STATISTICS IN TRANSITION-new series, December 2009
Vol. 10, No. 3, pp. 397—414

ESTIMATION OF MEAN UNDER IMPUTATION OF MISSING DATA USING FACTOR-TYPE ESTIMATOR IN TWO-PHASE SAMPLING

Diwakar Shukla, Narendra Singh Thakur, Sharad Pathak[1] and Dilip Singh Rajput[2]

ABSTRACT

In sample surveys, the problem of non-response is one of the most frequent and widely appearing, whose solution is required to obtain using relevant statistical techniques. The imputation is one such methodology, which uses available data as a source for replacement of missing observations. Two-phase sampling is useful when population parameter of auxiliary information is unknown. This paper presents the use of imputation for dealing with non-responding units in the setup of two-phase sampling. Two different two-phase sampling strategies (sub-sample and independent sample) are compared under imputed data setup. Factor-Type (F-T) estimators are used as tools of imputation and simulation study is performed over multiple samples showing the comparative strength of one over other. First imputation strategy is found better than second whereas second sampling design is better than first.

Key words: Estimation, Missing data, Imputation, Bias, Mean squared error (MSE), Factor Type (F-T) estimator, Two-phase sampling, Simple Random Sampling Without Replacement (SRSWOR), Compromised Imputation (C. I.).

1. Introduction

Let $\Omega = \{1, 2, ..., N\}$ be a finite population with Y_i as a variable of main interest and X_i $(i = 1,2,..., N)$ an auxiliary variable. As usual, $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$,

[1] Diwakar Shukla, Narendra Singh Thakur, Sharad Pathak, Deptt. of Mathematics and Statistics, H.S. Gour University of Sagar, Sagar,(M.P.) INDIA, Pin-470003.
e-mails:diwakarshukla@rediffmail.com, nst_stats@yahoo.co.in, s.pathak_stats@yahoo.co.in.
[2] Dilip Singh Rajput, Govt. College, Rehli, Sagar (M.P.), INDIA.

$\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$ are population means, $\overline{X}$ is assumed known and $\overline{Y}$ under investigation. Singh and Shukla (1987) proposed Factor Type (F-T) estimator to obtain the estimate of population mean under setup of SRSWOR. Some other contributions on Factor-Type estimator, in similar setup, are due to Singh and Shukla (1991) and Singh et al. (1993).

With $\overline{X}$ unknown, the two-phase sampling is used to obtain the estimate of population mean and Shukla (2002) suggested F-T estimator under this case. But when few of observations are missing in the sample, the F-T estimator fails to estimate. This paper undertakes the problem of Shukla (2002) with suggested imputation procedures for missing observations.

Rubin (1976) addressed three missing observation concepts: missing at random (MAR), observed at random (OAR) and parameter distribution (PD). Heitjan and Basu (1996) explained the concept of missing at random (MAR) and introduced the missing completely at random (MCAR). The present discussion is on MCAR wherever the non-response is quoted. Rao and Sitter (1995) discussed a new linearization variance estimator that makes more complete use of the sample data than a standard one. They have shown its application to 'mass' imputation under two-phase sampling and deterministic imputation for missing data. Singh and Horn (2000) suggested a Compromised Imputation (C-I) procedure in which the estimator of mean obtained through C-I remains better than obtained from ratio method of imputation and mean method of imputation. Ahmed et al. (2006) designed several generalized structure of imputation procedures and their corresponding estimators of the population mean. Motivation is derived from these and from Shukla (2002) to extend the content for the imputation setup.

Consider a preliminary large sample S' of size n' drawn from population Ω by SRSWOR and a secondary sample S of size n ($n < n'$) drawn in either of the following manners:

Case I: as a sub-sample from sample S' (denoted by design F_1) as in fig. 1(a),

Case II: independent to sample S' (denote by design F_2) as in fig. 1(b) without replacing S'.

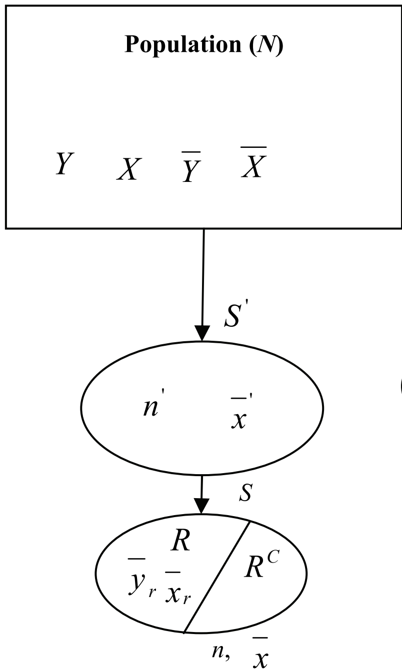

Fig. 1(a) [Case *I*, *F*$_1$]

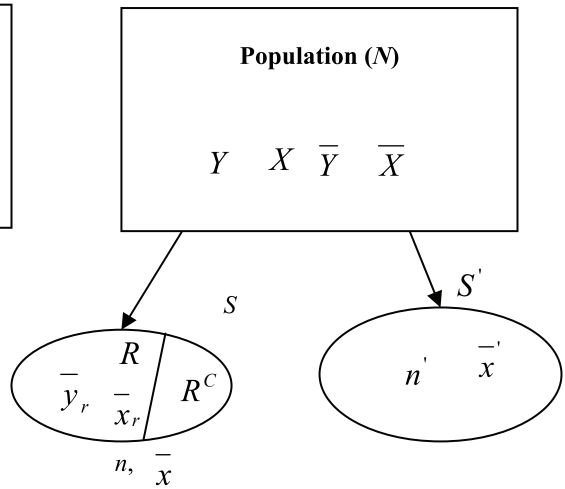

Fig. 1(b) [Case *II*, *F*$_2$]

Let sample *S* of *n* units contains *r* responding units ($r < n$) forming a subspace *R* and $(n - r)$ non-responding with sub-space R^C in $S = R \cup R^C$. For every $i \in R$, the y_i is observed available. For $i \in R^C$, the y_i values are missing and imputed values are computed. The i^{th} value x_i of auxiliary variate is used as a source of imputation for missing data when $i \in R^C$. Assume for *S*, the data $x_s = \{x_i : i \in S\}$ and $\{x_{i'} : i' \in S'\}$ are known with mean $\bar{x} = (n)^{-1} \sum_{i=1}^{n} x_i$ and $\bar{x}' = (n')^{-1} \sum_{i'=1}^{n'} x_{i'}$ respectively.

2. *F-T* Imputation Strategies

Two proposed strategies d_1 and d_2 for missing data under both cases are :

$$\boldsymbol{d_1}: \ (y_{d1})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r \left\{ \dfrac{(A+C)\overline{x'} + fB\bar{x}}{(A+fB)\overline{x'} + C\bar{x}} \right\} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{2.1}$$

$$\boldsymbol{d_2}: \ (y_{d2})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r \left\{ \dfrac{(A+C)\overline{x'} + fB\bar{x}_r}{(A+fB)\overline{x'} + C\bar{x}_r} \right\} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{2.2}$$

Under (2.1) and (2.2) point estimators of $\bar{Y}$ are :

$$(\bar{y}_{d1}) = \bar{y}_r \left[\frac{(A+C)\overline{x'} + fB\bar{x}}{(A+fB)\overline{x'} + C\bar{x}} \right]; \tag{2.3}$$

$$(\bar{y}_{d2}) = \bar{y}_r \left[\frac{(A+C)\overline{x'} + fB\overline{x_r}}{(A+fB)\overline{x'} + C\bar{x}_r} \right]; \tag{2.4}$$

where $A = (k-1)(k-2)$; $B = (k-1)(k-4)$; $C = (k-2)(k-3)(k-4)$ and k ($0 < k < \infty$) is a constant.

2.1. Some Special Cases

(i) At $k = 1$; $A = 0$; $B = 0$; $C = -6$

$$(\bar{y}_{d1}) = \bar{y}_r \left(\frac{\overline{x'}}{\bar{x}} \right) \tag{2.5}$$

$$(\bar{y}_{d2}) = \bar{y}_r \left(\frac{\overline{x'}}{\bar{x}_r} \right) \tag{2.6}$$

(ii) At $k = 2$; $A = 0$; $B = -2$; $C = 0$

$$(\bar{y}_{d1}) = \bar{y}_r \left(\frac{\bar{x}}{\overline{x'}} \right) \tag{2.7}$$

$$\left(\bar{y}_{d2}\right) = \bar{y}_r\left(\frac{\bar{x}_r}{\bar{\bar{x}}'}\right) \tag{2.8}$$

(iii) At $k = 3$; $A = 2$; $B = -2$; $C = 0$

$$\left(\bar{y}_{d1}\right) = \bar{y}_r\left(\frac{\bar{x}' + f\bar{x}}{(1-f)\bar{\bar{x}}'}\right) \tag{2.9}$$

$$\left(\bar{y}_{d2}\right) = \bar{y}_r\left(\frac{\bar{x}' - f\bar{x}}{(1-f)\bar{\bar{x}}'}\right) \tag{2.10}$$

(iv) At $k = 4$; $A = 6$; $B = 0$; $C = 0$

$$\left(\bar{y}_{d1}\right) = \bar{y}_r \tag{2.11}$$

$$\left(\bar{y}_{d2}\right) = \bar{y}_r \tag{2.12}$$

3. Properties of Imputation Strategies

Let $B(.)_t$ and $M(.)_t$ denote the bias and mean squared error (*M.S.E.*) of estimator under sampling design $t = I, II$ (or F_1, F_2). Large sample approximations are:

$$\bar{y}_r = \bar{Y}(1+e_1);\ \bar{x}_r = \bar{X}(1+e_2);\ \bar{x} = \bar{X}(1+e_3) \text{ and } \bar{x}' = \bar{X}(1+e_3')$$

Using two-phase sampling, following Rao and Sitter (1995) and the mechanism of MCAR, for given r, n and n', we write:

(i) Under design F_1 [Case *I*]:

$$E(e_1) = E(e_2) = E(e_3) = E(e_3') = 0;\ E(e_1^2) = \delta_1 C_Y^2;\ E(e_2^2) = \delta_1 C_X^2;$$
$$E(e_3^2) = \delta_2 C_X^2;\ E(e_3'^2) = \delta_3 C_x^2;\ E(e_1 e_2) = \delta_1 \rho\, C_Y\, C_X;$$
$$E(e_1 e_3) = \delta_2 \rho\, C_Y C_X;\ E(e_1 e_3') = \delta_3 \rho C_Y\, C_X;\ E(e_2 e_3) = \delta_2\, C_X^2;$$
$$E(e_2 e_3') = \delta_3\, C_X^2;\ E(e_3 e_3') = \delta_3\, C_X^2;$$

(*ii*) Under design F_2 [Case *II*]:

$$E(e_1) = E(e_2) = E(e_3) = E(e_3') = 0;\ E(e_1^2) = \delta_4 C_Y^2;\ E(e_2^2) = \delta_4 C_X^2;$$
$$E(e_3^2) = \delta_5 C_X^2;\ E(e_3'^2) = \delta_3 C_x^2;\ E(e_1 e_2) = \delta_4 \rho\, C_Y\, C_X;$$

$$E(e_1e_3)=\delta_5\rho\, C_YC_X;\quad E(e_1e_3^{'})=0\,;\ E(e_2e_3)=\delta_5\, C_X^2\,;\ E(e_2e_3^{'})=0\,;$$
$$E(e_3e_3^{'})=0$$

where

$$\delta_1=\left(\frac{1}{r}-\frac{1}{n^{'}}\right);\ \delta_2=\left(\frac{1}{n}-\frac{1}{n^{'}}\right);\ \delta_3=\left(\frac{1}{n^{'}}-\frac{1}{N}\right);$$
$$\delta_4=\left(\frac{1}{r}-\frac{1}{N-n^{'}}\right);\ \delta_5=\left(\frac{1}{n}-\frac{1}{N-n^{'}}\right)$$

Remark 3.1: Let

$$\theta_1=\frac{A+C}{A+fB+C};\ \theta_2=\frac{fB}{A+fB+C};\ \theta_3=\frac{A+fB}{A+fB+C};\ \ \theta_4=\frac{C}{A+fB+C};$$
$$P=-(\theta_1-\theta_3)=(\theta_2-\theta_4);\ (\theta_1\theta_4+\theta_2\theta_3-2\theta_3\theta_4)=P(\theta_3-\theta_4);\ V=\rho\frac{C_Y}{C_X}\ .$$

Theorem 3.1: Estimators $\left(\bar{y}_{d1}\right)$ and $\left(\bar{y}_{d2}\right)$, in terms of $e_i;\ \ i=1,2,3$ and $e_i^{'}$, could be expressed :

(i) $\bar{y}_{d1}=\bar{Y}\left[1+e_1+P\left\{e_3-e_3^{'}-\theta_4e_3^2+\theta_3e_3^{'2}+e_1e_3-e_1e_3^{'}-(\theta_3-\theta_4)e_3e_3^{'}\right\}\right]$ (3.1)

(ii) $\bar{y}_{d2}=\bar{Y}\left[1+e_1+P\left\{e_2-e_3^{'}-\theta_4e_2^2+\theta_3e_3^{'2}+e_1e_2-e_1e_3^{'}-(\theta_3-\theta_4)e_3e_3^{'}\right\}\right]$ (3.2)

While ignoring terms $E\left[e_i^re_j^s\right]$, $E\left[e_i^r\left(e_j^{'}\right)^s\right]$ for $r+s>2$, $r,s=0,1,2,...$ and $i,j=1,2,3,...$ which is first order of approximation [see Cochran (2005)].

Proof:

(i) $\bar{y}_{d1}=\bar{y}_r\left[\dfrac{(A+C)\overline{x^{'}}+fB\bar{x}}{(A+fB)\overline{x^{'}}+C\bar{x}}\right]$

$$=\bar{Y}(1+e_1)(1+\theta_1e_3^{'}+\theta_2e_3)(1+\theta_3e_3^{'}+\theta_4e_3)^{-1}$$
$$=\bar{Y}\left[1+e_1+P\left\{e_3-e_3^{'}-\theta_4e_3^2+\theta_3e_3^{'2}+e_1e_3-e_1e_3^{'}-(\theta_3-\theta_4)e_3e_3^{'}\right\}\right]$$

(ii) $\bar{y}_{d2}=\bar{y}_r\left[\dfrac{(A+C)\overline{x^{'}}+fB\bar{x}_r}{(A+fB)\overline{x^{'}}+C\bar{x}_r}\right]$

$$=\bar{Y}(1+e_1)(1+\theta_1e_3^{'}+\theta_2e_2)(1+\theta_3e_3^{'}+\theta_4e_2)^{-1}$$
$$=\bar{Y}\left[1+e_1+P\left\{e_2-e_3^{'}-\theta_4e_2^2+\theta_3e_3^{'2}+e_1e_2-e_1e_3^{'}-(\theta_3-\theta_4)e_2e_3^{'}\right\}\right]$$

Theorem 3.2: Biases of $\left(\bar{y}_{d1}\right)_t$ and $\left(\bar{y}_{d2}\right)_t$ under $t = I, II$ (or design F_1 and F_2), up to first order of approximation are:

(i) $B\left[\bar{y}_{d1}\right]_I = -\bar{Y}P(\delta_2 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$ (3.3)

(ii) $B\left[\bar{y}_{d1}\right]_{II} = \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right]$ (3.4)

(iii) $B\left[\bar{y}_{d2}\right]_I = -\bar{Y}P(\delta_1 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$ (3.5)

(iv) $B\left[\bar{y}_{d2}\right]_{II} = \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right]$ (3.6)

Proof:

(i) $B\left[\bar{y}_{d1}\right]_I = E\left[\bar{y}_{d1} - \bar{Y}\right]_I$

$= \bar{Y}E\left[1 + e_1 + P\left\{e_3 - e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2 + e_1 e_3 - e_1 e_3' - (\theta_3 - \theta_4)e_3 e_3'\right\} - 1\right]$

$= -\bar{Y}P(\delta_2 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$

(ii) $B\left[\bar{y}_{d1}\right]_{II} = E\left[\bar{y}_{d1} - \bar{Y}\right]_{II}$

$= \bar{Y}E\left[1 + e_1 + P\left\{e_3 - e_3' - \theta_4 e_3^2 + \theta_3 e_3'^2 + e_1 e_3 - e_1 e_3' - (\theta_3 - \theta_4)e_3 e_3'\right\} - 1\right]$

$= \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right]$

(iii) $B\left[\bar{y}_{d2}\right]_I = E\left[\bar{y}_{d2} - \bar{Y}\right]_I$

$= -\bar{Y}P(\delta_1 - \delta_3)\left(\theta_4 C_X^2 - \rho C_Y C_X\right)$

(iv) $B\left[\bar{y}_{d2}\right]_{II} = E\left[\bar{y}_{d2} - \bar{Y}\right]_{II}$

$= \bar{Y}P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right]$

Theorem 3.3: Mean squared errors of $\left(\bar{y}_{d1}\right)_t$ and $\left(\bar{y}_{d2}\right)_t$ under design F_1 and F_2, up to first order of approximation are:

(i) $M\left[\bar{y}_{d1}\right]_I = \bar{Y}^2\left[\delta_1 C_Y^2 + (\delta_2 - \delta_3)\left(P^2 C_X^2 + 2P\rho C_Y C_X\right)\right]$ (3.7)

(ii) $M\left[\bar{y}_{d1}\right]_{II} = -\bar{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_5)P^2 C_X^2 + 2P\delta_5\rho C_Y C_X\right]$ (3.8)

(iii) $M\left[\bar{y}_{d2}\right]_I = \bar{Y}^2\left[\delta_1 C_Y^2 + (\delta_1 - \delta_3)\left(P^2 C_X^2 + 2P\rho C_Y C_X\right)\right]$ (3.9)

(iv) $M\left[\bar{y}_{d2}\right]_{II} = \bar{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_4)P^2 C_X^2 + 2P\delta_4\rho C_Y C_X\right]$ (3.10)

Proof:

(i) $M\left[\bar{y}_{d1}\right]_I = E\left[\bar{y}_{d1} - \bar{Y}\right]_I^2$

$$=\overline{Y}^2 E\left[e_1 + P\left(e_3 - e_3^{'}\right)\right]^2$$

$$=\overline{Y}^2\left[\delta_1 C_Y^2 + (\delta_2 - \delta_3)\left(P^2 C_X^2 - 2P\rho C_Y C_X\right)\right]$$

(ii) $M\left[\overline{y}_{d1}\right]_{II} = E\left[\overline{y}_{d1} - \overline{Y}\right]_{II}^2$

$$=\overline{Y}^2 E\left[e_1 + P\left(e_3 - e_3^{'}\right)\right]^2$$

$$=\overline{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_5)P^2 C_X^2 + 2P\delta_5\rho C_Y C_X\right]$$

(iii) $M\left[\overline{y}_{d2}\right]_{I} = E\left[\overline{y}_{d2} - \overline{Y}\right]_{I}^2$

$$=\overline{Y}^2 E\left[e_1 + P\left(e_2 - e_3^{'}\right)\right]^2$$

$$=\overline{Y}^2\left[\delta_1 C_Y^2 + (\delta_1 - \delta_3)\left(P^2 C_X^2 - 2P\rho C_Y C_X\right)\right]$$

(iv) $M\left[\overline{y}_{d2}\right]_{II} = E\left[\overline{y}_{d2} - \overline{Y}\right]_{II}^2$

$$=\overline{Y}^2\left[\delta_4 C_Y^2 + (\delta_3 + \delta_4)P^2 C_X^2 + 2P\delta_4\rho C_Y C_X\right]$$

Theorem 3.4: Minimum mean squared errors of $\left(\overline{y}_{d1}\right)_t$ and $\left(\overline{y}_{d2}\right)_t$ under design F_1 and F_2 are :

(i) $Min\left[M\left(\overline{y}\right)_{d1}\right]_I = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right]S_Y^2$ when $P = -V$ (3.11)

(ii) $Min\left[M\left(\overline{y}_{d1}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_5)^{-1}\delta_5^2\rho^2\right]S_Y^2$

when $P = -\delta_5 V/\left(\delta_3 + \delta_5\right)$ (3.12)

(iii) $Min\left[M\left(\overline{y}_{d2}\right)\right]_I = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2$ when $P = -V$ (3.13)

(iv) $Min\left[M\left(\overline{y}_{d2}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_4)^{-1}\delta_4^2\rho^2\right]\ S_Y^2$

when $P = -\delta_4 V/\left(\delta_3 + \delta_4\right)$ (3.14)

Proof:

(i) $\dfrac{d}{dP}\left[M\left(\overline{y}_{d1}\right)_I\right] = 0 \Rightarrow P = -\rho\dfrac{C_Y}{C_X} = -V$ and using this in (3.7)

$$Min\left[M\left(\overline{y}_{d1}\right)\right]_I = \left[\delta_1 - (\delta_2 - \delta_3)\rho^2\right]S_Y^2$$

(ii) $\dfrac{d}{dP}\left[M\left(\overline{y}_{d1}\right)_{II}\right] = 0 \Rightarrow P = -\delta_5 V\left(\delta_3 + \delta_5\right)^{-1}$ and using this in (3.8)

$$Min\left[M\left(\overline{y}_{d1}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_5)^{-1}\delta_5^2\rho^2\right]S_Y^2$$

(iii) $\dfrac{d}{dP}\left[M\left(\overline{y}_{d2}\right)_I\right] = 0 \Rightarrow P = -\rho\dfrac{C_Y}{C_X} = -V$

$$Min\left[M\left(\overline{y}_{d2}\right)\right]_I = \left[\delta_1 - (\delta_1 - \delta_3)\rho^2\right]S_Y^2$$

(iv) $$\frac{d}{dP}\left[M\left(\overline{y}_{d2}\right)_{II}\right] = 0 \Rightarrow P = -\delta_4 V\left(\delta_3 + \delta_4\right)^{-1}$$

$$Min\left[M\left(\overline{y}_{d2}\right)\right]_{II} = \left[\delta_4 - (\delta_3 + \delta_4)^{-1}\delta_4^2\rho^2\right] S_Y^2$$

Lemma 3.0 [By Shukla (2002)]:
F-T estimator in two-phase sampling (without imputation) is

$$\left(\overline{y}_d\right)_w = \overline{y}\left[\frac{(A+C)\overline{x'} + fB\overline{x}}{(A+fB)\overline{x'} + C\overline{x}}\right] \tag{3.15}$$

With optimum MSE conditions
under design F_1 [Case I]: $P = -V$ (3.16)
under design F_2 [Case I]: $P = -V\left(1+\delta\right)^{-1}$ (3.17)
and optimum MSE expressions

$$opt\left[M\left(\overline{y}_d\right)_w\right]_I = \overline{Y}^2 V_{20}\left[1 - \rho^2\left(1-\delta\right)\right] \tag{3.18}$$

$$opt\left[M\left(\overline{y}_d\right)_w\right]_{II} = \overline{Y}^2 V_{20}\left[1 - \rho^2\left(1+\delta\right)^{-1}\right] \tag{3.19}$$

where $\delta = \left(\frac{1}{n'} - \frac{1}{N}\right)\left(\frac{1}{n} - \frac{1}{N}\right)^{-1}$;

$$V_{ij} = E\left[\left(\overline{y} - \overline{Y}\right)^i \left(\overline{x} - \overline{X}\right)^j\right] \Big/ \overline{Y}^i\,\overline{X}^j\;;\quad i = 0,1,2;\; j = 0,1,2$$

4. Appropriate Choice of *k* for Bias Reduction

(i) $B\left[\overline{y}_{d1}\right]_I = 0 \Rightarrow P\left(\theta_4 C_X^2 - \rho C_Y C_X\right) = 0$

If $P = 0 \Rightarrow (k-4)\left[k^2 - (5+f)k + (6+f)\right] = 0$ (4.1)

$$\left.\begin{aligned} \text{and}\quad k &= k_1 = 4 \\ k &= k_2 = \frac{1}{2}\left[(5+f) + \left(f^2 + 6f + 1\right)^{1/2}\right] \\ k &= k_3 = \frac{1}{2}\left[(5+f) - \left(f^2 + 6f + 1\right)^{1/2}\right] \end{aligned}\right\} \tag{4.2}$$

If $\theta_4 C_X^2 - \rho C_Y C_X = 0$

$$\Rightarrow AV + VfB + (V-1)C = 0 \tag{4.3}$$

$$\Rightarrow (V-1)k^3 - [(8-f)V - 9]k^2 - [(23+5f)V + 26]k \; - 2[(11-2f)V - 12] \; = 0 \tag{4.4}$$

(ii) $B\left[\bar{y}_{d1}\right]_{II} = 0 \;\Rightarrow P\left[(\theta_3\delta_3 - \theta_4\delta_5)C_X^2 + \delta_5\rho C_Y C_X\right] \; = 0$

If $P = 0$ we have solution as per (4.2) and

if $(\theta_3\delta_3 - \theta_4\delta_5)C_X + \delta_5\rho C_Y = 0$

$$\Rightarrow (V-1)k^3 + \left[\left(\frac{\delta_3}{\delta_5} + V\right)(1+f) - 9(V-1)\right]k^2 - \left[\left(\frac{\delta_3}{\delta_5} + V\right)(3+5f) - 26(V-1)\right]k$$

$$+2\left[\left(\frac{\delta_3}{\delta_5} + V\right)(1+2f) - 12(V-1)\right] \; = 0 \tag{4.5}$$

(iii) $B\left[\bar{y}_{d2}\right]_I = 0$ provides similar solution as in (i).

(iv) $B\left[\bar{y}_{d2}\right]_{II} = 0 \;\Rightarrow P\left[(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X\right] \; = 0$

if $P = 0$ we have solution as per (4.2) and

if $(\theta_3\delta_3 - \theta_4\delta_4)C_X^2 + \delta_4\rho C_Y C_X = 0$

$\Rightarrow [(A + fB)\delta_3 - C\delta_4] = -\delta_4 V(A + fB + C)$

$$\Rightarrow (V-1)k^3 + \left[\left(\frac{\delta_3}{\delta_4} + V\right)(1+f) - 9(V-1)\right]k^2 - \left[\left(\frac{\delta_3}{\delta_4} + V\right)(3+5f) - 26(V-1)\right]k$$

$$+2\left[\left(\frac{\delta_3}{\delta_4} + V\right)(1+2f) - 12(V-1)\right] \; = 0 \tag{4.6}$$

5. Comparison of the Estimators

(i) $\Delta_1 = \min\left[M\left(\bar{y}_{d1}\right)_I\right] - \min\left[M\left(\bar{y}_{d2}\right)_I\right] = \left(\frac{1}{r} - \frac{1}{N}\right)\rho^2 S_Y^2$

$\left(\bar{y}_{d2}\right)_{\mathrm{I}}$ is better than $\left(\bar{y}_{d1}\right)_{\mathrm{I}}$ if $\Delta_1 > 0 \Rightarrow N > r$ which is always true.

(ii) $\Delta_2 = \min\left[M\left(\bar{y}_{d1}\right)_{II}\right] - \min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$= \left(\frac{\delta_4^2}{(\delta_3 + \delta_4)} - \frac{\delta_5^2}{(\delta_3 + \delta_5)}\right)\rho^2 S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\mathrm{II}}$ is better than $\left(\bar{y}_{d1}\right)_{\mathrm{II}}$ if $\Delta_2 > 0$

$$\Rightarrow (n-r)\left[N^3 - \left(n'n + n'r + nr\right)N + 2n'nr\right] > 0$$

which generates two options as

(A) when $(n-r) > 0 \Rightarrow n > r$ and

(B) $\left[N^3 - \left(n'n + n'r + nr\right)N + 2n'nr\right] > 0$

if $n' \approx N$ $\left[i.e.\ n' \to N\right]$

then $N\left[N^2 - (n-r)N + nr\right] > 0$ (since $N > 0$ always)

$$\Rightarrow (N-n)(N-r) > 0$$

$$\Rightarrow (N-n) > 0 \Rightarrow N > n \quad \text{and} \quad N - r > 0 \Rightarrow N > r$$

The ultimate is $N > n > r$, which is always true.

(iii) $\Delta_3 = \min\left[M\left(\bar{y}_{d2}\right)_I\right] - \min\left[M\left(\bar{y}_{d2}\right)_{II}\right]$

$$= \frac{(\delta_1 - \delta_4)(\delta_3 + \delta_4) + \left(\delta_4^2 + \delta_3^2 - \delta_1\delta_3 - \delta_1\delta_4 + \delta_3\delta_4\right)}{(\delta_3 + \delta_4)} \rho^2 S_Y^2$$

$\left(\bar{y}_{d2}\right)_{\mathrm{II}}$ is better than $\left(\bar{y}_{d2}\right)_{\mathrm{I}}$, if $\Delta_3 > 0$

$$\Rightarrow \rho^2 > \left[\frac{1+m}{1+2m}\right] \text{where } m = \left[\frac{r\left(N-n'\right)}{n'(N-r)}\right]$$

$$\Rightarrow -1 < \rho < -\sqrt{\frac{1+m}{1+2m}} \quad \text{or} \quad \sqrt{\frac{1+m}{1+2m}} < \rho < 1$$

6. Empirical Study

The attached appendix A has an artificial population of size N = 200 [see Appendix A] containing values of main variable Y and auxiliary variable X. Parameters of this are given in table 6.1.

Table 6.1. Population Parameters

$\bar{Y}$	$\bar{X}$	S_Y^2	S_X^2	ρ	C_X	C_Y	$V = \rho \frac{C_Y}{C_X}$
42.485	18.515	199.0598	48.5375	0.8652	0.3763	0.3321	0.7635

Under design-I, we draw a preliminary random sample S' of size $n' = 110$ to compute $\bar{x}'$ and further draw a random sample S of size $n = 50$ such that

$S \subset S'$ by *SRSWOR*. The V is a stable quantity over time and assumed to be known [see Reddy (1978)].

Table 6.2.

Design	Optimum Condition for *MSE*	Three optimum values of *k* on one condition		
I	$P = -V$	$k_1 = 1.5206$	$k_2 = 2.4505$	$k_3 = 8.9456$
	$P = -\delta_5 V / (\delta_3 + \delta_5)$	$k_4 = 1.5880$	$k_5 = 2.8768$	$k_6 = 6.4279$
II	$P = -V$	$k_7 = 1.5206 = k_1$	$k_8 = 2.4505 = k_2$	$k_9 = 8.9456 = k_3$
	$P = -\delta_4 V / (\delta_3 + \delta_4)$	$k_{10} = 1.5645$	$k_{11} = 2.8572$	$k_{12} = 6.7221$

7. Simulation

The bias and optimum m.s.e. of proposed estimators under both designs are computed through 50,000 repeated samples n, n' as per design. Computations are in table 7.1 where efficiency loss measurement due to imputation is as $LI_t\left(\bar{y}_s\right) = \dfrac{Opt\left[M\left(\bar{y}_s\right)_t\right]}{Opt\left[M\left(\bar{y}_d\right)_w\right]}$ with $Opt\left[M\left(\bar{y}_s\right)_t\right]$ the optimum mean squared error of estimator $\bar{y}_s$,

$s = d, d_1, d_2; t = I, II, t = w =$ without imputation.

For design I and II the simulation procedure has following steps :

Step 1: Draw a random sample S' of size $n' = 110$ from the population of N = 200 by SRSWOR.

Step 2: Draw a random sub-sample of size $n = 50$ from S' for design *I* and independent random sample n= 50 from $\left(N - n'\right)$ for design *II*.

Step 3: Drop down 5 units randomly from each second sample corresponding to Y in both *I* and *II*.

Step 4: Impute dropped units of Y by proposed methods and available methods and compute the relevant statistic.

Step 5: Repeat the above steps 50,000 times, which provides multiple sample based estimates $(\hat{y}_{1s})_t, (\hat{y}_{2s})_t, (\hat{y}_{3s})_t, \ldots (\hat{y}_{50000s})_t$ for estimators $\left(\bar{y}_{d_1}\right)_t$, $\left(\bar{y}_{d_2}\right)_t$ and $\left(\bar{y}_d\right)_w$.

Step 6: Bias of $(\hat{y}_s)_t$ is $B(\hat{y}_s)_t = \dfrac{1}{50000}\sum_{i=1}^{50000}\left[(\hat{y}_{is})_t - \bar{Y}\right]$

Step 7: *M.S.E.* of $(\hat{y}_s)_t$ is $M(\hat{y}_s)_t = \frac{1}{50000}\sum_{i=1}^{50000}\left[(\hat{y}_{is})_t - \overline{Y}\right]^2$

Step 8: The efficiency comparisons are

Design efficiency $E_1 = \frac{M(\overline{y}_{d1})_I}{M(\overline{y}_{d1})_{II}} \times 100$;

Design efficiency $E_2 = \frac{M(\overline{y}_{d2})_I}{M(\overline{y}_{d2})_{II}} \times 100$

Estimator efficiency $E_3 = \frac{M(\overline{y}_{d1})_I}{M(\overline{y}_{d2})_I} \times 100$;

Estimator efficiency $E_4 = \frac{M(\overline{y}_{d1})_{II}}{M(\overline{y}_{d2})_{II}} \times 100$

Table 7.1. Bias and Mean Squared Error

Opt (k)	Design F_1				Design F_2			
	$(\overline{y}_{d_1})_{k_i}$		$(\overline{y}_{d_2})_{k_i}$		$(\overline{y}_{d_1})_{k_i}$		$(\overline{y}_{d_2})_{k_i}$	
	Bias	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*	*Bias*	*MSE*
$k_1 = 1.3813$	-0.1314	2.5947	-0.2855	2.1727	0.3665	2.7997	0.3680	3.0111
$k_2 = 2.7576$	-0.1780	2.6282	-0.3339	2.2192	0.1485	2.7801	0.1287	2.9542
$k_3 = 9.9538$	-0.1350	2.5968	-0.2892	2.1758	0.3498	2.7911	0.3497	2.9982
$k_4 = 1.5880$	-0.0895	2.2549	-0.2892	2.1758	0.4131	3.7278	0.3680	3.0110
$k_5 = 2.8768$	-0.0855	2.2706	-0.2268	1.9645	0.2664	3.5517	0.2499	3.4296
$k_6 = 6.4279$	-0.0951	2.2580	-0.2114	1.9333	0.3832	3.7016	0.3778	3.5901
$k_7 = 1.3813$	-0.1314	2.5947	-0.2855	2.1727	0.3665	2.7997	0.3680	3.0111
$k_8 = 2.7576$	-0.1780	2.6282	-0.3339	2.2192	0.1485	2.7801	0.1287	2.9542
$k_9 = 9.9538$	-0.1350	2.5968	-0.2892	2.1758	0.3498	2.7911	0.3497	2.9982
$k_{10} = 1.5645$	-0.0953	2.2743	-0.2082	1.9476	0.4044	3.4592	0.4024	3.3839
$k_{11} = 2.8572$	-0.1287	2.2961	-0.2429	1.9705	0.2473	3.2895	0.2301	3.1970
$k_{12} = 6.7221$	-0.1015	2.7780	-0.2146	1.9515	0.3756	3.4273	0.3707	3.3472

Table 7.2. Estimator without Imputation [Lemma 3.0]

	Optimum k - values			
	k_1	k_2	k_3	Optimum MSE
Case - I	1.3813	2.7576	9.9538	1.9021
Case - II	1.5311	2.8337	7.1604	2.7702

Table 7.3. Loss due to Imputation $LI_t[\ ]$

Opt (k)	Design F_1		Design F_2	
	$LI_I(\bar{y}_{d_1})_k$	$LI_I(\bar{y}_{d2})_k$	$LI_{II}(\bar{y}_{d_1})_k$	$LI_{II}(\bar{y}_{d_2})_k$
$k_1 = 1.3813$	1.364123863	1.142263814	1.01064905	1.08696123
$k_2 = 2.7576$	1.381735976	1.166710478	1.00357375	1.0664212
$k_3 = 9.9538$	1.365227906	1.143893591	1.00754458	1.08230453
$k_4 = 1.5880$	1.185479207	1.143893591	1.34567901	1.08692513
$k_5 = 2.8768$	1.193733242	1.032805846	1.28210959	1.23803335
$k_6 = 6.4279$	1.187108985	1.016402923	1.33622121	1.29597141
$k_7 = 1.3813$	1.364123863	1.142263814	1.01064905	1.08696123
$k_8 = 2.7576$	1.381735976	1.166710478	1.00357375	1.0664212
$k_9 = 9.9538$	1.365227906	1.143893591	1.00754458	1.08230453
$k_{10} = 1.5645$	1.195678461	1.023920929	1.2487185	1.22153635
$k_{11} = 2.8572$	1.207139477	1.035960254	1.18745939	1.1540683
$k_{12} = 6.7221$	1.460491036	1.025971295	1.23720309	1.20828821

Table 7.4. Efficiency Comparisons E_1, E_2, E_3, E_4

Opt (*k*)	E_1	E_2	E_3	E_4
$k_1 = 1.3813$	92.67%	72.15%	119.42%	92.97%
$k_2 = 2.7576$	94.53%	75.12%	118.43%	94.10%
$k_3 = 9.9538$	93.03%	72.57%	119.34%	93.09%
$k_4 = 1.5880$	60.48%	72.26%	103.63%	123.80%
$k_5 = 2.8768$	63.92%	57.28%	115.58%	103.56%
$k_6 = 6.4279$	61.00%	53.85%	116.79%	103.10%
$k_7 = 1.3813$	92.67%	72.15%	119.42%	92.97%
$k_8 = 2.7576$	94.53%	75.12%	118.43%	94.10%
$k_9 = 9.9538$	93.03%	72.57%	119.34%	93.09%
$k_{10} = 1.5645$	65.74%	54.24%	116.77%	102.22%
$k_{11} = 2.8572$	69.80%	61.63%	116.52%	102.89%
$k_{12} = 6.7221$	81.05%	58.30%	142.35%	102.39%

8. Almost Unbiased Imputation Methods

Using equations of section 4.0 we get

From (i) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 0.1812$.

From (ii) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 1.4236$; $k_5^{'} = 2.4469$; $k_6^{'} = 10.7864$. From (iii) similar to (i).

From (iv) $k_1^{'} = 4$; $k_2^{'} = 3.4253$; $k_3^{'} = 1.8246$; $k_4^{'} = 1.4339$; $k_5^{'} = 2.4488$; $k_6^{'} = 10.5426$

Using these *k*-values we can make proposed F-T imputation strategies almost unbiased. The best among them will be that having the lowest m.s.e. By this we have option to choose almost unbiased estimator with a control over mean squared error.

9. Discussion and Conclusion

The proposed estimators are found useful for situation when some observations are missing in the sample. As per table 7.3 for $\bar{y}_{d_1}$ and $\bar{y}_{d_2}$ under design F_1, the efficient performance of both is found when k = 1.5880, 1.5645, 6.4279 and 6.7221. On these specific choices the loss of efficiency with respect to without imputation is very low. Similarly, for $\bar{y}_{d_1}$ and $\bar{y}_{d_2}$ under design F_2, the efficient performance observed at k = 1.3813, 2.7576, 9.9538. It seems even by adopting imputation, the suggested estimators are loosing a little in terms of relative m.s.e. to the without imputation usual F-T estimator.

While mutual comparisons are in table 7.4, the design F_2 is uniformly efficient as F_1 at all the optimum k-values, over both suggested F-T strategies. Within F_1, the estimator $\bar{y}_{d_2}$ is more efficient than $\bar{y}_{d_1}$ whereas within F_2 it does not hold uniformly for all k-optimals. The $\bar{y}_{d_2}$ under F_2 found better when k = 1.2, 2.8, 6.4 and 6.7. One can get almost unbiased estimators also on choices k = 0.1812, 1.4236, 1.4339, 1.8246, 2.4469, 2.4488, 3.4253, 4, 10.5426, 10.7864. The most suitable will be that which has the lowest m.s.e. So these suggested strategies are almost unbiased with a reducing control over m.s.e. also.

Appendix A

Population (N = 200)

Y_i	45	50	39	60	42	38	28	42	38	35
X_i	15	20	23	35	18	12	8	15	17	13
Y_i	40	55	45	36	40	58	56	62	58	46
X_i	29	35	20	14	18	25	28	21	19	18
Y_i	36	43	68	70	50	56	45	32	30	38
X_i	15	20	38	42	23	25	18	11	09	17
Y_i	35	41	45	65	30	28	32	38	61	58
X_i	13	15	18	25	09	08	11	13	23	21
Y_i	65	62	68	85	40	32	60	57	47	55
X_i	27	25	30	45	15	12	22	19	17	21
Y_i	67	70	60	40	35	30	25	38	23	55
X_i	25	30	27	21	15	17	09	15	11	21
Y_i	50	69	53	55	71	74	55	39	43	45
X_i	15	23	29	30	33	31	17	14	17	19
Y_i	61	72	65	39	43	57	37	71	71	70
X_i	25	31	30	19	21	23	15	30	32	29
Y_i	73	63	67	47	53	51	54	57	59	39
X_i	28	23	23	17	19	17	18	21	23	20
Y_i	23	25	35	30	38	60	60	40	47	30
X_i	07	09	15	11	13	25	27	15	17	11
Y_i	57	54	60	51	26	32	30	45	55	54
X_i	31	23	25	17	09	11	13	19	25	27
Y_i	33	33	20	25	28	40	33	38	41	33
X_i	13	11	07	09	13	15	13	17	15	13
Y_i	30	35	20	18	20	27	23	42	37	45
X_i	11	15	08	07	09	13	12	25	21	22
Y_i	37	37	37	34	41	35	39	45	24	27
X_i	15	16	17	13	20	15	21	25	11	13
Y_i	23	20	26	26	40	56	41	47	43	33
X_i	09	08	11	12	15	25	15	25	21	15
Y_i	37	27	21	23	24	21	39	33	25	35
X_i	17	13	11	11	09	08	15	17	11	19
Y_i	45	40	31	20	40	50	45	35	30	35
X_i	21	23	15	11	20	25	23	17	16	18
Y_i	32	27	30	33	31	47	43	35	30	40
X_i	15	13	14	17	15	25	23	17	16	19
Y_i	35	35	46	39	35	30	31	53	63	41
X_i	19	19	23	15	17	13	19	25	35	21
Y_i	52	43	39	37	20	23	35	39	45	37
X_i	25	19	18	17	11	09	15	17	19	19

REFERENCES

AHMED, M. S., AL-TITI, O., AL-RAWI, Z. and ABU-DAYYEH, W. (2006): *Estimation of a population mean using different imputation methods,* Statistics in Transition, 7, 6, 1247–1264.

COCHRAN, W. G. (2005): *Sampling Techniques,* John Wiley and Sons, Fifth Edition New York.

HEITJAN, D. F. and BASU, S. (1996): *Distinguishing 'Missing at random' and 'missing completely at random'*, The American Statistician, 50, 207–213.

RAO, J. N. K. and SITTER, R. R. (1995): *Variance estimation under two-phase sampling with application to imputation for missing data*, Biometrica, 82, 453–460.

RUBIN, D. B. (1976): *Inference and missing data,* Biometrica, 63, 581–593.

SHUKLA, D. (2002): *F-T estimator under two-phase sampling*, Metron, 59, 1–2, 253–263.

SHUKLA, D., SINGH, V. K. and SINGH, G. N. (1991): *On the use of transformation in factor type estimator,* Metron, 49 (1–4), 359–361.

SINGH, S. and HORN, S. (2000): *Compromised imputation in survey sampling,* Metrika, 51, 266–276.

SINGH, V. K. and SHUKLA, D. (1987): *One parameter family of factor type ratio estimator,* Metron, 45, 1–2, 273–283.

SINGH, V. K. and SHUKLA, D. (1993): *An efficient one parameter family of factor - type estimator in sample survey*, Metron, 51, 1–2, 139–159.

SINGH, V. K. and SINGH, G. N. (1991): *Chain type estimator with two auxiliary variables under double sampling scheme*, Metron, 49, 279–289.

REDDY, V. N. (1978): *A study on the use of prior knowledge on certain population parameters in estimation*, Sankhya, C, 40, 29–37.