STATISTICA, anno LXXVIII, n. 3, 2018

ON THE ADJUSTMENT OF NON-RESPONSE THROUGH IMPUTATION FOR ESTIMATING CURRENT MEAN IN REPEATED SURVEYS

Priyanka Singh [1]
School of Management, SRM IST, Chennai, India
Ajeet K. Singh
Directorate of Economics and Statistics, Civil Lines, New Delhi, India
Vijay K. Singh
Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India

1. INTRODUCTION

The estimates of population parameters obtained through one-time surveys are usually relevant only up to a limited period of time and could not be used for populations which is dynamic, in the sense that population characteristics are subjected to changes over time. To overcome this limitation, the only way is to repeat the survey at regular intervals or even at random intervals over a period of time and, thus, the survey may be considered repetitive in character. Theory of successive sampling appears to have started with the work of Jessen (1942). Yates (1949) was the first to follow up the work of Jessen and to develop the theory of partial replacement for more than two occasions. Subsequently, Narain (1953), Tikkiwal (1951, 1953, 1956, 1958) published a series of interesting papers under same set up of estimation as given by Jessen. Utilizing different kinds of estimates and choice of samples, Sen (1971,1972,1973), Gupta (1979), Singh *et al.* (1991), Singh and Singh (2001) and Singh (2003) contributed a lot of researches towards the development of the theory of estimation of population mean in successive sampling. Feng and Zou (1997), Singh (2005), Choudhary *et al.* (2004) and Singh *et al.* (2012) considered the application of auxiliary information at both the occasions.

1.1. Non-response in successive sampling

Repeated surveys are generally more prone to the problem of non-response than single occasion surveys. Many authors have suggested different methods to deal with the

[1] Priyanka Singh. E-mail: priya.bhu.vns@gmail.com

problem on non-response where Sub-sampling and Imputation methods are predominant. Imputation is the technique of filling-in the incomplete sampled data in order to have a complete data set that can be analysed with traditional analysis methods. To deal with missing values effectively Kalton *et al.* (1981) and Sande (1979) suggested some imputation methods. Lee *et al.* (1994) used the information on an auxiliary variable for the purpose of imputation. Based on auxiliary variable, Singh and Horn (2000) suggested compromised methods of imputation. Several papers based on imputation techniques to deal with non-response in successive sampling have been suggested by Singh and Priyanka (2010), Singh *et al.* (2008), Singh *et al.* (2009), Singh *et al.* (2013) and Pandey *et al.* (2016).

In this paper we have proposed an imputation method based on a family of "factor-type estimator" for dealing with the problem of non-response assuming that the target population has been sampled at two different occasions. One parameter "factor-type estimators (FTE)" propounded by Singh and Shukla (1987) exhibits some nice properties, and includes sample mean estimator, ratio estimator, product estimator and dual to ratio estimator as particular cases.

2. THE PROBLEM AND NOTATIONS USED

Let Ω be the finite population of size N under consideration which has been sampled over two occasions. Let the characteristic under study be denoted by X (Y) on the first (second) occasion. Let the information on an auxiliary variable (with known population mean) Z be available such that Z_h; $(h = 1,2)$ stand for the auxiliary variable Z on h^{th} $h = 1,2$ occasion. We assume the presence of non-response at both the occasions. Let a simple random sample without replacement (SRSWOR) denoted by s_n of size n be selected at the first occasion, out of which r_1 units be respondents and $n - r_1$ be non-respondents. We denote the sets of respondent units and non-respondent units in this sample by R_1 and R_1^C respectively. Obviously then s_n=$R_1 \bigcup R_1^C$. Further, a random sub-sample s_m of $m = n\lambda$ units is retained (matched), for its use at the second occasion, from the r_1 units of the sample s_n and it is assumed that these matched units will respond at the second (current) occasion as well. A SRSWOR sample of size $u = (n-m)= n\mu$ units denoted by s_u is selected afresh at the second occasion from the entire population so that the overall sample size at the second occasion remain n. Let the number of responding units out of sampled u units, which are drawn afresh at the current occasion be denoted by r_2. Let us denote the sets of respondents and non-respondents in the sample s_u by R_2 and R_2^c respectively so that s_u=R_2 R_2^C. We observe that λ and μ; $(\lambda + \mu) = 1$ are the fractions of the matched and fresh samples, respectively, on the current occasion.

In what follows next, we shall use the following notations.

- $\bar{X}, \bar{Z}_1$: population means of the respective variables X and Z at the first occasion.
- $\bar{Y}, \bar{Z}_2$: population means of the respective variables Y and Z at the second occasion.

- $\bar{x}_n, \bar{y}_m, \bar{x}_m, \bar{x}_{r_1}, \bar{y}_{r_2}, \bar{z}_{r_1}, \bar{z}_{r_2}$: sample means of the respective variables based on the sample sizes shown in suffices.

- $\rho_{X,Y}, \rho_{X,Z_1}, \rho_{X,Z_2}, \rho_{Y,Z_1}, \rho_{Y,Z_2}, \rho_{Z_1,Z_2}$: correlation coefficients between the variables shown in suffice.

- $S_X^2, S_Y^2, S_{Z_1}^2, S_{Z_2}^2$: population mean squares of the variables X, Y, Z_1 and Z_2 respectively.

- $S_{X,Y}$: population covariance between the variables X and Y.

- C_p: coefficient of variation of the variable p.

- $s^2_{X(m)}, s^2_{X(n)}$: sample mean squares of the variables shown in suffice on the basis of sample of size given in the parenthesis.

- $s_{XY(m)}$: sample covariance between the variables shown in suffices in the sample given in the parenthesis.

- $f_1 = \frac{r_1}{n}, f_2 = \frac{r_2}{u}, f' = \frac{u}{N}, f'' = \frac{n}{N}, t_1 = 1 - f_1, t_2 = 1 - f_2.$

Other notations will be defined as and when required.

3. Proposed imputation methods

3.1. *For the fresh sample*

As we have assumed that non-response is present in the population at both the occasions, hence, the sample of size u, selected from the population of size N, which is a fresh sample, would exhibit some non-respondent units, which are, at the first, to be filled up through the method of imputation. Let us define the method as follows

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in R_2 \\ \frac{\bar{y}_{r_2}}{u - r_2}[u\phi_u(k) - r_2] & \text{if} \quad i \in R_2^c \end{cases} \tag{1}$$

where

$$\phi_u(k) = \frac{(A + C)\bar{Z}_2 + f' B \bar{z}_{r_2}}{(A + f' B)\bar{Z}_2 + C \bar{z}_{r_2}}. \tag{2}$$

Here we have $A=(k-1)(k-2), B=(k-1)(k-4)$ and $C=(k-2)(k-3)(k-4)$; $k > 0$ being the parameter involved in the FTE $\phi_u(k)$.

THEOREM 1. *The imputation method* (1) *gives rise to the point estimator T_u for estimating the population mean $\bar{Y}$ at the second occasion on the basis of fresh sample, as*

$$T_u = \bar{y}_{r_2} \frac{(A+C)\bar{Z}_2 + f' B\bar{z}_{r_2}}{(A+f'B)\bar{Z}_2 + C\bar{z}_{r_2}} = \bar{y}_{r_2} \phi_u(k). \tag{3}$$

The proof of the theorem is given in the Appendix.

3.2. *For the matched sample*

The second estimator based on the matched sample s_m of size m is common to both the occasions and utilizes the information from the first occasion. Since, there is non-response at the first occasion also; therefore, first of all the missing values in the sample s_n of size n will be filled-in by imputation for the computation of the sample mean $\bar{x}_n$ at the first occasion which would be an estimator at that occasion. For the purpose, we define the following imputation method

$$x_{.i} = \begin{cases} x_i & \text{if} \quad i \in R_1 \\ \frac{\bar{x}_{r_1}}{n-r_1}[n\phi_m(k) - r_1] & \text{if} \quad i \in R_1^c \end{cases} \tag{4}$$

where

$$\phi_m(k) = \frac{(A+C)\bar{Z}_1 + f'' B\bar{z}_{r_1}}{(A+f''B)\bar{Z}_1 + C\bar{z}_{r_1}}. \tag{5}$$

THEOREM 2. *Under the imputation method (4), the estimator $\bar{x}_n$ for estimating the population mean $\bar{X}$ at the first occasion is given by*

$$\bar{x}_n = \bar{x}_{r_1} \frac{(A+C)\bar{Z}_1 + f'' B\bar{z}_{r_1}}{(A+f''B)\bar{Z}_1 + C\bar{z}_{r_1}}. \tag{6}$$

Since our aim is to define an estimator for estimating the population mean at the second occasion, that is, $\bar{Y}$, on the basis of the matched sample; a double sampling regression estimator may be defined for the purpose, considering the sample s_n as a preliminary sample. A double sampling regression estimator for $\bar{Y}$ mean at the second occasion, utilizing the information gathered in the matched sample, is defined as

$$T_m = \bar{y}_m + b_{yx}(\bar{x}_n - \bar{x}_m) = \bar{y}_m + b_{yx}\left(\bar{x}_{r_1} \frac{(A+C)\bar{Z}_1 + f'' B\bar{z}_{r_1}}{(A+f''B)\bar{Z}_1 + C\bar{z}_{r_1}} - \bar{x}_m\right), \tag{7}$$

where b_{yx} is the estimate of population regression coefficient β_{yx} of Y on X.

4. THE PROPOSED POINT ESTIMATOR

In order to define an estimator for population mean $\bar{Y}$ at the second occasion, on the basis of both fresh and matched samples, we take a convex linear combination of T_u and T_m and hence we have the estimator as

$$T = \delta T_u + (1-\delta)T_m, \tag{8}$$

where δ is an unknown real constant to be determined under certain condition ($0 < \delta < 1$).

REMARK 3. *It is quite evident from* (8) *that may be considered as weight associated with the estimators defined on the basis of the unmatched (fresh) sample and the matched sample.*

4.1. *Some special members of the family T*

It is to be pointed out here that as the family of estimators T is a function of two different factor type estimators, some special cases are worthwhile to discuss herewith for assigned values of the parameter k. We consider here four values of k, namely, $k = 1,2,3$ and 4. Table 1 below depicts the estimators T_u and T_m for $k = 1,2,3,4$.

TABLE 1
Special cases of T for specific values of k.

k	T_u	T_m
1	$\bar{y}_{r_2}\frac{\bar{Z}_2}{\bar{z}_{r_2}}$	$\bar{y}_m + b_{yx}\left(\bar{x}_{r_1}\frac{\bar{Z}_1}{\bar{z}_{r_1}} - \bar{x}_m\right)$
2	$\bar{y}_{r_2}\frac{\bar{z}_{r_2}}{\bar{Z}_2}$	$\bar{y}_m + b_{yx}\left(\bar{x}_{r_1}\frac{\bar{z}_{r_1}}{\bar{Z}_1} - \bar{x}_m\right)$
3	$\bar{y}_{r_2}\frac{\bar{Z}_2 - f'\bar{z}_{r_2}}{(1-f')\bar{Z}_2}$	$\bar{y}_m + b_{yx}\left(\bar{x}_{r_1}\frac{\bar{Z}_1 - f''\bar{z}_{r_1}}{(1-f'')\bar{Z}_1} - \bar{x}_m\right)$
4	$\bar{y}_{r_2}$	$\bar{y}_m + b_{yx}\left(\bar{x}_{r_1} - \bar{x}_m\right)$

REMARK 4. *Table 1 shows that one gets ratio-type, product type, dual to ratio-type and usual mean estimators on the basis of matched sample as special cases of T_u for $k = 1,2,3,4$. Similarly, in the regression-type estimator T_m the similar estimators have been used for estimating the unknown population mean $\bar{X}$ respectively for $k = 1,2,3,4$ but the estimators are related to the first occasion estimators.*

REMARK 5. ***The convergence property of the FTE** which states that as $k \longrightarrow \infty$ the limiting estimator converges to the estimator which is obtained for $k = 1$. Letting $k \longrightarrow \infty$*

and taking limit of T_u and T_m, we observe that

$$limT_u = \bar{y}_{r_2}\frac{\bar{Z}_2}{\bar{z}_{r_2}} \tag{9}$$

and

$$T_m = \bar{y}_m + b_{yx}\left(\bar{x}_{r_1}\frac{\bar{Z}_1}{\bar{z}_{r_1}} - \bar{x}_m\right). \tag{10}$$

Hence, contrary to other one-parameter families of estimators for estimating population mean, such as, $\hat{\bar{Y}} = \bar{y}_n(\frac{\bar{X}}{\bar{x}_n})^{\beta}$; *$\beta$ being a constant, which does not exist for large values of β; the FTE possesses a novel property of convergence and existence for any arbitrarily chosen larger value of the parameter k.*

5. Bias and mean square error of the estimator T

The bias and MSE of the estimator is given below. The proof of the theorem is given in the Appendix.

THEOREM 6. *The bias of the estimator T, to the first order of approximation and for large population (ignoring finite population corrections) is given by*

$$B(T) = \delta B(T_u) + (1-\delta)B(T_m), \tag{11}$$

where

$$B(T_u) = -D'\frac{1}{r_2}\bar{Y}(\theta_2' C_{Z_2}^2 - \rho_{YZ_2}C_Y C_{Z_2}) \tag{12}$$

and

$$\begin{aligned} B(T_m) = {} & \bar{X}\beta_{YX}\left(\frac{1}{m} - \frac{1}{r_1}\right)\left(\frac{C_{300}}{\bar{X}S_X^2} - \frac{C_{210}}{\bar{X}S_{XY}}\right) + \bar{X}\beta_{YX}\frac{D''}{\bar{Z}_1 r_1}\left(\frac{C_{111}}{S_{XY}} - \frac{C_{201}}{S_X^2}\right) \\ & + \bar{X}\beta_{YX}\frac{1}{r_1}\left(D''\rho_{X,Z_1}C_X C_{Z_1} - \theta_1''\theta_2'' C_{Z_1}^2 + \theta_2'' C_{Z_1}^2\right). \end{aligned} \tag{13}$$

REMARK 7. *It is evident that $B[T]$ is a function of the parameter k. It is, therefore, easy to obtain the bias of the estimators T_u and T_m for the special cases as mentioned in Table* 1.

THEOREM 8. *The MSE of the estimator T, to the first order of approximation is given by*

$$M(T) = \delta^2 M(T_u) + (1-\delta)^2 M(T_m),$$

where

$$M(T_u) = \frac{1}{r_2}\bar{Y}^2(C_Y^2 + D'^2 C_{Z_2}^2 + 2D'\rho_{YZ_2}C_Y C_{Z_2}) \tag{14}$$

and

$$M(T_m) = S_Y^2 \left[\frac{1}{m}(1-\rho_{XY}^2) + \frac{\rho_{XY}^2}{r_1} + \frac{1}{r_1} D'' \rho_{XY} \frac{C_{Z_1}}{C_X} \left\{ D'' \rho_{XY} \frac{C_{Z_1}}{C_X} + 2\rho_{YZ_1} \right\} \right]. \quad (15)$$

Further since T_u and T_m are based on two different non-overlapping samples of size u and m respectively, therefore, for large population, we consider $Cov(T_u, T_m) = 0$. Hence the theorem.

REMARK 9. *It can be seen that if the coefficients of variation of the variables Y, X, Z_1, and Z_2 are all equal, that is, $C_X = C_Y = C_{Z_1} = C_{Z_2}$ then expression* (15) *can further be simplified as*

$$M(T_m) = \bar{Y}^2 C_Y^2 \left[\frac{1}{m}(1-\rho_{XY}^2) + \frac{1}{r_1} \left\{ \rho_{XY}^2 + D''^2 \rho_{XY}^2 + 2D'' \rho_{XY} \rho_{YZ_1} \right\} \right]. \quad (16)$$

REMARK 10. *We observe that in both the expressions* (14) *and* (15)*, the only terms which are function of k are D' and D''. Hence, the optimum value of the parameter k, say k_0 which minimizes $M(T_u)$ and $M(T_m)$ respectively can be obtained by solving the equations $\partial M(T_u)/\partial k = 0$ and $\partial M(T_m)/\partial k = 0$, which give*

$$\partial D' / \partial k = D_0' = -\rho_{YZ_2} \frac{C_Y}{C_{Z_2}} \quad (17)$$

and

$$\partial D'' / \partial k = D_0'' = -\frac{\rho_{YZ_1}}{\rho_{YX}} \frac{C_X}{C_{Z_1}}, \quad (18)$$

for which

$$M(T_u)_{min} = \frac{1}{r_2} \bar{Y}^2 C_Y^2 (1-\rho_{YZ_2}^2) \quad (19)$$

and

$$M(T_m)_{min} = \bar{Y}^2 C_Y^2 \left[\frac{1}{m}(1-\rho_{YX}^2) + \frac{1}{r_1} \left\{ \rho_{YX}^2 - \rho_{YZ_1}^2 \right\} \right]. \quad (20)$$

Therefore, for a given value of the constant δ the minimum MSE of the estimator T would be given by

$$\begin{aligned} M(T)_{min} = &\delta^2 \frac{1}{r_2} \bar{Y}^2 C_Y^2 (1-\rho_{YZ_2}^2) + (1-\delta)^2 \bar{Y}^2 C_Y^2 \\ &\times \left[\frac{1}{m}(1-\rho_{YX}^2) + \frac{1}{r_1} \left[\{ \rho_{YX}^2 - \rho_{YZ_1}^2 \right\} \right]. \end{aligned} \quad (21)$$

6. MINIMIZING $M(T)$ WITH RESPECT TO δ

Using the result

$$M(T) = \delta^2 M(T_u) + (1-\delta)^2 M(T_m),$$

we see that the optimum value of the constant δ, which minimizes the MSE of the estimator T for a specific choice of the parameter k would be

$$\delta_0 = \frac{M(T_m)}{M(T_u) + M(T_m)}. \tag{22}$$

The corresponding MSE of T then would be

$$M[T] = \frac{M(T_m)M(T_u)}{M(T_u) + M(T_m)}. \tag{23}$$

Further, the minimum $M[T]$ for the choice of $k = k_0$ would be

$$M[T]_{min} = \frac{M(T_m)_{min} M(T_u)_{min}}{M(T_u)_{min} + M(T_m)_{min}}. \tag{24}$$

REMARK 11. *It is seen that D' and D'' are two different functions of the parameter k, therefore equations* (17) *and* (18) *will yield possibly two different values of k which minimizes $M(T_u)$ and $M(T_m)$. Let k_1 and k_2 be the values of k satisfying equations* (17) *and* (18) *respectively. Therefore, expressions* (19) *and* (20) *are actually $M(T_u)_{k_1}$ and $M(T_m)_{k_2}$ respectively. Since $M[T]$ is*

$$M(T) = \delta^2 M(T_u) + (1-\delta)^2 M(T_m),$$

therefore, $\partial M(T)/\partial k = \delta^2 \partial M(T_u)/\partial k + (1-\delta)^2 \partial M(T_m)/\partial k = 0$ implies that necessarily $\partial M(T_u)/\partial k$ and $\partial M(T_m)/\partial k$ would be zero, which yield $k = k_1$ and $k = k_2$ respectively. Therefore, the optimum k for which $M(T)$ would be minimum, will be

$$k_0 = \delta^2 k_1 + (1-\delta)^2 k_2.$$

7. OPTIMUM REPLACEMENT POLICY

If the ultimate sample sizes at both the occasions is n, we know that the size of the fresh sample $u = (n-m) = n\mu$. Thus, $\mu = u/n$, that is, μ represents the fraction of the sample at the second occasion, which is replaced. It is, therefore, desirable to know that what must be the optimum replacement fraction of the sample of size n at the second occasion such that the estimate on the current occasion may have the maximum precision. In order to get the optimum values of μ, say μ_{opt}, we use the notations μ for u/n and $(1-\mu)$ for m/n respectively in the expression (24). We then have

$$M[T]_{min} = \frac{S_Y^2}{n} \frac{P(f_1 Q + R) - \mu P R}{f_1 P + \mu S - R f_2 \mu^2}, \tag{25}$$

where $P=(1-\rho^2_{YZ_2})$, $Q=(1-\rho^2_{YX})$, $R=(\rho^2_{YX}-\rho^2_{YZ_1})$, and $S=Qf_1f_2-f_1P+Rf_2$. Further, minimizing the expression (25), with respect to μ, we get

$$\mu^*_{opt}=\frac{PRf_2(f_1Q+R)\pm\sqrt{(PRf_2(f_1Q+R))^2-f_2PR^2(f_2RP^2+PS(f_1Q+R))}}{f_2PR^2}. \quad (26)$$

Therefore, the expression (25) will become

$$M[T]^*_{opt}=|M[T]_{min}|_{\mu_{opt}}=\frac{S_Y^2}{n}\frac{P(f_1Q+R)-\mu^*_{opt}PR}{f_1P+\mu^*_{opt}S-Rf_2\mu^{*2}_{opt}}. \quad (27)$$

REMARK 12. *As the value of μ should lie between 0 and 1, a negative and/or imaginary root as obtained from (26) would be inadmissible. Only a positive value of μ^*_{opt}, such that $0\leq\mu^*_{opt}\leq 1$ is admissible.*

8. SOME SPECIAL CASES

8.1. Case I: Non-response only at first occasion

If non-response is experienced only at the first occasion and it is not observed at the second occasion, then obviously

$$T_u=\bar{y}_u\frac{(A+C)\bar{Z}_2+f'B\bar{z}_u}{(A+f'B)\bar{Z}_2+C\bar{z}_u} \quad (28)$$

and $f_2=1$ since $r_2=u$. Accordingly, μ_{opt} for this case can be obtained from (26) by substituting $f_2=1$, after obtaining the MSE of T_u as given in (28). Obviously, the MSE of T_u would be similar as (14) with the replacement of r_2 by u.

8.2. Case II: Non-response only at second occasion

In this case $r_1=n$ implying that $f_1=1$. Further the estimator T_m would be

$$T_m=\bar{y}_m+b_{yx}\left(\bar{x}_n\frac{(A+C)\bar{Z}_1+f''B\bar{z}_n}{(A+f''B)\bar{Z}_1+C\bar{z}_n}-\bar{x}_m\right). \quad (29)$$

The corresponding μ_{opt} could be obtained from (26), letting $f_1=1$, when the MSE of T_m as given in (29) is obtained accordingly.

8.3. *Case III: Non-response at any occasion*

Under this case, we have $f_1 = f_2 = 1$, and are given by (28) and (29) respectively. The corresponding μ_{opt}, say μ^{**}_{opt} can, therefore, be easily obtained, letting $f_1 = f_2 = 1$ in (26), and using changes in the expressions of $M(T_u)$ and $M(T_m)$. We then have

$$\mu^{**}_{opt} = \frac{PR(Q+R) \pm \sqrt{(PR(Q+R))^2 - PR^2(RP^2 + PV(Q+R))}}{PR^2},$$

where $V = Q - P + R$. The corresponding $M[T]_{min}$ would be given by

$$M[T]^{**}_{opt} = \frac{S^2_Y}{n} \frac{P(Q+R) - \mu^{**}_{opt} PR}{P + \mu^{**}_{opt} V - R\mu^{**2}_{opt}}. \tag{30}$$

9. Effect of non-response on the precision of the estimators

The ideal situation in any kind of survey would be where there is no problem of non-response. It is, therefore, desirable to investigate the effect of presence of non-response with varying population parameters, on the performance of any estimator. With this view, we have tried to observe the percent relative loss in precision of the estimator T with respect to the estimator under the same circumstances but with complete information at both the occasions.

We define

$$L = \frac{M(T)^*_{opt} - M(T)^{**}_{opt}}{M(T)^*_{opt}} \cdot 100$$

as the percent relative loss in precision.

Since the MSE of T under optimality conditions involve correlations ρ_{YZ_1} and ρ_{YZ_2}, for simplicity in calculation of L, we assume that $\rho_{YZ_1} = \rho_{YZ_2} = \rho_0$. We have then computed the values of $\mu^*_{opt}, \mu^{**}_{opt}$ and L for different combinations of $\rho_0, \rho_{XY}, t_1 = \frac{(n-r_1)}{n}$ and $t_2 = \frac{(u-r_2)}{u}$ where t_1 and t_2, obviously are non-response rates in the samples selected at the first and second occasions respectively. Table 2 depicts the results.

REMARK 13. *From Table 2 the following conclusions can be drawn.*

(i) For the fixed values of ρ_{XY}, ρ_0 and t_2 (non-response rate at second occasion), the values of μ^*_{opt} increase while the values of L decrease with the increasing values of t_1 (non-response rate at first occasion). Thus, the higher the non- response rate at the first occasion, larger should be the fresh sample at the second occasion. Further, decrease in the values of L, indicates that the loss in precision of the estimator T (defined in the presence of non-response) would be smaller as compared to that of the estimator, defined in the case of absence of non-response, and sometimes T under non-response would be better than estimator T without non-response.

(ii) *For fixed values of t_1,ρ_{XY} and ρ_0, μ^*_{opt} and L increases for increasing values of t_2. That is, if non-response is more at the second occasion, the size of the fresh sample should be larger and the loss in precision of the proposed estimator T would be more.*

(iii) *For the fixed values of, t_1, t_2 and ρ_{XY}, the values of μ^*_{opt} and L decrease when ρ_0 increases, implying that higher the correlation between study and auxiliary variables, lower the amount of fresh sample required at the current occasion and the loss in precision will also decrease.*

(iv) *The overall comparison between μ^*_{opt} and μ^{**}_{opt} reveals that the replacement fraction is uniformly higher, when there exist non-response, than when the non-response is absent irrespective of values of other parameters.*

(v) *It is observed that the loss in precision of T reduces if there is a strong correlation between study and auxiliary variables.*

TABLE 2
Values of L,μ^*_{opt} and μ^{**}_{opt} for different values of ρ_{XY},ρ_0, t_1 and t_2.

ρ_0			0.7			0.8			0.9		
t_1	t_2	ρ_{XY}	μ^{**}_{opt}	μ^*_{opt}	L	μ^{**}_{opt}	μ^*_{opt}	L	μ^{**}_{opt}	μ^*_{opt}	L
0.05	0.05	0.3	0.43	0.49	1.94	0.38	0.44	1.30	0.31	0.36	-0.14
		0.5	0.45	0.54	2.29	0.40	0.47	1.67	0.33	0.38	-0.34
		0.7	-	-	-	0.45	0.55	2.36	0.37	0.43	1.18
	0.15	0.3	0.43	0.59	8.31	0.38	0.50	7.40	0.31	0.39	5.88
		0.5	0.45	0.68	9.07	0.40	0.54	7.87	0.33	0.42	6.36
		0.7	-	-	-	0.45	0.71	9.28	0.37	0.48	7.26
0.15	0.05	0.3	0.43	0.55	0.59	0.38	0.50	-1.25	0.31	0.42	-5.54
		0.5	0.45	0.59	1.51	0.40	0.52	-0.19	0.33	0.44	-4.10
		0.7	-	-	-	0.45	0.60	1.69	0.37	0.49	-1.60
	0.15	0.3	0.43	0.63	7.51	0.38	0.55	5.56	0.31	0.45	1.52
		0.5	0.45	0.71	8.74	0.40	0.59	6.64	0.33	0.48	2.85
		0.7	-	-	-	0.45	0.74	9.03	0.37	0.54	5.21

Note: Values of μ^*_{opt} and μ^{**}_{opt} Table 2 shown by dashes indicate that μ values do not exist.

10. Efficiency comparison

For the comparison of the proposed imputation strategy, we have chosen the estimator T^* proposed by Singh *et al.* (2013), which has been developed under the similar conditions.

10.1. Point estimator obtained on the fresh sample

Using the following imputation method in order to fill-in the missing data at the second occasion

$$y_{.i} = \begin{cases} y_i \frac{\bar{Z}_2}{\bar{z}_{2u}} & \text{if} \quad i \in R_2 \\ \frac{\bar{y}_{r_2}}{\bar{z}_{r_2}} \frac{\bar{Z}_2}{\bar{z}_{2u}} z_{2i} & \text{if} \quad i \in R_2^c \end{cases} \tag{31}$$

the estimator T_u^* for estimating the population mean $\bar{Y}$ on the basis of the fresh sample of size u, can be obtained as

$$T_u^* = \bar{Z}_2 \frac{\bar{y}_{r_2}}{\bar{z}_{r_2}}. \tag{32}$$

10.2. Point estimator obtained on the matched sample

The imputation method utilized in order to fill-in the missing data in the sample of size n was

$$x_{.i} = \begin{cases} x_i \frac{\bar{Z}_1}{\bar{z}_{1n}} & \text{if} \quad i \in R_1 \\ \frac{\bar{x}_{r_1}}{\bar{z}_{r_1}} \frac{\bar{Z}_1}{\bar{z}_{1n}} z_{1i} & \text{if} \quad i \in R_1^c \end{cases} \tag{33}$$

which yielded the point estimator for $\bar{X}$ as

$$\bar{x}_n^* = \bar{Z}_1 \frac{\bar{x}_{r_1}}{\bar{z}_{r_1}}. \tag{34}$$

Therefore, the estimator of $\bar{Y}$ on the basis of the matched sample was defined as

$$T_m^* = \bar{y}_m \frac{\bar{x}_n^*}{\bar{x}_m} \frac{\bar{Z}_2}{\bar{z}_{2m}}. \tag{35}$$

Finally, the estimator T^*, combining the two estimators T_u^* and T_m^*, was defined as

$$T^* = \delta^* T_u^* + (1 - \delta^*) T_m^*. \tag{36}$$

Using the expression $M[T^*]$, the optimum replacement policy, as obtained by Singh *et al.* (2013) was

$$\begin{aligned} \mu'_{opt} =& \frac{A^* C^* f_2 (f_1 B^* + C^*)}{f_2 A^* C^{*2}} \\ & \pm \frac{\sqrt{(A^* C^* f_2 (f_1 B^* + C^*))^2 - f_2 A^* C^{*2} (f_1 C^* A^{*2} + A^* D^* (f_1 B^* + C^*))}}{f_2 A^* C^{*2}}, \end{aligned} \tag{37}$$

where

$$M[T^*]_{\min} = \frac{S_Y^2}{n} \frac{A^*(f_1B^* + C^*) - \mu'_{opt}A^*C^*}{f_1A^* + \mu'_{opt}D^* - C^*f_2, \mu'^2_{opt}} \tag{38}$$

with

$$\begin{aligned} A^* &= 2(1 - \rho_{YZ_2}); B^* = 3 + 2(\rho_{XZ_2} - \rho_{YZ_2} - \rho_{XY}); \\ C^* &= 2(\rho_{Z_1Z_2} + \rho_{XY} - \rho_{XZ_2} - \rho_{YZ_1}); \\ D^* &= B^* f_1 f_2 - f_1 A^* + C^* f_2. \end{aligned}$$

For the comparison purpose, we have assumed that $\rho_{YZ_1} = \rho_{XZ_2} = \rho_{YZ_2} = \rho_0^*$. The efficiency of the proposed estimator T, under the optimality conditions, with $M[T]^*_{opt}$, given in (27), with respect to the estimator T^* under respective optimality conditions, with $M[T^*]_{opt}$, given in (38), is defined as

$$E = \frac{M[T^*]_{opt}}{M[T]^*_{opt}} * 100.$$

Table 3 depicts the values of E for some assumed values of $\rho_0, \rho_{z_1z_2}$ and ρ_{XY}.

TABLE 3
Values of E for some values of $\rho_0, \rho_{z_1z_2}$ and ρ_{XY}.

ρ_{z1z2}		0.5				0.7			
ρ_0		0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
ρ_{XY}	0.3	-	371.5	189.3	159.8	-	-	320.8	236.6
	0.5	354.5	-	200.9	161.4	504.3	-	1235.3	266.5
	0.7	204.7	110.6	-	170.9	258.3	205.8	-	363.4
	0.9	134.4	83.6	-	-	158.5	116.2	62.6	-

REMARK 14. *Table 3, which exhibits the performance of the proposed estimator T over the estimator T^*, proposed by Singh et al.* (2013), *reveals that T it is more efficient than T^* in almost all the combinations of different correlations, except for two choices of correlations. As we closely look into the table; for low or moderate values of ρ_0 the efficiency of the proposed estimator decreases as ρ_{XY} increases and for high correlation values of ρ_0 efficiency increases with the increase of ρ_{XY}. Although, the analysis depends upon a number of approximations related to correlation values, but since the selected values of correlations cover a larger range of their values, they may be generalized for most of the populations with positive correlations. As for some combinations, the optimum μ values do not exist, a clear-cut picture of the trend of the efficiency is hard to discuss herewith. However, it can be seen that the higher the correlation between the auxiliary variables Z_1 and Z_2 higher is the efficiency for all choices of ρ_{XY} value.*

11. Concluding remarks

The work presented suggested some imputation methods for the adjustment of non-response at both the occasions in rotation sampling when estimation of mean of the surveyed population at the current occasion was aimed at. It was observed that a combination of matched and fresh samples was taken into account for this purpose, the efficiency of the proposed estimator under non-response has been compared when there is no non-response. For some specific values of correlations, estimator in presence of non-response is found to be better than without non-response which validate the effectiveness of the proposed estimator. Apart from this the proposed estimator was proved to be better than Singh *et al.* (2013) imputation method for almost all correlation combinations. The presented work may also be used to estimate the changes in the performance of the estimator over time, which is another advantage of successive sampling scheme.

Appendix

A. Proofs

Proof (Theorem 1). It is clear that the mean of the fresh sample, say $\bar{y}_u$ will be an unbiased estimator of $\bar{Y}$ at the second occasion, where

$$\bar{y}_u = \frac{1}{u}\sum_{i\in s_u} y_i = \left[\frac{1}{u}\sum_{i\in R_2} y_i + \sum_{i\in R_2^c}\frac{\bar{y}_{r_2}}{(u-r_2)}(u\phi_u(k)-r_2)\right]$$

since

$$\sum_{i\in R_2} y_i = r_2\bar{y}_{r_2}.$$

Now, as there are $(u-r_2)$ units in R_2^c, hence we have

$$\bar{y}_u = \frac{r_2\bar{y}_{r_2}}{u} + \frac{(u-r_2)}{u}\frac{\bar{y}_{r_2}}{(u-r_2)}u\phi_u(k) - \frac{(u-r_2)}{u}\frac{\bar{y}_{r_2}}{(u-r_2)}r_2 = \phi_u(k)\bar{y}_{r_2}.$$

Therefore we get

$$T_u = \bar{y}_{r_2}\phi_u(k).$$

Hence the theorem. □

Proof (Theorem 2). On the same lines Theorem 2 can be proved. The large sample bias and MSE of T could be obtained up to the order $(O(n^{-1}))$, using the following large sample approximations

$$\bar{y}_m = \bar{Y}(1+e_0) \qquad \bar{x}_m = \bar{X}(1+e_1)$$
$$\bar{y}_{r_2} = \bar{Y}(1+e_2) \qquad \bar{z}_{r_2} = \bar{Z}(1+e_3)$$
$$\bar{x}_{r_1} = \bar{X}(1+e_4) \qquad \bar{z}_{r_1} = \bar{Z}_1(1+e_5$$
$$s_{xy_(m)} = S_{XY}(1+e_6) \qquad s_x^2 = S_X^2(1+e_7),$$

such that $E(e_g)=0$, $\left|e_g\right|<1$ for $g=0,1,2,3,4,5,6,7$ and letting $C_{abc}=E[(x-\bar{X})^a(y-\bar{Y})^b(z-\bar{Z})^c]$

Under the above mentioned large sample approximations,T_u takes the following form, retaining terms only up to the second degree of e_2 and e_3

$$T_u = \bar{Y}[1+e_2+D^{'}(e_3+e_2e_3-\theta_2^{'}e_3^2] \qquad \text{where}$$

$$D^{'}=(\theta_1^{'}-\theta_2^{'}); \quad f^{'}=\frac{u}{n}; \quad \theta_1^{'}=\frac{f^{'}B}{A+f^{'}B+C}; \quad \theta_2^{'}=\frac{C}{A+f^{'}B+C}.$$

Similarly, the estimator T_m, up to the order $O(n^{-1})$ is obtained.

$$T_m=\bar{Y}(1+e_0)+\bar{X}\beta_{YX}(e_4-e_1+D^{''}e_5+D^{''}e_4e_5-\theta_1^{''}\theta_2^{''}e_5^2+\theta_2^{''2}e_5^2+(e_4-e_1+D^{''}e_5)(e_6-e_7)],$$

where

$$D^{''}=(\theta_1^{''}-\theta_2^{''}); \quad f^{''}=\frac{n}{N};\theta_1^{''}=\frac{f^{''}B}{A+f^{''}B+C}; \quad \theta_2^{''}=\frac{C}{A+f^{''}B+C}. \qquad \square$$

PROOF (THEOREM 6). We have

$$B(T_u)=E[T_u]-\bar{Y}=\bar{Y}E[e_2+D^{'}(e_3+e_2e_3-\theta_2^{'}e_3^2)].$$

Thus

$$B(T_u)=\bar{Y}D^{'}E[e_2e_3-\theta_2^{'}e_3^2],$$

where

$$E(e_2e_3)=\frac{1}{r_2}\rho_{YZ_2}C_YC_{Z_2}$$

and

$$E[e_3^2]=\frac{1}{r_2}C_{Z_2}^2.$$

Hence we obtain as

$$B(T_u)=-D^{'}\frac{1}{r_2}\bar{Y}(\theta_2^{'}C_{Z_2}^2-\rho_{YZ_2}C_YC_{Z_2}).$$

On similar lines $B(T_m)$ can be obtained. $\square$

PROOF (THEOREM 8). We know that

$$M(T_u) = E[T_u - \bar{Y}]^2 = E[\bar{Y}(1 + e_2 + D^{'}(e_3 + e_2 e_3 - \theta_2^{'} e_3^2) - \bar{Y}]^2.$$

Therefore

$$\begin{aligned} M(T_u) =& E[\bar{Y}(e_2 + D^{'}(e_3 + e_2 e_3 - \theta_2^{'} e_3^2)]^2 \\ =& \frac{1}{r_2} \bar{Y}^2 (C_Y^2 + D^{'2} C_{Z_2}^2 + 2D^{'} \rho_{YZ_2} C_Y C_{Z_2}). \end{aligned}$$

Hence the result. Similarly $M[T_m]$ can be obtained. □

REFERENCES

R. K. CHOUDHARY, H. V. L. BATHLA, U. C. SUD (2004). *On non-response in sampling on two occasions*. Journal of the Indian Society of Agricultural Statistics, 58, no. 3, pp. 331–343.

S. FENG, G. ZOU (1997). *Sample rotation method with auxiliary variable*. Communications in Statistics: Theory and Methods, 26, no. 6, pp. 1497–1509.

P. C. GUPTA (1979). *Sampling on two successive occasions*. Journal of Statistical Research, 13, pp. 7–16.

R. J. JESSEN (1942). *Statistical investigation of a sample survey for obtaining farm Ffcts*. Iowa Agricultural Experiment Station Research Bulletin, 304, pp. 1–104.

G. KALTON, D. KASPRZYK, R. SANTOS (1981). *Issues of non-response and imputation in the survey of income and programme participation*. In D. KREWSKI, R. PLATEK, J. N. K. RAO (eds.) *Current Topics in Survey Sampling*, Academic Press, New York, pp. 455–480.

H. LEE, E. RANCOURT, C. E. SARNDAL (1994). *Experiments with variance estimation from survey data with imputed values*. Journal of Official Statistics, 10, no. 3, pp. 231–243.

R. D. NARAIN (1953). *On the recurrence formula in sampling on successive occasions*. Journal of the Indian Society of Agricultural Statistics, 5, pp. 96–99.

R. PANDEY, K. YADAV, N. S. THAKUR (2016). *Adapted factor-type imputation strategies*. Journal of Scientific Research, 8, no. 3, pp. 321–339.

I. G. SANDE (1979). *A personal view of hot deck imputation procedures*. Survey Methodology, 5, 238–246.

A. R. SEN (1971). *Successive sampling with two auxiliary variables*. Sankhya B, 33, 371–378.

A. R. SEN (1972). *Successive sampling with p ($p \geq 1$) auxiliary variables*. The Annals of Mathematical Statistics, 43, pp. 2031–2034.

A. R. SEN (1973). *Some theory of sampling on successive occasions*. Australian Journal of Statistics, 15, pp. 105–110.

G. N. SINGH (2003). *Estimation of population mean using auxiliary information on recent occasion in h-occasion successive sampling*. Statistics in Transition, 6, pp. 523–532.

G. N. SINGH (2005). *On the use of chain-type ratio estimator in successive sampling*. Statistics in Transition, 7, pp. 21–26.

G. N. SINGH, K. PRIYANKA (2010). *Use of imputation methods in two-occasion successive sampling*. Journal of Indian Society of Agricultural Statistics, 64, no. 3, 417–432.

G. N. SINGH, V. K. SINGH (2001). *On the use of auxiliary information in successive sampling*. Journal of the Indian Society of Agricultural Statistics, 54, no. 1, pp. 1–12.

G. N. SINGH, J. P. KARNA, L. N. SAHOO (2009). *Some imputation methods for nonresponse at current occasion in two-occasion rotation patterns*. Journal Statistical Research, 43, no. 2, pp. 37–54.

G. N. SINGH, D. MANJHI, S. PRASAD, F. HOMA (2013). *Assessment of non-response under ratio method of imputation in two occasion successive sampling*. Journal of Statistical Theory and Applications, 12, no. 4, pp. 403–418.

G. N. SINGH, K. PRIYANKA, M. KOZAK (2008). *Use of imputation methods at current occasion in two-occasion rotation patterns*. Model Assisted Statistics and Applications, 3, no. 2, 99–112.

G. N. SINGH, V. K. SINGH, K. PRIYANKA, S. PRASAD, J. P. KARNA (2012). *Rotation patterns under imputation of missing data over two occasion*. Communication in Statistics: Theory and Methods, 41, pp. 1857–1874.

S. SINGH, S. HORN (2000). *Compromised imputation in survey samplin*. Metrika, 51, pp. 267–276.

V. K. SINGH, D. SHUKLA (1987). *One parameter family of factor-type ratio estimators*. Metron, 45, no. 1-2, pp. 273–283.

V. K. SINGH, G. N. SINGH, D. SHUKLA (1991). *An efficient family of ratio-cum-difference type estimators in successive sampling over two occasions*. Journal of Scientific Research, 41C, pp. 149–159.

B. D. TIKKIWAL (1951). *Theory of Successive Sampling, Unpublished Diploma Dissertation*. Institute of Agricultural Research Statistics, New-Delhi, India.

B. D. TIKKIWAL (1953). *Optimum allocation in successive sampling*. Journal of the Indian Society of Agricultural Statistics, 5, pp. 100–102.

B. D. TIKKIWAL (1956). *A further contribution to the theory of univariate sampling on successive occasions*. Journal of the Indian Society of Agricultural Statistics, 8, pp. 84–90.

B. D. TIKKIWAL (1958). *An examination of the effect of matched samples on the efficiency of estimators in the theory of successive sampling*. Journal of the Indian Society of Agricultural Statistics, 10, pp. 16–22.

F. YATES (1949). *Sampling Methods for Censuses and Surveys*. Charles Griffin and Company Limited, London.

SUMMARY

In this paper we have proposed an imputation method based on a family of factor-type estimator to deal with the problem of non-response assuming that the target population has been sampled at two different occasions. The aim is to estimate the current population mean on the basis of matching the sample from the previous occasion and on the basis of fresh sample selected at the current occasion. It has been assumed that the non-response is exhibited by the population at both the occasions and, therefore, the imputation of missing values is required in both the samples, namely, matched sample and fresh sample. Accordingly, a combined point estimator has been suggested after imputation which generates a one-parameter family of estimators. The properties of the estimator have been investigated and the replacement policy has been discussed. Finally, the comparison of the proposed class has been made with another estimator for their performances.

Keywords: Non-response; Imputation; Repeated surveys, Factor type estimator.