Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 66(3) 2012 447-457

Two Stage Sampling for Estimation of Population Mean with Sub-sampling of Non-respondents

U.C. Sud*, Kaustav Aditya, Hukum Chandra and Rajender Parsad
Indian Agricultural Statistics Research Institute, New Delhi

Received 24 January 2011; Revised 26 August 2012; Accepted September 2012

SUMMARY

The estimation of population mean in the presence of non-response has been considered when the sampling design is two-stage. Considering, three different cases of nonresponse, the corresponding estimators based on sub-sampling of non-respondents, collecting data on the sub-sample through specialized efforts, are developed. Expressions for the variances of the estimators along with unbiased variance estimators are developed. Optimum values of sample sizes are obtained by considering a suitable cost function. The percentage reduction in the expected cost of the proposed estimators are studied empirically.

Keywords: Cost function, Non-response, Population mean, Sub-sampling, Two-stage sampling, Percentage reduction in the expected cost.

1. INTRODUCTION

For large or medium scale surveys we are often faced with the scenario that the sampling frame of ultimate stage units is not available and the cost of construction of the frame is very high. Sometimes the population elements are scattered over a wide area resulting in a widely scattered sample. Therefore, not only the cost of enumeration of units in such a sample may be very high, the supervision of field work may also be very difficult. For such situations, two-stage or multi-stage sampling designs are very effective.

It is also the case that, in many human surveys, information is not obtained from all the units in surveys. The problem of non-response persist even after call-backs. The estimates obtained from incomplete data may be biased particularly when the respondents differ from the non-respondents. Hansen and Hurwitz (1946) proposed a technique for adjusting for non-response to address the problem of bias. The technique consists of selecting a sub-sample of non-respondents. Through specialized efforts data are collected from the non-respondents so as to obtain an estimate of non-responding units in the population**.** Tripathi and Khare (1997) extended the sub-sampling of non-respondents approach to multivariate case. Okafor and Lee (2000) extended the approach to double sampling for ratio and regression estimation. Okafor (2001, 2005) further extended the approach in the context of element sampling and two-stage sampling respectively on two successive occasions. Chhikara and Sud (2009) used the sub-sampling of non-respondents approach for estimation of population and domain totals in the context of item non-response.

It may be mentioned that the weighting and imputation procedures aim at eliminating the bias caused by non-response. However, these procedures are based on certain assumptions on the response mechanism. When these assumptions do not hold good the resulting estimate may be seriously biased. Further, when the non-response is confounded, i.e. the response probability is dependent on the survey character, it

*Corresponding author : U.C. Sud
E-mail address : ucsud@iasri.res.in

becomes difficult to eliminate the bias entirely. Hansen and Hurwitz's sub-sampling approach although costly, is free from any assumptions. When the bias caused by non-response is serious this technique is very effective i.e. one does not have to go for 100 percent response, although this can be very expensive.

In what follows, three different cases of estimation of the population mean have been considered based on three different situations of non-response in two-stage sampling design. As for example, in case of socio-economic surveys, surveying villages can be taken as first-stage units and the households within the villages as the second-stage units. There may be situation, where some households within the selected villages do not respond at the first attempt of data collection through mail/postal enquiry creating a nonresponse situation in the ssus. This situation is termed as Case 1. Under Case 2, we have considered the situation where all the persons belonging to some of the selected villages respond at the first attempt of data collection whereas in the remaining selected villages some households do not responds at the first attempt. In Case 3 we have considered the situation where there is full response in some of the selected villages, partial non-response in some other selected villages and complete non-response in the remaining selected villages. The three non-response situations are frequently encountered in the mail surveys in the context of two-stage sampling designs. In these surveys the first attempt to collect information from the respondents is made through e-mail/postal. Many of the respondents may not send the required information through mails. To collect information from non-respondents, a sub-sample of non-respondents may be selected for data collection by specialised effort, say, personal interview. In these situations the methods available to tackle non-response i.e. call-backs and follow-ups may prove to be very costly and time consuming in practice and also provide biased estimates in most of the cases. To obtain unbiased estimates in these cases of non-response, the sub-sampling of the non-respondents technique developed by Hansen and Hurwitz (1946) is commonly used. Both in theory and in practice the method of sub-sampling of the non-respondents may provide better results than the other available methods to tackle the problem of nonresponse in mail surveys. Thus, corresponding to these cases of non-response, different estimators of population mean using two-stage sampling designs are developed in Section 2 based on the technique of sub-sampling the non-respondents. Also given are expressions for variance of the estimators and unbiased variance estimators. Optimum values of sample sizes are obtained by minimizing the expected cost for a fixed variance. The results are empirically illustrated in Section 3.

2. THEORETICAL DEVELOPMENTS

Let the finite population U under consideration consists of N known primary stage units (psus) labelled 1 through N. Let the i-th psu comprise M second stage units (ssus). Let y_{ij} be the value of study character pertaining to j-th ssu in the i-th psu, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, M$. The objective is to estimate the population mean which is defined as

$$\bar{Y} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$$

Case 1. Let n psus be selected by Simple Random Sampling without Replacement (srswor) design and within each selected psu, m ssus are also selected by srswor from M ssus. Further, let there be M_{i1} responding and M_{i2} non-responding units in the i-th psu, $M_{i1} + M_{i2} = M$, $i = 1, 2, \ldots, N$. Further, out of m ssus, m_{i1} ssus respond while m_{i2} ssus do not respond, $m_{i1} + m_{i2} = m$. A sub-sample of size h_{i2} is selected from m_{i2} by srswor and data are collected on the sub-sampled units through specialized efforts. Here $m_{i2} = h_{i2} f_{i2}$, $i = 1, 2, \ldots, n$.

Theorem 2.1 An unbiased estimator of $\bar{Y}$ is given by

$$\bar{y}' = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}}) \qquad (2.1)$$

where $\bar{y}_{m_{i1}} = \frac{1}{m_{i1}}\sum_{j=1}^{m_{i1}} y_{ij}$ and $\bar{y}_{h_{i2}} = \frac{1}{h_{i2}}\sum_{j=1}^{h_{i2}} y_{ij}$.

The variance of $\bar{y}'$ is

$$V(\bar{y}') = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$$

$$+ \frac{1}{Nmn}\sum_{i=1}^{N}\frac{M_{i2}}{M}(f_{i2} - 1)S_{M_{i2}}^2 \qquad (2.2)$$

where $S_b^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(\bar{Y}_{iM} - \bar{Y})^2$, $\bar{Y}_{iM} = \frac{1}{M}\sum_{j=1}^{M} Y_{ij}.$,

$$S_{iM}^2 = \frac{1}{(M-1)}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_{iM})^2,$$

$$S_{M_{i2}}^2 = \frac{1}{(M_{i2}-1)}\sum_{j=1}^{M_{i2}}(Y_{ij} - \bar{Y}_{M_{i2}})^2,$$

$$\bar{Y}_{M_{i2}} = \frac{1}{M_{i2}}\sum_{j=1}^{M_{i2}} Y_{ij}$$

Next, an unbiased variance estimator is given by

$$\hat{V}(\bar{y}') = (\frac{1}{n} - \frac{1}{N})s_b'^2 + \frac{1}{Nn}\sum_{i=1}^{n}(\frac{1}{m} - \frac{1}{M})s_{im}^2$$

$$+ \frac{1}{Nn}\frac{(M-1)m}{(m-1)M}\sum_{i=1}^{n}\frac{m_{i2}}{m^2}(f_{i2}-1)s_{h_{i2}}^2 \qquad (2.3)$$

where,

$$s_b'^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n}\bar{y}_{im}^2 - n\bar{y}'^2\right),$$

$$\bar{y}_{im} = \frac{1}{m}(m_{i1}\bar{y}_{mi1} + m_{i2}\bar{y}_{hi2}),$$

$$s_{im}^2 = \frac{1}{(m-1)}(\sum_{j=1}^{m_{il}} y_{ij}^2 + \frac{m_{i2}}{h_{i2}}\sum_{j=1}^{h_{i2}} y_{ij}^2 - m\bar{y}_{im}^2)$$

$$s_{h_{i2}}^2 = \frac{1}{(h_{i2}-1)}\sum_{j=1}^{h_{i2}}(y_{ij} - \bar{y}_{h_{i2}})^2, \bar{y}_{h_{i2}} = \frac{1}{h_{i2}}\sum_{j=1}^{h_{i2}} y_{ij}$$

Proof: By definition,

$$E(\bar{y}') = E_1E_2E_3\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})}{m}\right\}$$

$$= E_1E_2\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{m_{i2}})}{m}\right\}$$

$$= E_1E_2\frac{1}{n}\left\{\sum_{i=1}^{n}\bar{y}_{im}\right\} = E_1\frac{1}{n}\left\{\sum_{i=1}^{n}\bar{Y}_{iM}\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_{iM} = \bar{Y}$$

where E_3 represent conditional expectations of all possible samples of size h_{i2} drawn from a sample of size m_{i2}, E_2 is the conditional expectation of all possible samples of size m drawn from M while E_1 refers to expectation arising out of all possible samples of size n drawn from a population of size N.

By definition,

$$V(\bar{y}') = V_1E_1E_2(\bar{y}') + E_1V_2E_3(\bar{y}') + E_1E_2V_3(\bar{y}')$$

(V_1, V_2, V_3 are defined similarly as E_1, E_2, E_3)

where,

$$V_1E_2E_3(\bar{y}') = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2,$$

$$E_1V_2E_3(\bar{y}') = \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2,$$

$$E_1E_2V_3(\bar{y}') = \frac{1}{Nmn}\sum_{i=1}^{N}\frac{M_{i2}}{M}(f_{i2}-1)S_{M_{i2}}^2$$

Hence, $V(\bar{y}') = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$

$$+ \frac{1}{Nmn}\sum_{i=1}^{N}\frac{M_{i2}}{M}(f_{i2}-1)S_{M_{i2}}^2$$

To obtain an unbiased variance estimator, we proceed as follows

Consider, $s_b'^2 = \frac{1}{(n-1)}\left(\sum_{i=1}^{n}\bar{y}_{im}^2 - n\bar{y}'^2\right)$

It can be shown that

$$E_1E_2E_3E_4(s_b'^2) = S_b^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\frac{M_{i2}}{Mm}(f_{i2}-1)S_{M_{i2}}^2$$

and $E(s_{im}^2) = S_{iM}^2 - \frac{M_{i2}}{M(m-1)}(f_{i2}-1)S_{M_{i2}}^2$

E_4 represent conditional expectations of all possible samples of size h_{i2} drawn from a sample of size m_{i2}, E_3 is the conditional expectation of all possible samples of size m_{i1}, m_{i2} respectively drawn from M_{i1}, M_{i2} respectively by keeping m_{i1}, m_{i2} fixed. Here M_{i1} and

M_{i2} denote the number of responding and non-responding units in the i-th psu. E_2 refers to conditional expectation arising out of randomness of m_{i1}, m_{i2} while E_1 refers to expectation arising out of all possible samples of size n drawn from a population of size N.

Let,

$$\hat{S}_b^2 = s_b'^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m} - \frac{1}{M}\right)s_{im}^2 - \frac{1}{n}\sum_{i=1}^{n}\frac{m_{i2}}{m^2}(f_{i2}-1)s_{h_{i2}}^2$$

$$\hat{S}_{iM}^2 = s_{im}^2 + \frac{m_{i2}}{m(m-1)}((f_{i2}-1)s_{h_{i2}}^2$$

Substituting the above estimated values of S_b^2 and S_{iM}^2 in (2.2) we get the required result in (2.3).

We determine the optimum values of n, m and f_{i2} by minimising the expected cost for a fixed variance. To achieve this consider the following cost function

$$C = C_1 n + C_2\sum_{i=1}^{n} m_{i1} + C_3\sum_{i=1}^{n} h_{i2}$$

where,

C: Total cost

C_1: Per psu travel cost

C_2: Cost per ssu for collecting the information on the study character in the first attempt

C_3: Cost per ssu for collecting the information by expensive method after the first attempt has failed for obtaining information

It is envisioned that C_3 will be higher than C_1 and substantially higher than C_2 .

The expected cost in this case is,

$$C' = E(C) = n\left[C_1 + C_2\frac{m}{N}\sum_{i=1}^{n}\frac{M_{i1}}{M} + C_3\frac{m}{N}\sum_{i=1}^{n}\frac{M_{i2}}{Mf_{i2}}\right]$$

Consider the function $\phi = C' + \lambda\{V(\bar{y}') - V_0\}$

Here, λ is the Lagrangian multiplier. Also, V_0 can be determined by fixing the coefficient of variation, say equal to 5%. For the sake of simplicity we assumed f_2 in place of f_{i2}. Differentiation with respect to n, m, λ and, f_2 equating the resultant derivatives equal to '0' and simplifying gives the optimum values as,

$$n_{opt} = \frac{k}{(V_0 + \frac{S_b^2}{N})},\quad m_{opt} = \sqrt{\frac{C_1(A+B)}{(A_1+B_1)\left(NS_b^2 - \frac{B}{M}\right)}} \text{ and}$$

$$f_{2opt} = \sqrt{\frac{C_3\left[B - \sum_{i=1}^{N} DM_{i2}\right]}{C_2\sum_{i=1}^{N} DM_{i2}}}$$

where, $k = S_b^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$

$$+\frac{1}{N}\sum_{i=1}^{N}\frac{M_{i2}}{Mm}(f_2-1)S_{M_{i2}}^2,$$

$$D = \frac{S_{M_{i2}}^2}{M} \text{ and } A_1 = \frac{C_2}{N}\sum_{i=1}^{N}\frac{M_{i1}}{M},$$

$$B_1 = \frac{C_3}{N}\sum_{i=1}^{N}\frac{M_{i2}}{Mf_2},\quad B = \sum_{i=1}^{N} S_{iM}^2,$$

$$A = \sum_{i=1}^{N}\frac{M_{i2}}{M}(f_2-1)S_{M_{i2}}^2 \text{ and } V_0 = 0.0025\times\bar{Y}^2.$$

Case 2. Consider the situation that a sample of n psus is drawn from N and within each selected psu a sample of m ssus is drawn by srswor. Let there be no non-response in n_1 psus. In the remaining n_2 psus m_{i1} ssus respond while m_{i2} units do not respond. A sub-sample of h_{i2} units is selected by srswor from m_{i2} and from each selected psu data are collected through specialised efforts, $m_{i2}=h_{i2}f_{i2}$, $i=1, 2,\ldots, n_2$, $(n_1 + n_2 = n)$.

In this context we state the estimator and its variance as below:

Theorem 2.2. The estimator

$$\bar{y}'' = \frac{1}{n}\left\{\sum_{i=1}^{n_1}\bar{y}_{im} + \sum_{i=1}^{n_2}\frac{(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})}{m}\right\} \quad (2.4)$$

is unbiased for $\bar{Y}$ with variance

$$V(\bar{y}'') = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$$

$$+\frac{1}{Nn}\sum_{i=1}^{N_2}\frac{M_{i2}}{Mm}(f_{i2}-1)S_{M_{i2}}^2 \quad (2.5)$$

Its unbiased variance estimator is

$$\hat{V}(\bar{y}'') = \left(\frac{1}{n} - \frac{1}{N}\right)s_b''^2 + \frac{1}{Nn}\left(\frac{1}{m} - \frac{1}{M}\right)\left[\sum_{i}^{n_1} s_{1im}^2 + \sum_{i=1}^{n_2} s_{im}^2\right]$$

$$+ \frac{1}{Nn}\frac{(M-1)m}{(m-1)M}\sum_{i=1}^{n_2}\frac{m_{i2}}{m^2}(f_{i2}-1)s_{h_{i2}}^2 \qquad (2.6)$$

Here,

$$s_b''^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{n_1}\bar{y}_{im}^2 + \sum_{i=1}^{n_2}\frac{1}{m^2}(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})^2 - n\bar{y}''^2\right],$$

S_b^2, s_{im}^2, S_{iM}^2, $s_{h_{i2}}^2$, $S_{M_{i2}}^2$ etc. are defined earlier and

$$s_{1im}^2 = \frac{1}{(m-1)}\left(\sum_{j=1}^{m} y_{ij}^2 - m\bar{y}_{im}^2\right).$$

Proof: Consider,

$$E(\bar{y}'') = E_1E_2E_3\frac{1}{n}\left\{\sum_{i=1}^{n_1}\bar{y}_{im} + \sum_{i=1}^{n_2}\frac{(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})}{m}\right\}$$

$$= E_1E_2\frac{1}{n}\left\{\sum_{i=1}^{n_1}\bar{y}_{im} + \sum_{i=1}^{n_2}\bar{y}_{im}\right\}$$

$$= E_1\frac{1}{n}\left\{\sum_{i=1}^{n}\bar{Y}_{iM}\right\} = \frac{1}{N}\left\{\sum_{i=1}^{N}\bar{Y}_{iM}\right\} = \bar{Y}$$

E_1, E_2, E_3 have been defined earlier.

To obtain the variance we proceed as follows:

By definition,

$$V(\bar{y}'') = V_1E_2E_3(\bar{y}'') + E_1V_2E_3(\bar{y}'') + E_1E_2V_3(\bar{y}'')$$

where,

$$V_1E_2E_3(\bar{y}'') = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2,$$

$$E_1V_2E_3(\bar{y}'') = \frac{1}{Nn}\sum_{i=2}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$$

and $$E_1E_2V_3(\bar{y}'') = \frac{1}{Nmn}\sum_{i=2}^{N_2}\frac{M_{i2}}{M}(f_{i2}-1))S_{M_{i2}}^2$$

By adding the above three terms we get the required result. To obtain an unbiased variance estimator consider,

$$s_b''^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{n_1}\bar{y}_{im}^2 + \sum_{i=1}^{n_2}\frac{1}{m^2}(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})^2 - n\bar{y}''^2\right]$$

Taking the expectations and simplifying we get,

$$E_1E_2E_3E_4(s_b''^2) = S_b^2 + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2$$

$$+ \frac{1}{N}\sum_{i=1}^{N_2}\frac{M_{i2}}{Mm}(f_{i2}-1)S_{M_{i2}}^2$$

and $$E(s_{im}^2) = S_{iM}^2 - \frac{M_{i2}}{M(m-1)}(f_{i2}-1)S_{M_{i2}}^2.$$

Let,

$$\hat{S}_b^2 = \hat{s}_b''^2 - \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\left[\sum_{i=1}^{n_1}s_{1im}^2 + \sum_{i=1}^{n_2}s_{im}^2\right]$$

$$- \frac{1}{n}\sum_{i=1}^{n_2}\frac{m_{i2}}{m^2}(f_{i2}-1)s_{h_{i2}}^2$$

and $$\hat{S}_{iM}^2 = s_{im}^2 + \frac{m_{i2}}{m(m-1)}(f_{i2}-1)s_{h_{i2}}^2$$

and $$\hat{S}_{M_{i2}}^2 = s_{h_{i2}}^2$$

Substituting the estimated values in the variance expression (2.5) we obtain the required result.

To determine the optimum values of n, m and f_{i2} we proceed as earlier i.e. minimization of expected cost subject to fixed variance.

The relevant cost function in this case is,

$$C = C_1n_2 + C_2n_1m + C_2\sum_{i=1}^{n_2}m_{i1} + C_3\sum_{i=1}^{n_2}h_{i2}$$

where the different forms in the cost function are the same as defined earlier.

The expected cost is,

$$C'' = E(C)$$

$$= \frac{n}{N}\left[C_1N_2 + C_2N_1m + C_2\sum_{i=1}^{N_2}\frac{M_{i1}m}{M} + C_3\sum_{i=1}^{N_2}\frac{M_{i2}m}{Mf_{i2}}\right]$$

Consider the following function

$$\phi = C'' + \lambda\{V(\overline{y}'') - V_0\},$$

where λ is the Lagrangian multiplier and V_0 which is the fixed value of the variance can be determined by fixing the Coefficient of Variation (CV) value to 5%. While optimising we substituted f_2 in place of f_{i2}. Differentiation with respect to n, m, λ and f_2, equating the resultant derivatives equal to '0', we have the optimum values as follows:

$$n_{opt} = \frac{k}{(V + \frac{S_b^2}{N})}, m_{opt} = \frac{-b \pm \sqrt{b^2 - 4ae}}{2a} \text{ and}$$

$$f_{2opt} = \sqrt{\frac{C_3 \sum_{i=1}^{N_2} M_{i2} (\sum_{i=1}^{N} S_{iM}^2 + \sum_{i=1}^{N_2} \frac{M_{i2}}{M} S_{M_{i2}}^2)}{C_2 (N_1 + \sum_{i=1}^{N_2} \frac{M_{i1}}{M}) \sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2}}$$

where,

$$k = S_b^2 + \frac{1}{N} \sum_{i=1}^{N} (\frac{1}{m} - \frac{1}{M}) S_{iM}^2 + \frac{1}{N} \sum_{i=1}^{N_2} \frac{M_{i2}}{Mm} (f_2 - 1) S_{M_{i2}}^2$$

$$a = NC_3 \sum_{i=1}^{N_2} \frac{M_{i2}}{f_2^2} \left[S_b^2 + \frac{1}{N} \sum_{i=1}^{N} (\frac{1}{m} - \frac{1}{M}) S_{iM}^2 \right]$$

$$b = -\left[C_2 N_1 + C_2 \sum_{i=1}^{N_2} \frac{M_{i1}}{M} + C_3 \sum_{i=1}^{N_2} \frac{M_{i2}}{M f_2^2} \right] \sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2$$

$$e = -C_1 N_2 \sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2$$

Case 3. Here we consider the situation that a sample of n psus is drawn from N, within each selected psu a sample of m ssus is drawn by srswor design. Let there be no non-response in n_1 psus. In the n_2 psu m_{i1} ssus respond while m_{i2} ssus do not respond, let there be complete non-response in the n_3 psus, $n_1 + n_2 + n_3 = n$. A sub-sample of h_{i2} units is selected by srswor from m_{i2} and data are collected through specialized efforts, further a sub-sample of h_3 psus is drawn out of n_3 psus and data are collected through specialized efforts on each of m ssus in the selected h_3 psus. Let $n_3 = f_3 h_3$ and $m_{i2} = h_{i2} f_{i2}$, $i = 1, 2, \ldots, n_2$. Assume $N = N_1 + N_2 + N_3$ where N_1, N_2 and N_3 are the number of psus in the population representing the three nonresponse categories considered here.

In this context Theorem 2.3 as below state the estimator.

Theorem 2.3. The estimator

$$\overline{y}''' = \frac{1}{n} \left\{ \sum_{i=1}^{n_1} \overline{y}_{im} + \sum_{i=1}^{n_2} \frac{1}{m} (m_{i1} \overline{y}_{m_{i1}} + m_{i2} \overline{y}_{h_{i2}}) + \frac{n_3}{h_3} \sum_{i=1}^{h_3} \overline{y}_{im} \right\} \tag{2.7}$$

is unbiased for $\overline{Y}$ with the variance

$$V(\overline{y}''') = \left(\frac{1}{n} - \frac{1}{N} \right) S_b^2 + \frac{1}{Nn} \left(\frac{1}{m} - \frac{1}{M} \right) \left\{ \sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 + f_3 \sum_{i=1}^{N_3} S_{iM}^2 \right\} + \frac{1}{Nn} \sum_{i=1}^{N_2} \frac{M_{i2}}{Mm} (f_{i2} - 1) S_{M_{i2}}^2 + \frac{N_3}{Nn} (f_3 - 1) S_{bN_3}^2 \tag{2.8}$$

$$S_{bN_3}^2 = \frac{1}{N_3 - 1} \sum_{i=1}^{N_3} (\overline{Y}_{iM} - \overline{Y}_{N_3})^2, \text{ where } \overline{Y}_{N_3} = \frac{1}{N_3} \sum_{i=1}^{N_3} \overline{Y}_{iM}$$

and rest of the terms are defined earlier.

An unbiased estimate of variance is,

$$\hat{V}(\overline{y}''') = \left(\frac{1}{n} - \frac{1}{N} \right) s_b'''^2 + \frac{1}{Nn} \left(\frac{1}{m} - \frac{1}{M} \right) \left\{ \sum_{i=1}^{n_1} s_{1im}^2 + \sum_{i=1}^{n_2} s_{im}^2 + f_3 \frac{n_3}{h_3} \sum_{i=1}^{h_3} s_{1im}^2 \right\} + \frac{1}{Nn} \frac{(M-1)m}{(m-1)M} \sum_{i=1}^{n_2} \frac{m_{i2}}{m^2} (f_{i2} - 1) s_{h_{i2}}^2 + \frac{(Nn_3 - N_3)(f_3 - 1)}{N(n-1)n} \{ s_{bh_3}^2 - \frac{1}{h_3} \sum_{i=1}^{h_3} (\frac{1}{m} - \frac{1}{M}) s_{1im}^2 \} \tag{2.9}$$

where,

$$s_b'''^2 = \frac{1}{(n-1)} \left[\sum_{i=1}^{n_3} \overline{y}_{im}^2 + \sum_{i=1}^{n_2} \frac{1}{m^2} (m_{i1} \overline{y}_{m_{i1}} + m_{i2} \overline{y}_{h_{i2}})^2 + \frac{n_3}{h_3} \sum_{i=1}^{h_3} \overline{y}_{im}^2 - n \overline{y}'''^2 \right] \text{ and}$$

$$s_{bh_3}^2 = \frac{1}{(h_3 - 1)} \sum_{i=1}^{h_3} (\bar{y}_{im} - \bar{y}_{h_3})^2, \text{ where } \bar{y}_{h_3} = \frac{1}{h_3} \sum_{i=1}^{h_3} \bar{y}_{im}$$

Proof:

$$E(\bar{y}''') = E_1 E_2 E_3 E_4 \frac{1}{n} \{ \sum_{i=1}^{n_1} \frac{1}{m} \sum_{j=1}^{m} y_{ij} + \sum_{i=1}^{n_2} \frac{(m_{i1}\bar{y}_{m_{i1}} + m_{i2}\bar{y}_{h_{i2}})}{m} + \frac{n_3}{h_3} \sum_{i=1}^{h_3} \frac{1}{m} \sum_{j=1}^{m} y_{ij} \}$$

$$= E_1 E_2 E_3 \frac{1}{n} \left\{ \sum_{i=1}^{n_1} (\bar{y}_{im}) + \sum_{i=1}^{n_2} (\bar{y}_{im}) + \frac{n_3}{h_3} \sum_{i=1}^{n_3} (\bar{y}_{im}) \right\}$$

$$= E_1 E_2 \frac{1}{n} \left\{ \sum_{i=1}^{n_1} \frac{1}{M} \sum_{j=1}^{M} Y_{ij} + \sum_{i=1}^{n_2} \frac{1}{M} \sum_{j=1}^{M} Y_{ij} + \frac{n_3}{h_3} \sum_{i=1}^{h_3} \frac{1}{M} \sum_{j=1}^{M} Y_{ij} \right\}$$

$$= E_1 \frac{1}{n} \left[\sum_{i=1}^{n} \bar{Y}_{iM} \right] = \frac{1}{N} \sum_{i=1}^{N} \bar{Y}_{iM} = \bar{Y}$$

E_4 represent conditional expectations of all possible samples of size h_{i2} drawn from a sample of size m_{i2}, E_3 is the conditional expectation of all possible samples of size m, drawn from M, E_2 refers to conditional expectation arising out of selection of all possible samples of size h_3 drawn from n_3 while E_1 refers to expectation arising out of all possible samples of size n drawn from a population of size N.

To obtain the variance we proceed as follows:

By definition

$$V(\bar{y}''') = V_1 E_2 E_3 E_4 (\bar{y}''') + E_1 V_2 E_3 E_4 (\bar{y}''') + E_1 E_2 V_3 E_4 (\bar{y}''') + E_1 E_2 E_3 V_4 (\bar{y}''')$$

where,

$$V_1 E_2 E_3 E_4 (\bar{y}''') = (\frac{1}{n} - \frac{1}{N}) S_b^2,$$

$$E_1 V_2 E_3 E_4 (\bar{y}''') = \frac{N_3}{Nn} (f_3 - 1) S_{bN_3}^2,$$

$$E_1 E_2 V_3 E_4 (\bar{y}''') = \frac{1}{Nn} \left(\frac{1}{m} - \frac{1}{M} \right) \left\{ \sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 + f_3 \sum_{i=1}^{N_3} S_{iM}^2 \right\},$$

$$E_1 E_2 E_3 V_4 (\bar{y}''') = \frac{1}{Nn} \sum_{i=1}^{N_2} \frac{M_{i2}}{Mm} (f_{i2} - 1) S_{M_{i2}}^2$$

Thus by adding all the terms we obtain the required result.

$$V(\bar{y}''') = (\frac{1}{n} - \frac{1}{N}) S_b^2 + \frac{1}{Nn} (\frac{1}{m} - \frac{1}{M}) \left\{ \sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 + f_3 \sum_{i=1}^{N_3} S_{iM}^2 \right\} + \frac{1}{Nn} \sum_{i=1}^{N_2} \frac{M_{i2}}{Mm} (f_{i2} - 1) S_{Mi2}^2 + \frac{N_3}{Nn} (f_3 - 1) S_{bN_3}^2$$

Taking the expectations and simplifying we get,

$$E_1 E_2 E_3 E_4 E_5 E_6 E_7 (s_b'''^2) = S_b^2 + \frac{1}{N} \left(\frac{1}{m} - \frac{1}{M} \right) \left\{ \sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 + f_3 \sum_{i=1}^{N_3} S_{iM}^2 \right\} + \frac{1}{N} \sum_{i=1}^{N} \frac{M_{i2}}{Mm} (f_{i2} - 1) S_{M_{i2}}^2 - \frac{N_3}{N(n-1)} (f_3 - 1) S_{bN_3}^2$$

$$E(s_{im}^2) = S_{iM}^2 - \frac{M_{i2}}{M(m-1)} (f_{i2} - 1) S_{M_{i2}}^2$$

$$E(s_{bh_3}^2) = S_{bN_3}^2 + \frac{1}{N_3} \sum_{i=1}^{N_3} (\frac{1}{m} - \frac{1}{M}) S_{iM}^2 \text{ and}$$

$$E(s_{1im}^2) = S_{iM}^2$$

where E_7 represents conditional expectations of all possible samples of size h_{i2} drawn from a sample of sizeis m_{i2}, E_6 is the conditional expectation of all possible samples of size m_{i1}, m_{i2} respectively drawn from M_{i1}, M_{i2} respectively by keeping m_{i1}, m_{i2} fixed. HereM_{i1}, Mi_2 denote the number of responding and nonresponding units in the population, E_5 refers to conditional expectation arising out of randomness of m_{i1}, m_{i2}, E_4 refers to conditional expectation of all possible samples of size m drawn from M, E_3 refers to conditional expectation arising out of selection of all possible samples of size h_3 drawn from n_3, E_2 refers to expectation arising out of all possible samples of size n_1, n_2, n_3 drawn from N_1, N_2, N_3 keeping n_1, n_2, n_3 fixed while E_1 refers to expectation arising out of randomness of n_1, n_2, n_3.

Let,

$$\hat{S}_b^2 = s_b^{\prime\prime\prime 2} - \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right)\left\{\sum_{i=1}^{n_1} s_{1im}^2 + \sum_{i=1}^{n_2} s_{im}^2 + f_3 \sum_{i=1}^{n_3} s_{1im}^2\right\}$$

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{m_{i2}}{m^2}(f_{i2}-1)s_{h_{i2}}^2$$

$$+\frac{n_3}{n(n-1)}(f_3-1)\left(s_{bh_3}^2 - \frac{1}{h_3}\sum_{i=1}^{h_3}\left(\frac{1}{m}-\frac{1}{M}\right)s_{1im}^2\right)$$

$$\hat{S}_{iM}^2 = s_{im}^2 - \frac{m_{i2}}{m(m-1)}(f_{i2}-1)s_{h_{i2}}^2,$$

$$\hat{S}_{iM}^2 = s_{1im}^2 \text{ and}$$

$$\hat{S}_{bN_3}^2 = s_{bh_3}^2 - \frac{1}{h_3}\sum_{i=1}^{h_3}\left(\frac{1}{m}-\frac{1}{M}\right)s_{1im}^2$$

Substituting the estimated value in the variance expression (2.8) we get the required estimate.

To determine the optimum values of n, m, f_{i2} we make the following assumption *i.e.* $n_3 = f_3 h_3$ and $m_{i2} = h_{i2} f_{i2}$, $i = 1, 2, \ldots, n_2$.

The cost function in this case,

$$C = C_1(n_2 + h_3) + C_2 n_1 m + C_2 \sum_{i=1}^{n_2} m_{i1} + C_3 \left(\sum_{i=1}^{n_2} h_{i2} + h_3 m\right)$$

where C, C_1, C_2, C_3 are same as defined earlier.

The expected cost is,

$$C^{\prime\prime\prime} = E(C) = \frac{n}{N}\left[C_1 N_2 + C_1\frac{N_3}{f_3} + C_2 N_1 m + C_2\sum_{i=1}^{N_2}\frac{M_{i1}m}{M} + C_3\sum_{i=1}^{N_2}\frac{M_{i2}m}{Mf_{i2}} + C_3\frac{N_3 m}{f_3}\right]$$

To minimize the expected cost subject to fixed variance consider the function.

$$\phi = C^{\prime\prime\prime} + \lambda\{V(\bar{y}^{\prime\prime\prime}) - V_0\}.$$

During optimisation we have substituted f_2 in place of f_{i2}. To overcome the problem arising due to simultaneous minimization of n, m, f_2, f_3 we assume that $n_3 = f_2 h_3$. Thus minimization gives the optimum values as

$$n_{opt} = \frac{k}{(V_0 + \frac{S_b^2}{N})}$$

$$m_{opt} = \frac{-b_2 \pm \sqrt{b_2^2 - 4a_2 e_2}}{2a_2} \text{ and}$$

$$f_{2opt} = \frac{-b_1 \pm \sqrt{b_1^2 - 4a_2 e_1}}{2a_1}$$

where,

$$k = S_b^2 + \frac{N_3}{N}(f_2 - 1)S_{bN_3}^2$$

$$+\frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\left\{\sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 + f_2\sum_{i=1}^{N_3} S_{iM}^2\right\}$$

$$+\frac{1}{N}\sum_{i=1}^{N_2}\frac{M_{i2}}{Mm}(f_{i2}-1)S_{M_{i2}}^2$$

$$a_2 = C_3\sum_{i=1}^{N_2} M_{i2}\left(N_3 S_{bN_3}^2 + \sum_{i=1}^{N_3}\left(\frac{1}{m}-\frac{1}{M}\right)S_{iM}^2\right),$$

$$b_2 = -C_3 N_3 \sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2,$$

$$e_2 = -C_1 N_3 \sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2$$

$$a_1 = \left(C_2 N_1 + C_2\sum_{i=1}^{N_2}\frac{M_{i1}}{M}\right)\sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2,$$

$$b_1 = -C_3\left[\left(N_3 + \sum_{i=1}^{N_2}\frac{M_{i2}}{M}\right)\sum_{i=1}^{N_2} M_{i2} S_{M_{i2}}^2 - \left(\sum_{i=1}^{N_2}\frac{M_{i2}}{M}S_{M_{i2}}^2 + \sum_{i=1}^{N_3} S_{iM}^2\right)\sum_{i=1}^{N_2} M_{i2}\right],$$

$$e_1 = -C_3\left(\sum_{i=1}^{N_1} S_{iM}^2 + \sum_{i=1}^{N_2} S_{iM}^2 - \sum_{i=1}^{N_2}\frac{M_{i2}}{M}S_{M_{i2}}^2\right)\sum_{i=1}^{N_2} M_{i2}$$

Case 4. The following estimator was considered for efficiency comparison purpose. Here we assume that a srswor sample of n psus is selected from N and within each selected psu a srswor of m ssus is selected. Data are collected through specialised efforts i.e. there is no non-response. Then we have the following theorem for the estimator in this case.

Theorem 2.4. The estimator is unbiased for, where

$$\bar{y} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{i=1}^{m} y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{im} \quad (2.10)$$

With variance,

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S_b^2 + \frac{1}{Nn}\sum_{i=1}^{N}\left(\frac{1}{m} - \frac{1}{M}\right)S_{iM}^2 \quad (2.11)$$

and an unbiased estimate of variance,

$$\hat{V}(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)s_b^2 + \frac{1}{Nn}\sum_{i=1}^{n}\left(\frac{1}{m} - \frac{1}{M}\right)s_{1im}^2 \quad (2.12)$$

where

$$s_b^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(\bar{y}_{im} - \bar{y}_{nm})^2,$$

$$\bar{y}_{im} = \frac{1}{m}\sum_{j=1}^{m} y_{ij}$$

$$s_{1im}^2 = \frac{1}{(m-1)}\left(\sum_{i=1}^{m} y_{ij}^2 - m\bar{y}_{im}^2\right)$$

Proof: The proof of unbiasedness of the given estimator and its variance and unbiased variance estimator canbe found in Cochran (1999), p. 277-278.

The cost function in this case is,

$$C = C_1 n + C_3 nm$$

where C, C_1, C_3 have been defined earlier.

To obtain optimum values of n and m we minimize the cost by fixing the variance. The optimum values are as follows,

$$n_{opt} = \frac{S_b^2 + \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{m} - \frac{1}{M})S_{iM}^2}{(V_0 + \frac{S_b^2}{N})} \quad \text{and}$$

$$m_{opt} = \sqrt{\frac{C_1\sum_{i=1}^{N}\frac{S_{iM}^2}{N}}{C_3(S_b^2 - \frac{1}{MN}\sum_{i=1}^{N}S_{iM}^2)}}$$

3. EMPIRICAL ILLUSTRATION

For the purpose of empirical illustration we consider the population MU 284 as given in Sarndal *et al.* (1992). Using this data a population with N= 18 and M = 15 was generated by combining the adjacent 15 units and allocating them to the respective psus. The variable P85 was used for analysis purpose. For numerical evaluation we have used SAS software. For empirical illustration various combinations of C_1, C_2, and C_3 were considered. As stated earlier, we consider C_2 to be much smaller than C_1 and C_3. The percentage reduction in expected cost of Case 1, Case 2, Case 3 over Case 4 along with optimum values of sample sizes and C_1, C_2, C_3 are given in Table 3.1. The %RIEC in Case 1 is given as $\frac{(C - C')}{C} \times 100$ where as the %RIEC in Case 2 is given as $\frac{(C - C'')}{C} \times 100$ and the %RIEC in Case 3 is given as $\frac{(C - C''')}{C} \times 100$, where C is the total cost for Case 4.

A close perusal of Table 3.1 reveals that the percentage reduction in expected cost decreases with increase in cost of data collection by expensive method. The percentage reduction in expected cost (%RIEC) is maximum in Case 2 followed by Case 3, and it is least in Case 1. The percentage reduction in expected cost remains almost same with increase in travel cost (C_1) for all the three cases also it remains almost the same with increase in data collection cost at first attempt (C_2), but it decreases with respect to the other two cases. The percentage reduction in the expected cost (%RIEC) increases with the increase in the cost per unit for collecting the information by expensive method after the first attempt fails to obtain information (C_3) for Case 1 and Case 2 but decreases for Case 3. It may be seen that for same C_1 and C_3, the value of n decreases with increase in C_2 for Case 1 – Case 3. Interestingly, for same C_1 and C_2 the value of m does not increases with increase in C_3. Next, for same C_2 and C_3 the value of n and m both not change, in all the cases, with the increase in C_1. For Case 1 and Case 2 the value of f_2 remains constant with the increase in C_3 but it slightly varies in Case 3. With increase in the value of C_2 the value of f_2 decreases for all the cases and with the changes in the value of C_1 the value of f_2 remains

Table 3.1. The percentage reduction in expected cost of Case 1, Case 2, Case 3 over Case 4 (Complete response) along with optimum values of sample sizes

Cost			Case 4 $(\bar{y})$		Case 1 $(\bar{y}')$				Case 2 $(\bar{y}'')$				Case 3 $(\bar{y}''')$			
C_1	C_2	C_3	n	m	n	m	f_2	% RIEC	n	m	f_2	% RIEC	n	m	f_2	% RIEC
55	2	70	18	6	23	13	3	39.01	23	23	6	57.01	55	12	12	51.40
55	2	75	18	5	23	13	3	39.19	23	23	6	57.10	57	12	13	50.98
55	2	80	18	5	24	13	3	39.34	24	23	6	57.16	58	12	13	50.56
55	4	70	18	6	21	10	2	38.70	21	20	4	47.42	43	12	8.5	44.77
55	4	75	18	5	21	10	2	39.08	21	20	4	47.73	44	12	8.8	44.48
55	4`	80	18	5	21	10	2	39.41	22	20	4	48.00	45	12	9.1	44.18
50	2	70	18	5	23	12	3	39.40	23	23	6	56.22	55	12	12	50.22
50	2	75	18	5	23	12	3	39.57	23	23	6	56.31	57	12	13	49.77
50	2	80	18	5	24	12	3	39.71	24	23	6	56.37	58	12	13	49.32
50	4	70	18	5	21	10	2	39.02	21	20	4	45.98	43	12	8.5	43.04
50	4	75	18	5	21	9	2	39.39	21	20	4	46.30	44	12	8.8	42.72
50	4	80	18	5	21	9	2	39.72	21	20	4	46.58	45	12	9.1	42.39
45	2	70	18	5	23	12	3	39.80	23	22	6	55.27	55	12	12	48.79
45	2	75	18	5	23	11	3	39.97	23	22	6	55.36	57	12	13	48.31
45	2	80	18	5	24	11	3	40.11	24	22	6	55.42	58	12	13	47.82
45	4	70	18	5	21	9	2	39.35	21	20	4	44.27	43	12	8.5	40.97
45	4	75	18	5	21	9	2	39.72	21	20	4	44.60	44	12	8.8	40.61
45	4	80	18	5	21	9	2	40.05	22	20	4	44.89	45	12	9.1	40.25

constant for all the cases. It is noteworthy that the difference in the values of m for Case 1 and Case 2 are attributable to the difference in optimum values of f_2 in the two cases. Further, for the Cases 1 and Case 3 in order to obtain the closed form expression for the optimum values n, m, f_2 and f_3 we considered $f_2 = f_3$ where $f_2 = \frac{m_2}{h_2}$ and $f_3 = \frac{n_3}{h_3}$. Thus f_2 and f_3 pertain to the second and first stages respectively. In view of the approximation $f_2 = f_3$ the difference between the optimum 'n' values in Case 1 and Case 3 also arises due to the difference in optimum values of f_2 in the two cases.

ACKNOWLEDGEMENTS

We would like to thank anonymous referees for their constructive comments. Their comments have led to a substantially improved final version of the paper.

REFERENCES

Chhikara, Raj S., and Sud, U.C. (2009). Estimation of population and domain totals under two-phase sampling in the presence of non-response. *J. Ind. Soc. Agril. Statist.,* **63(3)**, 297-304.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. John Wiley and Sons, Inc., New York.

Foradori, G.T. (1961). Some non-response sampling theory for two stage designs. Institute of Statistics, North Carolina State College.

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.*, **41**, 517-529.

Okafor, F.C. (2001). Treatment of non-response in successive sampling. *Statistica*, **LXI(2)**, 195-204.

Okafor, F.C. (2005). Sub-sampling the non-respondents in two-stage sampling over successive occasions. *J. Ind. Statist. Assoc.*, **43(1)**, 33-49.

Okafor, F.C. and Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology*, **26(2)**, 183-188.

Rancourt, E., Lee, H. and Särndal, C.E. (1994). Bias corrections for survey estimates from data with ratio imputed values for confounded non-response. *Survey Methodology*, **20**, 137-147.

Singh, R. and Mangat, N.P.S. (1996). *Elements of Survey Sampling*. Kluwer Academic publishers.

Sarndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Tripathi, T.P. and Khare, B.B. (1997). Estimation of mean vector in presence of non-response. *Comm. Statist. -Theory Methods*, **26(9)**, 2255-2269.