

ESTIMATION OF VARIANCE FROM MISSING DATA

Raghunath Arnab and Sarjinder Singh¹

ABSTRACT

Any large scale survey may be prone to nonresponse problems. No exact formulation of the nature of nonresponse in surveys is available. So, several methods of handling nonresponse problems are proposed by survey statisticians. In this paper, the problems of estimation of population total and its variance have been studied in the presence of nonresponse. The proposed methodology is based on the assumption that the set of respondents (comprises response sample) is a Poisson sample elected by nature. The proposed method is more general than the method proposed by Särndal (1992) since one can use auxiliary information in estimating the unknown response probabilities. The proposed estimators are compared with Särndal 's (1992) estimator using similar simulation studies as prescribed by Särndal (1992). The simulation results reveal that the proposed estimators performs better than the usual estimator of variance.

KEY WORDS: Inclusion probabilities, Variance estimation, Superpopulation model, Nonresponse, Poisson sample.

RÉSUMÉ

N'importe quelle enquête à grande échelle peut être sujette aux problèmes de non réponse. Aucune formulation exacte de la nature de la non réponse pour l'étude n'est disponible. Ainsi, des statisticiens ont proposé plusieurs méthodes pour traiter ces problèmes de non réponse dans une étude. Dans cette présentation, les problèmes d'estimation de la taille de la population et de sa variance ont été étudiés en présence de non réponse. La méthodologie proposée est fondée sur l'hypothèse que l'ensemble de répondants (composé de l'échantillon ayant répondu) est un échantillon aléatoire de Poisson. La méthode proposée est plus générale que la méthode proposée par Sarndal (1992) puisqu'on peut utiliser de l'information auxiliaire en estimant les probabilités de réponse inconnues. Les estimateurs proposés sont comparés aux estimateurs semblables de Sarndal en utilisant des études de simulations similaires à celles décrites par Sarndal (1992). Les résultats de simulations indiquent que la méthode proposée performe mieux que les concurrents existants.

MOTS CLÉ : Poisson, estimation de la variance, model de superpopulation, non-réponse, probabilités d'inclusion.

1. INTRODUCTION

In almost all large scale surveys, non-responses are inevitable. Several methods for handling nonresponse problems in sample surveys are available in the literature. Good details are given by Rubin (1987). Särndal (1992) developed a method of estimation of the population total and its variance when a single imputation is used for estimating unobserved values under the superpopulation model described below. Let $U = \{1, 2, \dots, N\}$ be a finite population of N identifiable units and y_i be the value of the variate under study for the i th unit of the population U . It is assumed that the vector $\mathbf{y} = (y_1, \dots, y_i, \dots, y_N)$ is a random sample from a superpopulation ξ having the following distribution: $E_\xi(y_i) = \beta x_i$, $V_\xi(y_i) = \sigma^2 x_i^g$ and $C_\xi(y_i, y_j) = 0$ for $i \neq j$, $i, j = 1, \dots, N$, where E_ξ , V_ξ and C_ξ denote respectively expectation, variance and covariance operator with respect to the model ξ ;

$\beta, \sigma^2 (> 0)$ and $g (> 0)$ are unknown model parameters, and $x_i (> 0)$ is the value of the auxiliary variable for the i th unit. The objective is to estimate the finite population total $Y = \sum_{i \in U} y_i$ on the basis of a sample s , of size n , selected with probability $p(s)$ according to a sampling design p . Let π_i and π_{ij} be the inclusion probabilities for the i th, and i th and j th ($i \neq j$) unit of U . Let $s_r (\subset s)$ be the set of respondent units of size m from which responses y_i 's are observed and the complement $s - s_r$ (of size $n - m$) be the set of nonresponse units. Let \hat{y}_j for $j \in s - s_r$ be the imputed value of the j th unit computed according to a certain rule depending on the superpopulation model ξ (for details, see Särndal (1992)). Let $t = \sum_{i \in s} w_i y_i$ be an unbiased estimator used for estimating Y in case of 100% response (i.e. $m = n$) where w_i 's are suitably chosen weights independent of

¹ Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban-4000, South Africa and Sarjinder Singh, Department of Statistics, St. Cloud State University, St. Cloud, MN 56301-4498, USA

y_i 's. In the presence of nonresponse ($m < n$), Särndal (1992) modified the estimator t as $\hat{t} = \sum_{i \in s} w_i y_{i0}$ where

$$y_{i0} = \begin{cases} y_i & \text{for } i \in s_r \\ \hat{y}_i & \text{for } i \in s - s_r \end{cases}.$$

Denoting E_p (V_p) and E_r (V_r) as expectation (variance) with respect to the sampling design p and response mechanism respectively, Särndal (1992) derived the overall variance of \hat{t} as

$$V_{tot} = E_{\xi} V_p E_r (\hat{t} - Y)^2 \\ = E_{\xi} V_p (\hat{t}) + E_p V_r E_{\xi} (\hat{t}) + 2E_{\xi} E_p [(t - Y) \{E_{\xi}(\hat{t} - t) | s\}]$$

where $V_p(\hat{t}) = E_p(\hat{t} - Y)^2 =$ variance of \hat{t} with respect to the sampling design p . $V_{\xi c} = E_{\xi} \left\{ (\hat{t} - t)^2 | s, r \right\}$. Now writing $E_p E_r V_{\xi c}(\hat{t}) = V_{imp} =$ imputation variance and assuming $E_{\xi} E_p [(t - Y) \{E_{\xi}(\hat{t} - t) | s\}]$ is close to zero, Särndal (1992) derived an approximate expression for V_{tot} as

$$V_{tot} \cong V_{sam} + V_{imp} \quad (1.1)$$

Consider the situation where an SRSWOR sample s of size n is selected and m the size of the response sample s_r , then under the following superpopulation model

$$E_{\xi}(y_i) = \beta, \quad V_{\xi}(y_i) = \sigma^2 \text{ and} \\ C_{\xi}(y_i, y_j) = 0 \text{ for } i \neq j \quad (1.2)$$

the unobserved values of y_i 's may be imputed by using $\hat{y}_i = \hat{\beta} = \bar{y}_r = \sum_{i \in s_r} y_i / m$ for $i \in s - s_r$ i.e.

$y_{i0} = y_i$ for $i \in s$ and $y_{i0} = \bar{y}_r$ for $i \in s - s_r$. Särndal (1992) proposed an estimator for the population total Y as

$$\hat{t}_{01} = \frac{N}{n} \sum_{i \in s} y_{i0} = N \bar{y}_r \quad (1.3)$$

He proposed the estimator for variance of \hat{t}_{01} as

$$\hat{V}_{tot} = N^2 \left(\frac{1}{m} - \frac{1}{N} \right) S_{yr}^2 = \hat{V}_{t01} \text{ (say)} \quad (1.4)$$

where

$$(m-1)S_{yr}^2 = \sum_{i \in s_r} (y_i - \bar{y}_r)^2$$

For the super population model

$$E_{\xi}(y_i) = \beta x_i, \quad V_{\xi}(y_i) = \sigma^2 x_i \text{ and } C_{\xi}(y_i, y_j) = 0 \quad (1.5)$$

Särndal (1992) used:

$$y_{i0} = \begin{cases} y_i & \text{for } i \in s_r \\ \hat{\beta} x_i & \text{for } i \in s - s_r \end{cases}$$

with $\hat{\beta} = \sum_{i \in s_r} y_i / \sum_{i \in s_r} x_i$ and proposed an estimator for Y under SRSWOR as

$$t_{02} = N \bar{x}_s \frac{\bar{y}_r}{\bar{x}_r} \quad (1.6)$$

where

$$\bar{x}_s = \sum_{i \in s} x_i, \bar{x}_r = \sum_{i \in s_r} x_i \text{ and } \bar{y}_r = \sum_{i \in s_r} y_i.$$

Särndal (1992) estimated the variance of t_{02} by

$$\hat{V}_{02} = \hat{V}_{sam}(2) + \hat{V}_{imp}(2) \quad (1.7)$$

where $\hat{V}_{sam}(2) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \left\{ S_{yos}^2 + C_0 \hat{\sigma}^2 \right\}$,

$\hat{V}_{imp}(2) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) C_1 \hat{\sigma}^2$, $S_{yos}^2 = (n-1) \sum_{i \in s} (y_i - \bar{y}_{so})^2$,

$\bar{y}_{so} = \sum_{i \in s} y_{i0} / n$,

$\hat{\sigma}^2 = \sum_{i \in s_r} (y_i - \hat{\beta} x_i)^2 / \left\{ (m-1) \left[\bar{x}_r \left(1 - cv_{xr}^2 / m \right) \right] \right\}$,

$cv_{xr} = S_{xr} / \bar{x}_r$, $(m-1)S_{xr}^2 = \sum_{i \in s_r} (x_i - \bar{x}_r)^2$

$(n-1)C_0 = \sum_{i \in s-r} x_i - \frac{\sum_{i \in s-r} x_i^2}{\sum_{i \in s_r} x_i} + \frac{1}{n} \frac{\sum_{i \in s-r} x_i \sum_{i \in s} x_i}{\sum_{i \in s_r} x_i}$ and

$C_1 = \bar{x}_s \bar{x}_{s-r} / \bar{x}_r$.

Finally, in order to compare the relative efficiency of the proposed estimator \hat{t}_{02} , Särndal (1992) conducted a Monte Carlo study with 100, 000 repeated response sets s_r , $N = 100$, $n = 30$. Three different response mechanisms were used:

Mechanism 1: response probability $\theta_i = 1 - \exp(-a_1 y_i)$, decreases with y_i ;

Mechanism 2: $\theta_i = \exp(-a_2 y_i)$, increases with y_i and

Mechanism 3: $\theta_k = .7$, response probability is a constant.

The constants a_1 and a_2 are positive and so chosen to make the average of $\theta_i = \sum_i \theta_i / N = 0.7$. Särndal's (1992) estimators cannot be used gainfully when the response probabilities θ_k 's are known or can be estimated from the available data. So, in our present study, we would like to propose some alternative estimation procedures assuming that the response probabilities θ_k for the k th unit are either known or estimated through the log-it models:

$$(i)\log(\theta_i) = 1 - c_1 p_i$$

and

$$(ii)\log(\theta_i) = c_2 p_i$$

respectively, where $p_i = z_i/Z$ and c_1, c_2 are unknown positive constants which are appropriate when response probability increases or decreases with x .

2. PROPOSED METHOD OF ESTIMATION

Here, we assume that the response probabilities θ_i 's are independent. Under this assumption, we may consider that the response sample s_r (formed by the set of respondent units) is a sub-sample from s selected by the nature according to the Poisson sampling scheme with inclusion probabilities $\pi_{i|s} = \theta_i$ and $\pi_{ij|s} = \theta_{ij} = \theta_i \theta_j$ for $i \neq j$. The Horvitz-Thompson type estimator for the total Y is:

$$\hat{Y}_{ht} = \sum_{i \in s_r} \frac{y_i}{\pi_i \theta_i} \quad (2.1)$$

It can be easily checked that \hat{Y}_{ht} is unbiased for the total Y . The expression for the variance of \hat{Y}_{ht} is given in the following theorems:

Theorem 2.1. The variance of \hat{Y}_{ht} is $V(\hat{Y}_{ht}) = V_1 + V_2$ where

$$V_1 = \frac{1}{2} \sum_{i \neq j} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (2.2)$$

and,

$$V_2 = \sum_i \frac{y_i^2}{\pi_i} \left(\frac{1}{\theta_i} - 1 \right) \quad (2.3)$$

Proof. We have

$$\begin{aligned} V(\hat{Y}_{ht}) &= E\left\{V(\hat{Y}_{ht} | s)\right\} + V\left\{E(\hat{Y}_{ht} | s)\right\} \\ &= E\left\{V\left(\sum_{i \in s_r} \frac{y_i}{\pi_i \theta_i} \mid s\right)\right\} + V\left\{E\left(\sum_{i \in s_r} \frac{y_i}{\pi_i \theta_i} \mid s\right)\right\} \end{aligned}$$

Now

$$\begin{aligned} E\left\{V\left(\sum_{i \in s_r} \frac{y_i}{\pi_i \theta_i} \mid s\right)\right\} &= E\left\{\sum_{i \in s} z_i^2 \left(\frac{1}{\pi_{i|s}} - 1\right)\right\} \\ &\quad + \frac{1}{2} \sum_{i \neq j} \sum (\pi_{ij|s} - \pi_{i|s} \pi_{j|s}) \frac{z_i}{\pi_{i|s}} \frac{z_j}{\pi_{j|s}}, \\ &= \sum_i \frac{y_i^2}{\pi_i} \left(\frac{1}{\theta_i} - 1\right) = V_2 \end{aligned}$$

(since $\pi_{ij|s} = \theta_i \theta_j$ and $\pi_{i|s} = \theta_i$) and

$$V\left\{E\left(\sum_{i \in s_r} \frac{y_i}{\pi_i \theta_i} \mid s\right)\right\} = V\left(\sum_{i \in s} \frac{y_i}{\pi_i}\right) = V_1$$

(where $z_i = \frac{y_i}{\pi_i}$).

Theorem 2.2. If the sample size is large to ensure $\Pr\{ob\{m \geq 2\} \cong 1\}$, then the following two estimators:

$$\hat{v}_{ht}(1) = \hat{V}_{11} + \hat{V}_2 \quad \text{and} \quad \hat{v}_{ht}(2) = \hat{V}_{12} + \hat{V}_2$$

are unbiased for $V(\hat{Y}_{ht})$, where

$$\hat{V}_{11} = \frac{1}{2} \sum_{i \neq j} \sum_{j \in s_r} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij} \theta_i \theta_j} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

$$\hat{V}_2 = \sum_{i \in s_r} \frac{y_i^2}{\pi_i^2 \theta_i} \left(\frac{1}{\theta_i} - 1 \right)$$

and

$$\hat{V}_{12} = \sum_{i \in s_r} \frac{y_i^2}{\pi_i \theta_i} \left(\frac{1}{\pi_i} - 1 \right) - \sum_{i \neq j} \sum_{j \in s_r} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij} \pi_i \pi_j} \frac{y_i}{\theta_i} \frac{y_j}{\theta_j}.$$

Proof. Noting

$$\begin{aligned} V_1 &= \frac{1}{2} \sum_{i \neq j} \sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \\ &= \sum_i y_i^2 \left(\frac{1}{\pi_i} - 1 \right) + \sum_{i \neq j} \sum (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \end{aligned}$$

we can verify that both the estimators \hat{V}_{11} and \hat{V}_{12} are unbiased for V_1 . It can be easily checked that \hat{V}_2 is unbiased for V_2 .

Example 2.1. Consider an SRSWOR design where $\pi_i = n/N = \pi_{i0}$ and $\pi_{ij} = n(n-1)/N(N-1) = \pi_{ij0}$. In this

case, we get $\hat{Y}_{ht} = \frac{N}{n} \sum_{i \in s_r} \frac{y_i}{\theta_i}$. The variance of \hat{Y}_{ht} can

be estimated by putting $\pi_i = \pi_{i0}$ and $\pi_{ij} = \pi_{ij0}$ in the expression variance estimators given in the Theorem 2.2.

Example 2.2. Consider an SRSWOR design where $\pi_i = \pi_{i0}, \pi_{ij} = \pi_{ij0}$ and response probabilities θ_i 's are equal to θ for every i . Further if we estimate θ by $\hat{\theta} = m/n$, then

$$\hat{Y}_{ht} = \frac{N}{m} \sum_{i \in s_r} y_i = N \bar{y}_r = \hat{t}_{01} \quad (2.4)$$

The estimator \hat{t}_{01} was proposed by Särndal (1992) when the method of single imputation is used as described in (1.3). Putting $\theta_i = \hat{\theta} = \frac{m}{n}$, $\pi_i = \pi_{i0}$, $\pi_{ij} = \pi_{ij0}$ in the

Theorem 2.2, we get two approximate variance estimators for \hat{t}_{01} as

$$\hat{v}_{01} = \frac{N(N-n)(m-1)}{m(n-1)} s_{yr}^2 + \frac{N^2(n-m)}{nm^2} \sum_{i \in s_r} y_i^2$$

and

$$\hat{v}_{02} = \hat{v}_{01} + \frac{N(N-n)(n-m)}{n(n-1)m^2} \sum_{i \in s_r} y_i^2.$$

Finally, noting that $\hat{v}_o = N^2(\frac{1}{m} - \frac{1}{N})S_{yr}^2$, the expression

for the estimate of variance of \hat{t}_{01} given by Särndal in (1.4), we get the following theorem relating to the magnitude of the three variance estimators of \hat{t}_{01} .

Theorem 2.3: (i) $\hat{v}_{02} \geq \hat{v}_{01}$ for all values of y_i 's and (ii) $\hat{v}_{02} \geq \hat{v}_o$ whenever all y_i 's are positive. (The proof of the Theorem 2.3 is straight forward and hence omitted)

Remark 2.1: We cannot get any meaningful comparison between the magnitudes of \hat{v}_o and \hat{v}_{01} .

3. CALIBRATION ESTIMATION

3.1 Calibrated estimator and its variance : Suppose x_i 's, $i \in s$ are known. Then, following Deville and Särndal (1992), we propose a calibrated estimator for \hat{Y}_{ht} as

$$\hat{Y}_C = \sum_{i \in s_r} w_i \frac{y_i}{\pi_i} \quad (3.1)$$

where w_i 's are the calibrated weights obtained by minimizing the chi-square type distance function, $D = \sum_{i \in s_r} (w_i - 1/\theta_i)^2 q_i^{-1}$ subject to the calibrated

constraints $\sum_{i \in s_r} w_i \frac{x_i}{\pi_i} = \sum_{i \in s} \frac{x_i}{\pi_i}$. Here q_i 's are suitably

chosen weights. Minimization of D leads to calibrated weight:

$$w_i = \frac{1}{\theta_i} + \left(\sum_{i \in s_r} \frac{x_i^2 q_i}{\pi_i^2} \right)^{-1} \left(\sum_{i \in s} \frac{x_i}{\pi_i} - \sum_{i \in s_r} \frac{x_i}{\pi_i \theta_i} \right) \frac{x_i q_i}{\pi_i} \quad (3.2)$$

On putting (3.2) in (3.1), the calibrated estimator for \hat{Y}_{ht} emerges as

$$\hat{Y}_C = \hat{Y}_{ht} + B_r \left(\sum_{i \in s} \frac{x_i}{\pi_i} - \hat{X}_{ht} \right) \quad (3.3)$$

where

$$\hat{X}_{ht} = \sum_{i \in s_r} \frac{x_i}{\pi_i \theta_i}$$

and

$$\hat{B}_r = \sum_{i \in s_r} \frac{y_i x_i q_i}{\pi_i^2} / \sum_{i \in s_r} \frac{x_i^2 q_i}{\pi_i^2}. \quad (3.4)$$

Now writing

$$B_r = \left(E \sum_{i \in s_r} \frac{y_i x_i q_i}{\pi_i^2} \right) / \left(E \sum_{i \in s_r} \frac{x_i^2 q_i}{\pi_i^2} \right) \\ = \left(\sum_{i \in U} \frac{y_i x_i q_i \theta_i}{\pi_i} \right) / \left(\sum_{i \in U} \frac{x_i^2 q_i \theta_i}{\pi_i} \right)$$

and $E_i = y_i - B_r x_i$, an approximate expression for the variance of \hat{Y}_C is obtained as

$$V_C = V(\hat{Y}_C) \\ = E\{V(\sum_{i \in s_r} \frac{E_i}{\theta_i \pi_i} | s)\} + V\{E(\sum_{i \in s_r} \frac{y_i - B_r x_i}{\theta_i \pi_i} + B_r \sum_{i \in s} \frac{x_i}{\pi_i} | s)\} \\ = E\{ \sum_{i \in s} \frac{E_i^2}{\pi_i^2} (\frac{1}{\theta_i} - 1) \} + V(\sum_{i \in s} \frac{y_i}{\pi_i}) \\ = \sum_i \frac{E_i^2}{\pi_i} (\frac{1}{\theta_i} - 1) + \frac{1}{2} \sum_{i \neq j} (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$$

3.2 Estimation of variance of the calibrated estimator: It can be easily checked that

$$Q(r) = \sum_{i \in s_r} \left(e_i^2 / \pi_i^2 \theta_i \right) \left(\frac{1}{\theta_i} - 1 \right) \quad (3.5)$$

is an approximate unbiased estimator of $\sum_i \left(E_i^2 / \pi_i \right) \left(\frac{1}{\theta_i} - 1 \right)$ where $e_i = y_i - B_r x_i$ and B_r is as in (3.3). Hence using the **Theorem 2.2** we get the following two design consistent estimators for V_C as:

$$\hat{V}_C(1) = Q(r) + \hat{V}_{11} \quad (3.6)$$

and,

$$\hat{V}_C(2) = Q(r) + \hat{V}_{12} \quad (3.7)$$

where $Q(r)$ is given in (3.4), and \hat{V}_{11} , \hat{V}_{12} are as defined in the **Theorem 2.2**. Following Deville and Särndal (1992), we obtain two alternative model-based as well as design-based consistent estimators for V_C when we replace $1/\theta_i$ by the calibrated weight w_i in the expressions of $\hat{V}_C(1)$ and $\hat{V}_C(2)$. The estimators are as:

$$\hat{V}_C(3) = \tilde{Q}(r) + \hat{V}_{13} \quad (3.8)$$

and

$$\hat{V}_C(4) = \tilde{Q}(r) + \hat{V}_{14} \quad (3.9)$$

where

$$\tilde{Q}(r) = \sum_{i \in s_r} \frac{e_i^2 w_i}{\pi_i^2} (w_i - 1),$$

$$\hat{V}_{13} = \frac{1}{2} \sum_{i \neq j \in s_r} w_i w_j \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

and

$$\hat{V}_{14} = \sum_{i \in s_r} w_i \frac{y_i^2}{\pi_i} \left(\frac{1}{\pi_i} - 1 \right) - \sum_{i \neq j \in s_r} w_i w_j \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij} \pi_i \pi_j} y_i y_j.$$

Example 3.1. Consider an SRSWOR design where $\pi_i = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$ and $q_i = 1/x_i$, then

$$\hat{Y}_C = \frac{N}{n} \sum_{i \in s_r} \frac{y_i}{\theta_i} + b_r \left(\frac{N}{n} \sum_{i \in s_r} x_i - \frac{N}{n} \sum_{i \in s_r} \frac{x_i}{\theta_i} \right) \quad (3.10)$$

where $b_r = \bar{y}_r / \bar{x}_r$ and $\bar{x}_s = \sum_{i \in s} x_i / n$ and $\bar{x}_r = \sum_{i \in s_r} x_i / m$.

In particular, if we assume $\theta_i = \theta$ for every $i = 1, 2, \dots, N$,

then we obtain $w_i = \frac{n \bar{x}_s}{m \bar{x}_r} = w$ (say) and

$$\hat{Y}_C = N \frac{\bar{x}_s}{\bar{x}_r} \bar{y}_r = t_{02} \quad (3.11)$$

Remark 3.1. It is important to note that the estimator (3.11) is independent of the response probability θ . Hence, one can use t_{02} without estimating the response probability θ . The estimator (3.11) was obtained by Särndal (1992) under the single imputation method (given in (1.6)) in presence of the model (1.5). Now replacing θ by its estimate $\hat{\theta} = m/n$ and putting $\pi_i = n/N$, $\pi_{ij} = n(n-1)/N(N-1)$ and $w_i = w$ in the expression of $\hat{V}_C(j)$ for $j = 1, 2, 3, 4$ we get the following variance estimators for t_{02}

$$v_{oC}(1) = \frac{N^2}{n} \frac{n-m}{m^2} (m-1) s_{er}^2 + \frac{N(N-n)(m-1)}{m(n-1)} s_{yr}^2$$

$$v_{oC}(2) = \frac{N^2}{n} \frac{n-m}{m^2} (m-1) s_{er}^2 + \frac{N(N-n)}{m(n-1)} \left\{ (m-1) s_{yr}^2 + \left(\frac{1}{m} - \frac{1}{n} \right) \sum_{i \in s_r} y_i^2 \right\}$$

$$v_{oC}(3) = \frac{N^2}{n} \frac{(m-1)}{m^2} \left(\frac{\bar{x}_s}{\bar{x}_r} \right) \frac{(n\bar{x}_s - m\bar{x}_r)}{\bar{x}_r} s_{er}^2 + \frac{N(N-n)(m-1)}{m(n-1)} \left(\frac{\bar{x}_s}{\bar{x}_r} \right)^2 s_{yr}^2$$

$$v_{oC}(4) = \frac{N^2}{n} \frac{(m-1)}{m^2} \left(\frac{\bar{x}_s}{\bar{x}_r} \right) \frac{(n\bar{x}_s - m\bar{x}_r)}{\bar{x}_r} s_{er}^2 + \left(\frac{\bar{x}_s}{\bar{x}_r} \right)^2 \left[\frac{N(N-n)}{m(n-1)} \left\{ (m-1) s_{yr}^2 + \left(\frac{1}{m} - \frac{1}{n} \right) \sum_{i \in s_r} y_i^2 \right\} \right] + \frac{N(N-n)}{nm} \frac{\bar{x}_s}{\bar{x}_r} \left(1 - \frac{\bar{x}_s}{\bar{x}_r} \right) \left(\sum_{i \in s_r} y_i^2 \right)$$

where $(m-1) s_{er}^2 = \sum_{i \in s_r} (y_i - b_r x_i)^2$.

Example 3.2. Consider an SRSWOR sampling design, $q_i = 1$ and $\theta_i = \theta$ for every $i = 1, 2, \dots, N$ and if we replace θ by its estimate $\hat{\theta} = m/n$ in (3.1) and get

$$w_i = w_{i0} = \frac{n}{m} + n \left(\sum_{i \in s_r} x_i^2 \right)^{-1} (\bar{x}_s - \bar{x}_r) x_i$$

and $w_i = w_{i0}$ yields

$$\hat{Y}_C = N [\bar{y}_r + b(\bar{x}_s - \bar{x}_r)] = \hat{t}_{03} \text{ (say)} \quad (3.12)$$

where $b = \sum_{i \in s_r} y_i x_i / \sum_{i \in s_r} x_i^2$. Now writing

$(m-1) \tilde{s}_{er}^2 = \sum_{i \in s_r} (y_i - b_r x_i)^2$ and $\tilde{e}_i = y_i - b x_i$, we get the

approximate expressions for the variance of (3.12) as follows:-

$$\tilde{v}_{oC}(1) = \frac{N^2}{n} \left(\frac{n-m}{m^2} \right) (m-1) \tilde{s}_{er}^2 + \frac{N(N-n)(m-1)}{m(n-1)} s_{yr}^2$$

$$\tilde{v}_{oC}(2) = \frac{N^2}{n} \left(\frac{n-m}{m^2} \right) (m-1) \tilde{s}_{er}^2 + \frac{N(N-n)}{m(n-1)} \left[(m-1) s_{yr}^2 + \left(\frac{1}{m} - \frac{1}{n} \right) \sum_{i \in s_r} y_i^2 \right]$$

$$\tilde{v}_{oC}(3) = \frac{N^2}{n^2} \sum_{i \in s_r} w_{i0} (w_{i0} - 1) \tilde{e}_i^2 + \frac{N(N-n)}{n^2 (n-1)} \frac{1}{2} \sum_{i \neq j \in s_r} w_{i0} w_{j0} (y_i - y_j)^2$$

$$\tilde{v}_{oC}(4) = \frac{N^2}{n^2} \sum_{i \in s_r} w_{i0} (w_{i0} - 1) \tilde{e}_i^2 + \frac{N(N-n)}{n^2} \sum_{i \in s_r} w_{i0} y_i^2 - \frac{N(N-n)}{n^2 (n-1)} \sum_{i \neq j \in s_r} w_{i0} w_{j0} y_i y_j$$

The next section has been resorted to the empirical study to study the performance of these resultant four estimators of the variance of ratio and regression estimators.

4. EMPIRICAL STUDY

Although the proposed technique is useful for any kind of response probabilities, but we restricted to empirical

study to the situation of smooth response probability being equal to the ratio of respondents to the sample size. We generated a population of $N = 100$ units by using the following the model :

$$y_i = \beta x_i + \varepsilon_i$$

where $E_{\xi}(\varepsilon_i) = 0$ and $V_{\xi}(\varepsilon_i) = \sigma^2 x_i$. The variable ε_i is generated as normal with mean zero and variance $\sigma = 0.25$ and the variable x_i was generated as beta variate with both parameters equal to 0.2. We selected 50,000 samples each of size $n=30$ and in the first case, and then we randomly dropped 9 units from each sample, by keeping $m = 21$, so that the response probability $\hat{\theta}_i = m/n = 0.7$ in each one of the sample similar to the one considered by Sarndal (1992). The relative efficiency of the four estimators of the variance of ratio estimator was computed as:

$$RE(R(j)) = \frac{\sum_{k=1}^{50,000} \{v_{0C}(1)_k - V(t_{02})\}^2}{\sum_{k=1}^{50,000} \{v_{0C}(j)_k - V(t_{02})\}^2}$$

for $j=1,2,3,4$ and are presented in Table 1, and that of four estimators of regression estimator are:

$$RE(LR(j)) = \frac{\sum_{k=1}^{50,000} \{\tilde{v}_{0C}(1)_k - V(t_{03})\}^2}{\sum_{k=1}^{50,000} \{\tilde{v}_{0C}(j)_k - V(t_{03})\}^2}$$

for $j=1,2,3,4$ and are presented in Table 2. Then the values of $\hat{\theta}_i$ were changed to study the effect of response rate on the estimators of variance. As the response rate increases, the estimates are not changing much and the relative efficiency figures are not deviating much from 1.00, but the fourth estimator remains better than others. More simulation study remains to be done in which the response probability θ_i can be estimated with the help of third related auxiliary variable say, z_i .

REFERENCES

- Deville, J.C and Särndal, C.E (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 77, 66-96.
- Rubin, D.B. (1987): *Multiple imputation for Nonresponse in Surveys*. John Wiley, New York.
- Särndal, C.E.(1992): Methods for estimating the precision of survey estimates when imputation is used. *Survey Methodology*, 18, 241-252.

Table 1. Relative efficiency comparison of four estimators of variance of ratio estimator.

Response Probability	$RE(R(1))$	$RE(R(2))$	$RE(R(3))$	$RE(R(4))$
$\hat{\theta}_i = 0.7$	1.00	1.10	1.14	1.15
$\hat{\theta}_i = 0.83$	1.00	1.05	1.07	1.08
$\hat{\theta}_i = 0.90$	1.00	1.02	1.03	1.03

Table 2. Relative efficiency comparison of four estimators of variance of regression estimator.

Response Probability	$RE(LR(1))$	$RE(LR(2))$	$RE(LR(3))$	$RE(LR(4))$
$\hat{\theta}_i = 0.7$	1.00	1.08	1.12	1.11
$\hat{\theta}_i = 0.83$	1.00	1.03	1.06	1.07
$\hat{\theta}_i = 0.90$	1.00	1.01	1.03	1.03