Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 67(3) 2013 329-337

Calibration Estimator of Population Total with Sub-sampling of Non-respondents

Rohan Kumar Raman, U.C. Sud, Hukum Chandra and V.K. Gupta
Indian Agricultural Statistics Research Institute, New Delhi

Received 06 March 2013; Revised 06 May 2013; Accepted 29 May 2013

SUMMARY

Using the calibration approach, the Hansen and Hurwitz (1946) technique based estimator is developed for the situation where the information on auxiliary variable is assumed known for the entire sampled units. Expressions for the estimator of population total, its variance and variance estimator are developed. The theoretical results are illustrated with the help of simulation studies. Simulation results show that proposed calibration approach based estimator outperforms the Hansen and Hurwitz estimator.

Keywords: Calibration approach, Hansen and Hurwitz estimator, Non-response, Population total, Sub-sampling of non-respondents

1. INTRODUCTION

In many human surveys, it generally is not possible to obtain information from all the units in the surveyed population. The problem of non-response persists even after call backs. The estimates obtained from incomplete data may be biased particularly when the respondents differ from the non-respondents. To address the problem of bias, Hansen and Hurwitz (1946) proposed a technique essentially to adjust for non-response. The technique consists of selecting a sample from the population, identifying the non-respondents in the sample and selecting a sub-sample of non-respondents. Through specialized efforts data are collected from the sub-sampled non-respondents so as to obtain an estimate of non-responding units in the population**.** The data from the initial respondents and the respondents in the sub-sample is combined to produce an unbiased estimator of the population mean/ total. Foradori (1961) proposed the technique of sub-sampling of the non-respondents to estimate the population total in the context of two stage sampling design with unequal probability sampling at the first stage. Srinath (1971) used a different procedure for selecting the sub-sample of respondents where the sub-sampling procedure varied according to the non-response rates. Oh and Scheuren (1983) attempted to compensate for non-response 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 non-response.

It may be mentioned that the weighting and imputation procedures aim at elimination of 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: Rohan Kumar Raman
E-mail address: rohan@iasri.res.in

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. This technique is very effective when the bias caused by non-response is serious, *i.e.* it is possible to obtain unbiased estimators by observing only a sub-sample of non-respondents. In what follows, an estimator of population total has been proposed using the Hansen and Hurwitz (1946) technique through the calibration approach, described in Deville and Särndal (1992) and Särndal (2007), when information on a related variable is assumed known for the units selected in the sample. Different situations are considered. Besides, expressions for the variance and estimator of variance are also developed. The developed theory is illustrated with the help of a simulation study in section 3. Finally section 4 presents concluding remarks.

2. THEORETICAL DEVELOPMENTS

Let there be a finite population $U = (1, 2, ..., N)$ of N units. Let the population size of the responding stratum be N_1 and that of non-responding stratum be N_2 such that $N_1 + N_2 = N$. Let the study and the auxiliary variable be denoted by 'y' and 'x' respectively. The objective is to estimate the population total $T = \sum_U y_k$ on the basis of a probability sample selected from the population. We shall let π_i and π_{ij} denote the first order and second order inclusion probabilities for the unit i and pair of units i, j, respectively being included in a sample of size n drawn without replacement from a population U using a sampling design $p(.)$. Let us consider the following situations:

(a) In the first phase, a sample s_a of size n_a is drawn from the population of N units according to the design $p_a(.)$ with positive first order and second order inclusion probabilities, respectively as π_{ak} and π_{akl}, $k \neq l \in U$. Let $\Delta_{akl} = \pi_{akl} - \pi_{ak}\,\pi_{al}$, $k \neq l \in U$.

(b) Despite efforts to obtain responses y_k from all elements in s_a, some non-response occurs. However, the information on the auxiliary variable is available for all elements in s_a.

(c) We assume that the response is stochastic. In other words, there exists a response distribution (RD) that governs the dichotomization of the sample s_a into a responding subset s_{a1} of size n_{a1}, and the other non responding subset s_{a2}, of size n_{a2}. Thus, if a given s_a were surveyed repeatedly, the composition of the subsets would vary from one survey to the next.

(d) A sub-sample s_2 of size n_2 is drawn from s_{a2}, by a design $p(.|s_{a2})$ with positive first order and second order inclusion probabilities denoted by $\pi_{k/s_{a2}}$ and $\pi_{kl/s_{a2}}$, $k \neq l \in s_{a2}$.

Let $\Delta_{kl/s_{a2}} = \pi_{kl/s_{a2}} - \pi_{k/s_{a2}}\pi_{l/s_{a2}}$, $k \neq l \in s_{a2}$. The required efforts are made to record a response from every element in $s_2 \subseteq s_{a2}$. The set for which y is observed is denoted by $s = s_{a1} \cup s_2$. Hansen and Hurwitz (hereafter HH) estimator for population total T (Hansen and Hurwitz, 1946) is given by

$$\hat{T}_\pi = \sum_s y_k / \pi_k^*, \text{ with } \pi_k^* = \begin{cases} \pi_{ak} & \text{if } k \in s_{a1} \\ \pi_{ak}\pi_{k/s_{a2}} & \text{if } k \in s_2 \end{cases}.$$

We now consider two cases: (i) the stratum sizes of the population are unknown and (ii) the stratum sizes of the population are known. Then define the estimator for population total T for these two cases.

Case 1.

We consider that the size of the responding stratum is N_1 and that of non-responding stratum is N_2 such that $N_1 + N_2 = N$. We further assume that both N_1 and N_2 are unknown. In this case the HH estimator for the population total T is given by

$$\hat{T}_{\pi 1} = \sum_{s_{a1}} \hat{y}_k + \sum_{s_2} \hat{\hat{y}}_k, \tag{2.1}$$

where $\hat{y}_k = y_k / \pi_{ak}$ and $\hat{\hat{y}}_k = \dfrac{y_k}{\pi_{ak}\pi_{k/s_{a2}}}$. Note that $E_1\{E_{RD}(\hat{T}_{\pi 1} / s_a, s_{a2})\} = T$. Here $E_{RD}(./s_a)$ refers to expectation with respect to (unknown) response distribution, given s_a and E_1 refers to expectation of all possible samples of size n_a from N. This shows that the estimator $\hat{T}_{\pi 1}$ is unbiased estimator of T.

Further, under SRSWOR the estimator $\hat{T}_{\pi 1}$ reduces to

$$\hat{T}_{\pi 1} = N(w_{a1}\bar{y}_{n_{a1}} + w_{a2}\bar{y}_{n2}) \tag{2.2}$$

where

$$\bar{y}_{n_{a1}} = \frac{1}{n_{a1}} \sum_{k=1}^{n_{a1}} y_k, \bar{y}_{n_2} = \frac{1}{n_2} \sum_{k=1}^{n_2} y_k, w_{a1} = \frac{n_{a1}}{n_a}, w_{a2} = \frac{n_{a2}}{n_a}.$$

The variance of estimator for the population total $\hat{T}_{\pi 1}$ in (2.1) is given by

$$V(\hat{T}_{\pi 1}) = \sum_{k=1}^{N_1} \sum_{l=1}^{N_1} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{k=1}^{N_2} \sum_{l=1}^{N_2} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + 2 \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E_1 E_{RD} \left\{ \sum_{k=1}^{s_{a2}} \sum_{l=1}^{s_{a2}} \Delta_{kl/s_{a2}} \frac{y_k}{\pi_{ak} \pi_{k/s_{a2}}} \frac{y_l}{\pi_{al} \pi_{l/s_{a2}}} \right\} \quad (2.3)$$

Under SRSWOR the variance expression (2.3) reduce to

$$V(\hat{T}_{\pi 1}) = N(f_a - 1)S^2 + [f_a N_2 (f_2 - 1) S_2^2], \quad (2.4)$$

where

$$f_a = \frac{N}{n_a}, f_2 = \frac{n_{a2}}{n_2}, S^2 = \frac{1}{(N-1)} \sum_{k=1}^{N} (y_k - \bar{Y}_N)^2$$

and $S_2^2 = \frac{1}{(N_2 - 1)} \sum_{k=1}^{N_2} (y_k - \bar{Y}_{N_2})^2$.

An unbiased variance estimator of $\hat{T}_{\pi 1}$ in (2.3) is given as

$$\hat{V}(\hat{T}_{\pi 1}) = \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a1}} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{k=1}^{s_{a2}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + 2 \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \left\{ \sum_{k=1}^{s_2} \sum_{l=1}^{s_2} \frac{\Delta_{kl/s_{a2}}}{\pi_{kl/s_{a2}}} \frac{y_k}{\pi_{ak} \pi_{k/sa2}} \frac{y_l}{\pi_{al} \pi_{l/sa2}} \right\} \quad (2.5)$$

$$\pi^*_{kl} = \begin{cases} \pi_{akl} \pi_{kl/s_{a2}} & if\ k,l \in s_{a2} \\ \pi_{akl} \pi_{k/s_{a2}} & if\ k \in s_{a2}, l \in s_{a1} \\ \pi_{akl} \pi_{l/s_{a2}} & if\ k \in s_{a1}, l \in s_{a2} \\ \pi_{akl} & if\ k,l \in s_{a1} \end{cases}$$

The corresponding estimator of variance under SRSWOR is

$$\hat{V}(\hat{T}_{\pi 1}) = \frac{N(N-1)n_a}{(n_a - 1)} \left[\frac{(N - n_a)}{n_a (N-1)} (\bar{G}_w - \bar{y}_w^2) + (f_2 - 1) \frac{w_{a2}}{n_a} s_{n2}^2 \right], \quad (2.6)$$

where

$$\bar{G}_w = \frac{1}{n_a} \left(\sum_{k=1}^{n_{a1}} y_k^2 + \frac{n_{a2}}{n_2} \sum_{k=1}^{n_2} y_k^2, \right), \bar{y}_w = \frac{1}{n_a} (n_{a1} \bar{y}_{n_{a1}} + n_{a2} \bar{y}_{n_2})$$

and $s_{n_2}^2 = \frac{1}{(n_2 - 1)} \left(\sum_{k=1}^{n_2} y_k^2 - n_2 \bar{y}_{n_2}^2 \right)$.

Since non-response is assumed absent for the auxiliary variable, the variable x is known for both the subsets s_{a1} and s_{a2}. Let us define $d_{ak} = \frac{1}{\pi_{ak}}$, $d_{k/s_{a2}} = \frac{1}{\pi_{k/s_{a2}}}$ and $\hat{X} s_{a2} = \sum_{s_{a2}} d_{ak} x_k$. The calibration approach is used to modify the original weights, $d_{ak} d_{k/s_{a2}}$, by minimizing the chisquare type distance, $\sum_{k=1}^{s_2} \frac{(w_k - d_{ak} d_{k/s_{a2}})^2}{d_{ak} d_{k/s_{a2}} q_k}$ with respect to w_k and subject to the restriction $\sum_{s_2} w_k x_k = \hat{X} s_{a2}$, where q_k are suitably chosen known weights. In the literature, three commonly used values of q_k are 1, x_k and $\frac{1}{x_k}$. Depending upon the situation (or type of estimator) we choose different values. For example, calibrated estimator reduces to regression and ratio estimator if $q_k = 1$ and $q_k = \frac{1}{x_k}$ respectively.

To obtain the revised weights, we consider the Lagrangian multiplier technique and use the following function

$$\varphi = \sum_{s_2} \frac{(w_k - d_{ak} d_{k/s_{a2}})^2}{d_{ak} d_{k/s_{a2}} q_k} - \lambda_2 \left(\sum_{s_2} w_k x_k - \hat{X}_{s_{a2}} \right). \quad (2.7)$$

Minimization of (2.7) gives following weights as

$$w_k = d_{ak} d_{k/s_{a2}} + \left(\hat{X}_{s_{a2}} - \sum_{s_2} x_k d_{ak} d_{k/s_{a2}} \right) \times \frac{1}{\sum_{s_2} d_{ak} d_{k/s_{a2}} q_k x_k^2} d_{ak} d_{k/s_{a2}} q_k x_k. \quad (2.8)$$

Using the new set of weights given in (2.8), the proposed calibrated estimator of population total T is given by

$$\hat{T}_{cal1} = \sum_{s_{a1}} d_{ak} y_k + \sum_{s_2} w_k y_k. \quad (2.9)$$

Taking $q_k = 1/x_k$ and substituting the value of w_k in (2.9), we get

$$\hat{T}_{cal1} = \sum_{s_{a1}} y_k d_{ak} + \frac{\sum_{s_2} y_k d_{ak} d_{k/s_{a2}}}{\sum_{s_2} x_k d_{ak} d_{k/s_{a2}}} \hat{X}_{s_{a2}}. \quad (2.10)$$

Note that the estimator $\hat{T}_{cal1}$ is biased.

Under SRSWOR, the estimator $\hat{T}_{cal1}$ reduces to

$$\hat{T}_{cal1} = N \left(w_{a1} \bar{y}_{n_{a1}} + w_{a2} \frac{\bar{y}_{n_2}}{\bar{x}_{n_2}} \bar{x}_{n_{a2}} \right) \quad (2.11)$$

where

$$\bar{x}_{n_{a2}} = \frac{1}{n_{a2}} \sum_{k=1}^{n_{a2}} x_k, \bar{x}_{n_2} = \frac{1}{n_2} \sum_{k=1}^{n_2} x_k.$$

The variance of $\hat{T}_{cal1}$ to the first order of approximation is given by

$$V(\hat{T}_{cal1}) = \sum_{k=1}^{N_1} \sum_{l=1}^{N_1} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + 2 \sum_{k=1}^{N_2} \sum_{l=1}^{N_2} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + 2 \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} \Delta_{akl} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + E_1 E_{RD} \left\{ \sum_{k=1}^{s_{a2}} \sum_{l=1}^{s_{a2}} \Delta_{kl/s_{a2}} \frac{\in_{2k}}{\pi_{ak} \pi_{k/sa2}} \frac{\in_{2l}}{\pi_{al} \pi_{l/sa2}} \right\}, \quad (2.12)$$

where

$$\in_{2k} = y_k - \hat{\beta}_{s_{a2}} x_k \text{ and } \hat{\beta}_{s_{a2}} = \frac{\sum_{s_{a2}} d_{ak} y_k}{\sum_{s_{a2}} d_{ak} x_k}.$$

Under SRSWOR the variance expression reduces to

$$V(\hat{T}_{cal1}) = (f_a - 1) \frac{N}{(N-1)} \left[\sum_{k=1}^{N_1} y_k^2 + \sum_{k=1}^{N_2} y_k^2 - N \bar{Y}_N^2 \right] + E_1 E_{RD} \left\{ (f_2 - 1) \frac{n_{a2}}{(n_{a2} - 1)} \sum_{k=1}^{n_{a2}} \left(y_k - \frac{\bar{y}_{n_{a2}}}{\bar{x}_{n_{a2}}} x_k \right)^2 \right\}, \quad (2.13)$$

where

$$\bar{y}_{n_{a2}} = \frac{1}{n_{a2}} \sum_{k=1}^{n_{a2}} y_k \text{ and } \bar{Y}_N = \frac{1}{N} \sum_{k=1}^{N} Y_k.$$

Following Särndal (1992) the estimator of variance of $\hat{T}_{cal1}$ is

$$\hat{V}(\hat{T}_{cal1}) = \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a1}} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_{k=1}^{s_{a2}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + 2 \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \left(\frac{\hat{X}_{s_{a2}}}{\hat{\hat{x}}_{s_2}} \right)^4 \left\{ \sum_{k=1}^{s_2} \sum_{l=1}^{s_2} \frac{\Delta_{kl/s_{a2}}}{\pi_{kl/s_{a2}}} \frac{e_{2k}}{\pi_{ak} \pi_{kl/s_{a2}}} \frac{e_{2l}}{\pi_{al} \pi_{l/s_{a2}}} \right\}, \quad (2.14)$$

where

$$e_{2k} = y_k - \hat{\hat{\beta}}_{s_2} x_k, \hat{\hat{\beta}}_{s_2} = \frac{\sum_{k=1}^{s_2} d_{ak} d_{k/s_{a2}} y_k}{\sum_{k=1}^{s_2} d_{ak} d_{k/s_{a2}} x_k},$$

$$\sum_{s_2} d_{ak} d_{k/s_{a2}} x_k = \hat{\hat{x}}_{s_2}, \text{ and}$$

$$\pi_{kl}^{*}=\begin{cases}\pi_{akl}\pi_{kl/s_{a2}} & if\ \ k,l\in s_{a2}\\ \pi_{akl}\pi_{k/s_{a2}} & if\ \ k,\in s_{a2},l\in s_{a1}\\ \pi_{akl}\pi_{l/s_{a2}} & if\ \ k,\in s_{a1},l\in s_{a2}\\ \pi_{akl} & if\ \ k,l\in s_{al}\end{cases}$$

Under SRSWOR the variance estimator reduce to

$$\hat{V}(\hat{T}_{cal1})=\frac{(f_a-1)f_a}{(n_a-1)}\left\{n_a\sum_{k=1}^{n_{a1}}y_k^2-\left(\sum_{k=1}^{n_{a1}}y_k\right)^2\right\}$$

$$+\frac{f_2(f_a-1)f_a}{(n_a-1)}\left\{\left(n_a+\frac{(n_{a2}-n_2)}{(n_2-1)}\right)\sum_{k=1}^{n_{a2}}y_k^2\right.$$

$$\left.-(n_{a2}-1)\left(\sum_{K=1}^{n_{a2}}y_k\right)^2(n_2-1)^{-1}-2\sum_{K=1}^{n_{a1}}y_k\sum_{K=1}^{n_{a2}}y_l\right\}$$

$$+(\hat{X}_{n_{a2}})^4(\hat{x}_{n_2})^{-4}f_a^2(f_2-1)\frac{n_{a2}}{(n_2-1)}$$

$$\times\sum_{k=1}^{n_2}\left(y_k-\frac{\overline{y}_{n2}}{\overline{x}_{n2}}x_k\right)^2, \qquad (2.15)$$

with $\hat{X}_{n_{a2}}=\frac{N}{n_a}\sum_{k=1}^{n_{a2}}x_k,\hat{x}_{n_2}=\frac{N}{n_a}\frac{n_{a2}}{n_2}\sum_{k=1}^{n_2}x_k.$

Case 2.

We consider that size of the responding stratum is N_1 and that of non-responding stratum is N_2 such that $N_1 + N_2 = N$. Here we assume that both N_1 and N_2 are known. In this case the estimator of population total T is given by

$$\hat{T}_{\pi 2}=N_1\left(\sum_{k=1}^{s_{a1}}\frac{1}{\pi_{ak}}\right)^{-1}\sum_{k=1}^{s_{a1}}\hat{y}_k+N_2\left(\sum_{k=1}^{s_2}\frac{1}{\pi_{ak}\pi_{k/s_{a2}}}\right)^{-1}\sum_{k=1}^{s_2}\hat{\hat{y}}_k. \qquad (2.16)$$

Note that the estimator $\hat{T}_{\pi 2}$ is like häjek type estimator and therefore it is biased.

Under SRSWOR design the estimator $\hat{T}_{\pi 2}$ reduces to

$$\hat{T}_{\pi 2}=\frac{N_1}{n_{a1}}\sum_{k=1}^{n_{a1}}y_k+\frac{N_2}{n_2}\sum_{k=1}^{n_2}y_k. \qquad (2.17)$$

The variance of $V(\hat{T}_{\pi 2})$ is

$$V(\hat{T}_{\pi 2})=\sum_{k=1}^{N_1}\sum_{l=1}^{N_1}\Delta_{akl}\frac{(y_k-\overline{y}_{N_1})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_1})}{\pi_{al}}$$

$$+\sum_{k=1}^{N_2}\sum_{l=1}^{N_2}\Delta_{akl}\frac{(y_k-\overline{y}_{N_2})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_2})}{\pi_{al}}$$

$$+2\sum_{k=1}^{N_1}\sum_{l=1}^{N_2}\Delta_{akl}\frac{(y_k-\overline{y}_{N_1})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_2})}{\pi_{al}}$$

$$+E_1E_{RD}\left[N_2^2\left(\sum_{k=1}^{s_{a2}}\frac{1}{\pi_{ak}}\right)^{-2}\sum_{k=1}^{s_{a2}}\sum_{l=1}^{s_{a2}}\Delta_{kl/s_{a2}}c_{kl}\right], \qquad (2.18)$$

where

$$c_{kl}=\frac{\left(y_k-\left(\sum_{k=1}^{s_{a2}}\frac{1}{\pi_{ak}}\right)^{-1}\sum_{k=1}^{s_{a2}}\frac{y_k}{\pi_{ak}}\right)}{\pi_{ak}\pi_{k/s_{a2}}}\frac{\left(y_l-\left(\sum_{l=1}^{s_{a2}}\frac{1}{\pi_{al}}\right)^{-1}\sum_{l=1}^{s_{a2}}\frac{y_l}{\pi_{al}}\right)}{\pi_{al}\pi_{l/s_{a2}}},$$

$\overline{y}_{N_1}=\frac{1}{N_1}\sum_{k=1}^{N_1}y_k$ and $\overline{y}_{N_2}\frac{1}{N_2}\sum_{k=1}^{N_2}y_k.$

Under SRSWOR the variance expression reduce to

$$V(\hat{T}_{\pi 2})=\frac{(N-n_a)}{(N-1)}f_a[(N_1-1)S_1^2+(N_2-1)S_2^2]$$

$$+E_1E_{RD}\left[\frac{N_2^2}{n_{a2}}(f_2-1)s_2^2\right]. \qquad (2.19)$$

where

$$S_1^2=\frac{1}{(N_1-1)}\sum_{k=1}^{N_1}(y_k-\overline{Y}_{N_1})^2,S_2^2=\frac{1}{(N_2-1)}\sum_{k=1}^{N_2}(y_k-\overline{Y}_{N_2})^2,$$

$$\overline{y}_{n_{a2}}=\frac{1}{n_{a2}}\sum_{k=1}^{n_{a2}}y_k,s_2^2=\frac{1}{(n_{a2}-1)}\sum_{k=1}^{n_{a2}}(y_k-\overline{Y}_{n_{a2}}).$$

An variance estimator of $\hat{T}_{\pi 2}$ is given as

$$\hat{V}(\hat{T}_{\pi 2})=\sum_{k=1}^{s_{a1}}\sum_{l=1}^{s_{a1}}\frac{\Delta_{akl}}{\pi_{akl}^{*}}\frac{(y_k-\overline{y}_{s_{a1}})}{\pi_{ak}}\frac{(y_l-\overline{y}_{s_{a1}})}{\pi_{al}}$$

$$+\sum_{k=1}^{s_{a2}}\sum_{l=1}^{s_{a2}}\frac{\Delta_{akl}}{\pi^{*}_{kl}}\frac{(y_k-\overline{y}_{s_{a2}})}{\pi_{ak}}\frac{(y_k-\overline{y}_{s_{a2}})}{\pi_{al}}$$

$$+2\sum_{k=1}^{s_{a1}}\sum_{l=1}^{s_{a2}}\frac{\Delta_{akl}}{\pi^{*}_{kl}}\frac{(y_k-\overline{y}_{s_{a1}})}{\pi_{ak}}\frac{(y_l-\overline{y}_{s_{a2}})}{\pi_{al}}+D. \quad (2.20)$$

Here

$$D=N_2^2\left(\sum_{k=1}^{s_{a2}}\frac{1}{\pi_{ak}}\right)^{-2}\sum_{k=1}^{s_2}\sum_{l=1}^{s_2}\frac{\Delta_{kl/s_{a2}}}{\pi_{kl/s_{a2}}}\frac{\left(y_k-\dfrac{\sum_{k=1}^{s_2}\frac{y_k}{\pi_{ak}\pi_{k/s_{a2}}}}{\sum_{k=1}^{s_2}\frac{1}{\pi_{ak}\pi_{k/s_{a2}}}}\right)}{\pi_{ak}\pi_{k/s_{a2}}}$$

$$\times\frac{\left(y_l-\dfrac{\sum_{l=1}^{s_2}\frac{y_l}{\pi_{al}\pi_{l/s_{a2}}}}{\sum_{k=1}^{s_2}\frac{1}{\pi_{al}\pi_{l/s_{a2}}}}\right)}{\pi_{al}\pi_{l/s_{a2}}},$$

$$\overline{y}_{s_{a1}}=\frac{1}{n_{a1}}\sum_{k=1}^{n_{a1}}y_k \text{ and } \overline{y}_{s_{a2}}=\frac{1}{n_{a2}}\sum_{k=1}^{n_{a2}}y_k.$$

The corresponding estimator of variance under SRSWOR is given by

$$\hat{V}(\hat{T}_{\pi_2}) = f_a\left\{\frac{(N_a-n_a)}{(n_a-1)}\sum_{k=1}^{n_{a1}}(y_k-\overline{y}_{n_{a1}})^2\right.$$

$$+\frac{(f_a-1)f_2}{(n_a-1)}\left\{n_a+\frac{(n_{a2}-n_2)}{(n_2-1)}\sum_{k=1}^{n_{a2}}(y_k-\overline{y}n_{a2})^2\right\}$$

$$\left.+\frac{(n_{a2}-n_2)N_2^2}{n_2n_{a2}}s^2_{n_2}\right. \quad (2.21)$$

The proposed calibrated estimator of population total T for known population size is

$$\hat{T}_{cal2}= N_1\left(\sum_{k=1}^{s_{a1}}d_{ak}\right)^{-1}\sum_{k=1}^{s_{a1}}y_kd_{ak}$$

$$+N_2\left(\sum_{k=1}^{s_2}\frac{1}{d_{ak}d_{k/s_{a2}}}\right)^{-1}\frac{\sum_{k=1}^{s_2}y_kd_{ak}d_{k/s_{a2}}}{\sum_{k=1}^{s_2}x_kd_{ak}d_{k/s_{a2}}}\hat{X}s_{a2}. \quad (2.22)$$

Note that the estimator $\hat{T}_{cal2}$ is biased. The variance of $\hat{T}_{cal2}$ to the first order of approximation is

$$V(\hat{T}_{cal2}) = \sum_{k=1}^{N_1}\sum_{l=1}^{N_1}\Delta_{akl}\frac{(y_k-\overline{y}_{N_1})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_1})}{\pi_{al}}$$

$$+\sum_{k=1}^{N_2}\sum_{l=1}^{N_2}\Delta_{akl}\frac{(y_k-\overline{y}_{N_2})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_2})}{\pi_{al}}$$

$$+2\sum_{k=1}^{N_1}\sum_{l=1}^{N_2}\Delta_{akl}\frac{(y_k-\overline{y}_{N_1})}{\pi_{ak}}\frac{(y_l-\overline{y}_{N_2})}{\pi_{al}}$$

$$+E_1E_{RD}\left\{N_2^2\left(\sum_{k=1}^{s_{a2}}\frac{1}{\pi_{ak}}\right)^{-2}\right.$$

$$\left.\times\sum_{k=1}^{s_{a2}}\sum_{l=1}^{s_{a2}}\Delta_{kl/s_{a2}}\frac{E_{3k}}{\pi_{ak}\pi_{k/s_{a2}}}\frac{E_{3l}}{\pi_{al}\pi_{l/s_{a2}}}\right\}. \quad (2.23)$$

Here

$$E_{3k} = y_k-\frac{\sum_{k=1}^{s_{a2}}y_kd_{ak}}{\sum_{k=1}^{s_{a2}}d_{ak}}-\frac{\sum_{k=1}^{s_{a2}}y_kd_{ak}}{\sum_{k=1}^{s_{a2}}x_kd_{ak}}x_k.$$

Under SRSWOR the variance expression reduces to

$$V(\hat{T}_{cal2}) = \frac{(N-n_a)}{(N-1)}f_a[(N_1-1)S_1^2+(N_2-1)S_2^2]$$

$$+E_1E_{RD}\left[\frac{(f_2-1)N_2^2}{n_{a2}^2(n_{a2}-1)}\left\{n_{a2}\sum_{k=1}^{n_{a2}}E_{3k}^2-\left(\sum_{k=1}^{n_{a2}}E_{3k}\right)^2\right\}\right] \quad (2.24)$$

where

$$E_{3k} = y_k - \frac{\overline{y}_{n_{a2}}}{\overline{x}_{n_{a2}}} - \frac{\overline{y}_{n_{a2}}}{\overline{x}_{n_{a2}}} x_k, S_1^2 = \frac{1}{(N_1 - 1)} \sum_{k=1}^{N_1} (y_k - \overline{Y}_{N_1})^2.$$

Following Särndal (1992) the estimator of variance of $\hat{T}_{cal2}$ is given as

$$\hat{V}(\hat{T}_{cal2}) = N_1^2 \left(\sum_{k=1}^{s_{a1}} d_{ak} \right)^{-2} \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a1}} \frac{\Delta_{akl}}{\pi^*_{akl}} \frac{(y_k - \overline{y}_{s_{a1}})}{\pi_{ak}} \frac{(y_l - \overline{y}_{s_{a1}})}{\pi_{al}}$$

$$+ N_2^2 \left(\sum_{k=1}^{s_{a2}} d_{ak} \right)^{-2} \sum_{k=1}^{s_{a2}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{(y_k - \overline{y}_{s_{a2}})}{\pi_{ak}} \frac{(y_l - \overline{y}_{s_{a2}})}{\pi_{al}}$$

$$+ 2N_1 N_2 \left(\sum_{k=1}^{s_{a1}} d_{ak} \sum_{k=1}^{s_{a2}} d_{ak} \right)^{-1} \sum_{k=1}^{s_{a1}} \sum_{l=1}^{s_{a2}} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{(y_k - \overline{y}_{s_{a1}})}{\pi_{ak}} \frac{(y_l - \overline{y}_{s_{a2}})}{\pi_{al}}$$

$$+ \left\{ N_2^2 \left(\sum_{k=1}^{s_{a2}} d_{ak} \right)^{-2} \sum_{k=1}^{s_2} \sum_{l=1}^{s_2} \frac{\Delta_{kl/s_{a2}}}{\pi_{kl/s_{a2}}} \frac{e_{3k}}{\pi_{ak}\pi_{k/s_{a2}}} \frac{e_{3l}}{\pi_{al}\pi_{l/s_{a2}}} \right\}, \quad (2.25)$$

where

$$e_{3k} = y_k - \frac{\sum_{k=1}^{s_2} y_k d_{ak} d_{k/s_{a2}}}{\sum_{k=1}^{s_2} d_{ak} d_{k/s_{a2}}} - \frac{\sum_{k=1}^{s_2} y_k d_{ak} d_{k/s_{a2}}}{\sum_{k=1}^{s_2} x_k d_{ak} d_{k/s_{a2}}} x_k,$$

$$\overline{y}_{s_{a1}} = \frac{\sum_{k=1}^{s_{a1}} y_k d_{ak/s_{a1}}}{\sum_{k=1}^{s_{a1}} d_{ak}} \text{ and } \overline{y}_{s_{a2}} = \frac{\sum_{k=1}^{s_{a2}} y_k d_{ak/s_{a2}}}{\sum_{k=1}^{s_{a2}} d_{ak}}.$$

Under SRSWOR the variance estimator reduce to

$$\hat{V}(\hat{T}_{cal2}) = \frac{N_1^2}{n_{a1}^2} \frac{(N - n_a)(n_{a1} - 1)s_1^2}{(n_a - 1) f_a}$$

$$+ \frac{N_2^2}{n_{a2}^2} \frac{(N - n_a) f_2}{(n_a - 1)} \left\{ n_a + \frac{(n_{a2} - n_2)}{(n_2 - 1)} \right\} (n_{a2} - 1)s_2^2$$

$$+ \frac{N_2^2 (f_2 - 1)}{n_{a2} n_2 (n_2 - 1)} \left\{ n_2 \sum_{k=1}^{n_2} e_{3k}^2 - \left(\sum_{k=1}^{n_2} e_{3k} \right)^2 \right\}, \quad (2.26)$$

where

$$s_1^2 = \frac{1}{(n_{a1} - 1)} \sum_{k=1}^{n_{a1}} (y_k - \overline{y}_{n_{a1}})^2, s_2^2 = \frac{1}{(n_{a2} - 1)} \sum_{k=1}^{n_{a2}} (y_k - \overline{y}_{n_{a2}})^2$$

$$e_{3k} = y_k - \frac{\overline{y}_{n_2}}{\overline{x}_{n_2}} - \frac{\overline{y}_{n_2}}{\overline{x}_{n_2}} x_k.$$

3. SIMULATION STUDY

In this section we describe the simulation studies made to compare the performance of proposed calibration based estimators $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$ (denoted by CAL) with the Hansen and Hurwitz estimator, $\hat{T}_{\pi 1}$ and $\hat{T}_{\pi 2}$ (denoted by HH). The following two criteria were used for assessing the relative performance of these two estimators:

(i) Percent relative bias (% RB) defined as,

$$\%RB(\theta) = \frac{1}{L} \left(\sum_{l=1}^{L} \frac{\theta_l - \theta}{\theta} \right) \times 100.$$

(ii) Percent relative root mean square error (RRMSE) defined as,

$$\%RRMSE(\theta) = \left(\sqrt{\frac{1}{L} \sum_{l=1}^{L} \left(\frac{\theta_l - \theta}{\theta} \right)^2} \right) \times 100.$$

Here, θ_l is the value of estimator θ of θ in the l^{th} (l =1, ..., L = 500) simulation run.

For the simulation study, population size N = 1000 was considered. The population was assumed to be divided into two parts as respondents and non-respondents of respective sizes N_1 and N_2 such that N_1 + N_2 = N. In particular, we considered N_1 = 600 and N_2 = 400. The population data for the k^{th} (k = 1, 2) part was generated from the model: $y_k = x_k\beta_k + e_k$, with $e_k \sim N(0\ \sigma^2)$. Three population data sets were generated wherein correlation between x and y was 0.83, 0.73 and 0.56. For the first population data set N_1 values were generated from a normal population *i.e.* $x_k \sim$N(40, 16), β_1 = 3 and $e_1 \sim$ N(0, 6.35) and N_2 values were generated through $x_k \sim$ N(50, 25), β_2 = 4 and $e_2 \sim$ N(0, 14.21). For the second population data set N_1 values were generated through $x_k \sim$ N(40, 16) β_1 = 2 and $e_1 \sim$ N(0,

8.35) and N_2 values were generated through x_k (50, 25), $\beta_2 = 3$ and $e_2 \sim N(0, 23.91)$. For the third population data set N_1 values were generated through $x_k \sim N(40, 9)$, $\beta_1 = 3$ and $e_1 \sim N(0, 10.62)$ and N_2 values were generated same as in case of second population data.

For each of these three population data sets samples s_a of sizes $n_a = 200$ and 150 were selected by SRSWOR design. Further for each sample set s_a we considered two sets of samples s_2 from non-responding part, that is, size $n_2 = 20$ and 30. The HH estimator and the proposed estimator were worked out using values obtained for each of the samples. In all, 500 samples each were repeatedly drawn from the three generated populations. The percent relative bias and percent relative root mean squares error were calculated. The values of the percent relative biases (% RBs) and the percent relative root mean squares error (% RRMSEs) for two different estimators and various combinations of parameter sets are reported in Table 1. Table 2 presents the percent relative gains in RRMSE by proposed calibrated estimators $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$ over HH estimator $\hat{T}_{\pi 1}$ and $\hat{T}_{\pi 2}$.

The results in Table 1 show that in terms of their relative biases performance there is not much to choose among different estimators of population total considered in our simulation studies. It is noteworthy that the estimator $\hat{T}_{\pi 1}$ is unbiased and all other estimators are only asymptotically unbiased. There is no clear cut pattern of the relative biases for different estimators with respect to variation in overall sample sizes n_a, non-respondent sample size n_2 and correlation between x and y. However, as expected, the percent relative root mean squared error of all the estimators of population total decrease as the overall sample sizes n_a or non-respondent sample size n_2 increase. Further, the proposed (both $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$) CAL estimators outperform the corresponding HH estimators in terms of efficiency. However, the relative gain in percent relative root mean squares error of the proposed (both $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$) CAL estimators over the HH estimators does not show a distinct pattern with the increase in sample sizes (either n_a or n_2 or both). For known population sizes, the percent relative gain due to proposed calibrated estimator $\hat{T}_{cal1}$ over HH estimator $\hat{T}_{\pi 1}$ generally seem to increases with the increase in correlation $\rho(y, x)$ between the response variable and the auxiliary variable (see column 3, Table 2). This is also true when we compare the calibrated

Table 1. Percent relative bias (% RB) and percent relative root mean squared error (%RRMSE) of estimators of population total of y from simulation studies.

$\rho(y, x)$	Sample size	% Relative Bias				% RRMSE			
		HH		CAL		HH		CAL	
		$\hat{T}_{\pi 1}$	$\hat{T}_{\pi 2}$	$\hat{T}_{cal1}$	$\hat{T}_{cal2}$	$\hat{T}_{\pi 1}$	$\hat{T}_{\pi 2}$	$\hat{T}_{cal1}$	$\hat{T}_{cal2}$
0.56	$s_a = 200, s_2 = 20$	0.03	0.04	0.04	0.05	2.62	2.53	2.46	2.35
	$s_a = 200, s_2 = 30$	0.07	0.11	0.12	0.16	2.18	2.05	2.12	2.00
	$s_a = 150, s_2 = 20$	0.11	0.12	0.09	0.10	2.77	2.68	2.71	2.60
	$s_a = 150, s_2 = 30$	0.09	0.08	0.07	0.07	2.31	2.16	2.24	2.09
0.73	$s_a = 200, s_2 = 20$	0.47	0.26	0.44	0.24	3.78	3.09	3.63	2.94
	$s_a = 200, s_2 = 30$	0.04	0.08	0.05	0.15	3.30	2.71	3.16	2.54
	$s_a = 150, s_2 = 20$	0.03	0.20	0.02	0.19	4.10	3.45	3.66	2.67
	$s_a = 150, s_2 = 30$	0.05	0.18	0.10	0.24	3.84	3.19	3.75	3.08
0.83	$s_a = 200, s_2 = 20$	0.05	0.03	0.04	0.07	2.55	1.98	2.42	1.77
	$s_a = 200, s_2 = 30$	0.06	0.08	0.03	0.06	2.37	1.64	2.32	1.51
	$s_a = 150, s_2 = 20$	0.02	0.03	0.01	0.04	2.75	1.94	2.52	1.74
	$s_a = 150, s_2 = 30$	0.04	0.02	0.08	0.06	2.58	1.77	2.47	1.62

Table 2. Percent relative gain in RRMSE by calibrated estimators $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$ over Hensen Hurwitz estimators $\hat{T}_{\pi 1}$ and $\hat{T}_{\pi 2}$.

$\rho(y, x)$	Sample size	Relative Gain,%	
		$\hat{T}_{cal1}$ vs $\hat{T}_{\pi 1}$	$\hat{T}_{cal2}$ vs $\hat{T}_{\pi 2}$
0.56	$s_a = 200, s = 20$	6.50	7.66
	$s_a = 200, s = 30$	2.83	2.50
	$s_a = 150, s = 20$	2.21	3.08
	$s_a = 150, s = 30$	3.12	3.35
0.73	$s_a = 200, s = 20$	4.13	5.10
	$s_a = 200, s = 30$	4.43	6.69
	$s_a = 150, s = 20$	12.02	29.21
	$s_a = 150, s = 30$	2.4	3.57
0.83	$s_a = 200, s = 20$	5.37	11.86
	$s_a = 200, s = 30$	2.15	8.61
	$s_a = 150, s = 20$	9.12	11.49
	$s_a = 150, s = 30$	4.45	9.26

estimator $\hat{T}_{cal2}$ over HH estimator $\hat{T}_{\pi 2}$ (see column 4, Table 2). In general, it may be seen from the results given in the Tables 1-2 that the proposed CAL (both $\hat{T}_{cal1}$ and $\hat{T}_{cal2}$) estimators consistently outperform the corresponding HH estimators in terms of the criterion of percent relative root mean squares error.

4. CONCLUDING REMARKS

The calibration approach can be gainfully employed in non-response related situations in sample surveys. Substantial gain in relative root mean square error of the estimator can be obtained when the study and the auxiliary variables are highly correlated. Depending on the level of availability of auxiliary information different estimators can be developed. Work on these directions shall be reported in a separate paper.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable comments and suggestions of the referee. These led to a considerable improvement in 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.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.,* **87**, 376-382.

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

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.,* **47**, 663-685.

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non response in sample surveys. *J. Amer. Statist. Assoc.,* **46**, 147-190.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology,* **12**, 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. Academic press, New York, 143-184.

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.

Okafor, F.C. (2001). Treatment of non-response in successive sampling. *Statist.,* **61(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.

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.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology,* **33(2)**, 99-119.

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

Srinath, K.P. (1971). Multiphase sampling in non-response problems. *J. Amer. Statist. Assoc.,* **66,** 583-586.

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.