Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(3) 2009 297-304

Estimation of Population and Domain Totals under Two-phase Sampling in the Presence of Non-response

Raj S. Chhikara[1*] and U.C. Sud[2]

[1]*University of Houston-Clear Lake, Houston, Texas, USA*

[2]*Indian Agricultural Statistics Research Institute, New Delhi*

(Received: July 2009, Revised: November 2009, Accepted: November 2009)

SUMMARY

This paper considers estimation of both domain and population totals for an item of interest using a two-phase sampling where the domain identity is realized, but the item response is not necessarily available from a phase I sampled unit. The optimality of sample design is studied considering the probability of item response, the cost of phase I vs. phase II sampling, and the item variability in the domains. Numerical evaluations made using a simulation study show that the proposed sampling and estimation method is more efficient than an alternative method given by Agrawal and Midha (2007).

Key words : Cost function, Domain estimation, Optimal sample design, Probability of item response.

1. INTRODUCTION

A common practice in survey sampling is not only to estimate a population mean or total for an item of interest, but also its estimate is desired at the domain level. The domain size is invariably unknown and varies across domains. When the response for a sample unit is subject to a chance mechanism, a standard sampling approach is to employ a two phase sample design as discussed in Cochran (1977, p. 370) and Sarndal, *et al*. (1992, p. 566). It consists of a single or multistage phase I sample of a fixed number of units drawn from the population. For phase II sampling, the construction of an efficient sample design would depend upon whether or not the response for a sample unit in phase I lacks completely or partially. If the sampled unit has no response at all as in mail surveys, it is a complete non-response, termed as unit non-response, whereas if the unit identity for its domain is realized, but not its value for an item, it is a partial response, termed as item non-response. In this paper we assume the latter so that in phase I the sample unit response for its domain identity is realized, but not necessarily its value for the item of interest. Next, it is assumed that the item response would be ascertained for the phase II sample units. A subset of the item non-responding phase I sample units is considered for phase II sub-sampling.

It is worth while to mention here that the problem of estimation of domain means using two-phase sampling was considered as early as 1973 by Degraft-Johnson and Sedransk. However, the necessary theory was developed under the assumption of no non-response. Kun He (1995) obtained a minimax estimator under a squared error loss function for the domain totals when the number of sample units falling into the different domains is random.

Recently, Agrawal and Midha (2007) discussed estimation of domain total using a two-phase sample design that consisted using the phase I samples to estimate the domain size and the phase II sub-sample to estimate the domain total for the item observed. Since the item values may be available for some phase I

[*]*Corresponding author* : Raj S. Chhikara
E-mail address : chhikara@uhcl.edu

sample units, as in telephone surveys, their domain estimator can be improved upon by incorporating the item responses obtained during phase I sampling. Sarndal and Swensson (1987) employed basically the same two phase sample design as presently considered. Although they considered the item responses only from the phase II sub-sample drawn from strata (fixed or not) using stratified sampling, yet they assumed to have available the unit values for an auxiliary variable at the first phase sampling. As such they proposed and discussed regression estimation for both the population and strata/domains. The response homogeneity groups were assumed for domains so as to achieve higher precision for the estimator and reduce its bias due to item non-response. However, they did not discuss the optimality of their sample design as they did not raise the issue of cost of sampling in their study.

The estimator and its variance for the domain and population totals can be stated in terms of inclusion probabilities for the two phases of sampling under a general framework as given in Sarndal *et al*. (1992); but we skip it. Instead, we focus on the estimators and their variances for the case of simple random sampling without replacement, which is the frequently employed sampling method in sample surveys.

The two phase sample design and the estimation of domain totals are discussed in the next section. The optimality of phase II sample design is also developed. The estimator for population total as a sum of the domain total estimators is utilized to optimize phase I sample size and is described in Section 3. This estimator of the population total is contrasted with the standard estimator obtained directly without considering the domains as given in Sarndal *et al*. (1992) in the context of non-response. A simulation study is made to evaluate the properties of the proposed estimators of domain and population totals, and then these are compared with those given in Agrawal and Midha (2007).

2. DOMAIN ESTIMATION

2.1 Two-Phase Sampling Scheme

Consider a partitioning of the population U into sub-sets, $U_1,\ldots,U_d,\ldots,U_D$, called domains. Let N_d be the size of U_d, where N_d is assumed unknown, and the domain identity of units is also not known a priori. We have the partitioning equations

$$U = \bigcup_{d=1}^{D} U_d \ , \ N = \sum_{d=1}^{D} N_d$$

Let y denote the variate for the item of interest. The objective is to estimate the domain totals, $t_d = \sum_{i}^{N_d} y_{id}$, $d = 1,\ldots., D$, or the domain means $\bar{Y}_d = t_d / N_d$, $d = 1,\ldots., D$, as well as the population total $t = \sum_{1}^{D} t_d$ or mean $\bar{Y} = t/N$.

Assume that a survey is carried out for the population U where a sample s of size n is drawn from U according to SRSWOR sampling design. We assume that the selected unit is subject to item non-response; however the unit response in terms of its domain identity is obtained. The sample of size n is post-stratified into D domains on the basis of the unit domain identity as observed. Let s_d denote the part of s that happens to fall in U_d, that is, $s_d = s \cap U_d$. Denote by n_d the size of s_d. Hence

$$s = \bigcup_{d=1}^{D} s_d \ ; \ n = \sum_{d=1}^{D} n_d$$

Here n_d is random. We need to avoid for n_d to be quite small. As such a large sample of size n may have to be drawn if D is not small. Let n_{1d} of n_d units falling in the d-th domain respond for the item of interests, while the remaining n_{2d} units do not respond, $n_d = n_{1d} + n_{2d}$. Further, we assume that the item response set s_{d1} is generated as a result of n_d independent Bernoulli trials, one for each element k in s_d, with constant probability θ_d of 'success', i.e., item response. Thus for any $k \in s_d$, and every k and $l \in s_d$, $Pr(k \in s_{d1}/s_d) = \theta_d$; $Pr(k \ \& \ l \in s_{d1}/s_d) = \theta_d^2$.

For phase II sampling, a stratified sample is considered, where from the d-th domain ($d = 1, 2, \ldots, D$) a random sub-sample of size m_{2d} out of n_{2d} non-responding units is randomly drawn, and the responses are obtained for all m_{2d} sampled units through specialized efforts. Thus, the cost of obtaining response

in the second phase is expected to be much more than that in the first phase. With this limitation for the phase II sample of observations, the statistician's task is to minimize the phase II sampling, and to make the best possible use of the sample observations to estimate the domain totals or means.

2.2 Estimator of t_d

Define

$$y_i = \begin{cases} y_{id} & \text{if } i \in U_d \\ 0 & \text{otherwise} \end{cases}$$

Let $\bar{y}_{n_{1d}} = \dfrac{\sum^{n_{1d}} y_{id}}{n_{1d}}$ and $\bar{y}_{m_{2d}} = \dfrac{\sum^{m_{2d}} y_{id}}{m_{2d}}$

Then we propose

$$\hat{t}_d = \frac{N}{n}\left[n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{m_{2d}} \right] \qquad (2.1)$$

for an estimator of the item total for the d-th domain, d=1, 2,…,D.

Remark: The response mechanism is sometimes considered deterministic in the following way. The population consists of two groups U_1 and U_2. All elements in U_1 respond with probability 1 if selected, whereas for all elements in U_2 the response probability is 0. Thus the composition of the groups is fixed, once for all. This case has been dealt with by Sud *et al*. (2009).

The estimator $\hat{t}_d$ is subjected to variability due to (1) the first phase sampling of the population and the number of sample units falling in domain d, (2) the sample units drawn from the domain (3) the chance mechanism for item response for a unit as modeled by Bernoulli sampling, and (4) SRSWOR for the second phase sampling in each domain. Let E_1, E_2, E_3 and E_4 denote expectations with conditionals in the order of variability stated above. Then it easily follows that $\hat{t}_d$ is an unbiased estimator of t_d since

$$E(\hat{t}_d) = \frac{N}{n}E_1E_2E_3\left(n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{n_{2d}}\right)$$

$$= \frac{N}{n}E_1E_2\left(n_d\bar{y}_{n_d}\right) = N_d\bar{Y}_d$$

To compute the variance of $\hat{t}_d$, denote here V_1, V_2, V_3 and V_4 to be the variances with proper prior conditionals in the variability due to (2)-(4) as stated above. Then the variance of $\hat{t}_d$ can be obtained by considering

$$V(\hat{t}_d) = V_1E_2E_3E_4(\hat{t}_d) + E_1V_2E_3E_4(\hat{t}_d) + E_1E_2V_3E_4(\hat{t}_d) + E_1E_2E_3V_4(\hat{t}_d)$$

Let

$$P_d = \frac{N_d}{N};\ Q_d = 1 - P_d;\ f_{2d} = \frac{m_{2d}}{n_{2d}}$$

$$S_d^2 = \frac{1}{(N_d - 1)}\sum_i^{N_d}\left(y_i - \bar{Y}_d\right)^2$$

$$S_{n_{2d}}^2 = \frac{1}{(n_{2d} - 1)}\sum_i^{n_{2d}}\left(y_i - \bar{y}_{n_{2d}}\right)^2$$

$$S_{n_d}^2 = \frac{1}{(n_d - 1)}\sum_i^{n_d}\left(y_i - \bar{y}_{n_d}\right)^2$$

It follows that

$$E_1E_2E_3V_4(\hat{t}_d) = \frac{N^2}{n^2}E_1E_2E_3\left(n_{2d}\left(\frac{1}{f_{2d}} - 1\right)S_{n_{2d}}^2\right)$$

$$= \frac{N^2}{n^2}(1-\theta_d)E_1E_2\left(n_d\left(\frac{1}{f_{2d}} - 1\right)S_{n_d}^2\right)$$

$$= \frac{N^2}{n}P_d(1-\theta_d)\left(n_d\left(\frac{1}{f_{2d}} - 1\right)S_d^2\right)$$

$$E_1E_2V_3E_4(\hat{t}_d) = \frac{N^2}{n^2}E_1E_2V_3\left(n_d\bar{y}_d\right) = 0$$

$$E_1V_2E_3E_4(\hat{t}_d) = \frac{N^2}{n^2}E_1V_2\left(n_d\bar{y}_d\right)$$

$$= \frac{N^2}{n^2}E_1\left[n_d\left(\frac{N_d - n_d}{N_d - 1}\right)P_dS_d^2\right]$$

$$= \frac{N^2}{n}\left[1 - \frac{(n-1)}{(N_d - 1)}P_d\right]P_dS_d^2$$

$$V_1E_2E_3E_4(\hat{t}_d) = \frac{N^2}{n}V_1E_2\left(n_d\bar{y}_d\right)$$

$$= \frac{N^2}{n}\left(\frac{N-n}{N-1}\right)P_dQ_d\bar{Y}_d^2$$

Adding the above four terms and simplifying we get

$$V(\hat{t}_d) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N-1}\right)\left(\frac{1}{\left(CV_d\right)^2}-\frac{1}{N_d}\right)\right.$$

$$\left.+\left(1-\theta_d\right)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 \qquad (2.2)$$

where, $CV_d = S_d/\bar{Y}_d$ is the coefficient variation for the d-domain.

A variance estimator of $V(\hat{t}_d)$ is obtained by

$$\hat{V}(\hat{t}_d) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+q_d\left(\frac{N-n}{N-1}\right)\left[\frac{1}{\left(cv_d\right)^2}-\frac{n}{Nn_d}\right]\right.$$

$$\left.+(1-\hat{\theta}_d)\left(\frac{1}{f_{2d}}-1\right)\right]\left(\frac{n_d}{n}\right)s_d^2 \qquad (2.3)$$

Where $q_d = 1 - \frac{n_d}{n}$, $cv_d = \frac{s_d}{\bar{y}_d}$

$$s_d^2 = \frac{1}{(n_{1d}+m_{2d})-1}\sum_{1}^{n_{1d}+m_{2d}}(y_i-\bar{y}_d)^2$$

$$\bar{y}_d = \frac{1}{(n_{1d}+m_{2d})}\sum_{1}^{n_{1d}+m_{2d}}y_i \text{ and } \hat{\theta}_d = \frac{n_{1d}}{n_d}$$

If N and N_d are large, ignoring terms of order $1/N$ and $1/N_d$, $V(\hat{t}_d)$ is approximately equal to

$$V(\hat{t}_d) \sim \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N}\right)\left(\frac{1}{\left(CV_d\right)^2}\right)\right.$$

$$\left.+(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 \qquad (2.4)$$

A variance estimate can then be obtained by replacing the population quantities by their sample estimate as done in (2.3).

2.3 Optimization under a Cost Function

For domain d, consider the sample unit cost c_{1d} for the response at phase I and c_{2d} at phase II, where $c_{2d} >> c_{1d}$, possibly dominated by a large multiple. Then the cost function for domain d, except for a fixed overall cost, is

$$C_d = n_d c_{1d} + m_{2d}c_{2d}$$

Then its expected cost is given by

$$E(C_d) = nP_d[c_{1d} + f_{2d}(1-\theta_d)c_{2d}] \qquad (2.5)$$

The optimum value of f_{2d} is obtained by fixing the variance of $\hat{t}_d$ say equal to $N_d^2V_{0d}$ and minimizing the expected cost. This can be determined by minimizing the function

$$\phi = nP_d[c_{1d} + f_{2d}(1-\theta_d)c_{2d}]$$

$$+\lambda\left[\frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+Q_d\left(\frac{N-n}{N}\right)\left(\frac{1}{(CV_d)^2}\right)\right.\right.$$

$$\left.\left.+(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)\right]P_dS_d^2 - N_d^2V_{0d}\right]$$

where λ is the Lagrangian multiplier.

Differentiating with respect to n, λ, f_{2d}, equating the resultant equations to 0 and solving for f_{2d} gives the optimum value of f_{2d}, after ignoring $(1/N_d)$ term, as

$$f_{2d(\text{opt})} = \sqrt{\frac{c_{1d}/c_{2d}}{\theta_d + \frac{Q_d}{(CV_d)^2}}} \qquad (2.6)$$

To determine an optimum value of n one has to look at the objective of the sample survey. If the objective is simply to achieve a desired level of precision for the estimator of domain total or mean, then we need to determine n so that $V(\hat{t}_d) \le N_d^2V_{0d}$ as considered in the determination of $f_{2d(\text{opt})}$. The optimum value of n, say $n^{(d)}$, is given by

$$n^{(d)} = \frac{\left[N^2\left[(1-\theta_d)\left(\frac{1}{f_{2d(opt)}}-1\right)\right.\right.}{\left.\left.+1-N\frac{P_d}{Q_d}+Q_d(CV_d)^{-2}\right]\right]}$$

$$\frac{}{\left[N^2\frac{V_{0d}}{P_dS_d^2}+N-\frac{Q_d}{P_d}+NQ_d(CV_d)^{-2}\right]} \qquad (2.7)$$

Eq.(2.7) leads to an optimal value for n corresponding to each domain. One might choose $n = \max(n^{(d)})$ for the optimum phase I sample size. However, it is more appropriate and hence preferable to optimize the phase I sample size so that the estimator of the population total has a desired precision. The variance of the estimator of population total is discussed in the next section and so is determination of n.

3. ESTIMATOR OF POPULATION TOTAL

Theorem. The estimator

$$\hat{t} = \sum_1^D \hat{t}_d = \frac{N}{n}\sum_1^D \left[n_{1d}\bar{y}_{n_{1d}} + n_{2d}\bar{y}_{m_{2d}} \right]$$

is an unbiased estimator of population total t with variance

$$V(\hat{t}_d) = N^2 \frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D P_d\left(1-\theta_d\right)\left(\frac{1}{f_{2d}}-1\right)S_d^2 \quad (3.1)$$

where

$$S^2 = \frac{1}{(N-1)}\left[\sum_1^D N_d(\bar{Y}_d-\bar{Y})^2 + \sum_1^D (N_d-1)S_d^2\right]$$

$$\bar{Y} = \sum_1^D P_d\bar{Y}_d$$

Proof. From Section 2, $E_1E_2E_3E_4(\hat{t}_d) = t_d$. Therefore,

$$E_1E_2E_3E_4(\hat{t}) = \sum_1^D t_d = t$$

Next,

$$V(\hat{t}_d) = V_1E_2E_3E_4(\hat{t}) + E_1V_2E_3E_4(\hat{t}) + E_1E_2V_3E_4(\hat{t}) + E_1E_2E_3V_4(\hat{t})$$

where

$$E_1E_2E_3V_4(\hat{t}) = E_1E_2E_3\frac{N^2}{n^2}\sum_1^D n_{2d}\left(\frac{1}{f_{2d}}-1\right)S^2_{n_{2d}}$$

$$= \frac{N^2}{n^2}E_1E_2\sum_1^D(1-\theta_d)(n_d)\left(\frac{1}{f_{2d}}-1\right)S^2_{n_d}$$

$$= \frac{N^2}{n}\sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2$$

From Section 2, both $E_1E_2V_3E_4(\hat{t})$, $E_1V_2E_3E_4(\hat{t})$, can be shown to be equal to 0.

$$V_1E_2E_3E_4(\hat{t}) = V_1\frac{N}{n}\sum_1^D\left(n_d\bar{y}_{n_d}\right)$$

$$= V_1(N\bar{y})$$

$$= \frac{N}{n}\left(1-\frac{N}{n}\right)S^2$$

Here $\bar{y} = \dfrac{\sum_1^D n_d\bar{y}_{n_d}}{n}$

Adding the terms we obtain

$$V(\hat{t}) = N^2\frac{S^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2 \quad (3.1)$$

The optimum value of n is obtained by fixing the variance of $\hat{t}$, say, equal to N^2V_0, where V_0 is pre-specified. Thus, the optimum value of n, ignoring $1/N$ terms, is given by

$$n_{opt} = \frac{\left[S^2 + \sum_1^D P_d(1-\theta_d)\left(\frac{1}{f_{2d}}-1\right)S_d^2\right]}{V_0} \quad (3.2)$$

Next, a variance estimator is obtained by replacing the population quantities in (3.1) by the corresponding sample statistics

$$\hat{V}(\hat{t}) = N^2\frac{s^2}{n}\left(1-\frac{n}{N}\right) + \frac{N^2}{n}\sum_1^D p_d\left(1-\hat{\theta}_d\right)\left(\frac{1}{f_{2d}}-1\right)s_d^2$$

where

$$s^2 = \frac{1}{(n-1)}\left[\sum_1^D(n_{1d}+m_{2d})(\bar{y}_d-\bar{y})^2 + \sum_1^D(n_{1d}+m_{2d})s_d^2\right]$$

$$\bar{y} = \frac{1}{\sum_1^D(n_{1d}+m_{2d})}\sum_1^{n_{1d}}\sum_1^{m_{2d}} y_i$$

Alternatively, one may utilize the following unbiased estimator of t as given in Sarndal *et al*. (1992) obtained without consideration of domains

$$\hat{t}_0 = \frac{N}{n}\left[n_1\bar{y}_1 + n_2\bar{y}_{m_2}\right] \quad (3.4)$$

where

$$n_1 = \sum_1^D n_{1d}\ ,\ n_2 = n - n_1$$

$$m_2 = \sum_1^D m_{2d} \text{ and } \bar{y}_1 = \frac{1}{n_1}\sum_i^{n_1} y_{1i}$$

and $$\bar{y}_{m_2} = \frac{1}{m_2}\sum_i^{m_2} y_{2i}$$

Here y_{1i} (i = 1, 2, ..., n_1) and y_{2i} (i = 1, 2, ..., m_2) represent the observed values of y from the responses of units obtained using SRSWOR in phase I and II, respectively. Note that it ignores the partition of the population and hence the sample distribution by domain. Next, the variance of $\hat{t}_0$ is given by

$$V(\hat{t}_0) = \frac{N^2}{n}\left[\left(1-\frac{n}{N}\right)+(1-\theta)\left(\frac{1}{f_2}-1\right)\right]S^2 \quad (3.5)$$

where, $\theta = \sum_d^D P_d\theta_d$ and S^2 is the population variance. Comparing it with the variance given in Eq (3.1) and ignoring the first term which is common in both, we have $V(\hat{t}_0) > V(\hat{t})$ unless the domain partition of the population is completely random, in which case the two variances are equal.

Another alternative estimator is to utilize the one proposed by Agrawal and Midha (2007) as given by

$$\hat{t}^* = \sum_1^D \hat{t}_d^*$$

$$\hat{t}_d^* = \frac{N}{n} n_d \bar{y}_d \quad (3.6)$$

where

$$\bar{y}_d = \frac{1}{m_d}\sum_i^{m_d} y_{id} \quad (3.7)$$

Here in Eq (3.7), m_d denotes the number of sample units falling in domain d from the phase II sampling of $m' = n(m_2/n_2)$ out of n phase I sample units, drawn using SRSWOR. Note that $m' \geq m_2$, and so a larger sample may be required in phase II under this approach, resulting an increase in the cost of sampling. It assumes that n_d will be known from phase I sampling; however it does not exploit the known identity of phase I samples in selecting phase II samples by domain as well as the use of item responses that might be made available for some of the phase I sampled units, as for example in telephone surveys. Clearly, the estimator in (3.6) is unbiased; however, we demonstrate later using simulations that for the population total estimators, $V(\hat{t}^*) > V(\hat{t})$.

4. OPTIMAL SAMPLE DESIGN

Optimality for phase I and phase II sampling was investigated considering several populations with different domain configurations as listed in Table 1. The sample size n and the sampling fraction f_{2d} (d = 1, 2, 3) were found mainly influenced by the item response rate, the cost of phase I sampling relative to phase II sampling and the variability for the domains. Considering the cases of low, medium and high response probability θ_d, and a low value for the cost ratio, c_{1d}/c_{2d}, and a low to high coefficient of variation, CV_d, d = 1, 2, 3, the following conclusions were drawn: n_{opt} and $f_{2d,opt}$ (d =1, 2, 3) increase as θ_d (d = 1, 2, 3) decreases; $f_{2d,opt}$ increases as c_{1d}/c_{2d}

Table 1. Parametric values and optimal phase II sampling fractions for domains

Case 1						
d	N_d	$\bar{Y}_d$	S_d	θ_d	c_{1d}/c_{2d}	Sampling fraction, f_{2d}
1	2500	10	5	0.2	0.1	0.125
2	5000	20	15	0.5	0.1	0.268
3	2500	30	10	0.8	0.1	0.163
Case 2						
1	1500	10	5	0.1	0.1	0.169
2	3000	20	15	0.3	0.2	0.360
3	5500	30	10	0.6	0.3	0.254
Case 3						
1	4000	10	5	0.4	0.1	0.189
2	3000	20	15	0.6	0.2	0.329
3	3000	30	10	0.8	0.3	0.206

(d = 1, 2, 3) increases and as CV_d (d = 1, 2, 3) increases; the increase in $f_{2d,opt}$ is approximately proportional to CV_d (d = 1, 2, 3) and thus CV_d could substantially influence $f_{2d,opt}$ than perhaps the cost ratio c_{1d}/c_{2d} (d = 1, 2, 3).

The optimal values of the sampling fraction f_{2d} computed for the domains using Eq. (2.5) are listed in the last column of Table 1 that describes the three population cases considered. For the population, we have N = 10,000, $\bar{Y}$ = 20 and S = 13.32. The overall probability of response θ= 0.5, 0.435, 0.58 for the three cases considered. Letting the desired precision with CV = 0.05 in estimation of population total, the corresponding optimal phase I sample sizes computed from Eq (3.2) are given as n_{opt} = 409, 240, 311.

5. NUMERICAL EVALUATION USING SIMULATIONS

A simulation study was made to evaluate $\hat{t}_d$ (d = 1, 2, 3) and $\hat{t}$ for the three population cases considered above in Section 4. The simulations were performed using R package. For the three domains, item values for the units were simulated with $\bar{Y}_d$ and S_d (d = 1, 2, 3) as in Table 1. The normal distribution was used in generating the values for each domain. The domain means and standard deviations were chosen to have a substantial overlap between the middle domain and each of the remaining two. This is to reflect the most likely situation when the domains are considered in an ascending order of their means.

In each case of phase I sample size n = 409, 240, 311, sample units were drawn using SRSWOR. It was then followed by stratified sampling at phase II as described using the sample sizes, m_{2d}, d = 1, 2, 3, resulting from the sampling fractions given in Table 1 and the phase I non-response samples that were in domain d (d = 1, 2, 3). The estimates, $\hat{t}_d$ (d = 1, 2, 3) and $\hat{t}$ were computed for each simulation. The sampling and estimation process was repeated 2000 times. The empirical average, variance and root mean square error (RMSE) were determined for $\hat{t}_d$ and $\hat{t}$ from the 2000 estimates.

To compare the proposed estimator $\hat{t}$ to $\hat{t}_0$, the total phase II sample of size $m_2 = \sum_d m_{2d}$, was drawn using SRSWOR for computation of $\hat{t}_0$. Similarly, $\hat{t}^*$ was computed using the comparable phase II sample size of $m' = m_2\left(\frac{n}{n_2}\right)$ needed for the implementation of the Agrawal and Midha's (2007) SRSWOR sampling scheme at phase II. Agrawal and Midha (2007) provide estimates primarily at the domain level denoted by $\hat{t}_d^*$. The coefficient of variation was computed for each estimator from dividing the observed RMSE by the actual value of the corresponding total. The numerical results obtained are presented in Table 2 for the three cases discussed earlier.

These results show that the proposed estimator has much lower CV than the other two alternative estimators of the population total. Note that the CV for the proposed estimator is in close agreement with the specified CV= 0.05. For the domains, it consistently has smaller CV and hence performs better than the domain estimator of Agrawal and Midha (2007). One exception that occurs is for Domain 1 in Case 2. The reason for it is that this domain is relatively small and also has a much lower variability for the item value than the other two domains and hence, it gets allocated much fewer samples in phase II under the optimal subsampling than the sampling considered by Agrawal and Midha. Moreover, the overall sampling cost is expected to be higher under their approach, particularly if the item response in phase I sampling is moderate to high or the cost of sampling in phase II relative to that in phase I is high.

Table 2. CVs for Domain and Population Total Estimators

Case 1					
Domain	$CV(\hat{t}_d)$	$CV(\hat{t}_d^*)$	$CV(\hat{t})$	$CV(\hat{t}^*)$	$CV(\hat{t}_0)$
1	0.1366	0.1637	0.0489	0.1150	0.3427
2	0.0940	0.1147			
3	0.1001	0.1787			
Case 2					
1	0.2773	0.1726	0.0545	0.1119	0.2083
2	0.0877	0.1239			
3	0.1229	0.1864			
Case 3					
1	0.1565	0.1810	0.0532	0.1238	0.3673
2	0.1045	0.1211			
3	0.1097	0.1985			

ACKNOWLEDGEMENTS

The research work of Dr. Raj S. Chhikara was conducted during his visit to IASRI, New Delhi, while he was on Faculty Development Leave from the University of Houston-Clear Lake and was partially supported by research funding from the National Agricultural Statistics Service of the U.S. Department of Agriculture, Washington, D.C.

The authors are grateful to the referee for constructive suggestions which led to improvement in the paper.

REFERENCES

Agrawal, M.C. and Midha, C.K. (2007). Some efficient estimators of the domain parameters. *Statist. Probab. Lett*., **77**, 704-709.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd ed. Wiley, New York.

Degraft-Johnson, K.T. (1973). Estimation of domain means using two-phase sampling. *Biometrika*, **60(2)**, 387-393.

Kun He. (1995). On estimating domain totals over a subpopulation. *Ann. Inst. Statist. Math.*, **47(4)**, 637-643.

Rao, P.S.R.S. (2000). *Sampling Methodologies with Applications*. Chapman & Hall/CRC, New York.

Sarndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *Internat. Statist. Rev*., **55**, 279-294.

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

Sud, U.C., Chandra, Hukum and Chhikara, Raj S. (2009). Domain estimation in the presence of non-response. Paper submitted to *J. Ind. Soc. Agril. Statist*. for publication.