A NEW SAMPLING SCHEME FOR PARTIAL NON- RESPONSE SITUATIONS WITH MULTIPLE OBJECTIVES

S. Maqbool

Division of Agricultural Statistics,

Sher-e -Kashmir University of Agricultural Sciences & Technology of Kashmir

Shalimar, Srinagar-191121.India.

Email: showkatmaq@gmail.com

ABSTRACT

In multivariate surveys it is experienced that the data may not be obtained in the first attempt on all the characters, sometimes due to absentism and sometimes due to refusal of the subject for replying some (or all) of the questions. In this paper, we consider the problem of determining the initial sample size and sub sampling proportion in two variate surveys when some subjects refuse to reply a part of the questions i.e. their responses are partial. The population is divided into three groups: one of complete non- respondents, the second with response to only 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 , i.e., there may be some subjects who agree to reply the questions of category I but do not provide the information on the questions of category II unless some additional efforts are made. Accordingly new sampling scheme with the cost function is proposed.

Key Words: Sampling scheme, partial response, initial sample size, sub-sampling proportion, cost function.

1. Introduction:

Nonresponse is a severe problem in social statistics and survey research. It makes survey estimates questionable because of the potential and unmeasurable bias. It is a largely unknown factor and it is not known to what extent post survey adjustments can deal with this bias. As the nonresponse error is a function of the nonresponse rate and the difference between average scores among the nonrespondents and the respondents, increasing response rates do not always reduce the nonresponse error. Improvement of response rates might not be enough. Reducing the nonresponse bias must be the ultimate goal. Nonresponse is a phenomenon as old as survey-research itself. During the last twenty years quite a number of authors have devoted some time and energy to this problem. It seems however that there is a growing concern with respect to this problem during the last five years. The problem of non response in case of mail surveys was first considered by Hansen & Hurwitz (1946) and Srinath (1971) suggested the selection of sub samples by making several attempts in the non respondent group. Some useful contributions in the field of non response were made by Khare (1987), Lundstrom & Sarndal (2001), Maqbool (2003) and Shukla & Dubey (2007, 2010).

2. Sampling Scheme: Let y_{ji} be the measurement of j^{th} character on i^{th} individual of the population,($j = 1,.2, \quad i = 1,\ldots,N.$). The questionnaire is assumed to posses the questions of two categories. The questions of category I are designed to measure the character I and those of category II to

measure the character II. Thus the j^{th} category represents the j^{th} character. As an example we may consider an opinion survey where the subjects immediately respond to the questions related to their likings and dislikings (category I) but they avoid to respond to such questions as the total income or the various sources of income etc. (category II).

i. Select a random sample of size n in phase one.
ii. Send a mail questionnaire to all of the selected units.
iii. Identify the partial respondents (those who reply the questions of category I only) and denote their number by $n_1^{(1)}$. Also identify the complete respondents (those who reply the questions of category I and II both) and denote their number by $n_1^{(1,2)}$.
iv. Collect data from the selected non-respondents and the partial respondents in the sub sample by personnel interview (or through some additional efforts). We are assuming here that the data on the questions of category I is obtained with less efforts (resulting in less expenses) as compared to that on the questions of category II. Further we assume that a respondent to questions of category II always responds to questions of category I. The whole population is thus divided into three groups; one with non response, the second with partial response (those who respond to questions of category I only) and the third with complete response (those who respond to questions of category I and II both). We collect the information from non-respondents and partial respondents through extra efforts in second attempt and we assume that in the second attempt each unit of the sub sample 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. We designate the stages (attempts 1 and 2) by subscripts and the characters (categories I and II) by superscripts. The superscripts with bar will stand for the character under study corresponding to non-respondents. In our sampling scheme a random sample of size n is being selected using ordinary field method from the population, which is partitioned as $n = n_1^{(1)} + n_1^{(1,2)} + \bar{n}_1^{(1,2)}$,where $n_1^{(1)}$ is the number of respondents to questions of category I only, $n_1^{(1,2)}$ is the number of respondents to questions of categories I and II both and $\bar{n}_1^{(1,2)}$ is the number of complete non-respondents. By personal interviews or using other extensive method in phase 2, we collect the information from the complete non-respondents and partial respondents to questions of category I and II. A sampling scheme in which the information at the second attempt is collected only from the total non-respondents group is treated in Tripathi and Khare (1997). A sub-sample of size $n_2^{(1,2)}$ is attempted out of $\bar{n}_1^{(1,2)}$ non-respondent units at second attempt all of whom are supposed to respond to questions of both the categories due to additional efforts. Let K be the ratio to be sub sampled in the non-response class. The value of K depends upon the amount of additional expenses needed to convince the non-respondents for providing the required information. Then $K = \frac{\bar{n}_1^{(1,2)}}{n_2^{(1,2)}}$. Next a sub-sample of size $n_2^{(2)} = \frac{n_1^{(1)}}{K}$ is also attempted out of $n_1^{(1)}$ (partial respondents at first attempt, i.e. non-respondents to questions of category II at first attempt), all of whom now respond to questions of category II at second attempt. Note that we have assumed that the same proportion (K) of units is selected in the second attempt out of both- the partial non-respondents and the total non-respondents.

The number of units who responds to questions of category I at first attempt is $n_1^* = n_1^{(1)} + n_1^{(1,2)}$. Also the number of units who respond to questions of category I at second attempt is $n_2^* = n_2^{(1,2)}$. So the number

of respondents to the questions of category I is $\left(n_1^{(1)} + n_1^{(1,2)}\right) + n_2^{(1,2)} = n_1^* + n_2^*$. The number of respondents to questions of category II at first attempt is $n_1^{(1,2)}$ and those at second attempt is $n_2^{(2)} + n_2^{(1,2)}$ while the number of non-respondents to questions of category II only at first attempt is $n_1^{(1)}$.

3. Estimation Procedure: Let us denote the population means of characters I and II respectively by $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$. We define the estimators of $\bar{Y}^{(1)}$ and $\bar{Y}^{(2)}$ by

$$\bar{y}^{(1)} = \frac{1}{n}\left[n_1^* \bar{y}_1 + \bar{n}_1^{(1,2)} \bar{y}_1^{(1,2)*}\right], \quad (3.1) \qquad \bar{y}^{(2)} = \frac{1}{n}\left[n_1^{(1,2)} \bar{y}_2 + n_1^{(1)} \bar{y}_2^{(2)*}\right], \quad (3.2)$$

Where $n = n_1^{(1)} + n_1^{(1,2)} + \bar{n}_1^{(1,2)}$ (Total sample size), $n_1^* = n_1^{(1)} + n_1^{(1,2)}$. $\bar{y}_1$ = mean of respondents to questions of category I (character I) based on $n_1^{(1)} + n_1^{(1,2)}$ units at first attempt. $\bar{y}_1^{(1,2)*}$ = sub sample mean of respondents to questions of category I at second attempt based on $n_2^* = n_2^{(1,2)}$ units taken out of $\bar{n}_1^{(1,2)}$ non respondents at first attempt. $\bar{y}_2$ = mean of respondent to questions of category II (character II) based on $n_1^{(1,2)}$ at first attempt. $\bar{y}_2^{(2)*}$ = sub sample mean of respondents to questions of category II at second attempt based on $n_2^{(2)}$ units. Then

$$E(\bar{y}^{(1)}) = E_1 E_2\left[\bar{y}^{(1)} \mid n_1^*, \bar{n}_1^{(1,2)}\right] = E_1 E_2\left[\frac{1}{n}\left\{n_1^* \bar{y}_1 + \bar{n}_1^{(1,2)} \bar{y}_1^{(1,2)*}\right\} \mid n_1^*, \bar{n}_1^{(1,2)}\right]$$

Now $E_2(\bar{y}_1^{(1,2)*} \mid \bar{n}_1^{(1,2)}) = \bar{y}_1^{(1,2)}$, where $\bar{y}_1^{(1,2)}$ = mean of non- respondents to questions of category I based on $\bar{n}_1^{(1,2)}$ units at first attempt. Thus

$$E(\bar{y}^{(1)}) = E_1\left(\frac{n_1^*}{n}\bar{y}_1\right) + E_1\left(\frac{\bar{n}_1^{(1,2)}}{n}\bar{y}_1^{(1,2)}\right) = \frac{N_1^{(1)}}{N}\bar{Y}_1^{(1)} + \frac{\bar{N}_1^{(1,2)}}{N}\bar{Y}_1^{(1,2)} = \bar{Y}^{(1)}.$$

Similarly $E(\bar{y}^{(2)}) = E_1 E_2[\bar{y}^{(2)} \mid n_1^{(1,2)}, n_1^{(1)}] = E_1 E_2\left[\frac{1}{n}\left\{n_1^{(1,2)} \bar{y}_2 + n_1^{(1)} \bar{y}_2^{(2)*}\right\} \mid n_1^{(1,2)}, n_1^{(1)}\right]$

Now $E_2(\bar{y}_2^{(2)*} \mid n_1^{(1)}) = \bar{y}_2^{(2)}$, where $\bar{y}_2^{(2)}$ = mean of non-respondents to question of category II at second attempt. Thus

$$E(\bar{y}^{(2)}) = E_1\left(\frac{n_1^{(1,2)}}{n}\bar{y}_2\right) + E_1\left(\frac{n_1^{(1)}}{n}\bar{y}_2^{(2)}\right) = \frac{N_2^{(2)}}{N}\bar{Y}_2^{(2)} + \frac{\bar{N}_2^{(1,2)}}{N}\bar{Y}_2^{(1,2)} = \bar{Y}^{(2)}$$

Hence we get $E\left(\bar{y}^{(j)}\right) = \bar{Y}^{(j)}$, $j = 1,2$. we therefore find that the estimators defined in (3.1) and (3.2) are unbiased.

Theorem 3.1: The variances of the two estimators $\bar{y}^{(1)}$ and $\bar{y}^{(2)}$ corresponding to the categories I and II are given by

$$V\left(\bar{y}^{(1)}\right)=\left(1-f\right)\frac{S_1^2}{n}+\frac{K\ -1}{n}W_3\bar{S}_1^2 \quad (3.3) \qquad V\left(\bar{y}^{(2)}\right)=\left(1-f\right)\frac{S_2^2}{n}+\frac{K\ -1}{n}W_4\bar{S}_2^2 \ , \quad (3.4)$$

Where S_1^2, S_2^2 are the population variances, $\bar{S}_1^2, \bar{S}_2^2$ are the variances of the non-response classes, W_1, W_2 are the proportions of respondents and W_3, W_4 are the proportion of non-respondents for the characters I and II respectively such that $W_1 + W_3 = 1$ and $W_2 + W_4 = 1$. We assume that the population proportions W_1, W_2, W_3 and W_4 are known from the past data or experience.

Proof : $$V(\bar{y}^{(1)}) = V_1^{(1)}E_2(\bar{y}^{(1)}) + E_1V_1^{(2)}(\bar{y}^{(1)}) = (1-f)\frac{S_1^2}{n}E_1\left[V_2^{(1)}(\bar{y}^{(1)}) \mid n_1^*, \bar{n}_1^{(1,2)}\right]$$

$$= (1-f)\frac{S_1^2}{n} + \frac{K\ -1}{n}E_1\left[\frac{(\bar{n}_1^{(1,2)})}{n}\bar{s}_1^2 \mid \bar{n}_1^{(1,2)}\right], \text{where } \bar{s}_1^2 \text{ is the variance based on } \bar{n}_1^{(1,2)} \text{ units.}$$

Thus $V(\bar{y}^{(1)}) = (1-f)\frac{S_1^2}{n} + \frac{K\ -1}{n}W_3\bar{S}_1^2$ and similarly we can also obtain

$$V(\bar{y}^{(2)}) = (1-f)\frac{S_1^2}{n} + \frac{K\ -1}{n}W_4\bar{S}_2^2$$

4. Derivation of sample size and sub sampling proportion:

We define the cost function as

$$C' = C_0 + C\ n + C_1^{(1)}\left[n_1^{(1)} + n_1^{(1,2)}\right] + C_1^{(2)}\left[n_1^{(1,2)}\right] + C_2\left[n_2^{(2)} + n_2^{(1,2)}\right]$$

where C_0 = overhead cost., C = Cost of including a unit in the sample., $C_1^{(1)}$ = cost incurred per unit in enumerating questions of category I in first attempt., $C_1^{(2)}$ = cost incurred per unit in enumerating questions of category II in first attempt, and C_2 = Cost incurred per unit in enumerating both the characters in second attempt. As C' varies from sample to sample, the expected cost is used in planning the sample. The expected values of $n_1^*, n_1^{(1,2)}, n_2^{(1,2)}$ and $n_2^{(2)}$ are respectively $nW_1, nW_2, \frac{nW_3}{K}$ and $\frac{nW_4}{K}$. The total expected cost is given by

$$E(C') = C = C_0 + Cn + C_1^{(1)}E\left[n_1^*\right] + C_1^{(2)}\left[n_1^{(1,2)}\right] + C_2\left[E(n_2^{(1,2)} + n_2^{(2)})\right]$$

$$C = C_0 + n\left[C + C_1^{(1)}W_1 + C_1^{(2)}W_2 + C_2\left(\frac{W_3}{K}\right) + C_2\left(\frac{W_4}{K}\right)\right] \qquad (4.1)$$

To determine the optimal values of n and the sub sampling proportion k for a fixed budget, we consider the function. $\phi = \left[V(\bar{y}^{(1)}) + V(\bar{y}^{(1)})\right] + \lambda(C)$, (4.2)

Where λ is the lagrangian multiplier. Differentiating λ with respect to k & n and after simplification, we get the value of total sample size and sub sampling proportion using the variance and cost function defined in (3.3), (3.4) and (4.1) as

$$K = \sqrt{\frac{\left[(1-f)\left(S_1^2 + S_2^2\right) - \left(W_3\bar{S}_1^2 + W_4\bar{S}_2^2\right)\right] C_2 \left(W_3 + W_4\right)}{\left(W_3\bar{S}_1^2 + W_4\bar{S}_2^2\right)\left(C + C_1^{(1)}W_1 + C_1^{(2)}W_2\right)}} \qquad (4.3)$$

$$n = \frac{(C - C_0)}{\left[C + C_1^{(1)}\, W_1 + C_1^{(2)}\, W_2 + \frac{C_2}{K}\left(W_3 + W_4\right)\right]} \qquad (4.4)$$

ACKNOWLEDGEMENTS:

The author wishes to record his gratitude and thanks to Vice Chancellor and Director Research SKUAST-KASHMIR for providing the necessary facilities to carry out this research work.

References

Hansen, M .H., and Hurwitz, W. N.(1946)" The problem of non response in sample surveys", *Journal of American statistical association*,41,517-529.

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

S. Maqbool (2003) Optimization Techniques in sample Surveys, *Phd Thesis submitted to Aligarh Muslim University, Aligarh, India*.

Shukla, D and Dubey, Jayant (2007)" On PSPNR sampling scheme in post stratification", *Statistics In Transition*,7(5),1067-1085.

Shukla, D and Dubey, Jayant (2010)" A Generalized class of PSNR sampling scheme", *Journal of Reliability & Statistics*,3(1),75-94.

Sixten Lundstrom and Carl- Erik Sarndal (2001) *Estimation in presence of non response and frame imperfections, SCb*, Statistics Sweden.

Srinath, K.P.(1971)"Multiphase sampling in non response problems", *Journal of American statistical association*,66,167-194.

Tripathi,T.P and Khare, B. B.(1997)" Estimation of mean vector in presence of non response," *communications in statistics, theory methods* ,26,A,2255-2269.