 + n'_1}{2n'} \left[(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2 \right]^{-1} \left[V_{13} V_{44} (\bar{\bar{y}}_2 - \bar{\bar{y}}_1) + V_{44} (V_{11} + V_{22}) (\bar{\bar{y}}'_1) + \left\{ V_{33} (V_{11} + V_{22}) - V_{13}^2 \right\} \bar{\bar{y}}'_2 \right]$$

(4.8)

and has the variance

$$\frac{V_{44} \left[V_{33} (V_{11} + V_{22}) - V_{13}^2 \right]}{(V_{11} + V_{22})(V_{33} + V_{44}) - V_{13}^2}$$

(4.9)

where Y'_{ij}s are defined in (4.8)

Now if we assume that

$$\frac{S_{1b}}{\bar{\bar{Y}}_1} = \frac{S'_{1b}}{\bar{\bar{Y}}'_1} = C_{1b} \text{ and } \frac{\bar{S}_{1w}}{\bar{\bar{Y}}_1} = \frac{\bar{S}'_{1w}}{\bar{\bar{Y}}'_1} = C_{1w}$$

Then $\dfrac{V_{11}}{\bar{\bar{Y}}_1^2} = \dfrac{V_{33}}{\bar{\bar{Y}}'^2_1} = A, \quad \dfrac{V_{22}}{\bar{\bar{Y}}_1^2} = \dfrac{V_{44}}{\bar{\bar{Y}}'^2_1} = B$

and

$$\frac{V_{13}}{\bar{\bar{Y}}_1 \bar{\bar{Y}}'_1} = C \text{, say}$$

and the variance of the best linear unbiased estimate takes the simple form

$$V(\bar{\bar{y}}'_1) = \bar{\bar{Y}}'^2_1 \frac{B\left[A(A+B) - C^2 \right]}{(A+B)^2 - C^2}$$

.(4.11)

4.3. Two - Phase two - stage sampling on two occasions. Unequal first stage units.

Let n', n_1, n'_1, n, np and nq be defined in the same way as in section 4.2., and let

M_{1i} = number of second stage units in the i^{th} first stage units of the first stratum (i = 1, 2,......., N_1)

m_{1i} = number of second stage units sampled in the i^{th} first stage units of the first stratum (i = 1, 2, ...N_1)

and $\bar{\bar{y}}_1 = \frac{1}{np\,\bar{M}_1} \sum_{i=1}^{np} M_{1i}\,\bar{y}_{i(m_i)}$

$$\bar{\bar{y}}_2 = \frac{1}{nq\,\bar{M}_1} \sum_{i=1}^{nq} M_{1i}\,\bar{y}_{i(m_i)}$$

$$\bar{\bar{y}}'_1 = \frac{1}{np\,\bar{M}_1} \sum_{i=1}^{np} M_{1i}\,\bar{y}'_{i(m_i)}$$

$$\bar{\bar{y}}'_2 = \frac{1}{nq\,\bar{M}_1} \sum_{i=1}^{nq} M_{1i}\,\bar{y}'_{i(m_i)}$$

(4.12)

where $\bar{M}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} M_{1i}$

and $\bar{y}_{i(m_i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$, i = 1, 2,......, N_1

= Sample mean per second stage unit in the i^{th} first stage unit of first stratum on the first occasion.

$\bar{y}'_{i(m_i)} = \frac{1}{m_i} \sum_{j=1}^{m_i} y'_{ij}$, i = 1, 2,......., N_1

= Sample mean per second stage units in the i^{th} first stage unit from first stratum on the second occasion. Following the procedure adopted in (4.2) it is easy to see that the unbiased estimate of form (4.1) for the population mean on the second occasion is given by (4.8) and its variance by (4.9)

where

$\overset{*}{\bar{y}}_1$, $\overset{*}{\bar{y}}_2$, $\overset{*}{\bar{y}}'_1$ and $\overset{*}{\bar{y}}'_2$ are given in (4.12)

and

$$V_{11} = \frac{p_1(1-p_1)}{2n'} \overset{*}{\bar{Y}}{}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \times$$

$$\left[\frac{1}{np} s_{1b}^2 + \frac{1}{npN_1} \sum_{i=1}^{N_1} \frac{M_{1i}^2}{\bar{M}_1^2} \frac{s_{1i}^2}{m_1} \right]$$

$$V_{22} = \frac{p_1(1-p_1)}{2n'} \overset{*}{\bar{Y}}{}_1^2 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \times$$

$$\left[\frac{1}{nq} s_{1b}^2 + \frac{1}{nqN_1} \sum_{i=1}^{N_1} \frac{M_{1i}^2}{\bar{M}_1^2} \frac{s_{1i}^2}{m_1} \right]$$

$$V_{33} = \frac{p_1(1-p_1)}{2n'} \overset{*}{\bar{Y}}{}'^2_1 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \times$$

$$\left[\frac{1}{np} s'^2_{1b} + \frac{1}{npN_1} \sum_{i=1}^{N_1} \frac{M_{1i}^2}{\bar{M}_1^2} \frac{s'^2_{1i}}{m_1} \right]$$

$$V_{44} = \frac{p_1(1-p_1)}{2n'} \overset{*}{\bar{Y}}{}'^2_1 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \times$$

$$\left[\frac{1}{nq} s'^2_{1b} + \frac{1}{nqN_1} \sum_{i=1}^{N_1} \frac{M_{1i}^2}{\bar{M}_1^2} \frac{s'^2_{1i}}{m_1} \right]$$

and

$$V_{13} = \frac{p_1(1-p_1)}{2n'} \bar{Y}_1 \bar{Y}'_1 + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'} \right] \times$$

$$\left[\frac{1}{np} \rho_{1b} S_{1b} S'_{1b} + \frac{1}{npN_1} \sum_{i=1}^{N_1} \frac{M_{1i}^2}{\bar{M}_1^2} \frac{\rho_{1i} S_{1i} S'_{1i}}{m_i} \right]$$

(4.13)

where $S_{1b}^2 = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}_i - \bar{Y}_1 \right)^2$

and the population mean in first stratum on first occasion $\bar{Y}_1 = \frac{1}{N_1 \bar{M}_1} \sum_{i=1}^{N_1} M_{1i} \bar{y}_i$, $\bar{y}_i$ being the population mean per second stage unit on first occasion in the i th first stage unit of the first stratum (i = 1, 2, ...N_1)

$$S'^2_{1b} \frac{1}{N_1} \sum_{i=1}^{N_1} \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}'_i - \bar{Y}'_1 \right)^2$$

and the population mean in first stratum on first occasion $\bar{Y}'_1 = \frac{1}{N_1 \bar{M}_1} \sum_{i=1}^{N_1} M_{1i} \bar{y}'_i$, $\bar{y}'_i$ being the population mean per second stage units on the second occasion in the i th first stage unit of the first stratum, (i= 1, 2, ...N_1)

and $S_{1i}^2 = \frac{1}{M_{1i}} \sum_{j=1}^{M_{1i}} (y_{ij} - \bar{y}_i)^2$

= population variance within i th first stage unit from first stratum on the first occasion.

and

$$S'^{2}_{1i} = \frac{1}{M_{1i}} \sum_{j=1}^{M_{1i}} (y'_{ij} - \bar{y}'_{i})^{2}$$

= population variance withing i th first stage unit from first stratum on the second occasion.

and $$\rho_{1b} = \frac{\sum_{i=1}^{N_1} \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}_i - \bar{\bar{Y}}_1 \right) \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}'_i - \bar{\bar{Y}}'_1 \right)}{\left[\sum_{i=1}^{N_1} \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}_i - \bar{\bar{Y}}_1 \right)^{2} \sum_{i=1}^{N_1} \left(\frac{M_{1i}}{\bar{M}_1} \bar{y}'_i - \bar{\bar{Y}}'_1 \right)^{2} \right]^{\frac{1}{2}}}$$

and $$\rho_{1i} = \frac{\sum_{j=1}^{M_{1i}} (y_{ij} - \bar{y}_i)(y'_{ij} - \bar{y}'_i)}{\left[\sum_{j=1}^{M_{1i}} (y_{ij} - \bar{y}_i)^{2} \sum_{j=1}^{M_{1i}} (y'_{ij} - \bar{y}'_i)^{2} \right]^{1/2}}$$

(4.12)

: C H A P T E R V :

The Problem of Optimum Allocation

5.1 Introduction. The last two chapters were devoted for finding the best linear unbiased estimate for the mean on the second occasion and its variance when a particular type of double sampling is repeated on two occasions, with partial replacement of units (first stage units in the case of two- stage sampling) on second occasion. The variance was minimised with respect to the constants of the linear estimate. The size n' of the preliminary sample, n that of sub-sample in non-zero stratum, the fraction p of units retained for second occasion, the sample size m for second stage units were all assumed to be fixed. But these n', n, p and m can be chosen in a optimum way so as to minimise the variance for a given cost or minimise the cost for a given variance. In the present chapter we will study the problem of optimum allocation an obtain approximate solutions of n, n' and p.

5.2 Optimum allocation for single-stage double sampling on two occasions.

The variance of the estimate $\bar{Y}'_1$ for the mean on second occasion when a single stage double sampling is repeated on two occasion with partial replacement of units on second occasion as obtained under the assumption that the coefficients of variati in the non-zero stratum on both the occasions are same, is

$$\hat{\sigma}^2_{Y'_1} \quad \frac{B\left[A(A+B) - C^2\right]}{(A+B)^2 - C^2}$$

where
$$A = \frac{p_1(1-p_1)}{2n'} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right]\frac{C^2_{v1}}{np}$$

$$B = \frac{p_1(1-p_1)}{2n'} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right]\frac{C^2_{v1}}{nq}$$

and
$$C = \frac{p_1(1-p_1)}{2n'} + \left[p_1^2 + \frac{p_1(1-p_1)}{2n'}\right]\frac{\rho C^2_{v1}}{np}$$

Since the contribution of $\frac{p_1(1-p_1)}{2n'}$ C^2_{v1} will be very small as compared to p_1^2 C^2_{v1}, A, B and C can be approximated by

$$A \doteq \frac{p_1(1-p_1)}{2n'} + \frac{p_1^2 \; C^2_{v1}}{np}$$

$$B \doteq \frac{p_1(1-p_1)}{2n'} + \frac{p_1^2 \; C^2_{v1}}{nq}$$

$$C \doteq \frac{p_1(1-p_1)}{2n'} + \frac{p_1^2 \; C^2_{v1}}{np}$$

(5.1)

If we define
$$A' = \frac{1}{n'} + \frac{2p_1}{1-p_1} \; \frac{C^2_{v1}}{np}$$

$$B' = \frac{1}{n'} + \frac{2p_1}{1-p_1} \; \frac{C^2_{v1}}{nq}$$

and
$$C' = \frac{1}{n'} + \frac{2p_1}{1-p_1} \; \frac{C^2_{v1}}{np}$$

(5.2)

then
$$V(\overset{*}{\bar{Y}}{}'_1) = \frac{1}{2} p_1 (1-p_1) \bar{Y}'^{2}_1 \frac{B'\left[A'(A'+B') - C'^2\right]}{(A'+B')^2 - C'^2}$$

(5.3)

putting $\frac{2p_1}{1-p_1} C^2_{v1} = g$ we have

$$A' = \frac{1}{n'} + \frac{g}{np}, \quad B' = \frac{1}{np} + \frac{g}{nq}, \quad C' = \frac{1}{n'} + \frac{\rho g}{np}$$

(5.4)

The expression (5.3) is to be minimised for n, n' and q, subject to the cost considerations. The cost function for the first occasion can be written as

$$C'_0 = c_1 n' + c_2 n$$

(5.5)

where c_1 is the cost per unit in ascertaining whether a particular unit is zero unit or non-zero unit.
c_2 is the cost per unit for observing the variate in non-zero stratum.
and C'_0 is the total cost of the survey on the first occasion minus the overall expenditure on that occasion on stationery, contingencies and maintenance of statistical staff not directly related to the field work.
The cost function for the second occasion will be

$$C''_0 = c_1 n' + c_2 nq + c_3 np$$

(5.6)

where c_1 and c_2 are as before and c_3 is the cost per unit for observing the variate on the units already sampled of first occasion, while C_o'' is defined in the same way as C' except that it refers to second occasion.

Now the combined cost function can be written as

$$C_o = C_o' + C_o'' = 2c_1n' + c_2n(1+q) + c_3(1-q)n$$

$$= 2_c \qquad = 2c_1n' + (c_2+c_3)n + (c_2-c_3)nq$$

which can be written as

$$C_o = c_1'n' + c_2'n + c_3'nq \qquad (5.7)$$

where $c_1' = 2c_1$

$c_2' = c_2 + c_3$

$c_3' = c_2 - c_3$ (5.8)

Now consider the function $V + (C_o - C)$

$$= \frac{1}{4}p_1(1-p_1)Y_1'^2 \frac{B'\left[A'(A'+B') - C'^2\right]}{(A'+B')^2 - C'^2} + \lambda\left[c_1'n' + c_2'n + c_3'nq - C_o\right] \qquad (5.9)$$

which is to be minimised with respect to n, n' q and

Differentiating (5.9) partially with respect to n, n', q and λ, we have

$$\frac{\partial V}{\partial n} + \lambda\frac{\partial C_o}{\partial n} = 0$$

$$\frac{\partial V}{\partial n'} + \lambda\frac{\partial C_o}{\partial n'} = 0$$

$$\frac{\partial V}{\partial q} + \lambda\frac{\partial C_o}{\partial q} = 0 \quad \text{and } c_1'n' + c_2'n + c_3'nq = C_o \qquad (5.10)$$

But

$$\frac{\partial V}{\partial n} = \frac{\partial V}{\partial A'}\frac{\partial A'}{\partial n} + \frac{\partial V}{\partial B'}\frac{\partial B'}{\partial n} + \frac{\partial V}{\partial C'}\frac{\partial C'}{\partial n}$$

$$\frac{\partial V}{\partial n'} = \frac{\partial V}{\partial A'}\frac{\partial A'}{\partial n'} + \frac{\partial V}{\partial B'}\frac{\partial B'}{\partial n'} + \frac{\partial V}{\partial C'}\frac{\partial C'}{\partial n'}$$

and $$\frac{\partial V_o}{\partial q} = \frac{\partial V}{\partial A'}\frac{\partial A'}{\partial q} + \frac{\partial V}{\partial B'}\frac{\partial B'}{\partial q} + \frac{\partial V}{\partial C'}\frac{\partial C'}{\partial q}$$

(6.11)

It is easy to verify that

$$\frac{\partial V}{A'} = \tfrac{1}{2}\, p_1(1-p_1)\ \bar{Y}_1'^{\,2}\ \frac{B'^2\left[(A'+B')^2+C'^2\right]}{\left[(A'+B')^2-C'^2\right]^2}$$

$$\frac{\partial V}{\partial B'} = \tfrac{1}{2}\, p_1(1-p_1)\ \bar{Y}_1'^{\,2}\ \frac{(A'^2-C'^2)\left\{(A'+B')^2-C'^2\right\}+2B'^2C'^2}{\left[(A'+B')^2-C'^2\right]^2}$$

and

$$\frac{\partial V}{\partial C'} = \tfrac{1}{2}\, p_1(1-p_1)\ \bar{Y}_1'^{\,2}\ \frac{-2B'^2C'(A'+B')}{\left[(A'+B')^2-C'^2\right]^2}$$

while (6.12)

$$\frac{\partial A'}{\partial n'} = \frac{\partial B'}{\partial n'} = \frac{\partial C'}{\partial n'} = -\frac{1}{n'^2}$$

$$\frac{\partial A'}{\partial n} = \frac{-g}{n^2p},\ \frac{\partial B'}{\partial n} = \frac{-g}{n^2q},\ \frac{\partial C'}{\partial n} = \frac{-\rho g}{n^2p}$$

and $$\frac{\partial A'}{\partial q} = \frac{g}{np^2}\quad \frac{\partial B'}{\partial q} = \frac{-g}{nq^2},\ \frac{\partial C'}{\partial q} = \frac{\rho g}{np^2}$$

(6.13)

Again

$$B'^2\left[(A'+B')^2+C'^2\right]$$

$$= (n')^{-4}\left[a_0+a_1t+a_2t^2+a_3t^3+a_4t^4\right]$$

$$(A'^2-C'^2)\left[(A'+B')^2-C'^2\right]+2B'^2C'^2$$

$$=(n')^{-4}\left[b_0+b_1t+b_2t^2+b_3t^3+b_4t^4\right]$$

$$-2B'^2C'\quad(A'+B')$$

$$=(n')^{-4}\left[d_0+d_1t+d_2t^2+d_3t^3+d_4t^4\right]$$

and

$$\left[(A'+B')^2-C'^2\right]^2$$

$$=(n')^{-4}\left[e_0+e_1t+e_2t^2+e_3t^3+e_4t^4\right]$$

$$(5.14)$$

where $t = \dfrac{2n'}{n}$, and

$a_0 = 5$

$a_1 = 2(2+\rho q)\, p^{-1}q^{-1} + 10q^{-1}$

$a_2 = 5q^{-2} + 4q^{-2}p^{-1}(2+\rho q) + (1+\rho^2q^2)\, p^{-2}\, q^{-2}$

$a_3 = 2q^{-2}(2+\rho q)\, p^{-1}\, q^{-1} + 2q^{-1}(1+q^2\rho^2)\, p^{-2}\, q^{-2}$

$a_4 = (1+\rho^2q^2)\, p^{-2}\, q^{-4}$

$b_0 = 2$

$b_1 = 6(1-\rho)\, p^{-1} + 4\ p^{-1} + 4q^{-1}$

$b_2 = 2(1-\rho^2)\, p^{-2} + 4(1-\rho)(2-\ \rho q)\, p^{-2}\, q^{-1} + 2q^{-2} + 6q^{-1}\rho\ p^{-1}$

$+ 2\rho^2 p^{-2}$

$b_3 = 2(1-\rho^2)\ (2-\rho q)\ p^{-3}\ q^{-1} + 2(1-\rho)(1-\rho^2 q^2)\ p^{-3}\ q^{-2} +$

$4\ p^{-1} q^{-2}\rho + 4\rho^2 p^{-2} q^{-1}$

$b_4 = (1-\rho^2)(1-\ \rho^2 q^2)\ p^{-4}\ q^{-2} + 2\ q^{-2} p^{-2} \rho^2$

$d_0 = -4$

$d_1 = -2(p^{-1}q^{-1} + 2p^{-1}\rho\ ) - 2\ q^{-1}$

$d_2 = -4\ q^{-2}\ -4q^{-1}\ (p^{-1}q^{-1} + 2p^{-1}\rho\ ) - 2\ \ \rho\ \ p^{-2}q^{-1}$

$d_3 = -2q^{-2}\ (p^{-1}q^{-1} + 2p^{-1}\rho\ ) - 4q^{-1}\rho\ p^{-2}q^{-1}$

$d_4 = -2\rho\, p^{-2}q^{-3}$

$e_0 = 9$

$e_1\ =\ 12\ (2-\rho q)\ p^{-1}q^{-1}$

$e_2\ =\ 6(1-\rho^2 q^2) p^{-2}\ q^{-2} + 4(2-\rho q)^2 p^{-2}\ q^{-2}$

$e_3 =\ 4(2-\rho q)(1-\ \rho^2 q^2)\ p^{-3}\ q^{-3}$

$e_4 = (1-\ \rho^2 q^2)^2\ p^{-4}\ q^{-4}$

(5.15)

Again

$$\frac{\partial C_0}{\partial n'} = c'_1$$

$$\frac{\partial C}{\partial n} = c'_2 + c'_3 q$$

and

$$\frac{\partial C}{\partial q} = c'_3 n$$

(5.16)

Hence the equations giving optimum n, n' and q are

$$-\frac{g}{n^2 p} \sum_{i=0}^{4} a_i t^i - \frac{g}{n^2 q} \sum_{i=4}^{4} b_i t^i - \frac{g\rho}{n^2 p} \sum_{i=0}^{4} d_i t^i + \lambda (c'_2 + c'_3 q) \sum_{i=0}^{4} e_i t^i = 0$$

(5.17)

$$-\frac{1}{n'^2} \sum_{i=0}^{4} (a_i + b_i + d_i) t^i + \lambda c'_1 \sum_{i=0}^{4} e_i t^i = 0$$

(5.18)

and

$$\frac{g}{np^2} \sum_{i=0}^{4} a_i t^i - \frac{g}{nq^2} \sum_{i=0}^{4} b_i t^i + \frac{\rho g}{np^2} \sum_{i=0}^{4} d_i t^i + \lambda c'_3 n \sum_{i=0}^{4} e_i t^i = 0$$

(5.19)

It is not easy to solve these equations for n, n' and p. However if we assume p to be fixed equations (5.17) and (5.18) give the optimum value for t and hence optimum value for n'/n.

Eliminating   from (5.17) and (5.18) we have

$$t^2 \left[\frac{\sum a_i t^i}{p} + \frac{\sum b_i t^i}{q} + \frac{\sum d_i t^i}{p} \right] = h \sum (a_i + b_i + d_i) t^i$$

where $h = (c'_2 + c'_3 q)\, g / c'_1$

i.e.

$$(a_4/p + b_4/2 + \rho d_4/p)\ t^6 + (a_3/q + b_3/q + \rho d_3/p)\ t^5 +$$

$$(a_2/p + b_2/q + \rho d_2/p - h\ a_4 - hb_4 - hd_4)\ t^4 + (a_1/p + b_1/q +$$

$$\rho d_1/p - ha_3 - hb_3 - hd_3)\ t^3 + (a_0/p + b_0/q + \rho d_0/p - ha_2 - hb_2 - hd_2)\ t^2$$

$$+ h(a_1 + b_1 + d_1)\ t + h(a_0 + b_0 + d_0) = 0 \qquad (5.20)$$

which is a sixth degree equation in t and can be solved without much dificulty. Let the optimum solution be t_o.

then $gn'/n = to$, therefore $n' = (n/g)\ t_o$

substituting this in

$$C_o = c_1' n' + c_2' n + c_3' nq$$

we have $C_o = (c_1'\ t_o/g + c'_2 + c_3'\ q)\ n$

or $n = \dfrac{C_o}{c_1'\ g^{-1}\ t_o + c_2' + c_3' q}$

and $n' = \dfrac{t_o C_o}{c_1' t_o + g\ c_2' + g c_3' q}$

$$(5.21)$$

However since $t = gn'/n = \left(2p_1/1-p_1\right) C^2_{v1}\ n'/n$

Equation (5.20) can be approximated by a quadratic or cubic, if n' is very large compared to n, $C^2 v_1 > 1$ and $p_1 > (1-p_1)$.

If we approximate by quadratic, we have

$$L_o t^2 + L_1 t + L_2 = 0 \qquad (5.22)$$

Where $L_o = p^{-1} a_4 + q^{-1} b_4 + p^{-1} \rho\, d_4$

$$L_1 = p^{-1} a_3 + q^{-1} b_3 + p^{-1} \rho\, d_3$$

$$L_2 = p^{-1} a_2 + q^{-1} b_2 + p^{-1} \rho\, d_2 - M(a_4 + b_4 + d_4) \qquad (5.23)$$

and,

$$t = gn'/n = \frac{(L_1^2 - 4L_o L_2)^{\frac{1}{2}} - L_1}{2L_o} \qquad (5.24)$$

Substituting this in $C_o = c_1' n' + (c_2' + c_3' q)\, n$, we have

$$2g L_o C_o = c_1' \left[(L_1^2 - 4L_o L_2)^{\frac{1}{2}} - L_1\right] + 2g L_o (c_2' + c_3' q)$$

Hence

$$n = \frac{2g L_o C_o}{2g L_o (c_2' + c_3' q) + c_1' (L_1^2 - 4L_o L_2)^{\frac{1}{2}}}$$

While

$$n' = \frac{(L_1^2 - 4L_o L_2)^{\frac{1}{2}} - L_1 \quad C_o}{2g L_o (c_2' + c_3' q) + c_1' (L_1^2 - 4L_o L_2)^{\frac{1}{2}}} \qquad (5.25)$$

The optimum q for simple random sampling repeated on two occasions is given by

$$q_{opt} = \frac{1}{1 + (1 - \rho)^{\frac{1}{2}}} \qquad (5.26)$$

In graph No.3 page 64, the variances for both double and simple random sampling repeated on two occasions are plotted against p. It is seen that in both the cases, the variance is minimum for the same value of p. Four pairs of graphs corresponding to p = 0.2, 0.4, 0.6 and 0.8 respectively are drawn. The upper ones are for double sampling and lower ones for simple random sampling. The values of p, $C_{v1}^2 = \delta^2$, $K = \lambda n'/n$ and c_2/c_1 have been fixed as 0.6, 1.2, 2 and 10 respectively. From these graphs, it is evident that q from (5.26) is equal to or at least very close to the optimum value for q for double sampling repeated on two occasions.

However if we are interested in obtaining more exact solutions for n', n and p, we can obtain them with help of approximately optimum values obtained from (5.25) and (5.26). Let these be denoted by n_o', n_o and p_o, and the expressions on right hand side of the following three equations

$$t^2\left[\; p^{-1}\sum_{i=0}^{4} a_i t^i + q^{-1}\sum_{i=0}^{4} b_i t^i + p^{-1}\sum_{i=0}^{4} d_i t^i\right] - h\sum_{i=0}^{4}(a_i+b_i+d_i)\,t^i = 0$$

$$p^{-1}\sum_{i=0}^{4} a_i t^i + q^{-1}\sum_{i=0}^{4} b_i t^i + p^{-1}\sum_{i=0}^{4} d_i t^i + c_3'^{-1}(c_2'+c_3'q)\times$$

$$\left[p^{-2}\sum_{i=0}^{4} a_i t^i - q^{-2}\sum_{i=0}^{4} b_i t^i + p^{-2}\sum_{i=0}^{4} d_i t^i\right] = 0$$

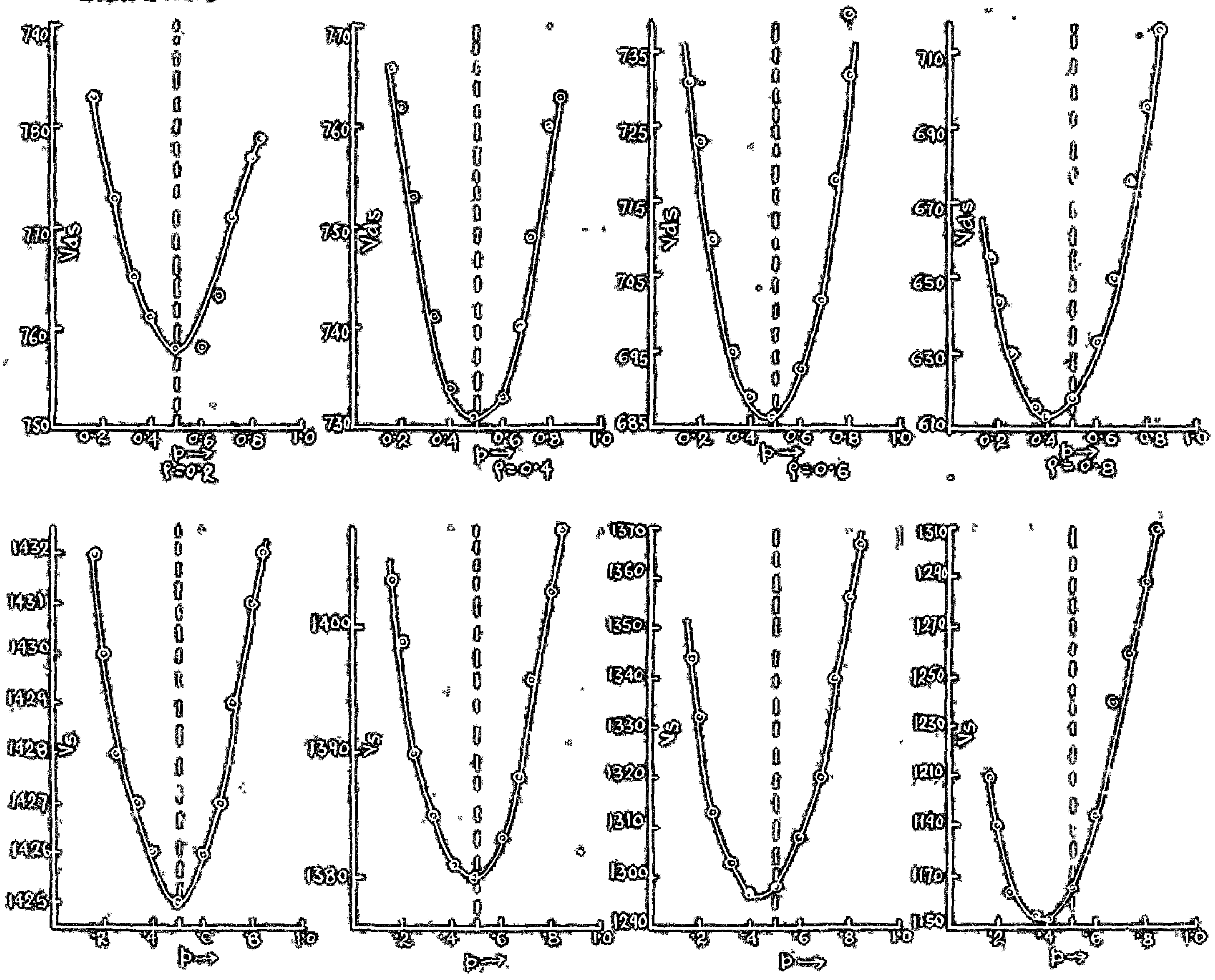
and

$$c_1'n' + c_2'n + c_3'nq - C_o = 0$$

(5.27)

by $\phi_1(n', n, q)$, $\phi_2(n', n, q)$ and $\phi_3(n', n, q)$, respectively.

GRAPH No. 3


790
780
770
760
750
Vds
0·2
0·4
0·6
0·8
1·0
p→
ρ=0·2
770
760
750
740
730
Vds
0·2
0·4
0·6
0·8
1·0
p→
ρ=0·4
735
725
715
705
695
685
Vds
0·2
0·4
0·6
0·8
1·0
p→
ρ=0·6
710
690
670
650
630
610
Vds
0·2
0·4
0·6
0·8
1·0
p→
ρ=0·8
1432
1431
1430
1429
1428
1427
1426
1425
Vs
·2
·4
·6
·8
1·0
p→
1400
1390
1380
Vs
·2
·4
·6
·8
1·0
p→
1370
1360
1350
1340
1330
1320
1310
1300
1290
Vs
·2
·4
·6
·8
1·0
p→
1310
1290
1270
1250
1230
1210
1190
1170
1150
Vs
·2
·4
·6
·8
1·0
p→

The three equations $\phi_i = 0$, $i = 1,2,3$ give the exact optimum solutions for $n', n,$ and q.

Let $n_o' + \delta n_o'$, $n_o + \delta n_o$ and $q_o + \delta q_o$ be the exact solutions. Then

$$0 = \phi_i (n_o' + \delta n_o', n_0 + \delta n_o, q_o + \delta q_o)$$

$$= \phi_i (n_o', n_o, q_o) + \delta n_o' \frac{\partial \phi_i}{\partial n_o'} + \delta n_o \frac{\partial \phi_i}{\partial n_o} + \delta q_o \frac{\partial \phi_i}{\partial q_o}$$

approximately. (2.28)

($i = 1, 2, 3.$)

Hence

$$(\delta n_o', \delta n_o, \delta q_o) = J_o^{-1} \left[\phi_1 (n_o', n_o, q_o), \phi_2(n_o', n_o, q_o), \phi_3(n_o', n_o, q_o) \right]$$

(5.29)

where J_o^{-1} is the inverse of the matrix J_o given by

$$J_o = \begin{bmatrix} \frac{\partial \phi_1}{\partial n_o'} & \frac{\partial \phi_1}{\partial n_o} & \frac{\partial \phi_1}{\partial q_o} \\ \frac{\partial \phi_2}{\partial n_o'} & \frac{\partial \phi_2}{\partial n_o} & \frac{\partial \phi_2}{\partial q_o} \\ \frac{\partial \phi_3}{\partial n_o'} & \frac{\partial \phi_3}{\partial n_o} & \frac{\partial \phi_3}{\partial q_o} \end{bmatrix}$$

(5.30)

The operation (5.29) can be repeated successively to obtain the exact solutions for n', n and q.

5.3 Optimum allocation for two-phase two-stage sampling on two occasions. The variance of the estimate of the population mean on second occasion, of the form (4.3), for the two-phase two-stage sampling on two occasions of section 4.2 is

$$\dot{Y}'^{\,2}_{1} \quad \frac{2\left[(A+B)-C^{2}\right]}{(A+B)^{2}-C^{2}}$$

where A, B and C are given in (4.10)

A simple cost function for this type of double sampling will be $C'_o = c_1 n' + c_2 n + c_3 nm$, for the first occasion

and

$C''_o = c_1 n' + c_2 nq + c_3 nm$, for the second occasion,

where

c_1 = cost of ascertaining whether a particular first stage unit belongs to first stratum or not,

c_2 = cost of making frame for a first stage unit to be used for selecting second-stage units,

c_3 = cost per unit of observing the variate y_{ij} on a second stage unit,

and C'_o and C''_o are total costs on first and second occasion respectivly .

The combined cost function for both the occasions can be written as

$$C_o = c_1 n' + \tfrac{1}{2} c_2 (1+q)\, n + c_3 nm \qquad (5.31)$$

We want to determine the optimum values of n', n and m, q that will minimise the variance for a given cost.

Consider the function

$$\phi = Y_1^{*2} \quad \frac{B\left[A(A+B) - C^2\right]}{(A+B)^2 - C^2} - \lambda\left[c_1 n' + \tfrac{1}{2}\, c_2(1+q)n + c_3 nm - C_0\right]$$

If we approximate A, B and C by

$$\frac{p_1(1-p_1)}{2n'} + \frac{p_1^2}{np}\left(C_{1b}^2 + \frac{C_{1w}^2}{m}\right),$$

$$\frac{p_1(1-p_1)}{2n'} + \frac{p_1^2}{nq}\left(C_{1b}^2 + \frac{C_{1w}^2}{m}\right)$$

and

$$\frac{p_1(1-p_1)}{2n'} + \frac{p_1^2}{np}\left(\rho_{1b}C_{1b}^2 + \frac{\rho_{1w}C_{1w}^2}{m}\right)$$ respectively,

then the final equations which give optimum values for n, n', m and q are:

$$-\frac{p_1^2}{n^2p}\left(C_{1b}^2 + C_{1w}^2/m\right) B^2\left[(A+B)^2 + C^2\right] - \frac{p_1^2}{n^2q}\left(C_{1b}^2 + C_{1w}^2/m\right)\times$$

$$\left[(A^2 - C^2)\left[-c^2+(A+B)^2\right] + 2B^2C^2\right] + \frac{2p_1^2}{n^2p}\left(\rho_{1b}C_{1b}^2 + \rho_{1w}C_{1w}^2/m\right) B^2C(A+B)$$

$$-\ \lambda\left[(A+B)^2 - C^2\right]^2\left[\tfrac{1}{2}\, c_2(1+q) + m\right] = 0$$

(5.33)

$$-\frac{p_1(1-p_1)}{n'^2}\left[\ B^2\left\{(A+B)^2 + C^2\right\} + (A^2 - C^2)\left\{(A+B)^2 - C^2\right\} + 2B^2C^2\right.$$

$$\left. -\ 2B^2C(A+B)\right] - \lambda\left[(A+B)^2 - C^2\right]^2 c_1 = 0$$

(5.34)

$$- \frac{C_{1w}^2 p_1^2}{nqm^2} B^2 \left\{ (A+B)^2 + C^2 \right\} - \frac{C_{1w}^2 p_1^2}{nqm^2} \left[(A^2 - C^2) \left\{ (A+B)^2 - C^2 \right\} \right.$$

$$\left. + 2B^2C^2 \right] + \frac{2p_1^2 \quad C_{1w}^2}{npm^2} B^2C(A+B) - \lambda\left[(A+B)^2 - C^2 \right]^2 c_3 m = 0$$

(5.35)

$$\frac{p_1^2}{np^2} (C_{1b}^2 + C_{1w}^2/m) B^2 \left\{ (A+B)^2 + C^2 \right\} \frac{p_1^2}{nq^2} (C_{1b}^2 + C_{1w}^2/m) \times$$

$$\left[(A^2 - C^2) \left\{ (A+B)^2 - C^2 \right\} + 2B^2C^2 \right] - \frac{2p_1^2}{np^2} (C_{1b}^2 + C_{1w}^2/m) \times$$

$$B^2C(A+B) - \lambda \left[(A+B)^2 - C^2 \right]^2 \frac{1}{2} c_2 n = 0$$

(5.36)

and

$$C_0 = c_1 nq + \frac{1}{2} c_2 (1+q) + c_3 m.$$

However, when expressions like

$$A = \frac{p_1(1-p_1)}{n'_2} + \frac{C_{1b}^2 p_1^2}{np} + \frac{C_{1w}^2 p_1^2}{mnp},$$

$$B = \frac{p_1(1-p_1)}{n'_2} + \frac{C_{1b}^2 p_1^2}{nq} + \frac{C_{2w}^2 p_1^2}{mnq}$$

and $C = \frac{p_1(1-p_1)}{n'_2} + \frac{\rho_{1b} C_{1b}^2 p_1^2}{np} + \frac{\rho_{1w} C_{1w}^2 p_1^2}{mnp}$

are used to replace A, B and C in equations (5.33) to (5.36) it is not easy to solve them even by the method of trial and error.

If we assume q and m to be fixed then equation (5.20) gives the optimum value of gn'/n, where

$$g = \frac{2p_1}{1-p_1} \left(C^2_{1b} + C^2_{1w}/m \right)$$

(5.37)

$$h = \frac{\frac{1}{2}c_2(1 + q) + c_3 m}{c_1}$$

(5.38)

$$\rho = \frac{\rho_{u} C^2_{1b} + \rho_{w} C^2_{1w}/m}{C^2_{1b} + C^2_{1w}/m}$$

(5.39)

and constants a_i, b_i, d_i and e_i, i = 0, 1, 2, 3, 4, are same as in section 5.2.

Let the optimum value for $t = gn'/n$ be denoted by t_0. Then

$$n = \frac{C_0}{c_1 g^{-1} t_0 + \frac{1}{2} c_2(1+q)+c_3 m}$$

and

$$n' = \frac{t_0 C_0}{c_1 t_0 + \frac{1}{2}g\, c_2(1+q) + g c_3 m}$$

(5.40)

Now q at first instance can be fixed as

$$q_0 = \frac{\alpha^2 - \alpha\sqrt{\alpha^2 - \gamma^2}}{\gamma^2} \qquad (5.41)$$

where $\alpha = S^2_{1b} + S^2_{1w}/m$, and $\gamma = \rho_u S^2_{1b} + \rho_u S^2_{1w}/m$

(5.42)

q in (5.41) is the value which minimises the variance of the mean on second occasion for a two-stage design repeated on two occasions (with no double sampling, and strata sizes supposed to be known).

Let n and n' obtained from (5.40) corresponding to the value of q_o be denoted by n_o and n'_o.
With g, h and p as defined in (5.37), (5.38) and (5.39) and denoting the expressions on right hand side of (5.27) with ϕ_1, ϕ_2 and ϕ_3 as before, we can solve (5.28) to obtain δn_o, $\delta n'_o$ and δq_o. Repeating this operation successively, we can obtain the exact solutions for n', n and q corresponding to a fixed, m.

Thus for a fixed m, optimum values for n', n and q can be obtained in an exact way, and hence the minimum variance coresponding to that fixed value for m. A graph between m and the conditional optimum variance opt (V/m) can be drawn. If opt(V/m) is minimum at m_{opt} and n'_{opt} , n_{opt} and q_{opt} are the optimum values that minimise the variance for $m = m_{opt}$ then n'_{opt}, n_{opt}, m_{opt} and q_{opt} give the optimum soluotions for n', n, m and q respectively.

—————————

CHAPTER VI

Applications and Conclusions.

6.1 The sample survey to estimate area under production of Coconut in Assam. The problems already discussed in the Chapters III, IV and V arose mainly in connection with a survey planned in Assam State for estimating area under production of coconut in Assam. The survey was first planned in 1968 with the following objects.

(i) To estimate the total number of coconut palms in the State with District-wise break down of the number of trees, under classifications 'bearing and non-bearing.'

(ii) To estimate the area under coconut.

(iii) To estimate the average yield per bearing palm and total production of coconut for each agricultural year, etc.

The survey is being conducted under the administrative and Techniqual control of the Director of Statistics, Assam, in consultation with the Statistical Adviser, Indian Council of Agricultural Research.

The initial sampling design adopted for the survey was one of stratified multistage random sampling, with sub-divisions of districts as strata. For the purpose of the first phase work, i.e. enumeration of number of palms and estimation of area under coconut crop a total of 800 villages was selected in the seven plain districts, the size of the sample in different sub-divisions being approximately in proportion to the area under "homestead and others miscellaneous crop". Within each stratum, the villages were selected by simple random sampling. In the villages thus selected complete enumeration of the trees was carried out. The subsequent operations, viz. harvesting, collection of data on cultivation practices, diseases, were confined to only a sub sample of 100 villages. For the purpose of selection of trees for these operations it was proposed to introduce, further stratification according to trees grown in gardens as well as the scattered ones. But as there was no coconut garden available in any one of the sample villages, this part of the work was confined to the scattered trees only. For collection of information on cultivation practices, diseases etc., 10 scattered trees from the total number of scattered trees were selected at random, out of these 10 scattered trees 5 were selected for harvesting experiments.

After two rounds of this survey were over, it was observed that more than ½ of the villages in Assam do not grow coconut crop, and inclusion of such villages in the sample increased the sampling error of the estimate considerably. To reduce the sampling error, the meeting of State Statisticians Sub-Committee held at Bombay, decided that the sample of villages to be retained from earlier rounds should be a sub-sample of those which have been found to grow coconut crop, whereas for the fresh selection of sample of villages in the new round a number of clusters, each of five villages would be selected, the number of such clusters being equal to the number of fresh villages to be selected for enumerating coconut palms. For enumeration of palms one village growing coconut crop from each cluster was selected at random from those growing coconut crop.

Under certain limitations the modified design recommended by the Sub-Committee can be treated as a double sampling, the size of the preliminary large sample for estimating the proportion of villages growing the coconut crop being five times that of the final sample in non-zero stratum for enumeration of coconut palms.

The following table gives the variance for double sampling and simple random sampling and the gain in efficiency of the former over latter, for the estimate of mean on second occasion, when both the sampling designs are repeated on two occasions. The results are given for each sub-division seperately, and are based on data giving number of trees in second and third round of Assam survey.

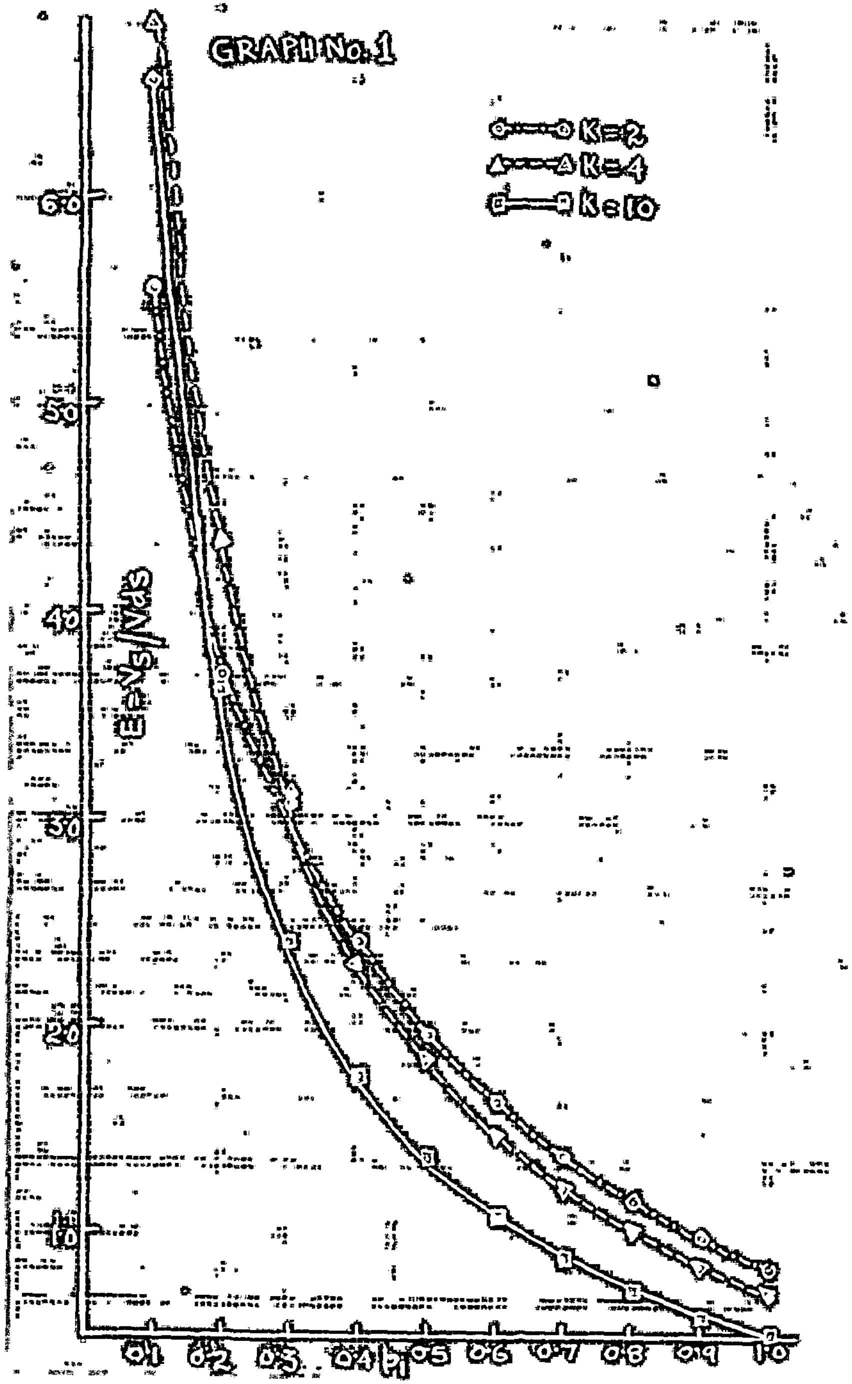
Sub - division.	Variance for estimating the mean on second occasion.		Percentage gain in efficiency. = $\frac{(i)-(ii)}{(ii)}$ x 100
	(i) Simple random sampling.	(ii) Double sampling.	
Dhubri	2.64	2.56	3.12
Goalpara	60.37	61.03	-1.02
Barpeta	344.56	325.72	5.78
Gauhati	351.88	326.13	7.90
Tezpur	106.23	100.33	5.88
Darrang	149.79	140.81	6.38
Nowgong	816.08	768.78	6.15
Jorhat	11.26	10.52	7.03
Golaghat	5.08	4.76	6.72
Sibsagar	6.89	6.66	5.03
Dibrugarh	2.60	2.72	-4.41
N.Lakhimpur	3.10	2.79	11.11
Haila Kadi	2.93	2.66	10.15
Silchar	1.47	1.42	3.52
Karimganj	2.91	2.71	7.38

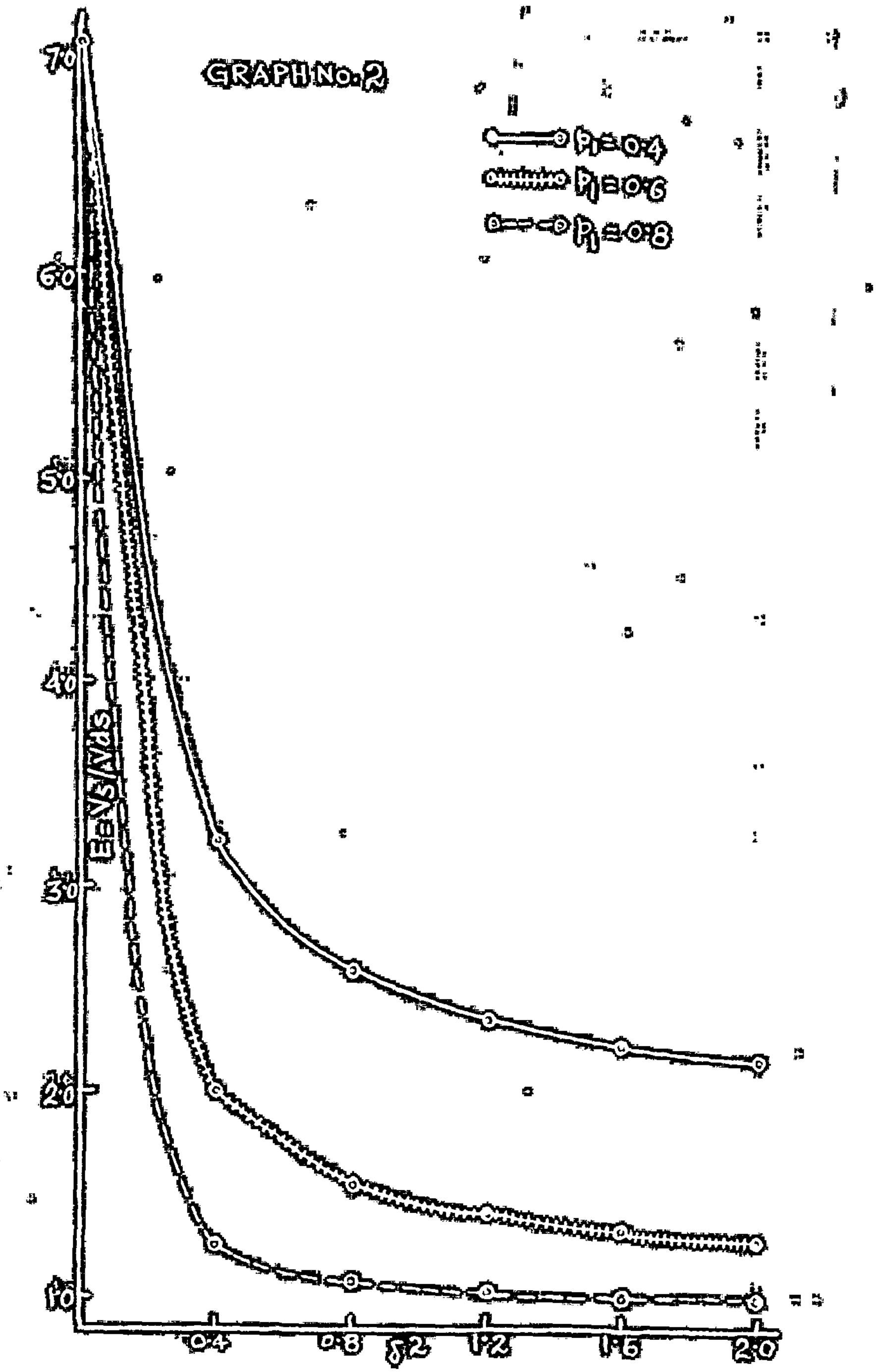
Thus it is seen from the above table that except from two sub-divisions Goalpara and Dibrugarh double sampling results in more precision than the corresponding simple random sampling. The ratio c_2/c_1 has been assumed to be 10. If this ration is more, the gain in efficiency will also be more.

6.2 Efficiency of double sampling. The behaviour of the efficiency of double sampling as compared to simple random sampling both repeated on two occasions for estimating the mean on second occasion has been studied in graph no.I and II on page 66.

In graph no.I, the efficiency is plotted against p_1, the size of non-zero stratum for k = 2, 4 and 10 (k = 2n'/n) p, δ^2 and ρ have been fixed as 0.4, 1.2 and 0.8 respectively It is seen from this graph that smaller the size of the non-zero stratum, more is the gain in efficiency of double sampling over simple random sampling. As p_1 comes closer to 1 the efficiency of double sampling decreases and finally becomes less than that of simple random sampling. It is obvious otherwise also, because if the proportionof zero - units is very small, it is a mere waste of a part of resources, if we take a preliminary large sample without achieving any significant reduction in sampling variance.

It is also clear from this graph that if p_1 is not very small then k should be small i.e. the size of the preliminary sample for estimating strata sizes should be moderate. In graph, the curve for k = 2 is consistently above those for k = 4 and k = 10, while that for k=4 is consistently above that for k = 10, after the point p_1 = 0.3. In other words given a total cost, the part of resources devoted for taking preliminary sample should be as small as possible.


GRAPH No.1


GRAPH No.2

In graph no 2, the efficiency is plotted against $C_{v1}^{2} = \delta^2$, the square of the coefficient of variation in non-zero stratum. Three curves corresponding to $p_1 = 0.4$, 0.6 and 0.8 respectively are drawn. The graph shows that for higher p_1, double sampling is more efficient only if δ^2 is small, say less than 1. When p_1 is not very large the double sampling results in more precision even if δ^2 is as large as 2.0 or more. It is also evident from this graph that as δ^2 increases the rate of decrease in efficiency is very slow. i.e. after a certain limit the efficiency is not affected very much by a further increase in δ^2.

CHAPTER VII

Sampling on two occasions with varying Probabilities of Selection.

Let the population consist of N units, and the probability of drawing the i^{th} unit be P_i. On the first occasion a sample of n units is drawn, with replacement and with varying probabilities of selection. We retain a sample of size pn units of these n units for the second occasion and supplement it by an independent sample of qn units $(p + q = 1)$ selected with varying probabilities, the probabilities remaining the same as on the first occasion.

Define $z_i = y_i/NP_i$ and $z'_i = y'_i/NP_i$ (7.1)

where y_i is the measurement on the first occasion and y'_i measurement on the second occasion, on the i^{th} unit of the population.

Let $\bar{z}_{pn} = \frac{1}{np} \sum_{1}^{np} \frac{y_i}{NP_i}$ where $y_1, y_2 \ldots\ldots y_{pn}$ are the measurements on the pn units that are common on both the occasions.

$\bar{z}_{qn} = \frac{1}{nq} \sum_{np+1}^{n} \frac{y_i}{NP_i}$, where $y_{pn+1}, \ldots\ldots y_n$ are the measurements on qn units that are in the sample on first occasion only.

Similarly, we have

$$\bar{z}'_{pn} = \frac{1}{np} \sum_{1}^{np} \frac{y'_i}{NP_i} \text{, and } \bar{z}'_{qn} = \frac{1}{nq} \sum_{np+1}^{n} \frac{y'_i}{NP_i} \qquad (7.2)$$

We will estimate $\bar{Y}'$ the mean, for the second period by a linear estimate of the form

$$\hat{\bar{Y}}' = a\bar{z}_{pn} + b\bar{z}_{qn} + c\bar{z}'_{pn} + d\,\bar{z}'_{qn} \tag{7.3}$$

$$E\,\hat{\bar{Y}}' = a\,E(\bar{z}_{pn}) + b\,E(\bar{z}_{qn}) + c\,E(\bar{z}'_{pn}) + d\,E(\bar{z}'_{qn})$$

$$= (a+b)\,\bar{Y} + (c+d)\,\bar{Y}' . \tag{7.4}$$

If we want that $\hat{\bar{Y}}'$ should be an unbiassed estimate for $\bar{Y}'$ we must have

$$a + b = 0 \quad \text{and} \quad c + d = 1.$$

Thus

$$\hat{\bar{Y}}' = a(\bar{z}_{qn} - \bar{z}_{pn}) + c\,\bar{z}'_{pn} + (1-c)\,\bar{z}'_{qn} . \tag{7.5}$$

The variance of $\hat{\bar{Y}}'$ is

$$V(\hat{\bar{Y}}') = a^2(1/pn + 1/qn)\,\sigma_z^2 + (c^2/pn)\,\sigma_z^2 + \frac{(1-c)}{qn}\,\sigma_{z'}^2 + \frac{2ac}{pn}\,\rho_{zz'}\sigma_{z'}\sigma_z \tag{7.6}$$

where

$$\sigma_z^2 = \sum_1^N P_i\,(z_i - \bar{z})^2$$

$$\sigma_{z'}^2 = \sum_1^N P_i\,(z'_i - \bar{z}')^2 \tag{7.7}$$

and

$$\rho_{zz'}\sigma_z\sigma_{z'} = \sum_1^N P_i(z_i - \bar{z})(z'_i - \bar{z})$$

where

$$\bar{z} = Ez = \sum_1^N y_i P_i/NP_i = \bar{Y}$$

and

$$\bar{z}' = Ez' = \sum_1^N y'_i P_i/NP_i = \bar{Y}'.$$

We wish to choose the values of a and c that minimise $V(\bar{Y}')$. Differentiating $V(\bar{Y}')$ with respect to a and c, we have

$$2a(1/pn + 1/qn)\sigma_z^2 + 2c\rho_{zz'}\sigma_z\sigma_{z'}/pn = 0$$

$$2c\sigma_z^2/pn - 2(1-c)\sigma_z^2/qn + 2a\rho_{zz'}\sigma_z\sigma_{z'}/pn = 0$$

Hence $$a = \frac{\rho_{zz'}\,pq}{1-q^2\rho^2_{zz'}}\;\frac{\sigma_{z'}}{\sigma_z}$$

and $$c = p/(1-q^2\rho^2_{zz'})$$

(7.8)

Thus the estimator with the optimum values for a and c may be written

$$\hat{\bar{Y}}' = \frac{\rho_{zz'}pq}{1-q^2\rho^2_{zz'}}\;\frac{\sigma_{z'}}{\sigma_z}(\bar{z}_{qn} - \bar{z}_{pn}) + \frac{p}{(1-q^2\rho^2_{zz'})}\bar{z}'_{pn}$$

$$+ \frac{q(1-q\rho^2_{zz'})}{1-q^2\rho^2_{zz'}}\;\bar{z}'_{qn}$$

(7.9)

and the variance of this estimate is

$$\frac{\sigma^2_{z'}}{n}\;\frac{1-\rho^2_{zz'}\,q}{1-\rho^2_{zz'}\,q^2}$$

(7.10)

Now we want to determine that value of q for which this variance is minimum. Therefore differentiating with respect to q and equating to zero, we have

$$\rho_{zz'}^{4}\, q^{2} - 2\,\rho_{zz'}^{2}\, q + \rho_{zz'}^{2} = 0$$

$$\text{or } \rho_{zz'}^{2} q^{2} - 2q + 1 = 0$$

Hence

$$q = \frac{1 \pm \sqrt{1 - \rho_{zz'}^{2}}}{\rho_{zz'}^{2}}$$

since + value is not admissible, we have

$$q = \frac{1 - \sqrt{1 - \rho_{zz'}^{2}}}{\rho_{zz'}^{2}} \qquad (7.11)$$

SUMMARY

In the present study, sampling on two occasions has been extended to cover the case of double sampling for regression and stratification. The particular case of double sampling for stratification when there are only two strata, one consisting of only zero units has been investigated in greater detail. The case of sampling with varying probabilities of selection has also been included.

The optimum solutions for n', n and q has been obtained by the method of successive approximation.

The behaviour of the efficiency of double sampling for variation of p_1 and δ^2 has been examined seperately. It has been found that for p_1 small, double sampling is highly efficient than simple random sampling. For p_1 very high double sampling is not very efficient and if it is applied at all, the size of preliminary sample should be kept as small as possible. Similarly, the efficiency of double sampling is very high compared to simple random sampling if δ^2 is small and decreases first rapidly and then gradually. After a certain limit, the fall in efficiency is negligible even if increase in δ^2 is significant.

The gain in efficiency of double sampling over simple random pling, has been estimated for 15 sub-divisions of coconut areas of Assam and it has been found that except in two ons, double sampling is more efficient than simple ng.

REFERENCES

Jessen, R.J. (1942) Statistical Investigation of a sample Survey for Obtaining Farm Facts; Iowa Agr. Exp. Sta. Res. Bull. 304.

Yates, F. (1949) Sampling Methods for Censuses and Surveys, Charless Griffin and Co., London.

Patterson, H.D. (1950) Sampling on Successive Occasions with Partial Replacement of Units, J. Roy. Stat. Soc., Series B,12.

Tikkiwal, B.D. (1953) Theory of Successive Sampling Jour. Ind. Soc. Agr. Stat.,5.

Sukhatme, P.V. (1953) Sampling Theory of Surveys with Applications. Indian Soc. of Agri. Stat., New Delhi. The Iowa State College Press, Ames, Iowa, U.S.A.

Tikkiwala, B.D. (1956) Some Further Contribution to the Theory of Univariate Sampling on Successive Occasions; Jour. Ind. Soc. Agr. Stat.8.

Hansen and Hurwitz (1956) Sample Survey Methods and Theory. John Wiley Sons, New York.

Singh, D. (1959) Paper read at the 12th Annual Meeting of the Indian Society of Agricultural Statistics held at Gwalior.

Kathuria O.P. (1959) Thesis submitted to I.C.A.R. for award of Diploma.

Department of Economics and Statistics Govt. of Assam. (1959) Report on the sample survey to estimate area under and production of coconut in Assam.