Sankhyā: The Indian Journal of Statistics
1993, Special Volume 55, Series B, Pt. 3, pp. 372-384
Dedicated to the memory of P. C. Mahalanobis

ON CERTAIN PROBLEMS OF SAMPLING DESIGNS AND ESTIMATION FOR MULTIPLE CHARACTERISTICS[1]

By T. J. RAO
Indian Statistical Institute, Calcutta

SUMMARY. We shall first review some of the sampling techniques involving multiple auxiliary variables. In certain situations, when we have many study variables, it is of interest to estimate parameters relating to these variables. However, some of the study variables may be poorly correlated with the selection probabilities when probability proportional to size sampling technique is being used. In this article, we shall also discuss how to provide alternative estimators in such cases. Next we consider multivariate stratified surveys and finally touch upon some of the recent developments including analysis of complex surveys based on several variables.

1. INTRODUCTION

The importance as well as intricacies of designing sample surveys for many variables were recognised by Mahalanobis even while methodology for univariate surveys was being developed. While discussing large scale sample surveys, he (1944) commented :

"The above examples sufficiently illustrate the general principles for uni-stage sampling in the case of a single variate. Solutions (with special forms for cost and variance functions) for multi-variate and uni-stage sampling or multi-stage univariate or multi-stage multi-variate sampling are not considered here. Material for satisfactory graduation of appropriate cost functions and variance or covariance functions is accumulating for certain crops, and it is hoped to deal with the problem in a later paper. Some simple artificial functions might have been discussed here, but for purposes of elucidation that is scarcely worthwhile, since the uni-variate uni-stage examples provide sufficient illustration of the general principles".

Subsequently, these observations were echoed in Mahalanobis (1952) where he visualised the following:

" In case of repeated surveys for single character (with added complications when several characters are to be estimated, as is usually the case) we need some sort of a 'composite' error defined in terms of standard errors of the several estimates calculated on the basis of the sample design."

[1]Based on the talk given at the International Symposium on Multivariate Analysis held in December 1992 at New Delhi.

Also while discussing the possibility of using the concept of analysis of dispersion in sample surveys for multiple variables, Chakravarti (1954) observed that *practical experimenters fight shy of the methods of multivariate analysis and application has failed to keep pace with the development of the theory*. However, by the sixties and seventies there has been tremendous growth in the literature which encompasses the use of multiple auxiliary variables for designing and estimation in sample surveys. On the other hand, inference for sample surveys of many variables is relatively of recent origin, born out of the increasing need for the analysis of complex surveys. In this paper, we shall review some of these topics, highlighting only certain important aspects.

2. Multipurpose surveys

The integrated multi-subject nature of the National Sample Survey (NSS) initiated by Mahalanobis in India in 1950 goes beyond the realm of designing of sample surveys for several variables in the sense that not only data is collected on different variables for the sample unit but data on different subjects from sample units which are necessarily different is obtained for certain surveys (Lahiri (1954, 1964), Murthy (1964)). In multipurpose surveys where estimation of parameters of several variables is involved, it is usually found that different sets of selection probabilities would suit different characters better. For example, in the NSS, for household enquiries probabilities of selection of units use data on the variate 'population' while for land utilization surveys the variate 'area' is considered.

In order to reduce field costs, in such situations, it is desirable to have a selection scheme which makes the sample units (villages) for both enquiries more or less identical. Lahiri (1954) has suggested 'serpentine method' and 'two-dimensional method' to achieve this while Des Raj (1956) developed the methodology for this, following Dantzig's solution for the transportation problem. Maczynski and Pathak (1984) studied the more general problem of integration of $k \geq 2$ surveys, while Mitra and Pathak (1984) provided algorithms for optimal integration for the case of two and three survey variables (see also Mitra (1988)).

3. Multiple auxiliary information

Historically, data collected on several auxiliary variables for inference purposes has been utilized by users and advocates of purposive selection method. Jensen (1926) gave an example from a study in Denmark where ten auxiliary variates were utilized successfully. At the same time, examples of complete failure of purposive selection method used in Gini and Galvani's (1929) balanced samples based on seven variables are too well known. For measuring the yield of Cinchona bark, Mahalanobis in 1940 suggested the use of three simple physical measurements for regressing (1946).

An important milestone with regard to the utilization of multivariate auxiliary information was due to Olkin (1958) who extended the ratio estimator to the case when data on p auxiliary variables is available. However, it was Ghosh (1947) who in a small note first envisaged the concept of double sampling with many auxiliary variables. Olkin (1958) considered the multivariate ratio estimator

$$\hat{Y}_{MR} = \sum_{i=1}^{p} \omega_i \hat{Y}_{R_i}$$

where $\hat{Y}_{R_i}$ is the ratio estimator of the population total Y of the study variable using the i-th auxiliary variable, $i = 1, 2, \cdots, p$ and the weights ω_i's are obtained such that $V(\hat{Y}_{MR})$ is minimised.

This approach has been followed by several authors who developed multivariate product, multivariate difference and multivariate regression estimators as well as weighted combinations of ratio (product) estimators for auxiliary variables positively (negatively) correlated with study variate (cf. Des Raj (1965), Singh (1965, 1967a, 1967b), Srivastava (1965), Rao and Mudholkar (1967)).

It may be pointed out here (also see Rao (1991a)) that certain authors erroneously combine a ratio estimator and a product estimator both based on the same single auxiliary variable or combine ratio and product estimators based on several auxiliary variables without regard to their correlation with the study variable. The question of determining optimum weights in multivariate ratio, product and regression estimators was discussed by Tripathi (1978) while Bedi (1985) considered the two-phase multivariate estimator. A class of estimators based on general sampling design and multivariate auxiliary information was given in Tripathi (1987). For a review of the use of auxiliary information which includes the multivariate situation, we refer to Tripathi *et al.* (1990) and Adhvaryu (1986).

Srivastava's (1967) estimator $\hat{Y}_R^{(\alpha)} = \hat{Y}(X/\hat{X})^{\alpha}$ obtained by 'exponentiation' has been interpreted by Rao (1991b) as a 'repeated substitution method' where starting from $\hat{Y}_R^{(1)} = \hat{Y}(X/\hat{X})$, the *better* estimator $\hat{Y}(X/\hat{X})$ is substituted in the place of $\hat{Y}$, thereby obtaining $\hat{Y}_R^{(\alpha)} = \hat{Y}(X/\hat{X})^{\alpha}$ after α interations. Rao (1991b) has demonstrated that in most populations, the optimum value of α which is equal to β/R, where β is the population regression coefficient and R is the ratio of the totals of study and auxiliary variables, is close to unity and hence one does not effectively gain anything by this 'exponentiation'. However, several authors tend to use this technique without any motivation. On the positive side, the 'repeated substitution method' leads to an improvement in estimation as can be seen in the following :

Suppose that y is the study variable and x and z are two auxiliary variates with respective population totals Y, X and Z. Consider the ratio estimator for Y given by

$$\hat{Y}_R^{(2)} = \{\hat{Y}(X/\hat{X})\}(Z/\hat{Z}).$$

Writing, as usual $\hat{Y} = Y(1 + e_1)$, $\hat{X} = X(1 + e_2)$, $\hat{Z} = Z(1 + e_3)$ where $E(e_1) = E(e_2) = E(e_3) = 0$ it can be easily shown that upto second degree approximation,

$$V(\hat{Y}_R^{(2)}) = V(\hat{Y}(X/\hat{X})) + Y^2(c_z^2 - 2\rho_{yz}c_yc_z - 2\rho_{xz}c_xc_z) < V(\hat{Y}(X/\hat{X}))$$

$$\text{if } \rho_{yz}\frac{c_y}{c_z} - \rho_{xz}\frac{c_x}{c_z} > \frac{1}{2}, \text{ where } c \text{ stands for } c.v..$$

When $c_x \simeq c_y \simeq c_z$, the condition reduces to $\rho_{yz} - \rho_{xz} > \frac{1}{2}$ which is likely to be true when the correlation between y and z is very high while there is less correlation between x and z (also see Singh (1967b)).

Next consider the repeated substitution estimate for the difference method given by

$$\hat{Y}_D^{(2)} = \hat{Y}_D^{(1)} + \beta_2(X - \hat{X})$$

$$\text{with} \quad \hat{Y}_D^{(1)} = \hat{Y} + \beta_1(X - \hat{X})$$

where $\beta_1 = \text{Cov. } (\hat{Y}, \hat{X})/V(\hat{X}) = \beta_{yx}$ and $\beta_2 = \text{Cov } (\hat{Y}_D^{(1)}, \hat{X})/V(\hat{X})$.

It was shown in Rao (1991b) that $\beta_2 = 0$ and thus there is no possibility of improvement by this method. However, when auxiliary information on two variables, say x and z is available we consider

$$\hat{Y}_D'' = [\hat{Y} + \beta_1(X - \hat{X})] + \beta_2(Z - \hat{Z})$$

where this time

$$\beta_2 = \text{Cov}(\hat{Y}_D^{(1)}, \hat{Z})/V(\hat{Z}) = \beta_{yz} - \beta_{yx}\beta_{xz}.$$

Hence

$$V(\hat{Y}_D'') = V(\hat{Y})(1 - \rho_{yx}^2) + \Delta$$

where

$$\begin{aligned} \Delta &= \beta_2^2 V(\hat{Z}) - 2\beta_2 \text{ Cov } (\hat{Y}, \hat{Z}) + 2\beta_1\beta_2 \text{ Cov } (\hat{X}, \hat{Z}) \\ &= -\beta_2^2 V(\hat{Z}) \\ &< 0. \end{aligned}$$

Thus $V(\hat{Y}_D'') < V(\hat{Y}_D^{(1)})$ and there is a definite improvement.

Alternatively, one can show that

$$\begin{aligned} V(\hat{Y}_D'') &= V(\hat{Y}_D{}^{(1)})(1 - \rho_{\hat{Y}_{D_1},z}^2) \\ &< V(\hat{Y}_d^{(1)}). \end{aligned}$$

The above can be extended to regression estimation on the lines of Rao (1991b) and to the situation when we have $p > 2$ auxiliary variables.

4. ALTERNATIVE ESTIMATORS FOR PPS SAMPLING

Auxiliary information on a variable x related to the study variable y can be profitably used for selecting the sample units with probabilities proportional to size (PPS) x. However, it was demonstrated by Des Raj (1954) that a PPS sampling estimation could be worse than SRS estimation if the regression line of y on x is far from the origin. Reddy and Rao (1977) considered a transformed variable $X_i' = X_i + d\overline{X}$, use of which produced very efficient estimators for PPX′ sampling.

Next consider the situation where we have information on two auxiliary variables leading to selection probabilities p_i and p_i' with $\sum p_i = \sum p_i' = 1$. Let $p_i'' = \alpha p_i + (1-\alpha)p_i', 0 \le \alpha \le 1$. For PPS With Replacement sampling, we have

$$\begin{aligned} V(\hat{Y}'') - V(\hat{Y}) &= \frac{1}{n}\sum_1^N Y_i^2\left(\frac{1}{\alpha p_i + (1-\alpha)p_i'} - \frac{1}{p_i}\right) \\ &= \frac{(1-\alpha)}{n}\sum_1^N Y_i^2\left\{\frac{p_i - p_i'}{p_i(\alpha p_i + (1-\alpha)p_i')}\right\}. \end{aligned}$$

Also

$$V(\hat{Y}') - V(\hat{Y}) = \frac{1}{n}\sum_1^N \frac{Y_i^2(p_i - p_i')}{p_i p_i'}$$

where $\hat{Y}, \hat{Y}', \hat{Y}''$ denote the estimators with the corresponding probabilities of selection p_i, p'_i, p_i'' respectively. Now

$$\begin{aligned} V(\hat{Y}'') - V(\hat{Y}) &= \frac{(1-\alpha)}{n}\left\{\sum_1^N \frac{Y_i^2(p_i - p'_i)}{p_i p'_i}\right. \\ &\qquad \left. + \sum_1^N \frac{Y_i^2(p_i - p'_i)}{p_i}\left(\frac{1}{\alpha p_i + (1-\alpha)p'_i} - \frac{1}{p_i'}\right)\right\} \\ &= (1-\alpha)\{V(\hat{Y}') - V(\hat{Y})\} \\ &\qquad - \frac{(1-\alpha)\alpha}{n}\sum_1^N \frac{(p_i - p'_i)^2 Y_i^2}{p_i p'_i(\alpha p_i + (1-\alpha)p'_i)} \\ &\le (1-\alpha)\{V(\hat{Y}') - V(\hat{Y})\} \end{aligned}$$

or

$$V(\hat{Y}'') \le \alpha V(\hat{Y}) + (1-\alpha)V(\hat{Y}').$$

Remark 4.1 : If $V(\hat{Y}) < V(\hat{Y}')$, then $V(\hat{Y}'') < V(\hat{Y}')$ and if $V(\hat{Y}') < V(\hat{Y})$, then $V(\hat{Y}'') < V(\hat{Y})$. This shows that selection with probabilities p''_i is better than the worse of selection with p_i and p'_i.

Remark 4.2 : When $p_i' = 1/N$, i.e., SRS is used, Reddy and Rao's (1977) Theorem 3.1 follows.

Remark 4.3 : The above result can easily be extended to the case of multivariate ($q > 1$) auxiliary information leading to

$$V(\hat{Y}_{PPS}{}^{(q)}) < \sum_{i=1}^{q} \alpha_i V(\hat{Y}^i{}_{PPS})$$

where $(\hat{Y}^i)$ is based on probabilities of selection p_j^i's, $j = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, q$ and $\hat{Y}_{PPS}^{(q)}$ is based on probabilities of selection $\sum_{i=1}^{q} \alpha_i p_j^i$ with $\sum_1^q \alpha_i = 1$. Agrawal and Singh (1980) and Tripathi and Chaubey (1990) considered the use of a function of the multivariate auxiliary information as size for obtaining optimum probabilities of selection.

In sample surveys of many variables, some of the study variables may be poorly correlated with the selection probabilities used for PPS sampling. J.N.K. Rao (1966) has provided alternative estimators when the study variable and size measure are unrelated and demonstrated that these alternative estimators are more efficient though biased. Bansal and Singh (1985) noticed that J.N.K. Rao's model deals with only zero correlation and hence developed a new estimator of the population total for characteristics that are poorly correlated with the selection probabilities. Simple alternatives were suggested by Amahia, Chaubey and Rao (1989) who considered the class of estimators

$$\begin{aligned} \hat{Y}_p &= \frac{1}{n} \sum_{1}^{n} \frac{y_i p_i / p_i^*}{p_i} \\ &= \frac{1}{n} \sum y_i / p_i^* \end{aligned}$$

where $p^* = (1 - \alpha)\frac{1}{N} + \rho p_i$, $i = 1, 2, \cdots, N$ which takes into account the correlation ρ between y and x.

Amahia *et al.* (1989) also considered another alternative

$$\hat{Y}'_p = \frac{1}{n} \sum_{1}^{n} \frac{y_i}{p_i'}$$

where

$$p_i' = [\,(1 - \rho)N + \frac{\rho}{p_i}\,]^{-1}$$

and determined the efficiency and robustness of these estimators. For PPS sampling without replacement for multiple characteristics, Rao (1987) discussed estimators alternative to Horvitz-Thompson estimator, Rao-Hartley-Cochran estimator and Murthy's estimator.

5. STRATIFIED SAMPLING AND OPTIMIZATION

Neyman (1934) in his celebrated paper discussed multi-variate stratified sampling. When several highly correlated variables are considered the minimum variance allocation for a particular variable will itself yield a compromise allocation for the other variables of the survey. Neyman's prescription was to 'sample proportionately to the size of strata' since 'in many cases the characters sought are not likely to be highly correlated'. He further (1938) recommended that if there are many variables of equal importance a 'basic characteristic' which is correlated with the ones we are interested in should be found and adjustments be made according to this 'basic characteristic'.

Peters and Bucher (1940) derived an allocation n_i maximizing $\sum_1^p e_j$ where e_j is the relative efficiency for the j-th variate defined as the ratio of variances with optimum allocation V_0^j and compromise allocation V_a^j. Dalenius (1953) proposed the minimisation of $\sum_{j=1}^p (V_a^j - V_0^j)/V_o^j$ subject to $\sum n_i^a = n$ where n_i^a is a compromise allocation. Hansen, Hurwitz and Madow (1953) discussed some practical problems. Under SRS, Chatterjee (1968) suggested minimisation of average relative increase in variance due to the use of actual allocation instead of optimum allocation, averaged over p variables, *i.e.*,

$$\text{minimise} \quad \frac{1}{p}\sum_{j=1}^{p}[\frac{1}{n}\sum_{i=1}^{k}(n_{ij}^0 - n_i^a)^2/n_i^a]$$

where n_{ij}^0 is the optimum allocation for the i-th stratum using variable j, which leads to

$$n_i^a = n(\sum_j (n_{ij}^0)^2)^{1/2} / \sum_i (\sum_j (n_{ij}^0)^2)^{1/2}.$$

Yates (1960) considered minimisation of $L = \sum a_j V_j$ subject to $C = C_0 + \sum C_i n_i$ where a_j are known constants and minimisation of C subject to $V_j \leq v_j$, $j = 1, 2, \cdots, p$ and $0 \leq n_i \leq N_i$. Overall measures of deviation for the p-variate case given by

$$D_1 = \sum_{i=1}^{k} \mathbf{d}_i' \mathbf{d}_i \text{ and } D_2 = \text{ Det. } (D'D)$$

have been suggested by Rao (1984) where

$$\begin{aligned} \mathbf{d}_i' &= (d_{i1}, d_{i2}, \cdots, d_{ip}) \\ D &= ((d_{ij})) \text{ and } d_{ij} = (n_{ij}^0 - n_i^a)/\sqrt{n_i^a}. \end{aligned}$$

Rao (1984) as well as Mukerjee and Rao (1985) considered the situation when costs are given and illustrated the efficiencies of all the above allocations.

One of the early users of programming techniques was Kokan (1963) who minimised the cost $C = C_0 + \sum_{i=1}^{k} C_i n_i$ or equivalently maximised

$$\phi = -\sum_{i=1}^{k} \frac{N_i c_i}{1 + N_i x_i}$$

subject to $V_j \leq v_j$ and $0 \leq x_i \leq 1 - \frac{1}{N_i}$, where $x_i = \frac{1}{n_i} - \frac{1}{N_i}$ and V_j is the variance of the estimator (SRS) for the j-th character, by non-linear programming methods. For different types of constraints, Hughes and Rao (1979) have given algorithms for optimal allocation. For a recent paper, we refer to Bethel (1989).

In a different context, the problem of choosing optimum number of primary sampling units (psu's) and secondary sampling units (ssu's) under given costs was discussed by Chakravarti (1953). He considered multistage sampling design for estimating the mean vector $\boldsymbol{\mu}' = (\mu_1, \mu_2, \cdots, \mu_p)$ of p variables. Let $\hat{\mu}_j = \bar{y}_j$ based on n_1 psu's and n_2 ssu's per selected psu. Consider the model

$$\boldsymbol{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}$$

where

$$\begin{aligned}
\boldsymbol{y}_{ij} &= (y_{1(ij)}, y_{2(ij)}, \cdots, y_{p(ij)}) \\
\boldsymbol{\mu}' &= (\mu_1, \mu_2, \cdots, \mu_p) \\
\boldsymbol{\beta}_i' &= (\beta_{1i}, \beta_{2i}, \cdots, \beta_{pi}) \\
\boldsymbol{\varepsilon}_{ij}' &= (\varepsilon_{1(ij)}, \varepsilon_{2(ij)}, \cdots, \varepsilon_{p(ij)})
\end{aligned}$$

for $i = 1, 2, \cdots, n_1$ psu's and $j = 1, 2, \cdots, n_2$ ssu's in each selected psu. Assuming

$$\mathrm{E}\,(\boldsymbol{\beta}_i) = 0, \quad \mathrm{E}\,(\boldsymbol{\varepsilon}_{ij}) = 0, \quad \mathrm{E}\,(\boldsymbol{\beta}_i \boldsymbol{\beta}_i') = \Lambda_1$$

$$\mathrm{E}\,(\boldsymbol{\varepsilon}_{ij} \boldsymbol{\varepsilon}_{ij}') = \Lambda_2 \forall i, j$$

$$\mathrm{E}\,(\boldsymbol{\beta}_i \boldsymbol{\varepsilon}_{ij}') = 0 \ \forall \ i, j,$$

we have the dispersion matrix given by

$$\mathrm{Cov.}\ (\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_p) = \frac{\Lambda_1}{n_1} + \frac{\Lambda_2}{n_1 n_2}.$$

Chakravarti then obtained allocation to two stages of sampling such that the efficiency of estimated mean vector is maximized, subject to a fixed cost $C = a + bn_1 + cn_1 n_2$.

It is interesting to note that even in the early forties, Mahalanobis (1944) was interested in determining the 'optimum size and density of grids' in surveys

dealing with several crops by attaching to the crops arbitrary weights 'determined by, say, the money value of the different variates...'. Alternatively, he formulated the problem as one of requiring the optimum distribution at given (different) levels of error for the different variates – the optimum having reference to minimum total cost.

6. Some theoretical developments

Suppose that we are interested in estimating the total $T_i = \sum_{j=1}^{N} y_{ij}$ of the i-th study variable taking values y_{ij} on the j-th unit of the i-th variable, $j = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, p$ based on a sample selected using the design $p(s)$. Our sample vector consists of

$\boldsymbol{y}_j' = (y_{1j}, y_{2j}, \cdots, y_{pj})$ for each $j \varepsilon s$ and $\hat{\boldsymbol{T}} = (\hat{T}_1, \hat{T}_2, \cdots, \hat{T}_p)$ where $\hat{T}_i = \sum_{j \varepsilon s} \beta_{sij} y_{ij}$. $\hat{\boldsymbol{T}}$ is unbiased for $\boldsymbol{T}$ if $\sum_{s \ni j} \beta_{sij} p(s) = 1 \; \forall \; i = 1, 2, \cdots, p$. The variance covariance matrix of $\hat{\boldsymbol{T}}$ is

$$V(\hat{\boldsymbol{T}}) = \mathrm{E}\,\{(\hat{\boldsymbol{T}} - \boldsymbol{T})(\hat{\boldsymbol{T}} - \boldsymbol{T})'\}.$$

This is symmetric and the i-th diagonal element is equal to the variance of $\hat{T}_i$ and the off-diagonal element $(i\,i')$ is equal to the covariance between $\hat{T}_i$ and $\hat{T}_{i'}$.

Following Godambe (1955), Godambe and Joshi (1965) and Basu (1971), one can visualize a result parallel to non existence theorem which is simply stated below.

Remark 6.1 : There does not exist a best unbiased estimator of the population vector $\hat{\boldsymbol{T}}' = (T_1, T_2, \cdots, T_p)$ when minimization of elements of the variance-covariance matrix of $\hat{\boldsymbol{T}}$ is chosen as the criterion of bestness.

Remark 6.2 : It is easy to see that any criterion which involves a linear combination of variances and covariances results in the non-existence of a best unbiased estimator. Criteria such as generalized variance can also be tried out.

Extending the theory further, it is possible to consider the choice of optimum sampling strategy for estimating the parametric vector $\boldsymbol{T}$ under certain suitably chosen super population models (see among others, Godambe (1955), Cassel *et al.* (1977)).

For notational convenience, assume that $p = 2$ and we wish to estimate the vector $\hat{\boldsymbol{T}}' = (\sum_{i=1}^{N} Y_i, \sum_{i=1}^{N} Z_i)$. To be more specific, consider the vector of Horvitz-Thompson (1952) estimators for the estimation of $\boldsymbol{T}$. We then have

$$\hat{\boldsymbol{T}}' = (\hat{Y}, \hat{Z}) = \left(\sum_{i \varepsilon j} \frac{y_i}{\pi_i}, \sum_{i \varepsilon s} \frac{z_i}{\pi_i}\right) \tag{6.1}$$

Now

$$V(\hat{\boldsymbol{T}}') = \begin{pmatrix} V(\hat{Y}) & \text{Cov. } (\hat{Y}, \hat{Z}) \\ & V((\hat{Z}) \end{pmatrix} \tag{6.2}$$

where

$$V(\hat{Y}) = \sum_1^N (\frac{1}{\pi_i} - 1) Y_i^2 + \sum_{i \neq j}^N \sum^N (\frac{\pi_{ij}}{\pi_i \pi_j} - 1) Y_i Y_j \tag{6.3}$$

and

$$\text{Cov.}(\hat{Y}, \hat{Z}) = \sum_1^N (\frac{1}{\pi_i} - 1) Y_i Z_i + \sum_{i \neq j}^N \sum^N (\frac{\pi_{ij}}{\pi_i \pi_j} - 1) Y_i Z_j. \tag{6.4}$$

When auxiliary information on a variable x is available, related to y and z assume a simple super population model δ of the following type :

$$\left.\begin{array}{lll} \mathcal{E}(y_i|x_i) & = & a_1 x_i \\ \mathcal{E}(z_i|x_i) & = & a_2 x_i \\ \vartheta(y_i|x_i) & = & \sigma_1^2 x_i^2 \\ \vartheta(z_i|x_i) & = & \sigma_2^2 x_i^2 \\ \mathcal{C}(y_i, z_i|x_i) & = & \sigma_3^2 x_i^2 \\ \text{and} \quad \mathcal{C}(y_i, y_j) & = & \mathcal{C}(y_i, z_j) = \mathcal{C}(z_i, z_j) = 0 \end{array}\right\} \tag{6.5}$$

Under the model (6.5) we have the Expected Dispersion Matrix given by :

$$\mathcal{E}_\delta V(\hat{\boldsymbol{T}}) = \begin{pmatrix} \mathcal{E}_\delta V(\hat{Y}) & \mathcal{E}_\delta \text{ Cov. } (\hat{Y}, \hat{Z}) \\ & \mathcal{E}_\delta V(\hat{Z}) \end{pmatrix} \tag{6.6}$$

where

$$\left.\begin{array}{lll} \mathcal{E}_\delta V(\hat{Y}) & = & \sigma_1^2 \sum_1^N (\frac{1}{\pi_i} - 1) x_i^2 + a_1^2 V(\sum_{i \epsilon s} \frac{x_i}{\pi_i}) \\ \mathcal{E}_\delta V(\hat{Z}) & = & \sigma_2^2 \sum_1^N (\frac{1}{\pi_i} - 1) x_i^2 + a_2^2 V(\sum \frac{x_i}{\pi_i}) \\ \mathcal{E}_\delta \text{ Cov. } (\hat{Y}, \hat{Z}) & = & \sigma_3^2 \sum_1^N (\frac{1}{\pi_i} - 1) x_i^2 + a_1 a_2 V(\sum \frac{x_i}{\pi_i}) \end{array}\right\} \tag{6.7}$$

which are obtained from (6.3) and (6.4) using the model assumptions (6.5).

Under the criterion of element-wise minimum or minimum generalized variance, it follows that the strategy $\{\pi Px, \hat{\boldsymbol{T}}' = (\hat{Y}_{HT}, \hat{Z}_{HT})\}$ is δ-optimum for $\hat{\boldsymbol{T}}'$.

Remark 6.3 : If in the above model $\mathcal{E}(z_i|x_i)$ is taken as $a_2 g(x_i)$, $\vartheta(z_i|x_i) = \sigma_2^2 g^2(x_i)$, while πPx design minimizes $\mathcal{E}V(\hat{Y}_{HT})$ and a $\pi Pg(x)$ design minimizes $\mathcal{E}V(\hat{Z}_{HT})$, either a πPx or a $\pi Pg(x)$ design minimizes $\mathcal{E}$ Cov. $(\hat{Y}, \hat{Z})$ according as $\mathcal{C}(y_i, z_i) = \sigma_3^2 x_1^2$ or $\sigma_3^2 (g(x_i))^2$. A relevant choice would perhaps be $\mathcal{C}(y_i, z_i) = \sigma_3^2 x_i g(x_i)$. Also see the comment by Hedayat and Sinha (1991, p. 305). For further details, Holt (1977) and Mukerjee and Sengupta (1989) may be referred to.

If one is interested in a population parameter Θ which is a non-linear function of p totals of the study variables $y_1, y_2, \cdots, y_p$ i.e. $\Theta = f(T_1, T_2, \cdots, T_p)$

then Taylor Linearization technique of variance estimation can be used. Also when the parameters of interest are the regression coefficients $\beta_1, \beta_2, \cdots, \beta_{p-1}$ obtained by fitting $y_p = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_{p-1} y_{p-1}$, a Horvitz-Thompson type π -estimator can be suggested for which variance estimation is done by using Taylor Linearization method. Fur further details on this and related problems we refer to Särndal, Swensson and Wretman (1992).

Most of the large scale sample surveys are complex with stratification, clustering etc.. Koch *et al.* (1975) discussed various strategies for multivariate analysis of data from complex surveys. Bebbington and Smith (1977) studied the effect of survey design on multivariate analytical techniques with particular reference to estimation of correlation matrix and principal component analysis. We also refer to Binder *et al.* (1984) for a detailed bibliography on complex survey data analysis. Further, Smith and Holmes (1989) exhibited a variety of studies to include regression analysis as well. For an excellent reference on these and related topics we refer to Skinner, Holt and Smith (1989).

References

ADHVARYU, D. (1986). Use of auxiliary information in survey sampling at the estimation stage : a review. In *Proceedings of the VIII Annual Conference of ISPS*, Kolhapur, pp. 25–44.

AGRAWAL, S.K. and SINGH, M. (1980). Use of multivariate auxiliary information in selection of units in probability proportional to size sampling with replacement. *J. Ind. Soc. Agr. Stat.*, **32**, 71–81.

AMAHIA, G.N., CHAUBEY, Y.P. and RAO, T. J. (1989). Efficiency of a new estimator in PPS sampling for multiple characteristics, *J. Stat. Plann. and Inf.*, **21**, 75–84.

BANSAL, M.L. and SINGH, R. (1985). An alternative estimator for multiple characteristics. *J. Stat. Plann. and Inf.*, **11**, 313–320.

BASU, D. (1971). An essay on the logical foundations of survey sampling, Part I. In *Foundations of Statistical Inference*, V.P. Godambe and D. A. Sprott (Eds.). Toronto : Holt, Rinehart and Winston, pp. 203–242.

BEBBINGTON, A.C. and SMITH, T.M.F. (1977). The effect of survey design on multivariate analysis. In *The Analysis of Survey data, Vol. 2, Model Fitting* , C. A. O'Muircheartaigh and C. Payne (Eds.). Wiley, New York. pp. 175–192.

BEDI, P.K. (1985). On two-phase multivariate sampling estimator. *J. Ind. Soc. Agr. Stat.*, **37**, 158–162.

BETHEL, J. (1989). Sampling allocation in multivariate surveys. *Survey Methodology*, **15**, 47–57.

BINDER, D. A., GRATTON, M., JEAYS, M., KRIGER, G., KUMAR, S., PATON, D., PATRIK C. and VAN BAAREN, A. (1984). Selected bibliography of data analysis for complex surveys. *Survey Methodology*, **10**, 119–125.

CASSEL, C. M., SÄRNDAL, C.E. and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling.* Wiley, New York.

CHAKRAVARTI, I. M. (1954). On a problem of planning a multistage survey for multiple correlated characters. *Sankhyā*, **14**, 211–216.

CHATTERJEE, S. (1968). Multivariate stratified surveys. *J. Amer. Stat. Assoc.*, **63**. 530–534.

DALENIUS, T. (1953). The multivariate sampling problem. *Skand. Akt.*, **36**, 92–102.

DES RAJ (1954). On sampling with probabilities proportionate to size. *Ganita*, **5**, 175–182.

——— (1956). On the method of overlapping maps in sample surveys. *Sankhyā*, **17**, 89–98.

——— (1965). On a method of using multiauxiliary information in sample surveys. *J. Amer. Stat. Assoc.*, **60**, 270–277.

GHOSH, B. (1947). Double sampling with many auxiliary variables. *Cal. Stat. Assoc. Bull.*, **1**, 91–93.

GINI, C. and GALVANI, L. (1929) : Di una applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione (1 December, 1921). *Annali di Statistica, Series 6*, **4**, 1–107.

GODAMBE, V. P. (1955). A unified theory of sampling from finite populations, *J. R. Statist. Soc., Ser. B*, **17**, 269–278.

GODAMBE, V.P. and JOSHI, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Ann. Math. Stat.*, **36**, 1707–1722.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory*, Vol. I. Wiley, New York.

HEDAYAT, A. S. and SINHA, B. K. (1991). *Design and Inference in Finite Population Sampling*. Wiley, New York.

HOLT, D. (1977). Correlation analysis using survey data. *Bull. Int. Stat. Inst.*, **47**, 4, 228–231.

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Stat. Assoc.*, **47**, 663–685.

HUGHES, E. and RAO, J. N. K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Comm. in. Stat., Part A*, **8**, 1551–1574.

JENSEN, A. (1926). The representative method in practice. *Bull. Int. Stat. Inst.*, **22**, 359–380.

KOCH, G. G., FREEMAN, D. H. (JR.) and FREEMAN, J. L. (1975). Strategies in the multivariate analysis of data from complex surveys. *Int. Stat. Rev.*, **43**, 59–78.

KOKAN, A. R. (1963). Optimum allocation in multivariate surveys. *J. Roy. Stat. Soc., Ser. A*, **126**, 557–565.

LAHIRI, D. B. (1954). Technical Paper on Some Aspects of the Development of Sample Design. National Sample Survey, No. 5, Government of India and *Sankhyā*, **14**, 264–316.

——— (1964). Multi-subject sample survey system. In *Contributions to Statistics*, C. R. Rao (Ed.). Statistical Publishsing Society, Calcutta, pp. 283–316.

MACZYNSKI, M. J. and PATHAK, P. K. (1980). Integration of Surveys. *Scand. J. Stat.*, **7**, 130–138.

MAHALANOBIS, P. C. (1944). On large scale sample surveys. *Phil. Trans. Roy. Soc.*, London, Series B, **231**, 329–451.

——— (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Stat. Soc.*, **109**, 325–378.

——— (1952). Some aspects of the design of sample surveys. *Sankhyā*, **12**, 1–7.

MITRA, S. K. (1988). On the method of overlapping maps in survey sampling. *Presidential address : Statistics Section*, Indian Science Congress Association.

MITRA, S. K. and PATHAK, P. K. (1984). Algorithims for optimal integration of two or three surveys. *Scand. J. Statist.*, **11**, 257–263.

MUKERJEE, R. and RAO, T. J. (1985). On a problem of allocation of sample size in stratified random sampling. *Biom. Journal*, **27**, 327–331.

MUKERJEE, R. and SENGUPTA, S. (1989). Optimal estimation of finite population total under a general correlated model. *Biometrika*, **76**, 789–794.

MURTHY, M.N. (1964). On Mahalanobis' contributions to the development of sample survey theory and methods. In *Contributions to Statistics*, C. R. Rao (Ed.). Statistical Publishing Society, Calcutta, pp. 283–316.

NEYMAN, J. (1934). On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection. *J. Roy. Stat. Soc.* , **97**, 558–625.

——— (1938). Lectures and Conferences on Mathematical Statistics. Washington, D. C.

OLKIN, I. (1958). Multivariate ratio estimators for finite populations. *Biometrika*, **45**, 154–165.

PETERS, J. H. and BUCHER, M. L . (1940). *The 1940 Section Sample Survey of Crop Acreage in Indiana and Iowa.* U.S. Dept. of Agriculture.

RAO, T. J. (1987). On certain alternative estimators for multiple characteristics in varying probability sampling. Tech. Report. No. 21/87, Stat-Math Division, I.S.I., Calcutta.

——— (1984). Allocation of sample size to strata and related problems. *Biom. Journal*, **26**, 517–526.

——— (1991a). Some aspects of recent trends in survey sampling. In *Recent Advances in Agricultural Statistical Research.* Prem Narain, O.P. Kathuria, V.K. Sharma and Prajneshu (Eds.). Wiley Eastern, pp. 242–255.

——— (1991b). On certain methods of improving ratio and regression estimators. *Comm. Stat. - Theory and Meth.*, **20** , 3325–3340.

RAO, J. N. K. (1966). Alternative estimators in p.p.s. sampling for multiple characteristics. *Sankhyā, Ser. A*, **28**, 47–60.

RAO, P. S. R. S. and MUDHOLKAR, G. (1967). Generalized multivariate estimator for the mean of finite population. *J. Amer. Stat. Assoc.*, **62**, 1009–1012.

REDDY, V. N. and RAO, T. J. (1977). Modified PPS method of estimation. *Sankhyā, Ser. C*, **39**, 185–197.

SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag.

SINGH, M.P. (1965). On the estimation of ratio and product of population parameters. *Sankhyā, Ser. B*, **27**, 321–328.

——— (1967a). Multivariate product method of estimation for finite populations. *J. Ind. Soc. Agr. Stat.*, **19**, 1–10.

——— (1967b). Ratio cum product method of estimation. *Metrika*, **12**, 34–42.

SKINNER, C. J., HOLT, D. and SMITH, T. M. F. (1989). *Analysis of Complex Surveys.* Wiley, New York.

SMITH, T. M. F. and HOLMES, D. J. (1989). Multivariate analysis. In *Analysis of Complex Surveys.* C. J. Skinner, D. Holt and T. M. F. Smith (Eds.). Wiley, pp. 165–190.

SRIVASTAVA, S. K. (1965). An estimate of mean of a finite poopulation using several auxiliary characteristics. *J. Ind. Stat. Assoc.*, **3**, 189–194

——— (1967). An estimator using auxiliary information in sample surveys. *Cal. Stat. Assoc. Bull.* , **16**, 121–132.

TRIPATHI, T. P. (1978). A note on optimum weights in multivariate ratio, product and regression estimators. *J. Ind. Soc. Agr. Stat.*, **30**, 101–109.

——— (1987). A class of estimators for population mean using multivariate auxiliary information under general sampling designs. *Aligarh J. Stat.*, **7**, 49–62.

TRIPATHI, T. P., DAS, A. K. and KHARE, B.B. (1990). *A Review of the Research Work on the Use of Auxiliary Information in Sample Surveys.* Sample Survey Theory and Methods Research Group, India.

TRIPATHI, T. P. and CHAUBEY, Y. P. (1990). Use of multivariate information for obtaining 'near-optimum' probabilities of selection in varying probability sampling. *Technical Report.*

YATES, F. (1960). *Sampling Methods for Censuses and Surveys.* Charles Griffin and Co., London.

STAT. MATH. UNIT
INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD,
CALCUTTA 700 035
INDIA