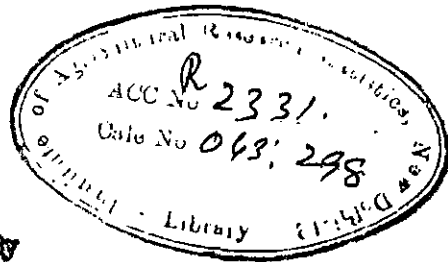


V.88 184 195
188

**Successive Sampling with Varying Probabilities
and
Optimal Use of Ancillary Variates**



By

Prem Prakash

Dissertation submitted in fulfilment of the requirements
for the award of Diploma in Agricultural and Animal
Husbandry Statistics of the Institute of Agricultural
Research Statistics (I.C.A.R.)

New Delhi

1966

DEC-04
4

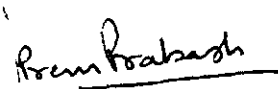
A_C_K_N_O_W_L_E_D_G_E_M_E_N_T

I have grate pleasure in expressing my deepest sense of gratitude to Dr. D. Singh, Deputy Statistical Adviser, Institute of Agricultural Research Statistics (I.C.A.R.), New Delhi for his valuable guidance, keen interest, constant encouragement and constructive criticism throughout the course of investigation and preparation of manuscripts.

I also wish to express my thanks to Dr. G.R. Seth, Statistical Adviser (I.C.A.R.), for providing me adequate facilities during the course of investigation.

I.A.R.S.

New Delhi
1966


(PREM PRAKASH)

C_O_N_T_E_N_T_S

	Page
(1) Introduction	1
(2) Use of Varying Probabilities at the 2nd Occasion.	11
(3) Use of Varying Probabilities at the 1st Occasion.	24
(4) Use of Varying Probabilities at both the Occasions.	43
(5) Optimal Use of Ancillary Variables	46
(6) Summary	63
(7) References	

INTRODUCTION

In the case of dynamic population i.e. the one, which is subject to change from time to time, a census at frequent intervals is of limited use, therefore the practice of relying on the samples for the collection of important series of data that are published at regular intervals is becoming more common. For example a highly precise information about the characteristics of a population in 1950 and 1960 may not help much in planning, that demands a knowledge of the population in 1970 or in any other year. A series of relatively small samples at annual or even shorter intervals may be more serviceable.

PLANNING OF REPEATED SURVEY:

In planning a sampling enquiry the entire relevant information available should be judiciously used. In repeated enquiries it is the use of cumulative information collected through the successive surveys that increases the reliability of the estimate. Besides using the field and organisation experience it is always advantageous to effectively use the past data for improving the estimate. In the case of dynamic population, such as extent of area under improved seeds, the number of unemployed persons in a country, prevalence of chronic diseases etc., (i) a single survey on a particular occasion gives information, about the properties of the population for that occasion only, and it does not give any information on the nature of, or rate of changes, which occur in the dynamic population,

but many a time, the interest of the experimenter 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 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. To meet these requirements, the survey will have to be repeated on several occasions.

If the experimenter is free to alter or retain the composition of the sample, and the total size of the sample is to be the same on all occasions there are several alternatives, one can choose in designing the sampling enquiry of this type such as

- 1) a new sample on each occasion (in the case of estimating overall occasions),
- 2) a fixed sample on all occasions (in the case of estimating the change)
- 3) a partial replacement of units from occasion to occasion (For current estimate).

Generally experimenter remains interested in alternative (3). This type of sampling is called sampling on successive occasions with partial replacement, (Patterson, Yates), a sampling for time series (Hansen, Hurwitz and Madow), and Rotation sampling (Wilks and A.R. Eckler). It refers to the process of eliminating some of the old elements from the sample and adding new elements to the sample, each time a new sample is drawn.

First attempt, on sampling, on successive occasions with partial replacement, was made by Jessen (1942). His work was confined to two occasions only. He considered two independent estimates for the mean on the 2nd occasion, one on the basis of the units common to both occasions and another based on the units selected a fresh. The information on the first occasion was utilized for the estimate at the 2nd occasion. 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 2nd occasion .

Yates (1942) contends that for estimating the value of the population mean on two successive occasions, it is better to treat each occasion separately, following whatever method of estimation is appropriate to the sample obtained on that occasion, regardless of the values obtained on the other occasions. Such estimates he termed as overall estimates. For two occasions he considered a sub-sample of the original sample as also a sample with some of the units retained from the previous occasion and some taken a fresh on the 2nd occasion. Yates has also extended his results to h occasions ($h > 2$) and has built an estimate for the population mean on the h th occasion, by taking into account the results upto and including the $(h - 1)$ th occasion, under the limitations

- 1) a given fraction of units is replaced on each occasion.

ii) the variability on the different occasions is constant i.e., $\sigma_h^2 = \sigma_{h-1}^2 = \sigma^2$ for all h and the correlation ρ between the units on successive occasions is constant,

iii) The correlation between the units occasions two apart is ρ^2 , and that between the units occasions three apart is ρ^3 etc.

He obtained the relation

$$\bar{Y}_h = (1 - \beta_h) \left\{ \bar{y}_h + \rho (\bar{Y}_{h-1} - \bar{y}_{h-1}) \right\} + \beta_h (\bar{y}_h)$$

where \bar{Y}_h is the most accurate estimate when can be obtained for occasion h , taking into account the result of sampling up to and including the occasion h , \bar{Y}_{h-1} is a similar estimate for the previous occasion, single dashes denote units common to occasion h and $(h-1)$, the mean on earlier occasion is distinguished by square brackets, and double dashes units occurring on h occasion only. The value of β_h varies from occasion to occasion and it depends upon the values of r and the fraction f replaced on each occasion.

Patterson (1950) approach for successive sampling was in a different and slightly more general way. He first built an estimate as a suitable linear function of a set of variates and then developed a set of conditions for this estimate to be the most efficient. Using these conditions he then determined an efficient estimate of the mean on the h th occasion, which comes out to be the same as given by Yates. With this

set of conditions he also established a recurrence relation between the weights β_h and β_{h-1} as

$$(1 - \beta_h) (1 - \beta_{h-1}) - (\alpha + \beta) (1 - \beta_h) + \alpha\beta = 0$$

where α, β are the roots of the quadratic equation obtained by putting $\beta_h = \beta_{h-1} = \beta$.

When h is very large, the limiting value of $\beta_h = \beta$ is given by

$$\beta = \frac{-(1 - \beta^2) + \sqrt{(1 - \beta^2)^2 / (2 - \beta^2) (1 - 4\lambda\mu)}}{2\lambda\beta^2}$$

where $\lambda + \mu = 1$.

Patterson also gave efficient estimates 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. Patterson also showed that the change $Y_h - h Y_{h-1}$ (where $h Y_{h-1}$ be the refined estimate for the $(h - 1)$ th occasion) is more efficient than the one given by Yates.

Tikkival (1953) approach was the same as given by Yates and Patterson, but his approach was more general than his predecessors. He allowed the correlation between units taken on two successive occasions to vary, but assumed the correlation between units two or more than two occasions apart is equal to the product of correlations between units on all

pairs of consecutive occasions formed by these. In case, the sample size and all the correlations were assumed to be equal on all occasions, he proved that with limiting ϕ , the replacement to be affected on different occasions is 50 percent from the above, i.e. under the conditions imposed the replacement fraction is always $\geq 1/2$.

Ecklar (1955) clarified Patterson's fundamental method for finding minimum variance linear unbiased estimates and extended his method to two level and three level rotation sampling. He compared three methods of rotation sampling on a cost basis and showed how the one level rotation sampling estimate of greatest practical interest could be derived from the two level estimate. He also extended Cochran's work in determining optimum patterns for the one level rotation sampling estimate.

Tikkiwal (1956) has shown that when the correlation and regression coefficients are estimated from the common units between two consecutive occasions, Y_h is still a consistent estimator of the population mean μ_h on the h th occasion, and its bias tends to zero with increasing sample sizes on h occasions. Its variance will in general be greater than the variance of the estimator where the correlations are known in advance and the weights (ϕ_h) themselves become functions of parameters to be estimated from the sample.

D. Singh (1959) investigated the problem of partial replacement in Multi-stage design and gave the expression,

for the mean at 2nd and 3rd occasion and its variances and mean over a period of time, when a two stage design is repeated on two occasions with partial replacement of first stage units only. This aspect of the problem apart from statistical point of view, has practical advantages, since in actual practice, 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.

Kathuria (1960) extended the case of two stage sampling repeated on two occasion with partial replacement of first stage units, to h occasions. He also considered the case of sampling on two occasions with replacement, among 2nd stage units also. He investigated the problem of optimum allocation for a given cost function.

B.D. Singh (1962) considered the case when double sampling is repeated on two occasions with partial replacement of units on 2nd occasion.

In a series of papers (1958, 1960, 1964, 1965) Tikkiwal has extensively studied various aspects of successive sampling.

B.D. Singh and D. Singh (1965) considered the case when the information on the ancillary variate in the preliminary sample is used for selecting the units for sub-sample with unequal probabilities.

Information obtained from the previous occasions can be used for forming ratio and regression estimates, for stratifying the population, and for selecting the sampling units with unequal probability. But generally in the literature,

a ratio or regression coefficient is calculated, with the help of the information available from the entire sample on the previous occasions, and a ratio or regression estimate is obtained for the character, under study, at the current occasion. This estimate is pooled with the one, obtained on the basis of the sample selected afresh on the current occasion. Unless the sample size on each occasion is large, the sample size retained on subsequent occasions may not be adequate. The application of regression estimate may not be valid as the use of regression method estimation is based on the concept of large sample theory. Besides this regression estimates give better results under the linear model only, and if the population is changing from time to time the change may not be always linear.

Hence to avoid these limitations the information on the previous occasion may be utilized for selecting the sampling units with probability proportional to their size as measured by the values of the character observed on the previous occasions. The estimates obtained in this way are unbiased and have neat expression for variance.

In general in a sampling design the selection of units with varying probabilities without replacement leads to more efficient estimate than with replacement, because if a unit is selected twice it will not give any additional information as compared to the situation when it might have been sampled

only once. Narain (1951) compared the two systems of sampling in a two stage sampling design and found the necessary condition. Durbin (1953) stated (without proof) that it is not difficult to find out some situation in which sampling without replacement will be less efficient as compared to that with replacement. D. Singh (1954) also studied the problem.

In the case of sampling with unequal probability with replacement, estimates and their variances can be easily obtained. But the theory of unequal probability without replacement involves mathematical complexity, computational difficulties and some times the estimate of error variance is negative. A number of research workers (Midsuno (1950), Narain (1951), Hurwits and Thompson (1952), Durbin (1953), Yates and Grundy (1953), D. Singh (1954, 1959), Des Raj (1958), Hartley and Rao (1962), Hartley, Rao and Cochran (1962) have attempted to simplify the theory, so that it can be easily be adopted in practice.

In successive sampling when we use the information, on the previous occasion, for unequal probability. If the units are selected with replacement it presents not so much difficulty, but when the units are selected without replacement, this difficulty may be overcome by using the Hartley, Rao and Cochran method.

The technique of unequal probability, in the successive sampling, can be used in three ways. (1) On the 1st occasion units are selected with simple random sampling, and on the 2nd occasion the sub-sample to be

retained may be selected with probability proportional to the sizes of the units as observed on previous occasion.

(2) When some information on auxiliary variable is given, on the 1st occasion the units can be selected with probability proportional to the sizes of the auxiliary variables, and at the second occasion, the units may be retained with

S.R.S. (3) When some information on auxiliary variable is given, on the 1st occasion the units can be selected with probability proportional to the sizes of the auxiliary variable and at the 2nd occasion, the sub-sample to be retained may be selected with probability proportional to the sizes of the units as observed on previous occasion.

Here all these three cases have been studied and the comparison is made with the regression estimate.

In the last the problem on the optimal use of auxiliary variates, has been tried, when there are more than one auxiliary characters correlated with the variate under consideration, the techniques of ratio, regression, unequal probability, stratification, can be combined to produce considerably more efficient estimates.

Use of Varying Probabilities at the 2nd Occasion.

2.1 In successive sampling, it is generally preferable to repeat a fraction of the units observed on the earlier occasion. The information collected on the previous occasion can be utilized for the succeeding occasions. The usual practice is to use the ratio or regression method for the utilization of the information available on the previous occasion. But if the sample size retained on the subsequent occasion is small, the bias in these estimators may be considerable and it often outweighs the gain in precision. When a fraction of units have been repeated in succeeding occasions, it is always desirable to make use of the entire information available from the enquiry on earlier occasions, instead of ratio or regression technique the information collected from the previous occasion can be utilized for selecting the units at the 2nd occasion i.e. the sub-sample retained may be selected with probability proportional to the sizes of the units observed on the previous occasion.

2.2 Sampling scheme:- For sampling on successive occasions in a unistage sampling, let us select n units with s. r. s. at first occasion, and on the 2nd occasion we select np units out of n units without replacement with varying probabilities, the probabilities being proportional to the sizes of the units observed on the 1st occasion. Remaining $n - np = nq$ units are selected with s. r. s. from the population itself.

2.3 Let y_i be the value of the character for the i th unit of the population. If the sample on the first occasion is observed as $y_1^{(1)}, y_2^{(1)} \dots y_i^{(1)} \dots y_n^{(1)}$ having mean as $\bar{y}_n^{(1)}$ on the second occasion for selecting np units out of n , we divide the n units into np groups, randomly.

The size of the r th group being n_r , and select one unit from each group with varying probability, the probability of selection of i th unit is $p_i = y_i^{(1)} / \sum_{i=1}^n y_i^{(1)}$.

The effective probability for the i th unit in the r th group will be p_i / π_r where $\pi_r = \sum_{i=1}^n p_i$

$$\text{also } p_i / \pi_r = y_i^{(1)} / \sum_{i=1}^n y_i^{(1)}$$

The effective probability is known in terms of sample values.

2.4 Now if we define

$$(1) s_i^{(2)} = y_i^{(2)} / n_r p_i / \pi_r$$

suffix (1) and (2) denote the occasion, 1st and 2nd.

The estimate on the basis of these np units is given

by

$$\hat{\bar{y}}_{np}^{(2)} = \bar{s}_{np}^{(2)} = \frac{1}{np} \sum_{r=1}^{np} s_r^{(2)}$$

Expected value of $\bar{y}_{np}^{(2)}$ is given by

$$E(\bar{y}_{np}^{(2)}) = E(\bar{s}_{np}^{(2)}) = E\left\{\frac{1}{np} \sum_{r=1}^{np} s_r^{(2)}\right\}$$

$$E\left\{\bar{s}_{np}^{(2)}\right\} = E_1\left\{E_2\left(\bar{s}_{np}^{(2)}\right)/n\right\}$$

$$= E_1\left[E_2\left\{E_0\left(\frac{1}{np} \sum_{r=1}^{np} s_r^{(2)}\right)/n_1, n_2, \dots, n_r, \dots, n_{np}\right\}\right]$$

where E_0 denotes the conditional expectation when the population is grouped into np groups of sizes $n_1, n_2, \dots, n_r, \dots, n_{np}$.

Therefore

$$\begin{aligned} E(\bar{y}_{np}^{(2)}) &= E_1\left[E_2\left\{E_0\left(\frac{1}{np} \sum_{r=1}^{np} \frac{y_r^{(2)}}{\left(\frac{n_r}{n} p_r / \pi_r\right)}\right)/n_1, \dots, n_r, \dots, n_{np}\right\}\right] \\ &= E_1\left\{E_2 \frac{1}{np} \sum_{r=1}^{np} (\bar{y}_{n_r}^{(2)})/n\right\} \end{aligned}$$

where $\bar{y}_{n_r}^{(2)}$ is the mean of the r th group:

$$= E_1\left\{\frac{1}{np} \sum_{r=1}^{np} E_2\left(\bar{y}_{n_r}^{(2)}\right)/n\right\}$$

$$= E_1\left\{\frac{1}{np} \sum_{r=1}^{np} (\bar{y}_n^{(2)})\right\} = \frac{1}{np} \sum_{r=1}^{np} E_1(\bar{y}_n^{(2)})$$

$$= \bar{y}_n^{(2)} \dots \dots \dots (2.41)$$

where $\bar{y}_n^{(2)}$ is the mean of the units on the 2nd occasion which are sampled on 1st occasion.

2.5 Variance of the estimator

$$\begin{aligned}
 V(\bar{y}_{np}^{(2)}) &= V_1 E_2 \left\{ \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} / . n \right\} \\
 &\quad + E_1 V_2 \left\{ \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} / . n \right\} \\
 &= V_1 E_2 \left\{ E_0 \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} \right\} + E_1 \left\{ V_2 E_0 \left\{ \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} \right\} \right. \\
 &\quad \left. + E_2 V_0 \left\{ \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} \right\} \right\}
 \end{aligned}$$

where the suffix c denotes the conditional variance and expectation under the condition that n is divided into np groups.

1st term:

$$\begin{aligned}
 &V_1 E_2 \left\{ E_0 \frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} / n_1 \dots n_r \dots n_{np} \right\} \\
 &= V_1 E_2 \left\{ \frac{1}{np} \sum_{r=1}^{np} \bar{y}_{n_r}^{(2)} / . n \right\} \\
 &= V_1 \left(\frac{1}{np} \sum_{r=1}^{np} \bar{y}_n^{(2)} \right) = V_1 \left(\bar{y}_n^{(2)} \right) \\
 &= \frac{N \sigma^2}{N n} S_y^2(2) \dots \dots \dots (2.51)
 \end{aligned}$$

2nd term:

$$\begin{aligned}
 &E_1 \left\{ V_2 E_0 \left(\frac{1}{np} \sum_{r=1}^{np} z_r^{(2)} \right) / n_1, n_2 \dots n_r \dots n_{np} \right\} \\
 &= E_1 \left[\left\{ V_2 \left(\frac{1}{np} \sum_{r=1}^{np} \bar{y}_{n_r}^{(2)} \right) \right\} / . n \right]
 \end{aligned}$$

$$\begin{aligned}
 &= E_1 \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} V_2 \left(\bar{y}_{n_r}^{(2)} \right) + \sum_{r \neq k} \sum_k \text{Cov.} \left(\bar{y}_{n_r}^{(2)}, \bar{y}_{n_k}^{(2)} \right) \right\} \\
 &= \frac{1}{(np)^2} E_1 \left\{ \sum_{r=1}^{np} \left(\frac{1}{n_r} - \frac{1}{n} \right) s_{y(2)}^2 \right. \\
 &\quad \left. + \sum_{r \neq k} \sum_k \text{Cov.} \left(\frac{1}{n_r} \sum_{i=1}^{n_r} y_i^{(2)}, \frac{1}{n_k} \sum_{i=1}^{n_k} y_i^{(2)} \right) \right\}
 \end{aligned}$$

where $r y_i^{(2)}$: i th unit in the r th group

$k y_i^{(2)}$: i th unit in the k th group

$$\text{and } s_{y(2)}^2 = \frac{1}{n-1} \sum_{i=1}^n \left(y_i^{(2)} - \bar{y}_n^{(2)} \right)^2$$

$$= \frac{1}{(np)^2} E_1 \left\{ \sum_{r=1}^{np} \left(\frac{1}{n_r} - \frac{1}{n} \right) s_{y(2)}^2 \right.$$

$$\left. + \sum_{r \neq k} \sum_k \frac{1}{n_r n_k} \text{Cov.} \left(r y_i^{(2)}, k y_i^{(2)} \right) \right\}$$

$$= \frac{1}{(np)^2} E_1 \left\{ \sum_{r=1}^{np} \left(\frac{1}{n_r} - \frac{1}{n} \right) s_{y(2)}^2 + \sum_{r \neq k} \sum_k \left(-\frac{1}{n} \right) s_{y(2)}^2 \right\}$$

$$= \frac{1}{(np)^2} E_1 \left\{ \sum_{r=1}^{np} \left(\frac{1}{n_r} - \frac{1}{n} \right) s_{y(2)}^2 - \frac{np(np-1)}{n} s_{y(2)}^2 \right\}$$

$$= E_1 \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{n_r} - \frac{1}{n} \right\} s_{y(2)}^2$$

$$= \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{n_r} - \frac{1}{n} \right\} \sum_{j=1}^R y_j^{(2)} \quad \dots (2.52)$$

where $S_{y_j^{(2)}} = \frac{1}{n-1} \sum_{i=1}^n (y_i^{(2)} - \bar{y}_n^{(2)})^2$

3rd Term:

$$E_1 \left[E_2 \left\{ V_0 \frac{1}{np} \sum_{r=1}^{np} (n_r^{(2)}) / n_1, n_2, \dots, n_r, \dots, n_{np} \right\} \right]$$

Now

$$V_0 (n_r^{(2)}) = \sum_{i=1}^{n_r} \frac{p_i}{\pi_i} \left(\frac{y_i^{(2)}}{p_i \pi_i} - \bar{y}_{n_r}^{(2)} \right)^2$$

$$= \frac{1}{(n_r)^2} \left\{ \sum_{i=1}^{n_r} \pi_i \frac{y_i^{(2)2}}{p_i} - n_r \bar{y}_{n_r}^{(2)2} \right\}$$

$$= \frac{1}{(n_r)^2} \sum_{i < i'}^{n_r} p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2$$

The probability that a pair of i 'th and i' 'th unit will be in the same group of size n_r out of n is $\frac{n_r(n_r-1)}{n(n-1)}$

Therefore $E_1 E_2 V_0 (n_r^{(2)}) =$

$$E_1 \left[E_2 \left\{ \frac{1}{(n_r)^2} \sum_{i < i'}^{n_r} p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2 \right\} / n \right]$$

$$= E_1 \frac{1}{(n_r)^2} \frac{n_r(n_r-1)}{n(n-1)} \sum_{i < i'}^n p_i p_{i'} \left\{ \frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right\}^2$$

Now the probability of selecting a pair of units in the sample of n out of N is given by $\frac{n(n-1)}{N(N-1)}$

Therefore

$$\begin{aligned}
 E_1 &= \frac{1}{(n_r)^2} \frac{n_r (n_r - 1)}{n(n-1)} \sum_{i < i'}^N P_i P_{i'} \left(\frac{y_i^{(2)}}{P_i} - \frac{y_{i'}^{(2)}}{P_{i'}} \right)^2 \\
 &= \frac{1}{(n_r)^2} \frac{n_r (n_r - 1)}{n(n-1)} \frac{n(n-1)}{N(N-1)} \sum_{i < i'}^N P_i P_{i'} \left(\frac{y_i^{(2)}}{P_i} - \frac{y_{i'}^{(2)}}{P_{i'}} \right)^2 \\
 &= \frac{(n_r - 1)}{(n_r)N(N-1)} \left(\sum_{i=1}^N \frac{y_i^{(2)2}}{P_i} - y_N^{(2)2} \right) \\
 E_1 \left[E_2 \left\{ V_0 \left(\sum_{r=1}^{np} n_r \right) \right\} / n \right] \\
 &= \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{(n_r)^2} \frac{n_r(n_r-1)}{n(n-1)} \frac{n(n-1)}{N(N-1)} \left\{ \sum_{i=1}^N \frac{y_i^{(2)2}}{P_i} - y_N^{(2)2} \right\} \\
 &= \frac{1}{(np)^2 N(N-1)} \sum_{r=1}^{np} \frac{(n_r-1)}{n_r} \sum_{i=1}^N P_i \left\{ \frac{y_i^{(2)}}{P_i} - y_N^{(2)} \right\}^2 \\
 &\dots\dots(2.53)
 \end{aligned}$$

Therefore

$$\begin{aligned}
 V \left(\frac{y}{n_p} \right)^{(2)} &= \frac{N-n}{Nn} S_{y^{(2)}}^2 + S_{y^{(2)}}^2 \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{n_r} - \frac{1}{N} \right\} \\
 &+ \frac{1}{(np)^2 N(N-1)} \sum_{r=1}^{np} \frac{(n_r-1)}{n_r} \sum_{i=1}^N P_i \left\{ \frac{y_i^{(2)}}{P_i} - y_N^{(2)} \right\}^2 \\
 &= V_1 \text{ (say)} \dots\dots(2.54)
 \end{aligned}$$

2.6 Now if we define a second estimator:

$$(11) \quad s'_1(2) = \frac{y_1^{(2)}}{\frac{1}{n} \frac{p_1}{\prod r}} \quad \text{where} \quad \bar{n} = \frac{n}{np} = \frac{np}{r} \frac{n_r}{np}$$

$$\bar{y}'_{np}(2) = \bar{s}'_{np}(2) = \frac{1}{np} \sum_{r=1}^{np} s'_r(2)$$

Expected value of $\bar{y}'_{np}(2)$ is given by

$$E(\bar{y}'_{np}(2)) = E(\bar{s}'_{np}(2)) = E_1 \{ E_2(\bar{s}'_{np}(2)) \} / . n \}$$

$$= E_1 \left\{ E_2 \left(\frac{1}{np} \sum_{r=1}^{np} E_3 \left(\frac{y_1^{(2)}}{\frac{1}{n} \frac{p_1}{\prod r}} \right) / n_1, n_2, \dots, n_r, \dots, n_{np} \right) \right\}$$

$$= E_1 \left\{ E_2 \left(\frac{1}{np} \sum_{r=1}^{np} \frac{1}{\bar{n}} \sum_{i=1}^{n_r} y_i \right) / . n \right\}$$

$$= E_1(\bar{y}'_n(2)) = \bar{y}'_N(2) \quad \dots \dots \dots (2.61)$$

$$V(\bar{y}'_{np}(2)) = E_1 \left\{ V_2 \left(\frac{1}{np} \sum_{r=1}^{np} s'_r(2) \right) / n \right\}$$

$$= V_1 \left\{ E_2 \left(\frac{1}{np} \sum_{r=1}^{np} s'_r(2) \right) / . n \right\}$$

1st term:

$$\begin{aligned}
 & E_1 \left\{ V_2 \left(\frac{1}{np} \sum_{r=1}^{np} s_r^{(2)} \right) / n \right\} \\
 &= E \left[E_2 \left\{ N_2 \left(\frac{1}{np} \sum_{r=1}^{np} s_r^{(2)} \right) / n_1 \cdot n_2 \dots n_r \dots n_{np} \right\} \right] \\
 & V_0 (s_r^{(2)}) = \frac{1}{(n_r)^2} \sum_{i < i'}^{n_r} p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2 \\
 & E_1 \left[E_2 \left\{ \frac{1}{(n_r)^2} \sum_{i < i'}^{n_r} p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2 / n \right\} \right] \\
 &= E_1 \left\{ \frac{n_r (n_r - 1)}{(n_r)^2 n(n-1)} \sum_{i < i'}^n p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2 \right\} \\
 &= \frac{n(n-1)}{N(N-1)} \frac{n_r (n_r - 1)}{(n_r)^2 n(n-1)} \sum_{i < i'}^N p_i p_{i'} \left(\frac{y_i^{(2)}}{p_i} - \frac{y_{i'}^{(2)}}{p_{i'}} \right)^2 \\
 &= \frac{(n_r - 1)}{(n_r)^N N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - N \bar{y}_N^{(2)} \right)^2 \\
 & E_1 \left[E_2 \left\{ V_0 \left(\frac{1}{np} \sum_{r=1}^{np} s_r^{(2)} \right) / n_1 \cdot n_2 \dots n_r \dots n_{np} \right\} \right] \\
 &= \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{(n_r - 1)}{n_r N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - y_N^{(2)} \right)^2 \\
 &= \sum_{r=1}^{np} \frac{n_r (n_r - 1)}{n^2 N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - y_N^{(2)} \right)^2
 \end{aligned}$$

$$= \frac{\sum_{r=1}^{np} n_r^2 - n}{n^2 N (N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2 \dots (2.62)$$

2nd term

$$\begin{aligned} & V_1 \left(E_2 \frac{1}{np} \sum_{r=1}^{np} n_r^{(2)} / n \right) \\ &= V_1 \left(\frac{1}{np} \sum_{r=1}^{np} \bar{y}_n^{(2)} \right) = V_1 \left(\bar{y}_n^{(2)} \right) \\ &= \frac{N-n}{Nn} S_y^2(2) \dots (2.63) \end{aligned}$$

Therefore

$$\begin{aligned} V \left(\bar{y}'_{np} \right) &= \frac{\sum_{r=1}^{np} n_r^2 - n}{n^2 N (N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2 \\ &+ \frac{N-n}{Nn} S_y^2(2) \\ &= V'_1 \text{ (say) } \dots (2.64) \end{aligned}$$

2.7 Comparison of the two estimates, $\bar{y}_{np}^{(2)}$ and $\bar{y}'_{np}^{(2)}$.

The estimate $\bar{y}'_{np}^{(2)}$ will be more efficient as compared to the estimate $\bar{y}_{np}^{(2)}$ if,

$$V'_1 < V_1$$

$$\text{or } \frac{N-n}{Nn} s_y^2(2) + \sum_{r=1}^{np} \frac{n_r^2 - n}{n^2 N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - N \bar{y}_N^{(2)} \right)^2$$

$$< \frac{N-n}{Nn} s_y^2(2) + s_y^2(2) \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{n_r} - \frac{1}{n} \right\}$$

$$\bullet \frac{1}{(np)^2 N(N-1)} \sum_{r=1}^{np} \frac{n_r - 1}{n_r} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$$

$$\text{or } \sum_{r=1}^{np} \frac{n_r^2 - n}{n^2 N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - N \bar{y}_N^{(2)} \right)^2$$

$$< s_y^2(2) \left\{ \frac{1}{(np)^2} \sum_{r=1}^{np} \frac{1}{n_r} - \frac{1}{n} \right\}$$

$$\bullet \frac{1}{(np)^2 N(N-1)} \sum_{r=1}^{np} \frac{n_r - 1}{n_r} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$$

If n is a multiple of np i.e. all n_r are equal and $n_r = \frac{n}{np} = \bar{n}$.

$$\text{Then L.H.S.} = \text{R.H.S.} = \frac{n - np}{n \cdot (np) N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$$

$$\text{or } V_1' = V_1 = V_A \text{ (say)} = \frac{N-n}{Nn} s_y^2(2)$$

$$\bullet \frac{n - np}{n \cdot np \cdot N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$$

When n is not a multiple of np , we have $n = np \cdot R + k$,
 where $0 < k < np$ and R is a positive integer. Then
 we choose $n_1 = n_2 = \dots = n_k = R + 1$; $n_{k+1} = n_{k+2} = \dots = n_{np} = R$

$$\text{Let } \frac{2}{N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - y_N^{(2)} \right)^2 = U^2$$

$$\text{Then } \left(\frac{1}{np} = \frac{1}{n} + \frac{k(np-k)}{np \cdot n^2} \right) U^2$$

$$< s_y^2(2) \left\{ \frac{1}{(np)^2} \left(\frac{k}{1} \frac{1}{R+1} + \frac{np}{k+1} \frac{1}{R} \right) - \frac{1}{n} \right\}$$

$$+ \frac{U^2}{(np)^2} \left\{ \frac{k(R+1-1)}{R+1} - \frac{(np-k)(R-1)}{R} \right\}$$

$$\text{or } \left\{ \frac{(n-k)(n-np+k)}{np \cdot n^2} \right\} U^2 < s_y^2(2) \left\{ \frac{k(np-k)}{n(n-k)(n+np-k)} \right\}$$

$$+ \frac{U^2}{np} \left\{ \frac{(n-np-k)}{n-k} + \frac{k \cdot np}{(n-k)(n+np-k)} \right\}$$

$$\text{or } \frac{U^2}{np} \left\{ \frac{(n-k)^2 \left\{ n^2 - (np-k)^2 \right\} - n^2 \left\{ n^2 - n^2 p^2 - 2nk + k^2 + knp \right\}}{n^2 (n-k)(n+np-k)} \right\}$$

$$< s_y^2(2) \left\{ \frac{k(np-k)}{n(n-k)(n+np-k)} \right\}$$

$$\text{or } \frac{U^2}{np \cdot n} \left\{ k (np-k) \left\{ n^2 + (2n-k)(np-k) \right\} \right\}$$

$$< s_y^2(2) \left\{ k (np-k) \right\}$$

$$\text{or } \frac{U^2}{np \cdot n} \left\{ n^2 + (2n-k)(np-k) \right\} < s_y^2(2)$$

$$\text{or } \frac{U^2}{s_y^2(2)} < \frac{np \cdot n}{n^2 + (2n-k)(np-k)} < 1. \dots\dots (2.62)$$

If we assume that $y_1^{(2)}, y_2^{(2)}, \dots, y_N^{(2)}$, the finite

population is a random sample from an infinite population, then under the linear model, Desh Raj (1959) has shown that

$$E(U^2) < E(s_y^2(2))$$

$$\text{or } \frac{E(U^2)}{E(s_y^2(2))} < 1.$$

Taking the expectations in (2.62) we see that the inequality holds good.

$$\text{Hence } V_1' < V_1$$

The variance of the estimates will have minimum value when all n_r 's are equal i.e. when $V_1' = V_1 = V_A$.

2.7 Estimate on the basis of n - np = nq units

The estimate of the mean, based on nq units is

$$\bar{y}_{n-np}^{(2)} = \bar{y}_{nq}^{(2)} = \frac{1}{n-np} \sum_{i=1}^{n-np} y_i^{(2)} = \frac{1}{nq} \sum_{i=1}^{nq} y_i^{(2)} \dots\dots\dots(2.71)$$

$$E \left(\bar{y}_{nq}^{(2)} \right) = \bar{y}_N^{(2)}$$

and the variance of the estimate is

$$V \left(\bar{y}_{nq}^{(2)} \right) = \frac{N - nq}{n \cdot nq} s^2_{y^{(2)}} = V_B \text{ (Say)} \dots\dots\dots(2.72)$$

2.8 Joint Estimate on the 2nd occasion.

Since both the estimates given by $\bar{y}_{np}^{(2)}$ and $\bar{y}_{nq}^{(2)}$

are independent and unbiased estimate of the population mean

$\bar{y}_N^{(2)}$, therefore the combined estimate for the population mean on the 2nd occasion will be

$$2\bar{y}_n^{(2)} = Q \left(\bar{y}_{np}^{(2)} \right) + (1 - Q) \bar{y}_{nq}^{(2)} \dots\dots(2.81)$$

the value of Q is determined, such that $V \left(2\bar{y}_n^{(2)} \right)$ is minimum.

$$V \left(2\bar{y}_n^{(2)} \right) = Q^2 V_A + (1-Q)^2 V_B \dots\dots\dots(2.82)$$

By differentiating (2.82) with respect to Q and equating to zero we get

$$Q = \frac{V_B}{V_A + V_B} \dots\dots\dots(2.83)$$

Substituting this value of q in (2.82) we get.

$$V(2\bar{y}_n^{(2)}) = \frac{V_A \cdot V_B}{V_A + V_B} = \frac{\left(\frac{N-n}{Nn} s_y^2(2) + \frac{n-np}{n \cdot np} U^2 \right) \left(\frac{N-nq}{N \cdot nq} s_y^2(2) \right)}{\frac{N-n}{Nn} s_y^2(2) + \frac{n-np}{n \cdot np} U^2 + \frac{N-nq}{N \cdot nq} s_y^2(2)}$$

.....(2.84)

2.9 Fraction of units to be replaced on the 2nd occasion i.e. optimum value of q .

We replace that fraction of units on the 2nd occasion which minimizes the variance of the estimated mean ($2\bar{y}_n$).

$$V(2\bar{y}_n^{(2)}) = \frac{V_A V_B}{V_A + V_B}$$

where $V_A = \frac{N-n}{Nn} s_y^2(2) + \frac{n-np}{np \cdot n \cdot N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$

$$= \frac{N-n}{Nn} s_y^2(2) + \left(\frac{1}{np} - \frac{1}{n} \right) U^2 \quad \dots\dots(2.91)$$

$$V_B = \frac{N-nq}{nq \cdot N} s_y^2(2) = \left(\frac{1}{nq} - \frac{1}{N} \right) s_y^2(2) \quad \dots\dots(2.92)$$

Therefore the minimum value of the variance with respect to q is found by differentiating $V(2\bar{y}_n^{(2)})$ with respect to q and equating to zero.

$$\frac{d(2\bar{y}_n^{(2)})}{dq} = 0,$$

$$\text{or } \frac{\left(\frac{dv_A}{dq} v_B + \frac{dv_B}{dq} v_A \right) (v_A + v_B) - \left\{ \frac{dv_A}{dq} \frac{v_B}{q} \right\} v_A v_B}{(v_A + v_B)^2} = 0$$

.....(2.98)

$$\text{or } \frac{dv_A}{dq} v_B^2 + \frac{dv_B}{dq} v_A^2 = 0$$

$$\frac{dv_A}{dq} = \frac{1}{n(1-q)^2} U^2$$

$$\frac{dv_B}{dq} = \frac{1}{nq^2} v_y^2(2)$$

$$\frac{dv_A}{dq} v_B^2 + \frac{dv_B}{dq} v_A^2 = \frac{1}{n(1-q)^2} U^2 v_B^2 - \frac{1}{nq^2} v_y^2(2) v_A^2 = 0$$

$$\text{or } \frac{q^2}{(1-q)^2} = \frac{v_y^2(2) v_A^2}{U^2 v_B^2}$$

$$\text{or } \frac{q}{1-q} = \frac{v_y(2) v_A}{U v_B}$$

$$q U v_B = (1-q) v_y(2) v_A$$

$$\text{or } q U \left(\frac{1}{nq} - \frac{1}{N} \right) v_y^2(2)$$

$$= (1 - q) s_y(2) \left\{ \frac{N-n}{Nn} s_y^2(2) + \left(\frac{1}{n(1-q)} - \frac{1}{n} \right) U^2 \right\}$$

or $\frac{U}{n} s_y(2) - \frac{q U s_y(2)}{N}$

$$= \frac{N-n}{Nn} s_y^2(2) - q \cdot \frac{N-n}{Nn} s_y^2(2) + \frac{1}{n} U^2 - \frac{1}{n} U^2 + \frac{q U^2}{n}$$

or $q = \frac{\frac{U}{n} s_y(2) - \frac{N-n}{Nn} s_y^2(2)}{\frac{1}{n} U^2 + \frac{U s_y(2)}{N} - \frac{N-n}{Nn} s_y(2)}$

$$= \frac{s_y(2) \left\{ \frac{U}{n} - \frac{s_y(2)}{n} + \frac{1}{N} s_y(2) \right\}}{(U + s_y(2)) \left(\frac{U}{n} - \frac{s_y(2)}{n} + \frac{1}{N} s_y(2) \right)}$$

$$= \frac{s_y(2)}{U + s_y(2)} \dots\dots\dots (2.94)$$

where $U^2 = \frac{1}{N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i^{(2)}}{p_i} - \bar{y}_N^{(2)} \right)^2$

and $s_y^2(2) = \frac{1}{N-1} \sum_{i=1}^N (y_i^{(2)} - \bar{y}_N^{(2)})^2$

If we assume that $y_1^{(2)}, y_2^{(2)}, \dots, y_N^{(2)}$, the finite

population is a random sample from an infinite population then under the linear model,

$$E(U^2) < E(s_y^2(2))$$

Therefore $q > \frac{1}{2}$; or $\frac{1}{2} \leq q \leq 1$

Thus replacement fraction should be more than 50%

$$\text{Therefore } V(\bar{y}_n^{(2)})_{\text{opt } q} = \frac{\left(\frac{N-n}{Nn} s_y^2(2) + \frac{U s_y(2)}{n} \right) \left(\frac{U}{n} s_y(2) + \frac{N \cdot n}{Nn} s_y^2(2) \right)}{2 \left\{ \frac{N-n}{Nn} s_y^2(2) + \frac{U s_y(2)}{n} \right\}}$$

$$= \frac{\frac{N-n}{Nn} s_y^2 + \frac{s_y(2) U}{n}}{2}$$

$$= \frac{N-n}{Nn} s_y^2 + \frac{s_y(2)}{n(N-1)N} \sum_{i=1}^N p_i \left(\frac{y_i}{p_i} - \bar{y}_N \right)^2$$

..... (2.9)

(2.10) Comparison with the estimate based on regression method

If the sub-sample n_p is selected with equal probability from the n units, obtained at the first occasion, the regression estimate is given by

$$\bar{y}_{n_p(\text{reg})}^{(2)} = \bar{y}_{n_p}^{(2)} + b \left(\bar{y}_n^{(1)} - \bar{y}_{n_p}^{(1)} \right)$$

where b is the regression coefficient

and

$$V(\bar{y}_{np(\text{reg})}^{(2)}) = \frac{S_y^2(2) (1 - \rho^2)}{np} \left(1 + \frac{n - np}{n} \frac{1}{np - 3} \right) + \frac{\rho^2 S_y^2(2)}{n}$$

$$= V_A' \text{ (say)} \quad \dots\dots(2.101)$$

where $\rho = \frac{\text{Cov}(y^{(1)}, y^{(2)})}{\sqrt{V(y^{(1)}) V(y^{(2)})}}$

The mean of the $n - np = nq$ units on the 2nd occasion, which are taken in addition to these np units, is given by

$$\bar{y}_{nq}^{(2)} = \frac{1}{nq} \sum_{i=1}^{nq} y_i^{(2)}$$

and

$$V(\bar{y}_{nq}^{(2)}) = \frac{n - nq}{n \cdot nq} S_y^2(2) = V_B \text{ (say)} \quad \dots\dots(2.102)$$

Therefore the combined estimate is given by

$$E(\bar{y}_B^{(2)}) = Q' (\bar{y}_{np(\text{reg})}^{(2)}) + (1 - Q') \bar{y}_{nq}^{(2)}$$

where $Q' = \frac{V_B}{V_A + V_B}$

and the variance of the estimate is given by

$$V(2 \bar{y}_{B(\text{reg})}^{(2)}) = \frac{V_A' V_B}{V_A' + V_B}$$

$$\frac{\left\{ \frac{S_y^2(2) (1 - \rho^2)}{np} \left\{ 1 + \frac{n - np}{n} \frac{1}{np-3} \right\} + \rho^2 \frac{S_y^2(2)}{n} \right\} \left\{ \frac{N - nq}{N \cdot nq} S_y^2(2) \right\}}{\left\{ \frac{S_y^2(2) (1 - \rho^2)}{np} \left\{ 1 + \frac{n - np}{n} \frac{1}{np-3} \right\} + \rho^2 \frac{S_y^2(2)}{n} \right\} + \left\{ \frac{N - nq}{N \cdot nq} S_y^2(2) \right\}}$$

.....(2.103)

The estimate based on varying probability selection is more efficient as compared to the estimate based on regression method, iff

$$\frac{V_{Ax} V_B}{V_A^* V_B} \leq \frac{V_A^* \times V_B}{V_A^* + V_B}$$

or $V_A \leq V_A^*$

or $\frac{N - n}{Nn} S_y^2(2) + \frac{n - np}{(n)(np)N(N-1)} \sum_{i=1}^N p_i \left(\frac{y_i(2)}{p_i} - \bar{y}_N \right)^2$

$$\leq \frac{S_y^2(2) (1 - \rho^2)}{np} \left\{ 1 + \frac{n - np}{n} \frac{1}{np-3} \right\} + \frac{\rho^2 S_y^2(2)}{n}$$

or $\frac{n - np}{n \cdot np} V^2 \leq \frac{S_y^2(2) (1 - \rho^2)}{np} \left\{ \frac{n - np}{n \cdot np} \left(1 + \frac{1}{np-3} \right) \right\}$

$$+ \frac{1}{n} S_y^2(2)$$

$$U^2 \leq S_y^2(2) (1 - \rho^2) \left\{ \frac{(np - 2)}{(np - 3)} \right\} + \frac{1 \cdot n \cdot np}{n \cdot np \cdot N} S_y^2(2)$$

.....(2.104)

Generally $S_y^2(2)$ is large as compared to U^2 and for the small np & N , the estimate based on varying probability selection is expected to be more efficient, as compared to the estimate based on regression method of estimation.

If N is large i.e. $\frac{1}{N} S_y^2(2)$ is negligible as compared to the other terms, then we have

$$U^2 \leq S_y^2(2) (1 - \rho^2) \frac{(np - 2)}{(np - 3)}$$

(2) (2) (2)

If we consider that the sample y_1, y_2, \dots, y_{np}

has been taken from an infinite population. Then under linear model, taking the expectation, we see that L.H.S. > 1 and R.H.S. < 1 . The inequality does not hold, therefore if N is large, the estimate based on regression method is more efficient as compared to the estimate based on varying probabilities.

2.11 Illustration: A sample survey for estimating the Milk yield of the cows was conducted in Gujarat state in the year 1963-64. The data given below gives the no. of cows in different seasons.

Village group	Census data No. of Cows x	Rainy Season No. of cows Enumerated y(1)	Summer Season No. of cows Enumerated y(2)
I	165	64	62
II	109	63	37
III	4	3	2
IV	6	4	2
V	48	43	32
VI	48	18	18
VII	82	24	30
VIII	14	14	11
IX	193	12	17
X	31	27	24
XI	35	22	22
XII	86	21	20
XIII	10	-	-
XIV	10	15	20
XV	240	133	100
XVI	116	123	116
XVII	119	156	132
XVIII	677	122	128
XIX	75	35	22
XX	36	18	23

Suppose a sample of 10 villages is taken at the rainy season with S.R.S. In the summer season 3 villages are retained from the previous year with varying probability. For selecting these 3 villages divide the sample of 10 villages into 5 groups of two villages each and from each group select one village with probability proportional to the no. of cows in that village in the rainy season.

Thus we have

$$N = 20 \quad ; \quad n = 10 \quad ; \quad np = 3$$

$$U^2 = 103.78 \quad ; \quad S_{y(2)}^2 = 1818.95,$$

$$S_{y(1)}^2 = 2174.63 \quad ; \quad \rho^2 = .9746$$

$$\text{Thus } V_A = 101.32$$

$$V_B = 90.94$$

$$\text{For regression method } V_A' = 188.53$$

Therefore

$$V(\bar{y}_n) = 47.92$$

$$V(\bar{y}_{n(\text{reg})}) = 61.33$$

The relative efficiency of estimate on the summer season using varying probability estimate as compared to estimate when regression method is used will be 128.02%.

Use of Varying Probabilities at the 1st Occasion.

3.1 Here the method of varying probabilities in the successive sampling will be studied when on the 1st occasion the units are drawn with varying probabilities of selection. On the 2nd occasion some units are retained with s.r.s. from the units sampled on the previous occasion and the remaining units are drawn afresh with varying probabilities of selection, from the population itself. The present scheme of sampling is particularly important where some information on an auxiliary variable is given and it is decided to assign the probabilities on the basis of this auxiliary variable.

3.2 Sampling with Replacements:

On the first occasion n units are drawn out of N , with varying probabilities of selection p_i and with replacement. On the 2nd occasion np units are retained with s.r.s. from the n units sampled on the 1st occasion and the remaining $n - np = nq$ units are drawn afresh, with varying probabilities of selection p_i , from the population itself.

Let

y_i = the value of the character for the i th sampling unit of the population.

x_i = the value of the auxiliary variable for the i th unit.

$$p_i = \frac{x_i}{\sum_{i=1}^N x_i}$$

define, $s_1 = y_1 / N p_1$; $s_1^{(1)} = y_1^{(1)} / N p_1$ and $s_1^{(2)} = y_1^{(2)} / N p_1$.

where the suffix (1) and (2) denotes the 1st and 2nd occasion.

$$E(s_1^{(1)}) = E\left(\frac{y_1}{N p_1}\right) = \sum_{i=1}^N \frac{y_i}{N p_1} p_i = \sum_{i=1}^N \frac{y_i}{N} = \bar{y}_N^{(1)}$$

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

$$E(\bar{s}_N^{(2)}) = \bar{y}_N^{(2)} ; E(\bar{s}_{np}^{(1)}) = \frac{1}{np} \sum_{i=1}^{np} E(s_i^{(1)}) = \bar{y}_N^{(1)}$$

$$E(\bar{s}_{nq}^{(1)}) = \frac{1}{nq} \sum_{i=1}^{nq} E(s_i^{(1)}) = \bar{y}_N^{(1)}$$

$$V(\bar{s}_N^{(1)}) = \frac{\sigma_s^2(1)}{N}$$

$$V(\bar{s}_{np}^{(1)}) = \frac{\sigma_s^2(1)}{np} \quad \text{and} \quad V(\bar{s}_{nq}^{(1)}) = \frac{\sigma_s^2(1)}{nq}$$

$$V(\bar{s}_{np}^{(2)}) = \frac{\sigma_s^2(2)}{np} \quad \text{and} \quad V(\bar{s}_{nq}^{(2)}) = \frac{\sigma_s^2(2)}{nq}$$

$$\text{Cov}(\bar{s}_{np}^{(1)}, \bar{s}_{np}^{(2)}) = \rho_s^{(1)(2)} \sigma_s^{(1)} \sigma_s^{(2)}$$

$$\text{where } \sigma_s^2(1) = \sum_{i=1}^N p_i (s_i^{(1)} - \bar{s}^{(1)})^2$$

$$\sigma_s^2(2) = \sum_{i=1}^N p_i (s_i^{(2)} - \bar{s}^{(2)})^2$$

$$\rho_s^{(1)(2)} \sigma_s^{(1)} \sigma_s^{(2)} = \sum_{i=1}^N (s_i^{(1)} - \bar{s}^{(1)})(s_i^{(2)} - \bar{s}^{(2)})$$

3.3 Now for the estimation of the parameters at the 2nd occasion we can take a linear function of the form.

$$\hat{y}_n^{(2)} = \bar{y}_n^{(2)} = a \bar{s}_{np}^{(1)} + b \bar{s}_{nq}^{(1)} + c \bar{s}_{np}^{(2)} + d \bar{s}_{nq}^{(2)}$$

$$\begin{aligned} E(\hat{y}_n^{(1)}) = E(\bar{y}_n^{(2)}) &= a E(\bar{s}_{np}^{(1)}) + b E(\bar{s}_{nq}^{(1)}) \\ &+ c E(\bar{s}_{np}^{(2)}) + d E(\bar{s}_{nq}^{(2)}) \\ &= (a + b) \bar{y}_n^{(1)} + (c + d) \bar{y}_n^{(2)} \end{aligned}$$

If $\hat{y}_n^{(2)}$ is an unbiased estimate for $\bar{y}_n^{(2)}$ we must have
 $a + b = 0$ and $c + d = 1$.

Therefore the equation can be written in the following form

$$\hat{y}_n^{(2)} = \bar{y}_n^{(2)} = a (\bar{s}_{np}^{(1)} - \bar{s}_{nq}^{(1)}) + c \bar{s}_{np}^{(2)} + (1 - c) \bar{s}_{nq}^{(2)}$$

where a and c are determined so that $V(\hat{y}_n^{(2)})$ is minimum

$$\begin{aligned} V(\hat{y}_n^{(2)}) &= a^2 V(\bar{s}_{np}^{(1)}) + a^2 V(\bar{s}_{nq}^{(1)}) + c^2 V(\bar{s}_{np}^{(2)}) \\ &+ (1-c)^2 V(\bar{s}_{nq}^{(2)}) + 2ac \text{Cov}(\bar{s}_{np}^{(1)}, \bar{s}_{np}^{(2)}) \\ &= a^2 \left(\frac{\sigma_s^2(1)}{np} + \frac{\sigma_s^2(1)}{nq} \right) + c^2 \left(\frac{\sigma_s^2(2)}{np} \right) \\ &+ (1-c)^2 \frac{\sigma_s^2(2)}{nq} + \frac{2ac}{np} \sqrt{\sigma_s^2(1)\sigma_s^2(2)} \end{aligned}$$

By differentiating and equating to zero we get

$$a = \frac{\rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)} \frac{1}{\sqrt{q}}}{1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}} \quad \frac{C_s^{(2)}}{C_s^{(1)}}$$

and

$$c = \frac{p}{1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}}$$

Thus the estimator with optimum values for a and c, can be written as

$$\bar{y}_n^{(2)} = \frac{\rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)} p q C_s^{(2)}}{(1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}) C_s^{(1)}} \left(\bar{y}_{np}^{(1)} - \bar{y}_{nq}^{(1)} \right)$$

$$+ \frac{p}{1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}} \bar{y}_{np}^{(2)} + \frac{q (1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)})}{1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}} \bar{y}_{nq}^{(2)} \quad \dots\dots\dots (3.31)$$

and $V(\bar{y}_n^{(2)}) = \frac{C_s^2(2)}{n} \frac{pq \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}}{1 - q^2 \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}} \quad \dots\dots\dots (3.32)$

Now to find the value of q, for which this variance is minimum.

Differentiate with respect to q and equating zero, we have

$$q = \frac{1 - \sqrt{1 - \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}}}{\rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}} \quad \dots\dots\dots (3.33)$$

and

$$V(\bar{y}_n^{(2)})_{opt\ q} = \frac{C_s^2(2) \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}}{2n (1 - \sqrt{1 - \rho_{s(1)}^{(1)} \rho_{s(2)}^{(2)}})} = \frac{C_s^2(2)}{2nq} \quad \dots\dots\dots (3.34)$$

3.4 Gain in Efficiency:

If there had been no repetition or the sample had been completely retained.

$$V(\bar{y}_n^{(2)})_{q=1 \text{ or } 0} = \frac{\sigma_y^2(2)}{n}$$

Therefore

$$\frac{V(\bar{y}_n^{(2)})}{V(\bar{y}_n^{(2)})_{\text{cpt } q}} = 2q \geq 1 \quad \dots\dots\dots(3.81)$$

Thus the efficiency of this procedure for estimating the mean at the 2nd occasion as compared to the one with independent sample each time or with completely retained sample will always be more, and its value will depend on replacement fraction of the sample.

3.5 Sampling without Replacements:

On the first occasion n units are drawn, out of N, with varying probabilities of selection p_1 and without replacement by random grouping procedure. On the 2nd occasion np units are retained with s.r.s. from the n units sampled on the 1st occasion and the remaining $n - np = nq$ units are drawn afresh from the population itself, with unequal probabilities without replacement by random group procedure.

$$\text{Let } s_1^{(1)} = \frac{y_1^{(1)}}{N \frac{p_1}{N_1}} \text{ and } s_1^{(2)} = \frac{y_1^{(2)}}{N \frac{p_1}{N_1}}$$

where

$$p_t = \frac{x_t}{\sum_{t=1}^N x_t} ; \quad \bar{\pi}_t = \frac{N_1}{\sum_{t=1}^N p_t}$$

$$\bar{N} = \frac{\sum_{t=1}^N N_1}{n} = \frac{N}{n}$$

$$E(\bar{s}_n^{(1)}) = \bar{y}_N^{(1)} ; \quad V(\bar{s}_n^{(1)}) = \frac{\sum_{t=1}^N N_1^2 - N}{N^2(N-1)} \left\{ \frac{N}{\sum_{t=1}^N p_t} \frac{y_t^2(1)}{p_t} - \frac{y_N^2(1)}{N} \right\}$$

Now for the estimation of the parameters at the 2nd occasion we can take a linear function of the form,

$$\bar{y}_n^{(2)} = \bar{s}_n^{(2)} = a(\bar{s}_{np}^{(1)}) + b(\bar{s}_{nq}^{(1)}) + c(\bar{s}_{np}^{(2)}) + d(\bar{s}_{nq}^{(2)})$$

If $\bar{y}_n^{(2)}$ is an unbiased estimate of $\bar{y}_N^{(2)}$, equation can be written as

$$\bar{y}_n^{(2)} = \bar{s}_n^{(2)} = a(\bar{s}_{np}^{(1)} - \bar{s}_{nq}^{(1)}) + c(\bar{s}_{np}^{(2)}) + (1-c)\bar{s}_{nq}^{(2)}$$

$$V(\bar{s}_n^{(2)}) = a^2 V(\bar{s}_{np}^{(1)}) + a^2 V(\bar{s}_{nq}^{(1)}) - 2a^2 \text{Cov}(\bar{s}_{np}^{(1)}, \bar{s}_{nq}^{(1)}) + c^2 V(\bar{s}_{np}^{(2)}) + (1-c)^2 V(\bar{s}_{nq}^{(2)}) + 2ac \text{Cov}(\bar{s}_{np}^{(1)}, \bar{s}_{np}^{(2)}) \dots\dots (3.51)$$

$$= a^2 \left(\frac{1}{np} + \frac{1}{nq} \right) \frac{N^2}{N^2(N-1)} \left(\frac{N}{\sum_{t=1}^N p_t} \frac{y_t^2(1)}{p_t} - \frac{y_N^2(1)}{N} \right) + c^2 \left(\frac{1}{np} - \frac{1}{n} \right) \left(\frac{N}{\sum_{t=1}^N p_t} \frac{y_t^2(2)}{p_t} - \frac{y_N^2(2)}{N} \right) +$$

$$2ac \frac{N}{\sum_{t=1}^N p_t} \frac{y_t^2(2)}{p_t} - \frac{y_N^2(2)}{N}$$

$$\bullet (1-c)^2 \frac{\frac{nq}{n} \frac{N_1^2 - N}{N^2 (N-1)}}{\frac{1}{np} - \frac{1}{n}} \left\{ \sum_{t=1}^N \frac{y_t^{(2)}}{pt} - y_N^{(2)} \right\}$$

$$\bullet 2ac \left(\frac{1}{np} - \frac{1}{n} \right) \rho_{s(1)(2)} \sigma_{s(1)} \sigma_{s(2)}$$

$$\bullet 2ac \frac{\frac{n}{1} \frac{N^2 - N}{N(N-1)}}{\frac{1}{np} - \frac{1}{n}} \rho_{s(1)(2)} \sigma_{s(1)} \sigma_{s(2)}$$

$$\rho_{s(1)(2)} = \frac{\sum_{t=1}^N pt \left(\frac{y_t^{(1)}}{pt} - y_N^{(1)} \right) \left(\frac{y_t^{(2)}}{pt} - y_N^{(2)} \right)}{\sqrt{\left[\sum_{t=1}^N pt \left(\frac{y_t^{(1)}}{pt} - y_N^{(1)} \right)^2 \right] \left[\sum_{t=1}^N pt \left(\frac{y_t^{(2)}}{pt} - y_N^{(2)} \right)^2 \right]}}$$

when $N_1 = \frac{N}{n}$ and $N_j = \frac{N}{nq}$

$$V\left(\frac{y_N^{(2)}}{n}\right) = a^2 \left(\frac{1}{np} - \frac{1}{nq} \right) \sigma_{s(1)}^2 + a^2 \left(\frac{1}{np} - \frac{1}{n} \right) \sigma_{s(2)}^2$$

$$\bullet a^2 \frac{N - n}{n N^2 (N-1)} \left[\sum_{t=1}^N pt \left(\frac{y_t^{(2)}}{pt} - y_N^{(2)} \right)^2 \right]$$

$$\bullet (1-c)^2 \frac{N - nq}{nq N^2 (N-1)} \left[\sum_{t=1}^N pt \left(\frac{y_t^{(2)}}{pt} - y_N^{(2)} \right) \right]$$

$$\bullet 2ac \frac{N - n}{n(N-1)} \sigma_{s(1)} \sigma_{s(2)} + 2ac \left(\frac{1}{np} - \frac{1}{n} \right) \sigma_{s(1)} \sigma_{s(2)}$$

The value of at C can be obtained by differentiating

$V(\bar{s}_n^{(2)})$ with respect to at c and equating to zero we get

$$a = \frac{\sigma_s^{(2)}}{\sigma_s^{(1)}} \left[\frac{N - nq}{q(N-1)} \left[\frac{N - np - q}{(N-1)p} - \frac{N - nq}{q^2(N-np-q)} - \frac{1}{pq} \right] \right] \dots\dots\dots(3.53)$$

$$c = \frac{1}{q^2} \frac{N - nq}{(N-np-q)} \left[\frac{1}{pq} + \frac{N - nq}{q^2(N-np-q)} - \frac{(N-np-q)}{(N-1)p} \right] \dots\dots\dots(3.54)$$

Thus the estimator with the optimum value of a and c can be written as:-

$$\begin{aligned} \bar{s}_n^{(2)} &= \left[\frac{\sigma_s^{(2)}}{\sigma_s^{(1)}} \frac{N-nq}{q(N-1)} \left[\frac{N-np-q}{(N-1)p} - \frac{N-nq}{q^2(N-np-q)} - \frac{1}{pq} \right] \right. \\ &\quad \left. \left(\bar{s}_{np}^{(1)} - \bar{s}_{nq}^{(1)} \right) + \frac{1}{q^2} \frac{N-nq}{(N-np-q)} \left[\frac{1}{pq} + \frac{N-nq}{q^2(N-np-q)} - \frac{(N-np-q)}{(N-1)p} \right] \right. \\ &\quad \left. \bar{s}_{np}^{(2)} + \left\{ 1 - \frac{1}{q^2} \frac{N-nq}{N-np-q} \left(\frac{1}{pq} + \frac{N-nq}{q^2(N-np-q)} - \frac{(N-np-q)}{(N-1)p} \right) \right\} \right. \\ &\quad \left. \bar{s}_{nq}^{(2)} \right] \dots\dots\dots(3.55) \end{aligned}$$

and the variance of this estimator is given by

$$\begin{aligned} V(\bar{s}_n^{(2)}) &= (1-c) \frac{\sigma_s^{(2)2}}{nq(N-1)} \\ &= \left[1 - \frac{1}{q^2} \frac{N-nq}{N-np-q} \left\{ \frac{1}{pq} + \frac{N-nq}{q^2(N-np-q)} - \frac{(N-np-q)}{(N-1)p} \right\} \right] \end{aligned}$$

$$\left\{ \frac{N - nq}{nq (N - 1)} \right\} \sigma_{\bar{x}}^2(2) \dots\dots\dots (3.58)$$

Use of Varying probabilities at Both the Occasions.

4.1 When some information on an auxiliary variate is given and it is desired to use the information, Units can be selected with probability proportional to the auxiliary variable. Now if we select units on the first occasion with probability proportional to the auxiliary variate. Again, on the 2nd occasion the information on the variable sampled on the first occasion is also given. It may be advantageous to use the information, collected on the previous occasion, for selecting the units to be retained at the successive occasions.

4.2 Sampling with Replacement

On the first occasion n units are selected from the given population containing N units with the probabilities of selection $p_t(N)$ and without replacement. The population is randomly divided into n groups and from each of these groups a unit is selected independently, following the sampling procedure of Rao, Hartley and Cochran. On the 2nd occasion, n_p units are selected out of n , with the probabilities of selection $p_1(n)$ and with replacement. Remaining n_q units are selected from the population itself with probabilities of selection $p_t(N)$ and without replacement.

Let

y_t = the value of the character for the t th sampling unit of the population

x_t = the value of the auxiliary variate for the i th unit

$$p_t(N) = \frac{x_t}{\sum_{t=1}^N x_t}, \quad \prod_1 = \sum_{t=1}^{N_1} p_t(N)$$

$$p_1(n) = \frac{y_1^{(1)}}{\sum_{i=1}^n y_i^{(1)}}$$

If the sample on the first occasion is observed as $y_1^{(1)}, y_2^{(1)}, \dots, y_n^{(1)}$ having mean as $\bar{y}_n^{(1)}$ on the 2nd occasion np units are selected, from the units selected on the first occasion, with probabilities of selection $p_1(n)$ and with replacement

Let

$$s_t = \frac{y_t^{(2)}}{N p_t} \quad t = 1, 2, \dots, N$$

$$i = 1, 2, \dots, n$$

$$\text{and } v_i = \frac{s_i}{n p_1(n)}$$

$$4.2.1 \quad \bar{v}_{np} = \bar{y}_{np}^{(2)} = \frac{1}{np} \sum_{i=1}^{np} v_i$$

Expected value of $\bar{y}_{np}^{(2)}$ is given as

$$E(\bar{v}_{np}) = E_1 E_2 (\bar{v}_{np}) = E_1 E_2 \left(\frac{1}{np} \sum_{i=1}^{np} v_i \right)$$

$$= E_1 \left(\frac{1}{np} \sum_{i=1}^{np} E_2 \left(\frac{s_i}{n p_1(n)} \right) \right) = E_1 \cdot \frac{1}{np} \sum_{i=1}^{np} \bar{s}_n$$

$$= E_1 (\bar{s}_n) = E_1 \left\{ \frac{1}{n} \sum_{i=1}^n E_0 \left(\frac{y_i^{(2)}}{N p_i^{(2)} / \pi_i} \right) / N_1 N_2 \dots N_n \right\}$$

$$= E_1 \left(\frac{1}{n} \sum_1^N \frac{1}{N} \sum_{t=1}^{N_1} y_t^{(2)} \right)$$

$$= E_1 \left(\frac{\bar{y}_N^{(2)}}{N} \right) = \frac{\bar{y}_N^{(2)}}{N}$$

4.2.2 Variance of the Estimate.

$$V(\bar{v}_{np}) = E_1 V_2 (\bar{v}_{np/n.}) + V_1 E_2 (\bar{v}_{np/n.})$$

$$= E_1 \left\{ \frac{1}{np} \sum_{i=1}^N p_{i2}(n) (v_{i1} - \bar{v}_{..})^2 \right\} + V(\bar{s}_n)$$

$$= E_1 \frac{1}{np n^2} \left\{ \sum_{i \neq j}^N \frac{p_{i1} y_{i1}^{(1)} y_{j1}^{(1)}}{y_{i1}^{(1)} y_{j1}^{(1)}} - \sum_{i \neq j}^N (s_{i1} s_{j1}) \right\} + V(\bar{s}_n)$$

$$= \frac{1}{np \cdot n^2} \sum_{t \neq j}^N \sum_{i=1}^{N_1} \frac{y_t^{(2)} y_j^{(1)}}{N^2 (p_t / \pi_1) y_t^{(1)}} - \frac{(n-1) \bar{y}_N^{(2)}}{n \cdot np}$$

$$+ \frac{N-n}{n N^2 (N-1)} \left\{ \sum_{t=1}^N \frac{y_t^{(2)^2}}{p_t} - Y_N^{(2)^2} \right\}$$

OPTIMAL USE OF ANCILLARY VARIATES

5.1 In theory of sampling techniques, the information on ancillary characters often play a very important part, namely reducing the variance of estimated character. They can be used, for unequal probability sampling, for forming ratio and regression estimators and finally for stratifying the population into relatively homogeneous strata. If there are more than one ancillary characters correlated with the variate under estimation, these techniques can be combined to produce considerably more efficient estimators.

The problem in its most general form may be stated thus; Given p ancillary variates $x_1, x_2 \dots x_p$. Let the population is first stratified on the basis of $S(x_1, x_2 \dots x_p)$ and then in any particular stratum i , the units are sampled with probability proportional to $P(x_1, x_2 \dots x_p)$ and finally the mean in that stratum is estimated by an estimator of the form.

$$\bar{y}'_{iR} = \frac{1}{n_1} \sum_{j=1}^{n_1} \left\{ y'_{1j} + \rho \frac{\sigma_{y_1}}{\sigma_{P_1}} \left(P'_{1j} - \bar{P}'_{1, N_1} \right) \right\}$$

where n_1 is the size of sample in that stratum and

$$P'_{1j} = \frac{y_{1j}}{N_1 \frac{P_{1j}(x_1, x_2 \dots x_p)}{\sum_{j=1}^{N_1} P_{1j}(x_1, x_2 \dots x_p)}} = \frac{y_{1j}}{N_1 P'_{1j}}$$

$$P'_{1j} = \frac{P_{1j}}{N_1 P'_{1j}}, \quad \bar{P}'_{1, N_1} = \frac{1}{N_1} \sum_{j=1}^{N_1} P'_{1j}$$

$$\sigma_{y'_i}^2 = \sum_{j=1}^{N_1} P_{1j} (y'_{1j} - \bar{y}_{1,N_1})^2 \quad \sigma_{F'_i}^2 = \sum_{j=1}^{N_1} P_{1j} (F_{1j} - \bar{F}_{1,N_1})^2$$

and

$$\rho_{y'_i, F'_i} = \frac{\sum_{j=1}^{N_1} P_{1j} (y'_{1j} - \bar{y}_{1,N_1}) (F_{1j} - \bar{F}_{1,N_1})}{\sigma_{y'_i} \sigma_{F'_i}}$$

F_{1j} being the numerical value of some function $F(x_{11}, x_{12}, \dots, x_{1p})$ on the j th unit of i th stratum.

The estimator for the population mean is then

$$\sum_{i=1}^k P_i \bar{y}'_{iR} \quad \text{where } P_i = \frac{N_i}{N}$$

Thus for optimal use of the ancillary variates the variance expression,

$$V\left(\sum_{i=1}^k P_i \bar{y}'_{iR}\right) = \sum_{i=1}^k P_i^2 \frac{1}{N_i} \sigma_{y'_i}^2 (1 - \rho_{y'_i, F'_i}^2)$$

should be minimized for the functional forms $S(x_1, x_2, \dots, x_p)$, $P(x_1, x_2, \dots, x_p)$ and $F(x_1, x_2, \dots, x_p)$ given that the multiple regression surface of y on x_1, x_2, \dots, x_p is given by

$$E(y) = R(x_1, x_2, \dots, x_p)$$

5.2 The Procedure:

Let $x_{11}, x_{12}, \dots, x_{1p}$ be the ancillary variates and y_i

the variate under estimation, $i = 1, 2, \dots, N$. If the

regression surface of y_{1j} on $x_{11}, x_{12}, \dots, x_{1p}$ be $E(y_{1j}) =$

$R(x_{11}, x_{12}, \dots, x_{1p})$ then we can write

$$y_{1j} = E (x_{11}, x_{12} \dots x_{1p}) + e_{1j} \dots \dots \dots (5.21)$$

where $E (e_{1j} / i) = 0$ and $E (e_{1j} e_{1j'} / i \neq i') = 0 \dots (5.22)$

and $E (e_{1j}^2 / i) = V(x_{11}, x_{12} \dots x_{1p}) \dots \dots (5.23)$

In relation (5.23) there is an implicit assumption that the conditional variance expressions of y_{1j} (for fixed i) are represented by the same functional form $V(x_{11}, x_{12} \dots x_{1p})$. No specific form need be assigned to this function at this stage.

The procedure adopted in subsequent pages, for evaluating the efficiency of different sampling procedures, and for determining the conditions under which a particular procedure may yield optimal results, will consist of finding the expectations of finite population sampling variances for different procedures and then determining the optimal conditions in terms of parameters belonging to super population.

5.3 Some possible estimates.

5.3.1 Sample mean \bar{y}_n

$$E (\bar{y}_n) = \bar{y}_N \quad V (\bar{y}_n) = \frac{1}{n} \sigma_y^2 \quad \sigma_y^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_N)^2$$

5.3.2 p. p. s. estimate \bar{y}'_n

Let $P_i = P (x_{11}, x_{12} \dots x_{1p}) ; P_i \geq 0$ for all i

and $\sum_{i=1}^N P_i = 1$

$$\text{Let } y'_1 = \frac{y_1}{N p_1}; \quad \bar{y}'_{..} = E \bar{y}'_1 = \sum_{i=1}^N p_1 y'_1 = \bar{y}_N$$

$$\text{and } \bar{y}'_N = \frac{1}{n} \sum_{i=1}^n y'_1$$

$$\text{Then } E(\bar{y}'_N) = \bar{y}_N; \quad V(\bar{y}'_N) = \frac{1}{n} \sigma_{y'}^2$$

$$\text{where } \sigma_{y'}^2 = \sum_{i=1}^N p_1 (y'_1 - \bar{y}'_{..})^2$$

5.3.2 Regression estimate \bar{y}_R

$$\bar{y}_R = \bar{y}_N + \rho_{y,F} \frac{\sigma_y}{\sigma_F} (\bar{F}_N - F_1); \quad E(\bar{y}_R) = \bar{y}_N$$

$$\text{and } V(\bar{y}_R) = \frac{1}{n} \sigma_y^2 (1 - \rho_{y,F}^2)$$

where $F_1 = F(x_{11}, x_{12}, \dots, x_{1p})$ is some arbitrary function

of $x_{11}, x_{12}, \dots, x_{1p}$ and

$$\bar{F}_N = \frac{1}{N} \sum_{i=1}^N F_1(x_{11}, x_{12}, \dots, x_{1p})$$

$$\sigma_F^2 = \frac{1}{N} \sum_{i=1}^N (F_1 - \bar{F}_N)^2$$

$$\rho_{y,F} = \frac{\sum_{i=1}^N (F_1 - \bar{F}_N)(y_1 - \bar{y}_N)}{\sigma_F \sigma_y}$$

3.3.4 p.p.s. regression estimate \bar{y}'_R

$$\bar{y}'_R = \frac{1}{N} \sum_{i=1}^N \frac{y_i + B (\bar{P}_N - P_i)}{N P_i}$$

where, for a minimum variance

$$B = \frac{\sum_{i=1}^N P_i \left\{ \frac{y_i}{N P_i} - \bar{y}_R \right\} \left\{ \frac{P_i - \bar{P}_N}{N P_i} \right\}}{\sum_{i=1}^N P_i \left\{ \frac{P_i - \bar{P}_N}{N P_i} \right\}^2}$$

$$= \rho_{y', P'} \frac{\sigma_{y'}}{\sigma_{P'}}$$

$$P' = \frac{P_i - \bar{P}_N}{N P_i}$$

$$E(\bar{y}'_R) = \bar{y}_R \text{ and } V(\bar{y}'_R) = \frac{1}{n} \sigma_{y'}^2 (1 - \rho_{y', P'}^2)$$

3.3.5 Stratified p.p.s. regression estimate.

$$\bar{y}'_{RH} = \frac{1}{N_1} \sum_{j=1}^{N_1} \left\{ \frac{y_{1j} + B' (\bar{P}_{1, N_1} - P_{1j})}{N_1 P'_{1j}} \right\}$$

$$P'_{1j} = \frac{P_{1j} (x_{11}, x_{12} \dots x_{1p})}{\sum_{j=1}^{N_1} P_{1j} (x_{11}, x_{12} \dots x_{1p})}$$

$$E' = \frac{\sum_{j=1}^{N_1} P'_{1j} \left\{ \frac{y_{1j}}{N_1 P'_{1j}} - \bar{y}_{1, N_1} \right\} \left\{ \frac{(F_{1j} - \bar{F}_{1, N_1})}{(N_1 P'_{1j})} \right\}}{\sum_{j=1}^{N_1} P'_{1j} \frac{(F_{1j} - \bar{F}_{1, N_1})}{N_1 P'_{1j}}}$$

$$= \rho_{y'_1, F'_1} \frac{\sigma_{y'_1}}{\sigma_{F'_1}}$$

$$\rho_{y'_1, F'_1} = \frac{\sum_{j=1}^{N_1} P'_{1j} (y'_{1j} - \bar{y}_{1, N_1})(F'_{1j} - \bar{F}_{1, N_1})}{\sigma_{y'_1} \sigma_{F'_1}}$$

$$\sigma_{y'_1}^2 = \sum_{j=1}^{N_1} P'_{1j} (y'_{1j} - \bar{y}_{1, N_1})^2, \quad \sigma_{F'_1}^2 = \sum_{j=1}^{N_1} P'_{1j} (F'_{1j} - \bar{F}_{1, N_1})^2$$

$$E(\bar{y}'_{1R}) = \bar{y}_{1, N_1} \text{ and } V(\bar{y}'_{1R}) = \frac{1}{N_1} \sigma_{y'_1}^2 (1 - \rho_{y'_1, F'_1}^2)$$

The estimate for the population mean is then

$$\sum_{i=1}^k p_i \bar{y}'_{iR}; \text{ where } p_i = \frac{N_i}{N}$$

and the variance expression is given by

$$V\left(\sum_{i=1}^k p_i \bar{y}_{iR}\right) = \sum_{i=1}^k p_i^2 \frac{1}{N_i} \left(\sigma_{y_i}^2 - \frac{\sigma_{y_i}^2}{N_i} \right).$$

5.4 Expectations of Finite Population Parameters.

Now we shall find the expectations of certain finite population parameters that will be needed for our purpose, expectations being evaluated under the assumptions (1.1) .. (1.2) and (1.3).

5.4.1 The parameter σ_y^2

$$\begin{aligned} \text{Now } \sigma_y^2 &= \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y}_N)^2 \\ &= \frac{1}{N} \sum_{i=1}^N y_i^2 - \frac{1}{N^2} \left(\sum_{i=1}^N y_i \right)^2 \\ &= \frac{N-1}{N^2} \sum_{i=1}^N y_i^2 - \frac{1}{N^2} \sum_{i \neq j} y_i y_j \\ &= \frac{N-1}{N^2} \sum_{i=1}^N (R_i^2 + 2 e_i R_i + e_i^2) \\ &\quad - \frac{1}{N^2} \sum_{i \neq j} (R_i R_j + e_i R_j + e_j R_i + e_i e_j) \end{aligned} \dots\dots(5.4.11)$$

$$\begin{aligned} \text{Hence } E \sigma_y^2 &= \frac{N-1}{N^2} \sum_{i=1}^N (R_i^2 + v_i) - \frac{1}{N^2} \sum_{i \neq j} R_i R_j \\ &= \sigma_R^2 + \frac{N-1}{N} \bar{v}_N \end{aligned} \dots\dots(5.4.12)$$

$$\begin{aligned} \text{where } \sigma_R^2 &= \frac{1}{N} \sum_{i=1}^N \left\{ R_i (x_{i1}, x_{i2} \dots x_{ip}) \right. \\ &\quad \left. - R_i R (x_{11}, x_{12} \dots x_{1p}) \right\}^2 \end{aligned}$$

and $\bar{V}_N = N_1 V(x_{11}, x_{12}, \dots, x_{1p}) = \frac{1}{N} \sum_{i=1}^N V_i \dots \quad (5.4.19)$

5.42

The parameter $\sigma_{y_1}^2$

$$\sigma_{y_1}^2 = \sum_{i=1}^N P_i (\bar{y}_1 - y_{1i})^2 ; y_{1i} = \frac{y_i}{N P_i}$$

$$P_i = P(x_{11}, x_{12}, \dots, x_{1p})$$

$$\begin{aligned} \text{R.H.S.} &= \sum_{i=1}^N P_i y_{1i}^2 - \left(\sum_{i=1}^N P_i y_{1i} \right)^2 \\ &= \frac{1}{N^2} \sum_{i=1}^N \frac{x_i^2}{P_i} - \frac{1}{N^2} \left(\sum_{i=1}^N y_i \right)^2 \\ &= \frac{1}{N^2} \sum_{i=1}^N \frac{y_i^2}{P_i} - \frac{1}{N^2} \sum_{i=1}^N y_i^2 - \frac{1}{N^2} \sum_{i \neq j} y_i y_j \\ &= \frac{1}{N^2} \sum_{i=1}^N \frac{R_i^2 + 2e_i R_i + e_i^2}{P_i} - \frac{1}{N^2} \sum_{i=1}^N (R_i^2 + 2e_i R_i + e_i^2) \\ &= \frac{1}{N^2} \sum_{i=1}^N (R_i + e_i) (R_i + e_i) \end{aligned}$$

$$\begin{aligned} \text{Hence } \sigma_{y_1}^2 &= \frac{1}{N^2} \sum_{i=1}^N \frac{R_i^2 + V_i}{P_i} - \frac{1}{N^2} \sum_{i=1}^N (R_i^2 + V_i) \\ &= \frac{1}{N^2} \sum_{i=1}^N \frac{R_i^2}{P_i} - \frac{1}{N^2} \left(\sum_{i=1}^N R_i \right)^2 + \frac{1}{N^2} \sum_{i=1}^N \frac{(1-P_i)V_i}{P_i} \\ &= \sigma_{R^2} + \frac{1}{N^2} \sum_{i=1}^N \frac{(1-P_i)V_i}{P_i} \dots \dots \dots (5.4.21) \end{aligned}$$

where,

$$\sigma_{R^2}^2 = \sum_{i=1}^N P_i (R_i' - \bar{R}_N)^2 \quad \dots\dots\dots(5.4.22)$$

and $R_i' = \frac{R_i}{NP_i}$

5.4.5

The Parameter $\sigma_{Y,F}^2$

$$\begin{aligned} \text{Cov}^2(Y,F) &= \left\{ \frac{1}{N} \sum_{i=1}^N P_i (Y_i - \bar{Y}_N) \right\}^2 \\ &= \left[\frac{1}{N} \sum_{i=1}^N P_i (R_i + \epsilon_i) - \frac{1}{N} \sum_{i=1}^N (R_i + \epsilon_i) \right]^2 \\ &= \left[\frac{1}{N} \sum_{i=1}^N P_i (R_i - \bar{R}_N) + \frac{1}{N} \sum_{i=1}^N P_i \epsilon_i \right]^2 \\ &= \frac{1}{N^2} \sum_{i=1}^N P_i \sum_{j=1}^N \epsilon_i \epsilon_j \\ &= \frac{1}{N^2} \left[\text{Cov}^2(P_i, R_i) + \frac{1}{N} \sum_{i=1}^N P_i \epsilon_i^2 - \frac{1}{N^2} \sum_{i=1}^N P_i \sum_{j=1}^N \epsilon_i \epsilon_j \right] \\ &= \text{Cov}^2(P_i, R_i) + \frac{1}{N^2} \sum_{i=1}^N P_i^2 \epsilon_i^2 + \frac{1}{N^2} \sum_{i=1}^N P_i P_j \epsilon_i \epsilon_j \\ &+ \frac{1}{N^2} \sum_{i=1}^N P_i^2 \left(\sum_{j=1}^N \epsilon_i^2 + \sum_{j=1}^N \epsilon_i \epsilon_j \right) \\ &+ \frac{2}{N} \text{Cov}(P_i, R_i) \sum_{i=1}^N P_i \epsilon_i - \frac{2}{N} \text{Cov}(P_i, R_i) \sum_{i=1}^N \epsilon_i \\ &= \frac{2}{N^2} \sum_{i=1}^N P_i \sum_{j=1}^N \epsilon_i \epsilon_j \end{aligned}$$

Hence,

$$\begin{aligned}
 E \text{Cov}^2 (y, P) &= \text{Cov}^2 (P_1, R_1) + \frac{1}{N^2} \sum_{i=1}^N P_i^2 V_i \\
 &+ \frac{1}{N} \bar{P}^2 \bar{V} - \frac{2 \bar{P}}{N} \sum_{i=1}^N P_i V_i \\
 &= \text{Cov}^2 (P_1, R_1) + \frac{1}{N^2} \sum_{i=1}^N (P_i - \bar{P})^2 V_i \quad \dots (5.4.51)
 \end{aligned}$$

but,

$$\text{Cov}^2 (y, P) = \rho_{y,P}^2 \sigma_y^2 \cdot \sigma_P^2$$

therefore,

$$\begin{aligned}
 E \sigma_y^2 \cdot \rho_{y,P}^2 &= \frac{1}{2} E \text{Cov}^2 (y, P) \\
 &= \rho_{R,P}^2 \sigma_R^2 + \frac{1}{N^2} \sum_{i=1}^N (P_i - \bar{P})^2 V_i \quad \dots (5.4.52)
 \end{aligned}$$

5.4.4.

The parameter $\rho_{y',P'}^2 \sigma_{y'}^2$

$$\begin{aligned}
 \text{Cov}^2 (y', P') &= \left[\sum_{i=1}^N P_i \left(\frac{Y_i}{NP_i} - \bar{Y} \right) \left(\frac{Y_i - \bar{Y}}{NP_i} \right) \right]^2 \\
 &= \left(\sum_{i=1}^N P_i \frac{Y_i}{NP_i} \cdot \frac{Y_i - \bar{Y}}{NP_i} \right)^2 \\
 &= \frac{1}{N^4} \left(\sum_{i=1}^N \frac{Y_i (Y_i - \bar{Y})}{P_i} \right)^2 \\
 &= \frac{1}{N^4} \left(\sum_{i=1}^N \frac{Y_i^2 (Y_i - \bar{Y})}{P_i} \right. \\
 &\quad \left. + \sum_{i \neq j} \frac{Y_i Y_j (Y_i - \bar{Y}) (Y_j - \bar{Y})}{P_i P_j} \right)
 \end{aligned}$$

$$= \frac{1}{N} \left\{ \sum_{i=1}^N \frac{(R_i^2 + 2 \cdot e_i R_i + e_i^2) (P_i - \bar{P}_N)^2}{P_i^2} + \sum_{i \neq j} \frac{(R_i R_j + R_i e_j + R_j e_i + e_i e_j) (P_i - \bar{P}_N) (P_j - \bar{P}_N)}{P_i P_j} \right\}$$

$$E \text{Cov}^2 (y', P') = \frac{1}{N^2} \left\{ \sum_{i=1}^N \frac{(R_i^2 + V_i) (P_i - \bar{P}_N)^2}{P_i^2} + \sum_{i \neq j} \frac{R_i R_j (P_i - \bar{P}_N) (P_j - \bar{P}_N)}{P_i P_j} \right\}$$

$$= \frac{1}{N^2} \left\{ \sum_{i=1}^N \frac{R_i (P_i - \bar{P}_N)}{P_i} \right\}^2$$

$$+ \frac{1}{N^2} \sum_{i=1}^N \frac{V_i (P_i - \bar{P}_N)^2}{P_i^2}$$

$$= \text{Cov}^2 (R', P') + \frac{1}{N^2} \sum_{i=1}^N \frac{V_i (P_i - \bar{P}_N)^2}{P_i^2}$$

.....(5.4.41)

Therefore

$$E \rho_{y', P'}^2 = \rho_{R', P'}^2 + \frac{1}{N^2} \sum_{i=1}^N \frac{V_i (P_i - \bar{P}_N)^2}{P_i^2}$$

.....(5.4.42)

Combining (5.4.12) and (5.4.51)

$$E \sigma_y^2 (1 - \rho_{y', P'}^2) = \sigma_R^2 (1 - \rho_{R', P'}^2) + \frac{N-1}{N} \bar{V}_N - \frac{1}{N^2} \sum_{i=1}^N V_i (P_i - \bar{P}_N)^2$$

.....(5.4.43)

and (5.4.21) with (5.4.42)

$$\begin{aligned}
 E \sigma_{y'}^2 (1 - \rho_{y', Y'})^2 &= \sigma_{R'}^2 (1 - \rho_{R', Y'})^2 + \frac{1}{N^2} \sum_{i=1}^N \frac{(1 - P_i) v_i}{P_i} \\
 &\quad - \frac{1}{N^2} \sum_{i=1}^N \frac{v_i (P_i - \bar{P}_N)^2}{P_i^2} \\
 &= \sigma_{R'}^2 (1 - \rho_{R', Y'})^2 + \frac{1}{N^2} \sum_{i=1}^N \frac{(1 - P_i) v_i}{P_i} \\
 &\quad - \frac{1}{N^2} \sum_{i=1}^N v_i \left(\frac{P_i}{N P_i} - \bar{P}_N \right)^2 \\
 &\dots\dots\dots(5.4.44)
 \end{aligned}$$

5.5 Determination of optimum P_1 and F_1

$$\begin{aligned}
 E V (\bar{y}'_n) &= \frac{1}{n} E \sigma_{y'}^2 \\
 \text{or } n E V (\bar{y}'_n) &= \frac{1}{N^2} \sum_{i=1}^N \frac{R_i^2}{P_i} + \frac{1}{N^2} \sum_{i=1}^N \frac{v_i}{P_i} \\
 &\quad - \frac{1}{N^2} \left(\sum_{i=1}^N R_i \right)^2 - \frac{1}{N^2} \sum_{i=1}^N v_i \\
 &= \frac{1}{N^2} \sum_{i=1}^N \frac{R_i^2}{P_i} + \frac{1}{N^2} \sum_{i=1}^N \frac{v_i}{P_i} - \frac{R_N^2}{N} - \frac{1}{N} \bar{V}_N \\
 &\dots\dots\dots(5.51)
 \end{aligned}$$

This has to be minimised with respect to P_1 , subject to the condition $\sum_{i=1}^N P_i = 1$.

By Schwarz's inequality, we have

$$(P_1)_{opt} = \frac{\sqrt{R_1^2 + v_1}}{\sum_{i=1}^N \sqrt{R_i^2 + v_i}} \dots\dots\dots (5.52)$$

and the minimum variance for p.p.s. estimate is

$$E V(\bar{y}'_N)_{min} = \frac{1}{nN^2} \left[\left(\sum_{i=1}^N \sqrt{R_i^2 + v_i} \right)^2 - N^2 \bar{R}_N^2 - N \bar{v}_N \right] \dots\dots\dots (5.53)$$

Further

$$\begin{aligned} & \frac{1}{n} \left[E \sigma_y^2 - (E \sigma_{y' opt}^2) \right] \\ &= \frac{N-1}{nN^2} \sum_{i=1}^N (R_i^2 + v_i) - \frac{1}{nN^2} \sum_{i \neq j} R_i R_j \\ & \quad - \frac{1}{nN^2} \left[\left(\sum_{i=1}^N \sqrt{R_i^2 + v_i} \right)^2 - \sum_{i=1}^N (R_i^2 + v_i) \right. \\ & \quad \left. - \sum_{i \neq j} R_i R_j \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{N-1}{nN^2} \sum_{i=1}^N (R_i^2 + v_i) - \frac{1}{nN^2} \sum_{i \neq j} R_i R_j \\
 &= \frac{1}{nN^2} \left[\sum_{i \neq j} (\sqrt{R_i^2 + v_i} \sqrt{R_j^2 + v_j} - R_i R_j) \right] \\
 &= \frac{1}{nN} \sum_{i=1}^N (R_i^2 + v_i) - \frac{1}{nN^2} \left(\sum_{i=1}^N \sqrt{R_i^2 + v_i} \right)^2 \\
 &= \frac{1}{n} \sigma^2 \sqrt{R_i^2 + v_i} \geq 0 \quad \dots\dots (5.54)
 \end{aligned}$$

Thus by taking p.p.s. sampling procedure with $P(x_{11}, x_{12}, \dots, x_{1p}) \propto \sqrt{R^2(x_{11}, x_{12}, \dots, x_{1p}) + V(x_{11}, x_{12}, \dots, x_{1p})}$

the average reduction in variance of simple random sampling is

$$\frac{1}{n} V(\sqrt{R_i^2 + v_i})$$

For obtaining optimum F_1 ; first assume that

$$V(x_{11}, x_{12}, \dots, x_{1p}) = \text{Const} = \sigma^2$$

Then by, (5.4.43)

$$\begin{aligned}
 E V(\bar{y}_R) &= \frac{1}{n} \sigma^2 \left(1 - \beta_{R,F}^2 \right) + \frac{N-1}{nN} \sigma^2 - \frac{1}{nN} \sigma^2 \\
 &= \frac{1}{n} \sigma^2 + \frac{N-2}{nN} \sigma^2 - \frac{1}{n} \beta_{R,F}^2 \sigma^2 \\
 &\dots\dots\dots (5.55)
 \end{aligned}$$

Now min EV (\bar{y}_R) will correspond to

$$\max \frac{\left[\sum (R_1 - \bar{R}_N)(F_1 - \bar{F}_N) \right]^2}{\sum (R_1 - \bar{R}_N)^2 \sum (F_1 - \bar{F}_N)^2}$$

whose minimum upper bound is 1 and is attained when

$$F_1 = a + b R_1 \quad \dots\dots\dots(5.56)$$

i.e. a linear function of R_1

when $v(x_{11}, x_{12}, \dots, x_{1p}) \neq \text{Constt.}$, we have to maximise

$$\phi = \frac{\left[\sum (R_1 - \bar{R}_N)(F_1 - \bar{F}_N) \right]^2 + \sum v_1 (F_1 - \bar{F}_N)^2}{\sum (F_1 - \bar{F}_N)^2}$$

$$\dots\dots\dots(5.57)$$

$$= \frac{(\sum R_1 G_1)^2 + \sum v_1 G_1^2}{\sum G_1^2} ; \sum G_1 = 0,$$

$$\frac{d\phi}{dG_1} = \frac{2(\sum R_1 G_1) R_1 + 2 v_1 G_1}{\sum G_1^2} - \frac{(\sum R_1 G_1)^2 + \sum v_1 G_1^2}{(\sum G_1^2)^2} \cdot 2 G_1$$

Therefore $\frac{d(\beta + 2 \sum Q_1)}{d Q_1} = 0$ gives.

$$(\sum Q_1^2) [v_1 Q_1 + R_1 \sum R_1 Q_1] - [(\sum R_1 Q_1)^2 + \sum v_1 Q_1^2] Q_1 \\ \cdot \lambda (\sum Q_1^2)^2 = 0$$

Summing over 1,

$$\sum R_1 \sum R_1 Q_1 \sum Q_1^2 + \sum Q_1^2 \sum v_1 Q_1 + N \lambda (\sum Q_1^2)^2 = 0$$

or $\sum R_1 \sum R_1 Q_1 + \sum v_1 Q_1 + N \lambda \sum Q_1^2 = 0$

or $\lambda = - \frac{\sum R_1 \sum R_1 Q_1 + \sum v_1 Q_1}{N \sum Q_1^2}$ (5.58)

5.6 Comparison between simple regression estimate and varying Probabilities regression estimate.

$$V(\bar{y}_R) = \frac{1}{n} \sigma_y^2 (1 - \rho_{R,F}^2) \dots\dots\dots(5.61)$$

$$V(\bar{y}'_R) = \frac{1}{n} \sigma_{y'}^2 (1 - \rho_{y',F'}^2) \dots\dots\dots(5.62)$$

For minimising $V(\bar{y}'_R)$, probability should be chosen

such that $\rho_{y',F'}^2$ is maximum i.e. 1;

$$\rho_{y',F'}^2 = 1 \text{ if } \frac{y_1}{NP_1} - \bar{y}_N = \frac{F_1}{NP_1} - \bar{F}$$

or $P_1 = \frac{F_1 - y_1}{\sum (F_1 - y_1)}$

Now by taking $P_1 = \frac{(P_1 - y_1)^2}{\sum_{i=1}^N (P_1 - y_1)^2}$

and obtaining expected probabilities

$$E(P_1) = \frac{E(P_1 - y_1)^2}{\sum E(P_1 - y_1)^2} = \frac{\sigma_1^2}{\sum \sigma_1^2} = \frac{V_1}{\sum V_1} \dots\dots\dots (5.63)$$

$$E[V(\bar{y}'_R)] = \frac{1}{n} \sigma_{R'}^2 (1 - \rho_{R', F'}^2) + \frac{1}{nN^2} \sum_{i=1}^N \frac{(1 - P_1)}{P_1} V_1 \cdot \frac{1}{nN^2} \frac{\sum V_1 (\frac{P_1}{P_1} - \bar{P})^2}{P_1} \dots\dots\dots (5.64)$$

Taking $R' = F'$ and $P_1 = \frac{V_1}{\sum V_1}$

we obtain

$$E[V(\bar{y}'_R)] = \frac{N - 2}{nN^2} \sum V_1 \dots\dots\dots (5.65)$$

If all $V_1 = \sigma^2$

$$E[V(\bar{y}'_R)] = \frac{N - 2}{nN} \sigma^2 = E[V(\bar{y}_R)] \dots\dots (5.66)$$

Hence we see that there is not much hope of improvement by introducing varying probability when all the information has already been used in regression estimate.

(5.7) In a particular case when $p = 2$, i.e. given two auxiliary variates x_{11} and x_{12} for each unit of the character under study say y_1 . We can form two types of estimates (i) regression estimate (ii) p.p.s. regression estimate.

For simplicity sake we assume that x_1 and x_2 are linearly related with y of the form

$$y_1 = a_1 x_{11} + a_2 x_{12} + e_1 \quad \dots\dots\dots (5.71)$$

Estimates are given by

$$(i) \bar{y}_R = \bar{y}_n + \rho_{y,F} (\bar{F}'_N - F_1) \frac{\sigma_y}{\sigma_F} \quad \dots\dots\dots (5.72)$$

where $F_1 = a_1 x_{11} + a_2 x_{12}$

$$V(\bar{y}_R) = \frac{\sigma_y^2}{n} (1 - \rho_{y,F}^2) \quad \dots\dots\dots (5.73)$$

(ii) In this estimate we use x_{11} for the regression and x_{12} for the varying probabilities.

$$\bar{y}'_R = \bar{y}'_n + B (\bar{F}'_N - F'_1) \quad \dots\dots\dots (5.74)$$

where $y'_1 = \frac{y_1}{NP_1}$; $\bar{y}'_n = \frac{1}{n} \sum_{i=1}^n \frac{y_1}{NP_1}$

$\bar{F}'_1 = \frac{x_{11}}{NP_1}$; $\bar{F}'_N = \bar{x}_1$

$$P_1 = \frac{x_{12}}{\sum_{i=1}^n x_{12}}; \quad B = \frac{\sum P_1 (\frac{y_1}{NP_1} - \bar{y}'_n) (\frac{x_{11}}{NP_1} - \bar{x}_1)}{\sum P_1 (\frac{x_{11}}{NP_1} - \bar{x}_1)^2}$$

(3.7.1) Comparison of expected variances of the two estimates.

$$E V (\bar{y}_R) = \frac{N-2}{nN} \sigma^2 \quad \text{where } \sigma^2 = \frac{\sum e_i^2}{N}$$

$$E V (\bar{y}'_R) = \frac{\sigma^2 R_1^2}{n} (1 - \rho_{R_1, P_1}^2) = \frac{1}{nN^2} \sum_{i=1}^N \frac{(1-p_1) V_1}{P_1} - \frac{1}{nN^2} \sum V_1 \left(\frac{P_1}{N P_1} - \bar{P}_N \right)^2$$

where,

$$R_1^2 = \frac{a_1 x_{11} + a_2 x_{12}}{N P_1}$$

and V_1 is the variance ignoring x_{12} .

Let $V_1 = \text{Const.} = \sigma'^2$

Then after simplification,

$$E V (\bar{y}'_R) = \frac{\sigma'^2}{nN^2} \left\{ \sum \frac{1}{P_1} - N - \frac{\sum \left(\frac{P_1}{N P_1} - \bar{P} \right)}{\sum P_1 \left(\frac{P_1}{N P_1} - \bar{P} \right)^2} \right\}$$

$$= \frac{\sigma'^2}{nN^2} \left\{ \sum \frac{1}{P_1} - N - \frac{\sum \left(\bar{X}_2 \frac{x_{11}}{x_{12}} - \bar{X}_2 \right)}{\sum x_{12} \left(\bar{X}_2 \frac{x_{11}}{x_{12}} - \bar{X}_2 \right)^2} \right\}$$

.....(3.7.11)

The estimate \bar{y}_R will be more efficient as compared to the estimate \bar{y}'_R if

$$V(\bar{y}_R) \leq V(\bar{y}'_R)$$

$$\text{or } \frac{N-2}{nN} \sigma^2 \leq \frac{\sigma^2}{nN^2} \left[\sum \frac{1}{P_1} - N \frac{\sum (\bar{x}_2 \frac{x_{11}}{x_{12}} - \bar{x}_1)}{\sum x_{12} (\bar{x}_2 \frac{x_{11}}{x_{12}} - \bar{x}_1)^2} \right] N \bar{x}_2$$

$$\text{or } N(N-2) \frac{\sigma^2}{N^2} \leq N \sum \frac{1}{P_1} - N \bar{x}_2 \frac{\sum (\bar{x}_2 \frac{x_{11}}{x_{12}} - \bar{x}_1)}{\sum x_{12} (\bar{x}_2 \frac{x_{11}}{x_{12}} - \bar{x}_1)^2}$$

.....(5.7.12)

Now under the assumptions $\sigma^2 \leq \sigma'^2$.

Hence L.H.S. is always lesser than $N(N-1)$ and in most of the cases R.H.S. will be greater than $N(N-1)$.

Because dominating term of the R.H.S. is $\sum \frac{1}{P_1}$ which

attains minimum when $P_1 = \frac{1}{N}$ and in that case the value of

R.H.S. in most of the cases will, be greater than $N(N-1)$.

Therefore in most of the cases it is advisable to use the information as regression estimate for obtaining precise estimates.

5.8 Comparison of regression estimate with (p.p.s.)_{opt} estimate, under the linear model.

$$V(\bar{y}_R) = \frac{N-2}{nN} \sigma^2 \quad \text{where } Y_1 = a + b R_1$$

$$V(\bar{y}'_n)_{opt} = \frac{1}{n} \left[\frac{1}{N^2} (\sum \alpha_1)^2 - \frac{R_N^2}{N} - \frac{\sigma^2}{N} \right]$$

where $\alpha_1 = \sqrt{R_1^2 + V_1}$

and $V_1 = \text{Const.} = \sigma^2$

$$V(\bar{y}_R) \leq V(\bar{y}'_n)_{opt}$$

If

$$\frac{N-2}{nN} \sigma^2 \leq \frac{1}{n} \left[\frac{1}{N^2} (\sum \alpha_1)^2 - \frac{R_N^2}{N} - \frac{\sigma^2}{N} \right]$$

$$\frac{N-2}{N} \sigma^2 \leq \frac{1}{N^2} (\sum \alpha_1)^2 - \frac{1}{N} \sum \alpha_1^2 + \frac{1}{N} \sum \alpha_1^2 - \frac{R_N^2}{N} - \frac{\sigma^2}{N}$$

$$\frac{N-1}{N} \sigma^2 \leq V(R_1) - V(\alpha_1) + \sigma^2$$

$$\frac{\sigma^2}{N} \leq V(R_1) - V(\alpha_1)$$

$$\text{or } V(\alpha_1) - V(R_1) \leq \frac{\sigma^2}{N}$$

$$\text{or } \frac{1}{N} \sum \alpha_1^2 - \frac{\sum \alpha_1 \cdot \sum \alpha_1}{N^2} - \frac{1}{N} \sum R_1^2 + \frac{\sum R_1^2 \cdot \sum R_1 R_j}{N^2}$$

$$\leq \frac{\sigma^2}{N}$$

$$\text{or } \frac{1}{N} \sum R_1^2 + \sigma^2 - \frac{1}{N^2} \sum R_1^2 - \frac{\sigma^2}{N} - \sum_{i \neq j} \frac{\alpha_i \alpha_j}{N^2} - \frac{1}{N} \sum R_1^2$$

$$\cdot \frac{\sum R_1^2}{N^2} + \frac{\sum_{i \neq j} R_1 R_j}{N^2} \leq \frac{\sigma^2}{N}$$

$$\text{or } \frac{1}{N^2} \sum_{i \neq j} (R_1 R_j - \alpha_i \alpha_j) \leq \frac{2-N}{N} \sigma^2$$

$$\text{or } N(N-2) \sigma^2 \leq \sum_{i \neq j} (\alpha_i \alpha_j - R_1 R_j)$$

$$\text{or } N(N-2) \sigma^2 \leq \sum_{i \neq j} \left[\sqrt{R_1^2 R_j^2 + \sigma^2 (R_1^2 + R_j^2) + \sigma^4} - R_1 R_j \right]$$

$$\text{R.H.S.} = \sum_{i \neq j} \left[\sqrt{R_1^2 R_j^2 + \sigma^2 (R_1^2 + R_j^2) + \sigma^4} - R_1 R_j \right]$$

$$> \sum_{i \neq j} \left[\sqrt{R_1^2 R_j^2 + 2\sigma^2 R_1 R_j + \sigma^4} - R_1 R_j \right]$$

$$> \sum_{i \neq j} \left[R_1 R_j + \sigma^2 - R_1 R_j \right]$$

$$> N(N-1) \sigma^2$$

Hence R.H.S. > L.H.S.

Therefore, regression estimate is always better than (p.p.s.) in the present model.

S_U_M_M_A_R_Y

In the present study, successive sampling with unequal probabilities has been considered in the first three chapters. And in the last the technique of combining the information on ancillary characters has also been studied.

In the successive sampling three cases have been considered i.e. (1) when the units at the first occasion some units are retained with probability proportional to the sizes of the units as observed on the previous occasion. The results are compared with regression estimate which shows that the regression estimate is better as compared to the estimate based on probability proportional selection, when N the number of units in the population is large which is evidently shown by Desh Raj (1958) in the Uniphase sampling. But if N and n_p the no. of units retained are small the estimate based on probability proportional is more efficient as compared to the estimate based on regression estimate.

In the 2nd case when some information on auxiliary variate is given on the 1st occasion the units are selected with probability proportional to the sizes of the auxiliary variate and at the 2nd occasionsub sample is selected by the method of simple random sampling. Whereas in the 3rd case at both the occasions units are selected with probability proportional to the sizes determined by the auxiliary variable.

In the last when more than one auxiliary characters correlated with the variate under consideration are given, the techniques of ratio, regression unequal probability stratification are combined to produce more efficient estimates and they are compared with the usual methods.

1

R E F E R E N C E S

- (1) Jenson, R.J. 1942 Statistical investigation of a Sample Survey for obtaining farm facts; Iowa Agri. Exp. Stat.; Res. Bull. 304
- (2) Patterson, H.B. 1950 Sampling on successive occasions with partial Replacement of units; J.R.S.S.; Series B.12. 241 - 255.
- (3) Tikkiwal, B.B. 1951 Theory of Successive Sampling Thesis for Diploma I.A.R.S. (I.C.A.E.)
- (4) Tikkiwal, B.B. 1953 Optimum allocation in Successive sampling Jour. Ind. Soc. Agr. Stat; 5. 100 - 102.
- (5) Singh, B 1954 On efficiency of the sampling with varying probabilities without replacement. J. Ind. Soc. Agr. Stat. 4. 48 - 57.
- (6) Eckler, A.R. 1955 Rotation sampling, A.M.S. Vol. 26. 1955 (664, 625)
- (7) Tikkiwal, B.B. 1955 Multiphase Sampling on Successive occasions. Ph.D. Thesis; North Carolina State College.
- (8) Tikkiwal, B.B. 1956 Some further contribution to the theory of univariate Sampling on Successive Occasion.
- (9) Dosh Raj 1958 On the Relative Accuracy of some Sampling Techniques Vol. 53, 98 - 101.
- (10) Singh, B 1959 Estimates in Successive Sampling of Multi-stage Design - Paper read at the 12th Annual Meeting of the Indian Society of Agricultural Statistics held at Gwalior.
- (11) Hartley, H.G. and 1962 Sampling with unequal probabilities and without replacement; A.M.S. 33, 350 - 374.
Rao, J.N.K.
- (12) Hartley, H.G. 1962 On a Simple procedure of unequal probability Sampling without replacement.
Rao, J.N.K. and J.R.S.S. (B) 24, 482 - 490.
Cochran, W.G.

- (13) Tikkiwal B. B. 1965 The Theory of Two-stage Sampling on Successive occasions, J. Ind. Stat. Ass. Vol. 2. No. 2 & 3 1965.
- (14) B. Singh and B. B. Singh 1965 Some contributions to two phase Sampling. Austral. J. Statist. 7(2). 1965, 45 - 47.