Nihonkai Math. J.
Vol.16(2005), 135-143

ALLOCATION TO RESPONSE AND NON-RESPONSE GROUPS IN TWO CHARACTER STRATIFIED SAMPLING

S. MAQBOOL AND S. PIRZADA

ABSTRACT. In this paper,we consider the problem of sample allocation in stratified sampling for two characters in presence of partial non-response. The population in each stratum is divided into three groups:one of complete non-respondents,the second with response to questions of category I and third with response to questions of both the categories. It is assumed that the respondents of the questions of category II always reply the questions of category I but not necessarily the vice versa. Using the Hansen and Hurwitz(1946)technique,we determine the sample sizes and the subsampling proportion of various strata.

1.Introduction. During the past several years, the number of surveys, as a means of collecting a variety of data, has greatly increased in most countries. Any survey, whatever its type and whatever the method of collecting data, will suffer from some non-response. Most practicing Statisticians or data analysts recognize non-response as an important measure of quality of data since it affects the estimates by introducing both a possible bias and an increase in sampling variance. In case of stratified population, the problem of determining the initial sample size to be drawn and the value of subsampling proportion for each stratum to be drawn on the second occasion was considered by Khare (1987) in case of fixed cost as well as in case of specified precision. Further improvement in the estimation of population mean in presence of non-response has been made by using information on auxiliary character. In this direction some conventional and alternate ratio, product and regression type estimators have been proposed by Rao (1986, 1987, 1990) and Khare and Srivastava (1993, 2000), when the population mean of the auxiliary character is known or unknown.

2.Sample size selection for single strata. Let $Y_{i1}, Y_{i2}, \ldots, Y_{iN_i}$ be the N_i units of the ith stratum $(i = 1, 2, \cdots, L)$ be independently identically distributed with mean $\overline{Y_i}$ and the variance S_i^2 . The population of each stratum is divided into two classes, those who will response at the first attempt and those who will not response, hence creates the problem of incomplete sample in the mail survey. We propose the following scheme for single character.

1) Select a random sample from each stratum.
2) Send a mail questionnaire to all the selected units in each stratum.
3) After the deadline is over, identify the non-respondents in each stratum.

Key words and phrases. Partial non-response, sampling scheme, estimation procedure, cost function, sample size, subsampling proportion.

4) Collect data from the selected non-respondents in the subsample by interview and combine data from the two parts of the survey in each stratum to provide the unbiased estimate of population mean. For detailed discussion, the readers are requested to go through the paper of Khare, B.B. (1987).

3.Sampling scheme for more than one stratum. Let $Y_{ij1}, Y_{ij2}, \ldots, Y_{ijN_i}$ be the measurements on N_i units who respond to jth character in ith stratum, $(i = 1, \ldots, L;\ j = 1, \ldots, p)$. It is assumed that the questions of category I provide the information on character one and the questions of category II measure the second character.

The sampling scheme is as follows.

i) Select a random sample from each stratum in phase one.

ii) Send a mail questionnaire to all of the selected units in each stratum.

iii) Identify the partial respondents in each stratum (those who reply the questions of category I only) and the complete respondents in each stratum (those who reply the questions of category I and II both).

iv) Collect data from the selected non-respondents and the partial respondents from each stratum in the subsample by personnel interview. We collect the information from non-respondents and partial respondents in each stratum through extra efforts in the second attempt and we assume that in the second attempt each unit of the subsample yields information on both the categories (i.e. questions of category I and II). This is possible due to higher expenditure on a unit in the second attempt.

Let us designate the stages (attempts 1 and 2) by subscripts and the characters by superscripts. The superscripts along with bar will stand for the character under study corresponding to non-respondents. The random sample of size n_i $(i = 1, \ldots, L)$ from i^{th} stratum is partitioned as

$$n_i = n_{i1}^{(1,2)} + \bar{n}_{i1}^{(1)} + \bar{n}_{i1}^{(1,2)}, \text{where}$$

$n_{i1}^{(1,2)}$ = the number of (complete) respondents to questions of category I and II both in i^{th} stratum at first phase.

$\bar{n}_{i1}^{(1)}$ = the number of respondents to only the questions of category I only in i^{th} stratum at first phase (that is non-respondents to questions of category II).

$\bar{n}_{i1}^{(1,2)}$ = the number of complete non-respondents to the questions of both the categories in i^{th} stratum at first phase.

$n_{i1}^{(1,2)}$ = the subsample size at second attempt in the i^{th} stratum out of the complete non-respondents $\bar{n}_{i1}^{(1,2)}$, all of whom respond to questions of both the categories. Let

$$k_i = \frac{\bar{n}_{i1}^{(1,2)}}{n_{i2}^{(1,2)}}. \qquad (3.1)$$

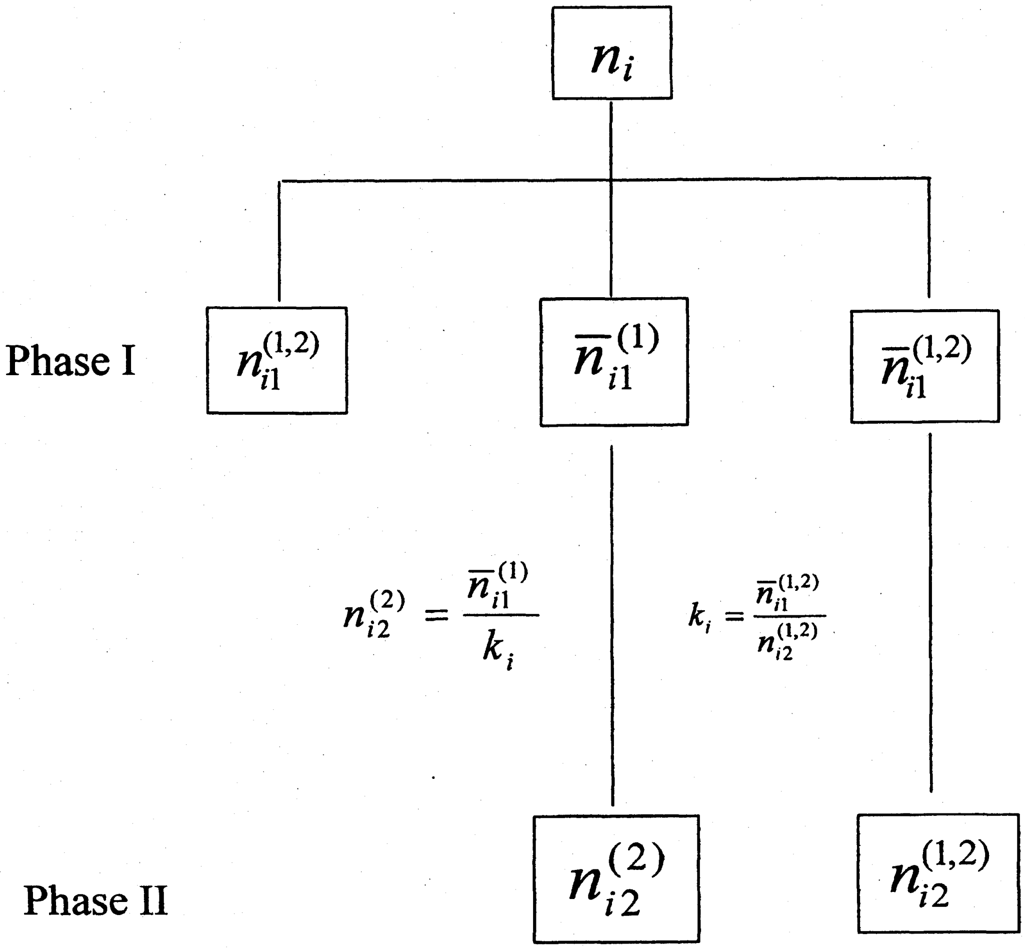
The value of k_i depends on the amount of additional expenses needed to convince the non-respondents for providing the required information in i^{th} stratum.

Then using the same k_i, we also select a subsample out of $\bar{n}_{i1}^{(1,2)}$ size

$$n_{i2}^{(1,2)} = \frac{\bar{n}_{i1}^{(1,2)}}{k_i},$$

all of whom are assumed to respond to the questions of category II at the second attempt.

SAMPLE PARTITION



In the i^{th} stratum the number of units who respond to questions of category I at first phase are $n_{i1}^{(1,2)} + \bar{n}_{i1}^{(1)} = n_{i1}^{*}$, say and the number of units who respond to questions of category I at second attempt are $n_{i2}^{(1,2)}$. Number of units who are non-respondents to questions of category I at first attempt are $\bar{n}_{i1}^{(1,2)}$.

The number of respondents to questions of category II at first phase are $n_{i1}^{(1,2)}$ and those at second attempt are $n_{i2}^{(2)} + n_{i2}^{(1,2)}$, while the number of non-respondents to question of category II only at first attempt are $n_{i1}^{(1,2)} + \bar{n}_{i1}^{(1)} = n_{i2}^{*}$, say.

4.Estimation procedure. Let us define the population means of Character I and II respectively by $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$. The estimators of $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$ are defined by

$$\bar{Y}^{(1)} = \sum_{i=1}^{L} \frac{P_i}{n_i}\left[n_{i1}^{*}\bar{y}_{i1} + \bar{n}_{i1}^{(1,2)}\bar{y}_{i1}^{(1,2)*}\right], \qquad (4.1)$$

$$\bar{Y}^{(2)} = \sum_{i=1}^{L} \frac{P_i}{n_i}\left[n_{i1}^{(1,2)}\bar{y}_{i2} + \bar{n}_{i2}^{*}\bar{y}_{i2}^{(2)*}\right], \qquad (4.2)$$

where p_i =population in the i^{th} stratum and

$\bar{y}_{i1}$ = mean of respondents to questions of category I for character I based on $n_{i1}^{(1,2)} + \bar{n}_{i1}^{(1)}$ units at first attempt.

$\bar{y}_{i1}^{(1,2)*}$ = subsample mean of respondents to question of category I at second attempt based on $n_{i2}^{(1,2)}$ units taken out of $\bar{n}_{i1}^{(1,2)}$ non-respondents.

$\bar{y}_{i2}$ =mean of respondents to questions of category II based on $n_{i1}^{(1,2)}$ units at first attempt.

$\bar{y}_{i2}^{(1,2)*}$ = subsample mean of respondents to questions of category II at second attempt based on $n_{i2}^{(1,2)}$ units.

Then,

$$\begin{aligned} E(\bar{y}^{(1)}) &= E_1E_2\left[\bar{y}^{(1)}/n_{i1}^{*}, \bar{n}_{i1}^{(1,2)}\right] \\ &= E_1E_2\frac{1}{n_i}\left[n_{i1}^{*}\bar{y}_{i1} + \bar{n}_{i1}^{(1,2)}\bar{y}_{i1}^{(1,2)*}/n_{i1}^{*}, \bar{n}_{i1}^{(1,2)}\right] \\ &= E_1E_2\left[\frac{n_{i1}^{*}}{n_i}\bar{y}_{i1}/n_{i1}^{*}\right] + E_1E_2\left[\frac{n_{i1}^{(1,2)}}{n_i}\bar{y}_{i1}^{(1,2)*}/\bar{n}_{i1}^{(1,2)}\right] \\ &= E_1\left[\frac{n_{i1}^{*}}{n_i}E_2(\bar{y}_{i1}/n_{i1}^{*})\right] + E_1\left[\frac{\bar{n}_{i1}^{(1,2)}}{n_i}E_2(\bar{y}_{i1}^{(1,2)*}/\bar{n}_{i1}^{(1,2)})\right]. \end{aligned}$$

Now,

$$E_2\left(\bar{y}_{i1}^{(1,2)*}/\bar{n}_{i1}^{(1,2)}\right) = \bar{y}_{i1}^{(1,2)},$$

where $\bar{y}_{i1}^{(1,2)}$ = mean of non-respondents to questions of category I based on $\bar{n}_{i1}^{(1,2)}$ units at first attempt. Thus,

$$E(\bar{y}^{(1)}) = E_1\left(\frac{n_{i1}^{*}}{n_i}\bar{y}_{i1}\right) + E_1\left(\frac{\bar{n}_{i1}^{(1,2)}}{n_i}\bar{y}_{i1}^{(1,2)}\right).$$

Similarly,

$$\begin{aligned} E(\bar{y}^{(2)}) &= E_1E_2\left[\bar{y}^{(2)}/n_{i1}^{(1,2)}, n_{i2}^{*}\right] \\ &= E_1E_2\frac{1}{n_i}\left[n_{i1}^{(1,2)}\bar{y}_{i2} + n_{i2}^{*}\bar{y}_{i2}^{(2)*}/n_{i1}^{(1,2)}, n_{i2}^{*}\right] \\ &= E_1\left[\frac{n_{i1}^{(1,2)}}{n_i}E_2(\bar{y}_{i2}/n_{i1}^{(1,2)})\right] + E_1\left[\frac{n_{i2}^{*}}{n_i}E_2(\bar{y}_{i2}^{(2)*}/n_{i2}^{*})\right]. \end{aligned}$$

Now,

$$E_2(\bar{y}_{i2}^{(2)*}/n_{i2}^{*}) = \bar{y}_{i2}^{(2)},$$

where $\bar{y}_{i2}^{(2)}$ = mean of non -respondents to questions of category II at second attempt. Thus,

$$E(\bar{y}^{(2)}) = E_1\left(\frac{n_{i1}^{(1,2)}}{n_i}\bar{y}_{i2}\right) + E_1\left(\frac{n_{i2}^{*}}{n_i}\bar{y}_{i2}^{(2)}\right)$$

$$E\left(\bar{y}^{(1)}\right) = \bar{Y}^{(1)}$$

$$E\left(\bar{y}^{(2)}\right) = \bar{Y}^{(2)}.$$

We therefore find that the estimators defined in (4.1) and (4.2) are unbiased.

Theorem 4.1. *The variances of two estimators $\bar{y}^{(1)}$ and $\bar{y}^{(2)}$ corresponding to the character I and II are given by*

$$V(\bar{y}^{(1)}) = \sum_{i=1}^{L}\left[\left(\frac{N_i - n_i}{N_i n_i}\right) + \left(\frac{k_i - 1}{n_i}\right)W_{i3}\right]p_i^2 S_{i1}^2, \qquad (4.3)$$

$$V(\bar{y}^{(2)}) = \sum_{i=1}^{L}\left[\left(\frac{N_i - n_i}{N_i n_i}\right) + \left(\frac{k_i - 1}{n_i}\right)W_{i4}\right]p_i^2 S_{i2}^2, \qquad (4.4)$$

where S_{i1}^2 and S_{i2}^2 are the variances of the non-response classes for the characters I and II respectively. Here we assume that in each stratum respondents and non-respondent population has mean square equal to the stratum mean square.

Proof.

$$\begin{aligned} V(\bar{y}^{(1)}) &= V_1^{(1)}E_2(\bar{y}^{(1)}) + E_1V_1^{(2)}(\bar{y}^{(1)}) \\ &= (1-f)\frac{S_{i1}^2}{n_i} + E_1\left[V_2^{(1)}(\bar{y}^{(1)}/n_{i1}^{*}, \bar{n}_{i1}^{(1,2)})\right] \\ &= (1-f)\frac{S_{i1}^2}{n_i} + E_1\left[V_2^{(1)}\left(\frac{\bar{n}_{i1}^{(1,2)}}{n_i}\bar{y}_{i1}^{(1,2)*}\right)\right] \\ &= (1-f)\frac{S_{i1}^2}{n_i} + E_1\left[\frac{(\bar{n}_{i1}^{(1,2)})^2}{n_i^2}\left(\frac{1}{\bar{n}_{i2}^{(1,2)}} - \frac{1}{\bar{n}_{i1}^{(1,2)}}\right)s_{i1}^2\right], \end{aligned}$$

where s_{i1}^2 is the variance based on $\bar{n}_{i1}^{(1,2)}$ units in i^{th} stratum.

$$V(\bar{y}^{(1)}) = (1-f)\frac{S_{i1}^2}{n_i} + E_1\left[\frac{\bar{n}_{i1}^{(1,2)}}{n_i^2}(k_i - 1)s_{i1}^2\right].$$

Thus,

$$V(\bar{y}^{(1)}) = \sum_{i=1}^{L}\left[\left(\frac{N_i - n_i}{N_i n_i}\right) + \left(\frac{K_i - 1}{n_i}\right) w_{i3}\right] p_i^2 S_{i1}^2.$$

Also,

$$\begin{aligned} V(\bar{y}^{(2)}) &= V_2^{(1)}(\bar{y}^{(2)}) + E_1 V_2^{(2)}(\bar{y}^{(2)}) \\ &= (1-f)\frac{S_{i2}^2}{n_i} + E_1\left[V_2^{(2)}(\bar{y}^{(2)}/n_{i1}, n_{i2}^*)\right] \\ &= (1-f)\frac{S_{i2}^2}{n_i} + E_1\left[V_2^{(2)}(\frac{n_{i2}^*}{n_i} y_{i2}^{(2)*})\right] \\ &= (1-f)\frac{S_{i2}^2}{n_i} + E_1\left[\frac{(n_{i2}^*)^2}{n_i^2}\left(\frac{1}{\bar{n}_{i2}^{(1,2)}} - \frac{1}{\bar{n}_{i1}^*}\right) s_{i2}^2\right], \end{aligned}$$

where s_{i2}^2 is the variance based on n_{i2}^* units in i^{th} stratum.

$$V(\bar{y}^{(2)}) = (1-f)\frac{S_{i2}^2}{n_i} + E_1\left[\frac{n_{i2}^*}{n_i^2}(k_i - 1)s_{i2}^2\right].$$

Thus,

$$V(\bar{y}^{(2)}) = \sum_{i=1}^{L}\left[\left(\frac{N_i - n_i}{N_i n_i}\right) + \left(\frac{k_1 - 1}{n_i}\right) w_{i4}\right] p_i^2 S_{i2}^2.$$

5. Definition of the Cost Function. We define the cost function as

$$\begin{aligned} C = C_0 + \sum_{i=1}^{L} C_i n_i + \sum_{i=1}^{L} C_{i1}^{(1)} + \sum_{i=1}^{L} C_{i1}^{(1)}(\bar{n}_{i1}^{(1)} + n_{i1}^{(1,2)}) \\ + \sum_{i=1}^{L} C_{i1}^{(2)}(n_{i1}^{(1,2)}) + \sum_{i=1}^{L} C_{i2}(n_{i2}^{(2)} + n_{i2}^{(1,2)}), \end{aligned}$$

where

C_0 = Overhead cost for the i^{th} stratum.

C_i = cost of including a unit in the sample in i^{th} stratum.

$C_{i1}^{(1)}$ = Cost incurred per unit in enumerating questions of catgory I in i^{th} stratum at first attempt.

$C_{i1}^{(2)}$ = Cost incurred per unit in enumerating questions of category II in i^{th} stratum at first attempt.

C_{i2} = Cost incurred per unit in i^{th} stratum in enumerating both the characters in second attempt.

Since the values of $\bar{n}_{i1}^{(1)}$ and $\bar{n}_{i1}^{(1,2)}$ are not known until the first attempt is made, the expected cost is used in planning the sample. The expected values of n_{i1}^{*}, $n_{i1}^{(1,2)}$, $n_{i2}^{(1,2)}$ and $n_{i2}^{(2)}$ are respectively $n_i w_{i1}$, $n_i w_{i2}$, $\frac{n_i w_{i3}}{k_i}$ and $\frac{n_i w_{i4}}{k_i}$. Thus, the expected cost is given by

$$C = C_0 + \sum_{i=1}^{L} C_i n_i + \sum_{i=1}^{L} C_{i1}^{(1)} n_i w_{i1} + \sum_{i=1}^{L} C_{i1}^{(2)} n_i w_{i2} + \sum_{i=1}^{L} n_i C_{i2}\left(\frac{w_{i3}}{k_i} + \frac{w_{i4}}{k_i}\right). \qquad (5.1)$$

6. Determination of n_i and k_i.

For determining the optimum value of n_i and k_i for the cost function given by (5.1), we consider the function

$$\phi = \left[V(\bar{y}^{(1)}) + V(\bar{y}^{(2)})\right] + \lambda\,[C], \qquad (6.1)$$

where λ is a Lagrange's multiplier. Differentiating ϕ with respect to k_i and equating to zero, we get

$$n_i = k_i p_i \sqrt{\frac{w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2}{\lambda C_{i2}(w_{i3} + w_{i4})}}. \qquad (6.2)$$

Again differentiating ϕ with respect to n_i and equating to zero, we get

$$\frac{\partial\phi}{\partial n_i} = -\frac{p_i^2}{n_i^2}\left[\{1 + (k_i - 1)w_{i3}\}S_{i1}^2 + \{1 + (k_i - 1)w_{i4}\}S_{i2}^2\right]$$
$$+ \lambda\left[C_i + C_{i1}^{(1)} w_{i1} + C_{i1}^{(2)} w_{i2} + \frac{C_{i2}}{k_i}(w_{i3} + w_{i4})\right] = 0. \qquad (6.3)$$

Eliminating λ from (6.2) and (6.3), we have

$$k_i = \sqrt{\frac{\{(S_{i1}^2 + S_{i2}^2) - (w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2)\}C_{i2}(w_{i3} + w_{i4})}{(w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2)(C_i + C_{i1}^{(1)} w_{i1} + C_{i1}^{(2)} w_{i2})}}. \qquad (6.4)$$

Now, substituting the value of n_i from (6.2) in (5.1), we get

$$\frac{1}{\sqrt{\lambda}} = \frac{(C - C_0)\sqrt{C_{i2}(w_{i3} + w_{i4})}}{\sum_{i=1}^{L} k_i p_i \sqrt{w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2}\left[C_i + C_{i1}^{(1)} w_{i1} + C_{i1}^{(2)} w_{i2}^{(2)} + \frac{C_{i2}}{k_i}(w_{i3} + w_{i4})\right]}. \qquad (6.5)$$

Again eliminating $\frac{1}{\sqrt{\lambda}}$ from (6.2) and (6.5), we get

$$n_i = \frac{k_i p_i (C - C_0)\sqrt{w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2}}{\sum_{i=1}^{L} k_i p_i \sqrt{w_{i3} S_{i1}^2 + w_{i4} S_{i2}^2}\left[C_i + C_{i1}^{(1)} w_{i1} + C_{i1} w_{i2} + \frac{C_{i2}}{k_i}(w_{i3} + w_{i4})\right]}. \qquad (6.6)$$

7.Numerical Illustration. Suppose a population is divided into four strata. having following values.

	Stratum			
	I	II	III	IV
P_i	0.2	0.3	0.4	0.1
S_{i1}	3.5	5.5	6.5	5.5
S_{i2}	3.2	4.8	6.2	5.3
C_1	0.5	0.7	0.4	0.6
$C_{i1}^{(1)}$	8.5	7.4	7.0	9.0
$C_{i1}^{(2)}$	8.7	7.6	7.2	9.2
C_{i2}	25.0	20.0	18.0	25.0

Stratum	Known Response rate		k_i	n_i (approx.)
	w_{i1}	w_{i2}		
1.	0.4	0.3	1.65	12
2.	0.4	0.3	1.55	27
3.	0.4	0.3	1.54	43
4.	0.4	0.3	1.59	10

References

[1] Hansen, M.H. and Hurwitz, W.N. (1946), *The problem of non-response in sampling surveys,* J.Amer.Stat.Assoc., **41**, 517–529

[2] Khare, B.B. (1987), *Allocation in stratified sampling in presence of non-response*, Metron, **45**(I/II), 213–221

[3] Khare, B.B. and Srivastava, S. (1993), *Estimation of population mean using auxiliary character in presence of non-response*, Not Acad. Sc. Letters, **16**(3), 111–114.

[4] Khare, B.B. and Srivastava, S. (2000), *Generalized estimators for population mean in presence of non-response*, Inter. J. Math. Stat. Sci., **9**(1), 75–87.

[5] Rao, P.S.R.S. (1986), *Estimation with sub-sampling the non-respondents*, Survey Methadology, **12**(2), 217–230.

[6] Rao, P.S.R.S. (1987), *Ratio and regression estimates with sub-sampling the non-respondents,* Paper presented at a special contributed session of the International Statistical Association meeting, Sept. 2–16,Tokyo.

[7] Rao, P.S.R.S. (1990), *Regression estimators with sub-sampling of non-respondents,* Data Quality Control, Theory and pragmentics (Eds. E.Gunar and V.R.Uppulari) Marcel Dekker, New York, 191–208.

[8] Rao, P.S.R.S (2000), *Sampling Methodologies with Applications*, Chapman and Hall, CRS Press.

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH, AMU, ALIGARH, INDIA

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF KASHIMIR, SRINAGAR, INDIA
E-mail address: sdpirzada@yahoo.co.in

Received 5 July, 2005 Revised 7 November, 2005