ESTIMATION OF DOMAIN MEAN USING TWO STAGE SAMPLING IN THE PRESENCE OF NONRESPONSE

Kaustav Aditya*

ABSTRACT

The problem of estimation of domain mean under random response mechanism has been considered when the sampling design is two-stage with two phases at the second stage. A estimator is developed based on the technique of sub sampling of non-respondents. Expressions for the variances of the estimators are developed. Optimum values of sample sizes are obtained by considering a suitable cost function. The percentage reduction in the expected cost of proposed estimators is studied empirically.

Keywords: Cost function; Optimum values; Random Response; Nonresponse; Sub-sampling; Two-stage sampling.

* **Indian Agricultural Statistics Research Institute, New Delhi**

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 nonresponse 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 nonresponse 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. Foradori (1961) studied the sub-sampling of the non-respondents technique to estimate the population total in two stages using unequal probability sampling. Srinath (1971) used a different procedure for selecting the sub-sample of respondents where the sub-sampling procedure varied according to the nonresponse rates. Oh and Schereun (1983) attempted to compensate for nonresponse by weighing adjustment. Kalton and Karsprzyk (1986) tried the imputation technique. 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 nonresponse. Again, Sud *et al.* (2012) considered the problem of estimation of finite population mean in the presence of nonresponse under two stage sampling design when the response mechanism was assumed to be deterministic. Further, Sud *et al.* (2013) considered the problem of estimation of finite population mean in the presence of nonresponse under two stage sampling design when the response mechanism was assumed to be random.

It may be mentioned that the weighting and imputation procedures aim at elimination of bias caused by nonresponse. 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 nonresponse is confounded i.e. the response probability is dependent on the survey character; it becomes difficult to eliminate the bias entirely. Rancourt *et al.* (1994) provided a partial correction for the situation. Hansen and Hurwitz's sub-sampling approach although costly, is free from any assumptions. When the bias caused by nonresponse is serious this technique is very effective i.e. one does not have to go for 100 percent response, which can be very expensive.

Besides the overall estimates, the estimates for different subgroups/domains of population are also required, (Sarndal *et al.*, 1992). The case of domain estimation is very common in agricultural surveys i.e. in case of crop area and yield estimation in case of mixed cropping there will be various types of mixtures prevailing in an area making it difficult to employ common large scale survey techniques and to perform crop cutting experiments. To solve this kind of problem we treat the various prevailing crop mixtures as domains of crops allocate the sample size randomly in those domains. Further, in case of household surveys we have different household income groups in a tehsil or district and to estimate the average household income at tehsil or district level we treat the various income groups as different domains and calculate the average income domain wise. Some time the sample sizes allocated in the domains randomly becomes very less making use of the design based estimators impossible and gives rise to a new chapter in survey sampling called as Small Area Estimation technique. Now, in the context of estimation of the domain parameters (mean/total), Agrawal and Midha (2007) proposed a two phase sampling design when the size of the domain was not known. Sud *et al.* (2010) considered the problem of estimation of finite population mean of a domain in the presence of nonresponse under a deterministic response mechanism. Chhikara and Sud (2009) considered the problem of estimation of finite population mean of a domain in the presence of item nonresponse under a random response mechanism. Aditya *et al.* (2013) considered the problem estimation of domain mean in the presence of nonresponse under two stage sampling design when the response mechanism was assumed to be deterministic. Again, Aditya *et al.* (2012, 2014) described the problem estimation of domain mean in the presence of nonresponse under two stage sampling

design. Now a days, most of the large scale surveys are based on multistage sampling designs with emphasis on estimation of various socioeconomic parameters and complex agricultural survey situations as this is an era of conservation agriculture when area under cultivation is decreasing each and every day and people are trying to obtain more produce from the same piece of land using mixed or intercropping. Now in large scale surveys nonresponse is very common. These days conduct of the surveys moved from paper based surveys to computer or mobile based surveys and interviews are conducted generally through emails or PDAs/Tablets using Computer Assisted Personal interview softwares. To tackle the problem of nonresponse in case of mail surveys, sub-sampling of nonrespondents technique was found very useful. In this paper we try to validate that is this technique of subsampling of nonrepondents is also a very efficeient technique to tackle nonresponse in modern day large scale surveys. In what follows, at estimator of domain mean using two-stage sampling designs are developed in Section 2 based on the technique of sub-sampling of the non-respondents. The response mechanism was assumed to follow the random response model where it was assumed that units go on missing in the population randomly. Also given are expressions for variance of the 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) labeled 1 through N. Let the i-th psu comprise M second stage units (ssus). Let us consider a population $U=(1,\ldots,k,\ldots,N)$ of size N partitioned into D sub-sets $U_1 \ldots U_d \ldots U_D$ (hereafter we refer them as domains) and let N_d (which is assumed large) be the size of $U_d (d=1,\ldots,D)$ such that $U=\bigcup_{d=1}^{D} U_d$ and $N=\sum_{d=1}^{D} N_d$. It is assumed that N_d is known. We assume that a sample s of n psus is drawn from the population under a simple random sampling without replacement (srswor) sampling design. Let s_d denote the part of sample s that happens to fall in U_d, that is, $s_d = s \cap U_d$. Let us denote by n_d the size of s_d such that $s=\bigcup_{d=1}^{D} s_d$ and $n=\sum_{d=1}^{D} n_d$. Note that throughout this article the sample size n drawn from N is fixed and known, however, the domain sample size n_d is random variable and unknown. As a consequence, the domain sample size n_d

needs to be estimated. When the domain sizes are small, n_d may turn out to be very small or it may be equal to '0' in some cases. In such cases small area estimation techniques are needed for reliable estimation at the domain level. However, we do not consider this case here. Let M_d be the size of the units in each psu belonging to the d-th domain and from each selected psu m_d ssus are selected by srswor and letters/mails containing questionnaires are sent to each unit in the sample. With the random sample of observations, the statistician's task is to make the best possible estimate for the domain. Let y_{dkj} be the value of study character pertaining to j-th ssu in the k-th psu in d-th domain, k=1, 2,…, N_d, j=1,2,…, M_d, d=1, 2,..., D. Our objective here is to estimate the domain mean $\bar{Y}_d = \frac{1}{N_d}\sum_{k=1}^{N_d}\frac{1}{M_d}\sum_{j=1}^{M_d} y_{dkj}$.

Let n psus be selected from N psus by simple random sampling without replacement (srswor) where n_d out of n psus fall in the d-th domain randomly and within each selected psu, m_d ssus are also selected from M_d ssus by srswor. Let out of a sample of m_d ssus selected from M_d ssus, m_{dk_1} units responds and m_{dk_2} units do not respond, $m_{dk_1} + m_{dk_2} = m_d$. From the m_{dk_2} nonresponding units a subsample of h_{dk_2} units is selected by srswor, $m_{dk_2} = h_{dk_2} f_{dk_2}$, k=1, 2,..., n_d. Let $\bar{y}_{m_{dk_1}}$ denote the mean of the sample from the response class for the d-th domain while $\bar{y}_{h_{dk_2}}$ denote the mean of the sample for the nonresponse class, where $\bar{y}_{m_{dk_1}} = \frac{1}{m_{dk_1}}\sum_{j=1}^{m_{dk_1}} y_{dkj}$ and $\bar{y}_{h_{dk_2}} = \frac{1}{h_{dk_2}}\sum_{k=1}^{h_{dk_2}} y_{dkj}$. Here at the second stage m_{dk_1} responding and m_{dk_2} nonresponding units are being generated as a result of m_d independent Bernoulli trials, one for each element k in m_d with constant probability θ_{dk} of "success", i.e. the response. So, we have

$$\Pr(k \in m_{dk_1} \mid m_d) = \theta_{k|m_d} = \theta_{dk} \text{ and } \Pr(k \,\&\, l \in m_{dk_1} \mid m_d) = \theta_{kl|m_d} = \theta_{dk}^2$$

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

$$\bar{y}_{1dr} = \frac{N}{nN_d}\sum_{k=1}^{n_d}\frac{1}{m_d}(m_{dk_1}\bar{y}_{m_{dk_1}} + m_{dk_2}\bar{y}_{h_{dk_2}}) \qquad (1)$$

with variance

$$V(\bar{y}_{1dr}) = \frac{N(N-n)(N_d-1)}{nN_d^2(N-1)} S_{bd}^2 + \frac{N(N-n)}{nN_d^2(N-1)} N_d Q_d \bar{Y}_d^2$$

$$+ \frac{N}{nN_d^2} \sum_{k=1}^{N_d} \left(\frac{1}{m_d} - \frac{1}{M_d}\right) S_{dk}^2 + \frac{N}{nN_d^2} \sum_{k=1}^{N_d} \frac{(1-\theta_{dk})}{m_d} (f_{dk_2} - 1) S_{dk}^2, \qquad (2)$$

where,

$$S_{bd}^2 = \frac{1}{(N_d-1)} \sum_{k=1}^{N_d} (\bar{Y}_{dk} - \bar{Y}_d)^2, \ \bar{Y}_{dk} = \frac{1}{M_d} \sum_{j=1}^{M_d} Y_{dk_j} \text{ and } \bar{Y}_d = \frac{1}{N_d} \sum_{k=1}^{N_d} \bar{Y}_{dk} = \frac{1}{N_d} \sum_{k=1}^{N} Z_{dk} \bar{Y}_k.$$

$$S_{dk}^2 = \frac{1}{(M_d-1)} \sum_{j=1}^{M_d} (Y_{dk_j} - \bar{Y}_{dk})^2, P_d = \frac{N_d}{N}, Q_d = 1 - P_d.$$

Proof: By definition,

$$E(\bar{y}_{1dr}) = E_1 E_2 E_3 E_4 E_5 \left[E_6 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} \frac{1}{m_d} (m_{dk_1} \bar{y}_{m_{dk_1}} + m_{dk_2} \bar{y}_{h_{dk_2}}) \right\} \right]$$

$$= E_1 E_2 E_3 E_4 \left[E_5 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} \frac{1}{m_d} (m_{dk_1} \bar{y}_{m_{dk_1}} + m_{dk_2} \bar{y}_{m_{dk_2}}) \right\} \right]$$

$$= E_1 E_2 E_3 \left[E_4 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} \frac{1}{m_d} (m_{dk_1} \bar{y}_{dk} + m_{dk_2} \bar{y}_{dk}) \right\} \right]$$

$$= E_1 E_2 \left[E_3 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} \frac{1}{m_d} (\theta_{dk} m_d \bar{y}_{dk} + (1-\theta_{dk}) m_d \bar{y}_{dk}) \right\} \right]$$

$$= E_1 E_2 \left[E_3 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} \bar{y}_{dk} \right\} \right]$$

$$= E_1 \left[E_2 \left\{ \frac{N}{nN_d} \sum_{k=1}^{n_d} (\bar{Y}_{dk}) \right\} \right]$$

$$= E_1 \left[\frac{N}{nN_d} \frac{n_d}{N_d} \sum_{k=1}^{N_d} \bar{Y}_{dk} \right] = \frac{N}{N_d} \frac{N_d}{NN_d} \sum_{k=1}^{N_d} \bar{Y}_{dk} = \bar{Y}_d.$$

Thus, $\bar{y}_{1dr}$ is an unbiased estimator of $\bar{Y}_d$. Here E_6 represents conditional expectations of all possible samples of size h_{dk_2} drawn from m_{dk_2}, E_5 is the conditional expectation of all possible samples of size m_{dk_1} and m_{dk_2} drawn respectively from M_{dk_1} and M_{dk_2} by keeping m_{dk_1} and m_{dk_2} fixed, E_4 refers to conditional expectation arises out of m_d independent Bernoulli trials resulting in m_{dk_1} success and m_{dk_2} failures, $m_{dk_1}+m_{dk_2}=m_d$, E_3 is the conditional expectation of all possible samples of size m_d drawn from M_d, E_2 refers to is the conditional expectation of all possible samples of size n_d drawn from a population of size N_d keeping n_d fixed and E_1 represents expectation arising out of randomness of n_d.

To find out the variance of the estimator we proceed as follows

$$V(\bar{y}_{1dr})=V_1E_2E_3E_4E_5E_6(\bar{y}_{1dr})+E_1V_2E_3E_4E_5E_6(\bar{y}_{1dr})+E_1E_2V_3E_4E_5E_6(\bar{y}_{1dr})$$
$$+E_1E_2E_3V_4E_5E_6(\bar{y}_{1dr})+E_1E_2E_3E_4V_5E_6(\bar{y}_{1dr})+E_1E_2E_3E_4E_5V_6(\bar{y}_{1dr}).$$

Here V_1,V_2,V_3,V_4,V_5,V_6 are defined similarly as E_1,E_2,E_3,E_4,E_5,E_6.

Hence,

$$V_1E_2E_3E_4E_5E_6(\bar{y}_{1dr})=\frac{N(N-n)}{nN_d^2}\left\{\frac{N_dQ_d\bar{Y}_d^2}{(N-1)}\right\},$$

$$E_1V_2E_3E_4E_5E_6(\bar{y}_{1dr})=\frac{N(N-n)(N_d-1)}{nN_d^2(N-1)}S_{bd}^2,$$

$$E_1E_2V_3E_4E_5E_6(\bar{y}_{1dr})=\frac{N}{n\ N_d^2}\sum_{k=1}^{N_d}\left(\frac{1}{m_d}-\frac{1}{m_d}\right)S_{dk}^2,$$

$$E_1E_2E_3V_4E_5E_6(\bar{y}_{1dr})=0,$$

$$E_1E_2E_3E_4V_5E_6(\bar{y}_{1dr})=0 \text{ and}$$

$$E_1E_2E_3E_4E_5V_6(\bar{y}_{1dr})=\frac{N}{nN_d^2}\sum_{k=1}^{N_d}\frac{(1-\theta_{dk})}{m_d}(f_{dk_2}-1)S_{dk}^2.$$

Hence, we get the required expression in eq. (2). We determine the optimum values of n, m_d and f_{dk_2} by minimizing the expected cost for a fixed variance. To achieve this consider the following cost function

$$C = C_{1d} n_d + C_{2d} \sum_{k=1}^{n_d} m_{dk_1} + C_{3d} \sum_{k=1}^{n_d} h_{dk_2},$$

where,

C: Total cost

C_{1d}: Per unit travel and miscellaneous cost in the d-th domain.

C_{2d}: Cost per unit of collecting the information on the study character in the first attempt in the d-th domain.

C_{3d}: Cost per unit of collecting the information by expensive method after the first attempt of collecting information failed in the d-th domain.

The expected cost in this case is,

$$E(C) = \frac{n}{N}[C_{1d} N_d + C_{2d} m_d \sum_{k=1}^{N_d} \theta_{dk} + C_{3d} m_d \sum_{k=1}^{N_d} \frac{(1-\theta_{dk})}{f_{dk_2}}].$$

Consider the function $\phi = E(C) + \lambda\{V(\bar{y}_{1dr}) - V_0\}$. Here, λ is the Lagrangian multiplier. To get closed form expression of the optimum value we assume that $m_{dk_2} = h_{dk_2} f_{2d}$, k=1, 2,..., n_d in place of $m_{dk_2} = h_{dk_2} f_{dk_2}$, k=1, 2,..., n_d. Also, V_0 can be determined by fixing the coefficient of variation, say equal to 5% or 10%. Differentiation with respect to n, m_d, λ and f_{2d}, equating the resultant derivatives to '0' and simplifying gives the optimum values as,

$$n_{opt} = \frac{K_{24}}{K_{23}}, \quad m_{dopt} = \frac{-b_{12} \pm \sqrt{b_{12}^2 - 4a_{12}e_{12}}}{2a_{12}}$$

and $$f_{2dopt} = \pm \sqrt{\frac{C_{3d} \sum_{k=1}^{N_d} (1-\theta_{dk}) \left(\sum_{k=1}^{N_d} S_{dk}^2 - \sum_{i=1}^{N_d} (1-\theta_{dk}) S_{dk}^2 \right)}{C_{2d} \sum_{k=1}^{N_{2d}} \theta_{dk} \sum_{k=1}^{N_d} (1-\theta_{dk}) S_{dk}^2}}.$$

To avoid negative values we consider,

$$m_{dopt} = \frac{-b_{12} + \sqrt{b_{12}^2 - 4a_{12}e_{12}}}{2a_{12}}$$

and $f_{2dopt} = \sqrt{\dfrac{C_{3d}\sum_{k=1}^{N_d}(1-\theta_{dk})\left(\sum_{k=1}^{N_d}S_{dk}^2 - \sum_{i=1}^{N_d}(1-\theta_{dk})S_{dk}^2\right)}{C_{2d}\sum_{k=1}^{N_{2d}}\theta_{dk}\sum_{k=1}^{N_d}(1-\theta_{dk})S_{dk}^2}}$.

Here,

$$K_{23} = V_0 + \frac{N}{N_d^2}\left\{\frac{N_d Q_d \bar{Y}_d^2}{(N-1)}\right\} + \frac{k_{12}}{N_d}S_{bd}^2, \left[k_{12} = \left[\frac{N(N_d-1)}{N_d(N-1)}\right]\right],$$

$$K_{24} = \left[\frac{N}{N_d}k_{12}S_{bd}^2 + \frac{N^2}{N_d^2}\left\{\frac{N_d Q_d \bar{Y}_d^2}{(N-1)}\right\} + \frac{N}{N_d^2}\sum_{k=1}^{N_d}\left(\frac{1}{m_d} - \frac{1}{M_d}\right)S_{dk}^2 + \frac{N}{N_d^2}\sum_{k=1}^{N_d}\frac{(1-\theta_{dk})}{m_d}(f_{2d}-1)S_{dk}^2\right],$$

$$a_{12} = \left(C_{3d}\sum_{k=1}^{N_d}\frac{(1-\theta_{dk})}{f_{2d}^2}\right)\left[N_d k_{12}S_{bd}^2 + N\left\{\frac{N_d Q_d \bar{Y}_d^2}{(N-1)}\right\} - \sum_{k=1}^{N_d}\frac{1}{M_d}S_{dk}^2\right],$$

$$b_{12} = -\left(C_{2d}\sum_{k=1}^{N_d}\theta_{dk}\sum_{k=1}^{N_d}(1-\theta_{dk})S_{dk}^2 - C_{3d}\sum_{k=1}^{N_d}\frac{(1-\theta_{dk})}{f_{2d}^2}\left\{\sum_{k=1}^{N_d}S_{dk}^2 - \sum_{k=1}^{N_d}(1-\theta_{dk})S_{dk}^2\right\}\right),$$

$$e_{12} = -C_{1d}N_d\sum_{k=1}^{N_d}(1-\theta_{dk})S_{dk}^2 \quad \text{and} \quad V_0 = 0.0025\times\bar{Y}_d^2.$$

Control Case. The following estimator was also considered for efficiency comparison purpose. Let n psus be selected from N psus by srswor design where n_d out of n psus fall in the d-th domain and within each selected psu, m_d ssus are also selected from M_d ssus by srswor. Data are collected through specialized efforts i.e. there is no nonresponse. Then we give the following theorem,

Theorem 2.4. The unbiased estimator for $\bar{Y}_d$ is given as,

$$\bar{y}_{dr} = \frac{N}{nN_d}\sum_{k=1}^{N_d}\bar{y}_{dk} \qquad (7)$$

with variance,

$$V(\bar{y}_{dr}) = \frac{N(N-n)(N_d-1)}{nN_d^2(N-1)}S_{bd}^2 + \frac{N(N-n)}{nN_d^2(N-1)}N_d Q_d \bar{Y}_d^2 + \frac{N}{nN_d^2}\sum_{k=1}^{N_d}\left(\frac{1}{m_d} - \frac{1}{M_d}\right)S_{dk}^2. \qquad (8)$$

S_{bd}^2, S_{dk}^2, Q_d etc. are defined earlier.

Proof: By definition,

$$E(\bar{y}_{dr}) = E_1E_2\left[E_3(\frac{N}{nN_d}\sum_{k=1}^{n_d}\bar{y}_{dk})\right]$$

$$= E_1\left[E_2(\frac{N}{nN_d}\sum_{k=1}^{n_d}\bar{Y}_{dk})\right]$$

$$= E_1\left[\frac{N}{nN_d}\frac{n_d}{N_d}\sum_{k=1}^{N_d}\bar{Y}_{dk}\right]$$

$$= \frac{1}{N_d}\sum_{k=1}^{N_d}\bar{Y}_{dk} = \bar{Y}_d .$$

Thus, $\bar{y}_{dr}$ is an unbiased estimator of $\bar{Y}_d$. Here, E_3 is conditional expectation pertaining to all possible samples of size m_d drawn from M_d and E_2 is expectation pertaining to all possible samples of size n_d drawn from N_d keeping n_d fixed and E_1 is the expectations arising out of randomness of n_d.

By definition,

$$V(\bar{y}_{dr}) = V_1E_2E_3(\bar{y}_{dr}) + E_1V_2E_3(\bar{y}_{dr}) + E_1E_2V_3(\bar{y}_{dr}).$$

Where,

$$V_1E_2E_3(\bar{y}_{dr}) = \frac{N(N-n)}{nN_d^2(N-1)}N_dQ_d\bar{Y}^2,$$

$$E_1V_2E_3(\bar{y}_{dr}) = \frac{N(N-n)(N_d-1)}{nN_d^2(N-1)}S_{bd}^2,$$

$$E_1E_2V_3(\bar{y}_{dr}) = \frac{N}{nN_d^2}\sum_{k=1}^{N_d}(\frac{1}{m_d}-\frac{1}{M_d})S_{dk}^2.$$

For optimization the relevant cost function in this case is,

$$C = C_{1d}n_d + C_{3d}n_dm_d .$$

Here expected cost is given as,

$$E(C) = \frac{n}{N}(C_{1d}N_d + C_{3d}N_dm_d).$$

where, C, C_{1d}, C_{3d} have been defined earlier.

To obtain optimum values of n and m_d we minimize the expected cost by fixing the variance. The optimum values are obtained in the same way as earlier is,

$$n_{opt} = \frac{\left[\frac{N^2}{N_d^2}\frac{(N_d-1)}{(N-1)}S_{bd}^2 + \frac{N^2}{N_d^2}\left\{\frac{N_d Q_d \bar{Y}_d^2}{N-1}\right\} + \frac{N}{N_d^2}\sum_{k=1}^{N_d}(\frac{1}{m_d}-\frac{1}{M_d})S_{dk}^2\right]}{(V_0 + \frac{N}{N_d^2}\left\{\frac{N_d Q_d \bar{Y}_d^2}{N-1}\right\} + \frac{N(N_d-1)}{N_d^2(N-1)}S_{bd}^2)} \quad \text{and}$$

$$m_{dopt} = \pm\sqrt{\frac{C_{1d}\sum_{k=1}^{N_d}S_{dk}^2}{C_{3d}(\frac{N(N_d-1)}{(N-1)}S_{bd}^2 + \left\{\frac{NN_d Q_d \bar{Y}_d^2}{N-1}\right\} - \frac{1}{M_d}\sum_{k=1}^{N_d}S_{dk}^2)}}$$

We avoid negative values, therefore, $m_{dopt} = \sqrt{\frac{C_{1d}\sum_{k=1}^{N_d}S_{dk}^2}{C_{3d}(\frac{N(N_d-1)}{(N-1)}S_{bd}^2 + \left\{\frac{NN_d Q_d \bar{Y}_d^2}{N-1}\right\} - \frac{1}{M_d}\sum_{k=1}^{N_d}S_{dk}^2)}}$

3. Empirical Illustration

For empirical illustration, first a population of size 1000 was generated from normal distribution with mean 22 and variance 2.5 and then from this population N=100 psus was formed by combining the adjacent 10 units and allocating them to the respective psus. From these N psus a sample of n = 50 psus each of size 10 ssus are drawn using srswor. Here the population was divided into three domains. The first domain is of size N_1= 25 psus each of size M_1=10 ssus, the second domain is of size N_2= 35 psus each of size M_2=10 ssus and the third domain is of size N_3=40 psus each of size M_3=10 ssus. The mean and variance of the character under study for the first, second and third domains are 21.90 and 4.36, 22.05 and 4.17, and 21.98 and 3.79 respectively. We considered $M_{dk_1} = M_{dk_2} = 5$, i.e. nonresponse rate as 50%. However, we reported the results for nonresponse rate of 50 per cent only. Moreover, as we expect the performance of various estimators improve when we reduce the nonresponse rate. Here we have considered different values of θ_{dk} for each psu in each domain. For empirical illustration various combinations of C_{1d}, C_{2d}, and C_{3d} were considered. The percentage reduction in expected cost of $\bar{y}_{1dr}$ over $\bar{y}_{dr}$ along with optimum values of sample sizes for different values of C_{1d}, C_{2d}, and C_{3d}

are given in **Table 1** for domain 1, in **Table 2** for domain 2 and in **Table 3** for domain 3. We have considered CV value as 5%.

4. Discussion and Conclusion

A close perusal of **Tables 1** reveals that there is a gain in %RIEC for the proposed estimator. The %RIEC decreases with increase in travel and miscellaneous cost (C_{1d}) and data collection cost at first attempt (C_{2d}) for the estimator and the %RIEC increases with the increase in the cost per unit of collecting the information by expensive method after the first attempt to obtain information failed (C_{3d}) for all the three domains. A close perusal of all the tables shows that the proposed estimator of domain mean in the presence of nonresponse, based on sub-sampling of the non-respondents, when domain size is known in advance, are better than an estimator based only on the interview method and having 100% response at the second stage. Further, the rate of increase of %RIEC is proportional to the cost of data collection by expensive method for the proposed estimator. Finally, the %RIEC is maximum for the first estimator in domain 3 followed by domain 2 and it is least in domain 1. Occasionally, the optimum value of sample in a domain may exceed population, i.e., the total units in the domain. In this situation, the best can be done is to take entire population as a sample for that domain (i.e., 100 per cent sampling), see Sukhatme *et al.* (1984).

References

- Aditya, K., Sud, U.C. and Hukum Chandra (2012). Estimation of domain total for unknown domain size in the presence of nonresponse. Statistics and Applications, 10, Nos.1 & 2, 13-25.
- Aditya, K., Sud, UC and Chandra, H (2014). Estimation of domain mean using two stage sampling with sub-sampling of non-respondents Journal of the Indian Society of Agricultural Statistics, 68(1), 39-54.
- Agrawal, M.C., and Midha, C.K. (2007). Some efficient estimators of the domain parameters. *Statistics and Probability Letters*, **77**, 704-709.
- Chhikara, Raj S., and Sud, U.C. (2009). Estimation of population and domain totals under two-phase sampling in the presence of non-response. *Journal of the Indian Society of Agricultural Statistics*, **63(**3), 297-304.

- Durbin, J. (1954). Nonresponse and call-backs in surveys. *Bulletin of International Statistical Institute*, **34,** 72-86.
- 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. *Journal of the American Statistical Association*, **41**, 517-529.
- Okafor, F.C. (2001). Treatment of non-response in successive sampling. *Statistica*, **61(2)**, 195-204.
- Okafor, F.C. (2005). Sub-sampling the non-respondents in two-stage sampling over successive occasions. *Journal of Indian Statistical Association*, **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.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, **12**, pp. 1-16.
- Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit non-response. In: W.G. Madow, I. Olkin and B. Rubin (eds.), *Incomplete data in sample surveys*, Vol.2. New York: Academic press, 143-184.
- 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.
- Sarndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Srinath, K.P. (1971). Multiphase sampling in non-response problems. *Journal of the American Statistical Association*, **66,** 583-586.
- Sud, U.C., Chandra H., and Chikkara, R.S. (2010). Domain estimation in the presence of non-response. *Journal of the Indian Society of Agricultural Statistics*. **64**, 343-347.
- Sud, U. C., Aditya, K., Chandra H., and Parsad, Rajender (2012). Two stage sampling for estimation of population mean with sub-sampling of non respondents, Journal of the Indian Society of Agricultural Statistics, 66(3), 447-457.

- Sud, U. C., Aditya, K., Chandra H., and Parsad, Rajender (2013). Two stage sampling with two-phases at the second stage of sampling for estimation of finite population mean under random response mechanism, Journal of the Indian Society of Agricultural Statistics, 67(3), 305-317.
- Sukhatme, P. V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Third Revised Edition, Iowa State University Press, USA.
- Tripathi, T. P., and Khare, B. B. (1997). Estimation of mean vector in presence of non-response. *Communications in Statistics - Theory and Methods*, **26(9)**, 2255- 2269.

Table 1. The optimum values of sample sizes along with percentage reduction in expected cost of $\bar{y}_{11r}$ over $\bar{y}_{1r}$ in domain 1.

cost			Control ($\bar{y}_{41r}$)		First estimator ($\bar{y}_{11r}$)			
C_{11}	C_{21}	C_{31}	n	m_1	n	m_1	f_{21}	% RIEC
25	1	45	46	4	46	9	4.28	74.32
25	1	50	46	4	46	9	4.51	74.99
25	1	55	46	4	46	9	4.73	75.57
25	2	45	46	4	46	8	3.02	67.98
25	2	50	46	4	46	8	3.19	68.77
25	2	55	46	4	46	8	3.34	69.46
30	1	45	46	4	46	9	4.28	74.31
30	1	50	46	4	46	9	4.51	75.01
30	1	55	46	4	46	9	4.73	75.61
30	2	45	46	4	46	9	3.02	68.63
30	2	50	46	4	46	9	3.19	69.43
30	2	55	46	4	46	9	3.34	70.13
35	1	45	46	5	46	9	4.28	74.18
35	1	50	46	4	46	9	4.51	74.89

35	1	55	46	4	46	9	4.73	75.51
35	2	45	46	5	46	9	3.02	68.99
35	2	50	46	4	46	9	3.19	69.81
35	2	55	46	4	46	9	3.34	70.52

Table 2. The optimum values of sample sizes along with percentage reduction in expected cost of $\bar{y}_{12r}$ over $\bar{y}_{42r}$ in domain 2.

cost			Control ($\bar{y}_{42r}$)		First estimator ($\bar{y}_{12r}$)			
C_{12}	C_{22}	C_{32}	n	m_2	n	m_2	f_{22}	% RIEC
25	1	45	52	9	52	3	1.52	77.96
25	1	50	52	8	52	3	1.60	78.83
25	1	55	52	8	52	3	1.68	79.58
25	2	45	52	9	52	2	1.07	76.63
25	2	50	52	8	52	2	1.13	77.51
25	2	55	52	8	52	2	1.18	78.27
30	1	45	52	10	52	3	1.52	76.63
30	1	50	52	9	52	3	1.60	77.56
30	1	55	52	9	52	3	1.68	78.37
30	2	45	52	10	52	3	1.07	75.43
30	2	50	52	9	52	3	1.13	76.38
30	2	55	52	9	52	3	1.18	77.20
35	1	45	52	11	52	3	1.52	75.38
35	1	50	52	10	52	3	1.60	76.38
35	1	55	52	10	52	3	1.68	77.24
35	2	45	52	11	52	3	1.07	74.28
35	2	50	52	10	52	3	1.13	75.29
35	2	55	52	10	52	3	1.18	76.17

Table 3. The optimum values of sample sizes along with percentage reduction in expected cost of $\bar{y}_{1dr}$ over $\bar{y}_{4dr}$ in domain 3.

cost			Control ($\bar{y}_{43r}$)		First estimator ($\bar{y}_{13r}$)			
C_{13}	C_{23}	C_{33}	n	m_3	n	m_3	f_{23}	% RIEC
25	1	45	50	12	50	5	1.52	81.54
25	1	50	50	12	50	5	1.60	82.20
25	1	55	50	11	50	5	1.68	82.78
25	2	45	50	12	50	4	1.07	79.63
25	2	50	50	12	50	4	1.13	80.32
25	2	55	50	11	50	4	1.18	80.91
30	1	45	50	13	50	5	1.52	80.69
30	1	50	50	13	50	5	1.60	81.41
30	1	55	50	12	50	5	1.68	82.03
30	2	45	50	13	50	4	1.07	78.98
30	2	50	50	13	50	4	1.13	79.71
30	2	55	50	12	50	4	1.18	80.35
35	1	45	50	14	50	5	1.52	79.87
35	1	50	50	14	50	5	1.60	80.63
35	1	55	50	13	50	5	1.68	81.29
35	2	45	50	14	50	4	1.07	78.30
35	2	50	50	14	50	4	1.13	79.08
35	2	55	50	13	50	4	1.18	79.76