African Journal of Mathematics and Computer Science Research Vol. 6(1), pp. 5-15, January 2013
Available online at http://www.academicjournals.org/AJMCSR
DOI:10.5897/AJMCSR11.136
ISSN 2006-9731 ©2013 Academic Journals

Full Length Research Paper

Some new aspects on imputation in sampling

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

[1]Department of Mathematics and Statistics, Dr. H.S. Gour University of Sagar, Sagar (M. P.), 470003, India.
[2]Centre for Mathematical Sciences (CMS), Banasthali University, Rajasthan, 304022, India.

Accepted 15 January, 2013

The number of causes that affect the quality of survey and missing data is one of such that keeps sample incomplete. Many imputation methods are available in literature, and are used to replace missing observations like Mean method, Ratio method, Compromised method, Ahmed's method, Factor-type (F-T) method etc. This paper suggests some new estimation aspects in imputation theory using both, F-T estimator and Compromised estimator. All these are derived from existing procedure of usual Compromised method and found efficient, bias controlled, having many properties. In support of derived results, an empirical study is incorporated over artificial data set showing the efficiency in favour of the suggested methods.

Key words: Estimation, missing data, imputation, bias, mean squared error (m.s.e.), compromised estimator, factor-type (F-T) estimator, factor-type compromised imputation (FTCI).

INTRODUCTION

Missing data is one of the most common problems in sample surveys and various imputation methodologies are frequently used to substitute these values. Statisticians have recognized that failure to account for the stochastic nature of incompleteness in the form of absence of data can spoil inference. There are three major incompleteness concepts: Missing at random (MAR), observed at random (OAR), and parametric distribution (PD). In what follows in this paper, missing completely at random (MCAR) is used. Some well known imputation methods in literature are: deductive imputation, mean imputation overall (MO), random imputation overall (RO), mean imputation within classes (MC), random imputation within classes (RC), hot-deck imputation, flexible matching imputation, predicted regression imputation (PR), random regression imputation (RR), distance function matching etc.

In a contribution, Shukla (2002) suggested Factor-type (F-T) estimator in two-phase sampling which incorporates a parameter *k* combining known auxiliary information survey components. But *k* used therein is in the form of factors $(k-1)(k-2)$; $(k-1)(k-4)$; $(k-2)(k-3)(k-4)$ which ultimately produces a noble property of getting multiple choices of *k*-values maintaining certain standard level of mean squared error (m.s.e). As suggested, best *k* among all is that also optimizing other characteristic of estimator. Singh and Horn (2000) proposed a compromised imputation procedure having a constant α in linear combination of "available main information" and "imputed auxiliary information". The α bears an optimum selection at the minimum level of m.s.e. but this choice does not have a control over bias factor. This paper derives motivation from this source and, using properties of F-T estimator, presents a new F-T compromised imputation procedure for survey sampling. Special feature of the suggested is the same parameter *k* used twice, in the imputation part for missing observations as well as in the linear combination with the known auxiliary part. Some other useful contributions are due to Rueda and Gonzalez (2008), Singh et al. (2009) etc. over imputation techniques.

Let $\overline{Y}$ be mean of a finite population for desired estimation $\left[\overline{Y} = N^{-1}\sum_{i-1}^{N} Y_i\right]$. A random sample *S* of size *n* drawn without replacement (SRSWOR) from population $\Omega = \{1,2,3,.....,N\}$ to estimate $\overline{Y}$. Sample *S* of *n* units contains *r* responding units $r < n$ forming a set *R* and $(n-r)$ non-responding units with the sub-space $(n-r)$ having symbol R^C. The variable *Y* is of main interest and

*Corresponding author. E-mail: nst_stats@yahoo.co.in.

X an auxiliary variable correlated with Y. For every unit $i \in R$, the value y_i observed is available. For $i \in R^C$, y_i values are missing and imputed values need to be derived. The i^{th} value x_i of auxiliary variate X is used as a source of imputation for missing data when $i \in R^C$. Assume for sample S, the data $x_s = \{x_i : i \in S\}$ are known and $S = R \cup R^C$. Under this setup, some well-known imputation methods in survey sampling literature are:

Mean method of imputation

For y_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \bar{y}_r & if \quad i \in R^C \end{cases} \tag{1}$$

The imputation-based estimator of population mean $\bar{Y}$ is:

$$\bar{y}_m = \frac{1}{r}\sum_{i \in R} y_i = \bar{y}_r \tag{2}$$

Ratio method of imputation

For y_i and x_i, define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b}x_i & if \quad i \in R^C \end{cases} \tag{3}$$

Where $\hat{b} = \sum_{i \in R} y_i \Big/ \sum_{i \in R} x_i$

The imputation-based estimator of $\bar{Y}$ is:

$$\bar{y}_S = \frac{1}{n}\sum_{i \in S} y_{\bullet i} = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r}\right) = \bar{y}_{RAT} \tag{4}$$

Where $\bar{y}_r = \frac{1}{r}\sum_{i \in R} y_i$, $\bar{x}_r = \frac{1}{r}\sum_{i \in R} x_i$ and $\bar{x}_n = \frac{1}{n}\sum_{i \in S} x_i$

Compromised method

Singh and Horn (2000) proposed compromised imputation procedure

$$y_{\bullet i} = \begin{cases} (\alpha n/r) y_i + (1-\alpha)\hat{b}x_i & if \quad i \in R \\ (1-\alpha)\hat{b}x_i & if \quad i \in R^C \end{cases} \tag{5}$$

The imputation-based estimator, in this case, is

$$\bar{y}_{COMP} = \left[\alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}\right] \tag{6}$$

Where α is a suitable constant value such that the resultant variance is minimum.

Lemma 1

The bias, m.s.e. and minimum m.s.e. of $\bar{y}_{COMP}$ is (Singh and Horn, 2000):

(i) $B(\bar{y}_{COM}) = (1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(C_X^2 + \rho C_Y C_X\right)$ (7)

(ii) $M(\bar{y}_{COM})$

$= \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left[(1-\alpha)^2 C_X^2 - 2(1-\alpha)\rho C_Y C_X\right]$ (8)

(iii) For optimum $\alpha = \left(1 - \rho\frac{C_Y}{C_X}\right)$, the minimum m.s.e. of $\bar{y}_{COMP}$ is given by the expression

$$M(\bar{y}_{COM})_{\min} = \left[\left(\frac{1}{r} - \frac{1}{N}\right) - \left(\frac{1}{r} - \frac{1}{n}\right)\rho^2\right] S_Y^2 \tag{9}$$

Ahmed methods

For case where y_{ji} denotes i^{th} available observation for the j^{th} imputation method, Ahmed et al. (2006) suggested:

$$\text{(A) } y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)}\left[n\bar{y}_r\left(\dfrac{\bar{X}}{\bar{x}_n}\right)^{\beta_1} - r\bar{y}_r\right] & if \quad i \in R^C \end{cases} \tag{10}$$

Under this, the point estimator of $\bar{Y}$ is

$$t_1 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} \tag{11}$$

(B)
$$y_{2i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{12}$$

Under this, the point estimator of $\bar{Y}$ is

$$t_2 = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} \tag{13}$$

(C)
$$y_{3i} = \begin{cases} y_i & if \quad i \in R \\ \dfrac{1}{(n-r)} \left[n\bar{y}_r \left(\dfrac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \tag{14}$$

Under this, the point estimator of $\bar{Y}$ is

$$t_3 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} \tag{15}$$

Terms β_1, β_2 and β_3 are suitably chosen constants, so as to keep the variance minimum.
As special cases:

When $\beta_3 = 1$, then

$$t_{Ratio} = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right) \tag{16}$$

and when $\beta_3 = -1$, then

$$t_{\mathrm{Product}} = \bar{y}_r \left(\frac{\bar{x}_r}{\bar{X}} \right) \tag{17}$$

This is natural analogue of ratio estimator called the product estimator used when an auxiliary variate *x* has negative correlation with *y*.

Factor-type methods of imputation

Shukla and Thakur (2008) have suggested F-T imputation procedure. For this case:

(D)
$$(y_{FT1})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)} \left[n\phi_1(k) - r \right] & if \quad i \in R^C \end{cases} \tag{15}$$

(E)
$$(y_{FT2})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)} \left[n\phi_2(k) - r \right] & if \quad i \in R^C \end{cases} \tag{16}$$

(F)
$$(y_{FT3})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)} \left[n\phi_3(k) - r \right] & if \quad i \in R^C \end{cases} \tag{17}$$

Where

$$\phi_1(k) = \left[\frac{(A+C)\bar{X} + fB\bar{x}_n}{(A+fB)\bar{X} + C\bar{x}_n} \right];$$

$$\phi_2(k) = \left[\frac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r} \right];$$

$$\phi_3(k) = \left[\frac{(A+C)\bar{X} + fB\bar{x}_r}{(A+fB)\bar{X} + C\bar{x}_r} \right];$$

$$A = (k-1)(k-2); \quad B = (k-1)(k-4);$$

$$C = (k-2)(k-3)(k-4); \quad f = \frac{n}{N} \; 0 \le \mathrm{k} \le \infty$$

Under Equations 15, 16 and 17, point estimators of $\bar{Y}$ is

$$\left.\begin{aligned} T_{FT1} &= \bar{y}_r \phi_1(k) \\ T_{FT2} &= \bar{y}_r \phi_2(k) \\ T_{FT3} &= \bar{y}_r \phi_3(k) \end{aligned}\right\} \tag{18}$$

As special cases, when $k = 1, \beta_l = 1$ then $T_{FTl} = t_l$

When $k=2,\beta_l=-1$ then $T_{FTl}=t_l$

When
$k=4,\beta_l=0$ then $T_{FTl}=t_l=\bar{y}_r;\qquad (l=1,2,3)$

PROPOSED ESTIMATOR

Using Singh and Horn (2000), F-T compromised imputation (FTCI) estimator is $(j=1,2,3)$:

$$(y_{\bullet i})_j=\begin{cases}\left(\frac{kn}{r}\right)y_i+(1-k)\phi_j(k) & \text{if } i\in R\\ (1-k)\phi_j(k) & \text{if } i\in R^C\end{cases} \quad (19)$$

Where, $\phi_1(k)=\bar{y}_r\left[\frac{(A+C)\bar{X}+fB\bar{x}_n}{(A+fB)\bar{X}+C\bar{x}_n}\right];$

$\phi_2(k)=\bar{y}_r\left[\frac{(A+C)\bar{x}_n+fB\bar{x}_r}{(A+fB)\bar{x}_n+C\bar{x}_r}\right];$

$\phi_3(k)=\bar{y}_r\left[\frac{(A+C)\bar{X}+fB\bar{x}_r}{(A+fB)\bar{X}+C\bar{x}_r}\right];$

$A=(k-1)(k-2);\ \ B=(k-1)(k-4);$

$C=(k-2)(k-3)(k-4);\ \ 0\le k\le\infty.$

Remark 1

Singh et al. (2001) have suggested a hybrid of calibration and imputation method in the presence of random non-response in survey sampling. Estimation of the population mean associated with the proposed hybrid method remains unbiased under a design based approach. Liu et al. (2005, 2006) have presented new imputation methods developed for the treatment of missing data, which remove the bias of usual ratio imputation. Singh (2009) has a useful contribution in the area of imputation techniques for missing data by proposing a new method of imputation in survey sampling.

Remark 2

In the contribution of Singh and Horn (2000), a point is raised by Singh and Deo (2003) that the adjustment in responding units y_i is not allowed. The present paper also bears this point. Some arguments in favour of proposed techniques are in Remark 3.

Remark 3

The available information y_i in sample of size *n* may be affected by some serious non-sampling errors like:

(i) Biasness in recording by investigator,
(ii) Under estimation or over estimation in reporting facts by respondents,
(iii) Partial unwillingness or lack of interest in answering question by respondent,
(iv) Deliberate miss reporting by respondents,
(v) False response/answer recording etc.

If available data is affected by these non-sampling errors, then there in need to also adjust available y_i values. Sometimes, auxiliary information correlated to main variable *Y* is almost error free, for example, in an economic survey of expenditure (*Y*) of military soldiers, the income (*X*) record is correctly available through salary pay roll but expenditure in houses may be affected by serious non-sampling errors even if response is made. So, one can adjust partially the available expenditure data by the auxiliary income data to get better picture of available sample. There are many smoothing statistical techniques available in the literature useful over available sample data. Moreover, it is a well proved fact that sample mean of available data is also highly affected by extreme observations, outliers which need to be tackled by correlated information. In light of these, one can also think of adjusting the responding units. The estimator of Singh and Horn (2000) has dual property that it smoothens the outliers in available data as well as imputes the missing observations.

Remark 4

The proposed estimator is in the setup of SRSWOR. One can think of extending the same in the general setup of sampling design with multivariate structure.

Properties of $\phi_j(k)$

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

$\phi_1(1)=\bar{y}_r\frac{\bar{X}}{\bar{x}_n};\quad \phi_2(1)=\bar{y}_r\frac{\bar{x}_n}{\bar{X}};\quad \phi_3(1)=\bar{y}_r\frac{\bar{X}}{\bar{x}_r}$

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

$\phi_1(2)=\bar{y}_r\frac{\bar{x}_n}{\bar{X}};\ \phi_2(2)=\bar{y}_r\frac{\bar{X}}{\bar{x}_n};\qquad \phi_3(2)=\bar{y}_r\frac{\bar{X}}{\bar{x}_r}$

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

Table 1. Some special cases of FTCI.

k	Estimators		
	$\left(\bar{y}_{FTCI}\right)_1$	$\left(\bar{y}_{FTCI}\right)_2$	$\left(\bar{y}_{FTCI}\right)_3$
$k = 1$	$\bar{y}_r$	$\bar{y}_r$	$\bar{y}_r$
$k = 2$	$\bar{y}_r\left(2-\frac{\bar{x}_n}{\bar{X}}\right)$	$\bar{y}_r\left(2-\frac{\bar{x}_r}{\bar{x}_n}\right)$	$\bar{y}_r\left(2-\frac{\bar{x}_r}{\bar{X}}\right)$
$k = 3$	$\bar{y}_r\left(3-\frac{2\left(\bar{X}-f\bar{x}_n\right)}{(1-f)\bar{X}}\right)$	$\bar{y}_r\left(3-\frac{2\left(\bar{x}_n-f\bar{x}_r\right)}{(1-f)\bar{x}_n}\right)$	$\bar{y}_r\left(3-\frac{2\left(\bar{X}-f\bar{x}_r\right)}{(1-f)\bar{X}}\right)$
$k = 4$	$\bar{y}_r$	$\bar{y}_r$	$\bar{y}_r$

$$\phi_1(3)=\bar{y}_r\left[\frac{\bar{X}-f\bar{x}_n}{(1-f)\bar{X}}\right];\ \phi_2(3)=\bar{y}_r\left[\frac{\bar{x}_n-f\bar{x}_r}{(1-f)\bar{x}_n}\right];$$

$$\phi_3(3)=\bar{y}_r\left[\frac{\bar{X}-f\bar{x}_r}{(1-f)\bar{X}}\right]$$

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

$$\phi_1(4)=\phi_2(4)=\phi_3(4)=\bar{y}_r \quad \text{(Table 1)}$$

Theorem 1

The point estimator of $\bar{Y}$ under FTCI is:

$$\left(\bar{y}_{FTCI}\right)_j=\left[k\bar{y}_r+(1-k)\phi_j(k)\right],\quad j=1,2,3 \qquad (20)$$

Proof

$$\left(\bar{y}_{FTCI}\right)_j=\left(\bar{y}_S\right)_j=\frac{1}{n}\sum_{i\in S}\left(y_{\bullet i}\right)_j$$

$$=\frac{1}{n}\left[\sum_{i\in R}\left(y_{\bullet i}\right)_j+\sum_{i\in R^C}\left(y_{\bullet i}\right)_j\right]$$

$$=\frac{1}{n}\left[\sum_{i\in R}\left\{\left(\frac{kn}{r}\right)y_i+(1-k)\phi_j(k)\right\}+\sum_{i\in R^C}(1-k)\phi_j(k)\right]$$

$$=\frac{1}{n}\left[\frac{kn}{r}\sum_{i\in R}y_i+(1-k)\left\{\sum_{i\in R}\phi_j(k)+\sum_{i\in R^C}\phi_j(k)\right\}\right]$$

$$=\frac{1}{n}\left[kn\bar{y}_r+(1-k)\sum_{i\in S}\phi_j(k)\right]$$

$$=\frac{1}{n}\left[kn\bar{y}_r+n(1-k)\phi_j(k)\right]$$

$$\left(\bar{y}_{FTCI}\right)_j=\left[k\bar{y}_r+(1-k)\phi_j(k)\right]$$

BIAS AND MEAN SQUARED ERROR

Let $B(.)$ and $M(.)$ denote, respectively the bias and m.s.e. of an estimator under a given sampling design. Under the large sample approximations, let

$$\bar{y}_r=\bar{Y}(1+\varepsilon),\ \bar{x}_r=\bar{X}(1+\delta),\ \bar{x}_n=\bar{X}(1+\eta)$$

Using the concept of two-phase sampling following, Cochran (2005), Heitjan and Basu (1996), and Rao and Sitter (1995), and with the mechanism of MCAR, for given r and n, we have

$$E(\varepsilon)=E(\delta)=E(\eta)=0,$$
$$E(\varepsilon^2)=M_1C_Y^2,\ \ E(\delta^2)=M_1C_X^2,\ \ E(\eta^2)=M_2C_X^2;$$

$$E(\varepsilon\delta)=M_1\rho C_YC_X,\ \ E(\delta\eta)=M_2C_X^2,\ \ E(\varepsilon\eta)=M_2\rho C_YC_X$$

Where

$$M_1 = \left(\frac{1}{r} - \frac{1}{N}\right), \quad M_2 = \left(\frac{1}{n} - \frac{1}{N}\right), \quad M = \left(M_1 - M_2\right)$$

Remark 5

Some specific symbols

$$\alpha_1 = \frac{fB}{A + fB + C}, \alpha_2 = \frac{C}{A + fB + C},$$
$$\alpha_3 = \frac{A + C}{A + fB + C}, \alpha_4 = \frac{A + fB}{A + fB + C},$$

$$P = \left(\alpha_1\delta + \alpha_3\eta\right), Q = \left(\alpha_2\delta + \alpha_4\eta\right),$$
$$\left(\alpha_1 + \alpha_3\right) = \left(\alpha_2 + \alpha_4\right) = 1,$$

$$\alpha = \left(\alpha_1 - \alpha_2\right) = -\left(\alpha_3 - \alpha_4\right), \theta = \left(1 - k\right)\alpha,$$
$$V = \rho\frac{C_Y}{C_X}.$$

Theorem 2

$[a_1]$: The estimator $\left(\bar{y}_{FTCI}\right)_1$ in terms of ε, δ and η upto first order of approximation is:

$$\left(\bar{y}_{FTCI}\right)_1 = \bar{Y}\left[1 + \varepsilon + \theta\left(\eta + \varepsilon\eta - \alpha_2\eta^2\right)\right] \quad (21)$$

Proof $\left(\bar{y}_{FTCI}\right)_1 = k\bar{y}_r + \left(1 - k\right)\phi_1\left(k\right)$

$$= \left[k\bar{y}_r + \left(1 - k\right)\bar{y}_r\frac{\left(A + C\right)\bar{X} + fB\bar{x}_n}{\left(A + fB\right)\bar{X} + C\bar{x}_n}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\frac{\left(A + C\right)\bar{X} + fB\bar{X}\left(1 + \eta\right)}{\left(A + fB\right)\bar{X} + C\bar{X}\left(1 + \eta\right)}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\frac{1 + \alpha_1\eta}{1 + \alpha_2\eta}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\left(1 + \alpha_1\eta\right)\left(1 + \alpha_2\eta\right)^{-1}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\left(1 + \alpha_1\eta\right)\left(1 - \alpha_2\eta + \alpha_2^2\eta^2 - \alpha_2^3\eta^3 + \ldots\right)\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\left(1 - \alpha_2\eta + \alpha_2^2\eta^2 + \alpha_1\eta - \alpha_1\alpha_2\eta^2\right)\right]$$

(by first order only)

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + \left(1 - k\right)\left\{1 + \left(\alpha_1 - \alpha_2\right)\eta - \left(\alpha_1 - \alpha_2\right)\alpha_2\eta^2\right\}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[k + 1 - k + \left(1 - k\right)\left\{\left(\alpha_1 - \alpha_2\right)\eta - \left(\alpha_1 - \alpha_2\right)\alpha_2\eta^2\right\}\right]$$

$$= \bar{Y}\left(1 + \varepsilon\right)\left[1 + \left(1 - k\right)\left(\alpha_1 - \alpha_2\right)\left(\eta - \alpha_2\eta^2\right)\right]$$

$$= \bar{Y}\left[1 + \varepsilon + \left(1 - k\right)\left(\alpha_1 - \alpha_2\right)\left(\eta - \alpha_2\eta^2\right)\left(1 + \varepsilon\right)\right]$$

$$\left(\bar{y}_{FTCI}\right)_1 = \bar{Y}\left[1 + \varepsilon + \theta\left(\eta + \varepsilon\eta - \alpha_2\eta^2\right)\right]$$

$[a_2]$: Bias of $\left(\bar{y}_{FTCI}\right)_1$ is up to first order approximation is:

$$B\left[\left(\bar{y}_{FTCI}\right)_1\right] = -\bar{Y}\theta M_2\left(\alpha_2 C_X^2 - \rho C_Y C_X\right) \quad (22)$$

Proof $B\left[\left(\bar{y}_{FTCI}\right)_1\right] = E\left[\left(\bar{y}_{FTCI}\right)_1 - \bar{Y}\right]$

Using Equation 21 and taking expectations

$$B\left[\left(\bar{y}_{FTCI}\right)_1\right] = \left[\bar{Y}\left\{1 + \varepsilon + \theta\left(\eta + \varepsilon\eta - \alpha_2\eta^2\right)\right\} - \bar{Y}\right]$$

$$= -\bar{Y}\theta M_2\left(\alpha_2 C_X^2 - \rho C_Y C_X\right)$$

$[a_3]$: The m.s.e of $\left(\bar{y}_{FTCI}\right)_1$ up to some level of approximations:

$$M\left[\left(\bar{y}_{FTCI}\right)_1\right] = \bar{Y}^2\left[M_1 C_Y^2 + M_2\left\{\theta^2 C_X^2 + 2\theta\rho C_Y C_X\right\}\right] \quad (23)$$

Proof $M\left[\left(\bar{y}_{FTCI}\right)_1\right] = E\left[\left(\bar{y}_{FTCI}\right)_1 - \bar{Y}\right]^2$

$$= E\left[\bar{Y}\left\{1 + \varepsilon + \theta\left(\eta + \varepsilon\eta - \alpha_2\eta^2\right)\right\} - \bar{Y}\right]^2$$

$$= \bar{Y}^2 E\left[\varepsilon + \theta\left(\eta + \varepsilon\eta - \alpha_2\eta^2\right)\right]^2$$

$$= \bar{Y}^2 E\left[\varepsilon + \theta\eta\right]^2$$

$= \bar{Y}^2\left[M_1 C_Y^2 + M_2\left\{\theta^2 C_X^2 + 2\theta\rho C_Y C_X\right\}\right]$

[a_4]: Minimum m.s.e. of the estimator $\left(\bar{y}_{FTCI}\right)_1$ holds at $\theta = -\rho\frac{C_Y}{C_X} = -V$ with expression: $M\left[\left(\bar{y}_{FTCI}\right)_1\right]_{\min}$

$= \left(M_1 - M_2\rho^2\right)S_Y^2$ (24)

Proof $\frac{d}{d\theta}\left[M\left(\bar{y}_{FTCI}\right)_1\right] = 0$

$\Rightarrow \theta = -\rho\frac{C_Y}{C_X}$

Substituting $\theta = -\rho\frac{C_Y}{C_X}$ in Equation 23, we get

$M\left[\left(\bar{y}_{FTCI}\right)_1\right]_{\min} = \left(M_1 - M_2\rho^2\right)S_Y^2$

Theorem 3

[a_5]: The estimator $\left(\bar{y}_{FTCI}\right)_2$ in terms of ε, δ and η up to first order of approximation is:

$\left(\bar{y}_{FTCI}\right)_2 =$
$\bar{Y}\left[1 + \varepsilon + \theta\left\{\delta - \eta + \varepsilon\delta - \varepsilon\eta + (\alpha_2 - \alpha_4)\delta\eta - \alpha_2\delta^2 + \alpha_4\eta^2\right\}\right]$ (25)

[a_6]: Bias of $\left(\bar{y}_{FTCI}\right)_2$ is

$B\left[\left(\bar{y}_{FTCI}\right)_2\right] = -\bar{Y}\theta M\left(\alpha_2 C_X^2 - \rho C_Y C_X\right)$ (26)

[a_7]: The m.s.e of $\left[\left(\bar{y}_{FTCI}\right)_2\right]$ is

$M\left[\left(\bar{y}_{FTCI}\right)_2\right] = \bar{Y}^2\left[M_1 C_Y^2 + M\left\{\theta^2 C_X^2 + 2\theta\rho C_Y C_X\right\}\right]$ (27)

[a_8]: Minimum m.s.e. of the estimator $\left[\left(\bar{y}_{FTCI}\right)_2\right]$ is on $\theta = \left(-\rho\frac{C_Y}{C_X}\right) = -V$ and expression is

$M\left[\left(\bar{y}_{FTCI}\right)_2\right]_{\min} = \left(M_1 - M\rho^2\right)S_Y^2$ (28)

Theorem 4

[a_9]: The estimator $\left(\bar{y}_{FTCI}\right)_3$ in terms of ε, δ and η upto first order of approximation is:

$\left(\bar{y}_{FTCI}\right)_3 = \bar{Y}\left[1 + \varepsilon + \theta\left(\delta + \varepsilon\delta - \alpha_2\delta^2\right)\right]$ (29)

[a_{10}]: Bias of $\left(\bar{y}_{FTCI}\right)_3$ is

$B\left[\left(\bar{y}_{FTCI}\right)_3\right] = -\bar{Y}\theta M_1\left(\alpha_2 C_X^2 - \rho C_Y C_X\right)$ (30)

[a_{11}]: The m.s.e of $\left(\bar{y}_{FTCI}\right)_3$ is

$M\left[\left(\bar{y}_{FTCI}\right)_3\right] = \bar{Y}^2 M_1\left[C_Y^2 + \theta^2 C_X^2 + 2\theta\rho C_Y C_X\right]$ (31)

[a_{12}]: Minimum m.s.e. of the estimator $\left(\bar{y}_{FTCI}\right)_3$ while $\theta = \left(-\rho\frac{C_Y}{C_X}\right) = -V$ holds and expression is given by

$M\left[\left(\bar{y}_{FTCI}\right)_3\right]_{\min} = M_1 S_Y^2\left(1 - \rho^2\right)$ (32)

Remark 6

For minimum m.s.e., we have

$\theta = -V = -\rho\frac{C_Y}{C_X} \Rightarrow (1-k)(\alpha_1 - \alpha_2) = -V$

$\Rightarrow (1-k)\frac{(fB - C)}{(A + fB + C)} = -V$

On simplification, we get polynomial of degree four:

$k^4 - (f - V)k^3 - [(4f + 15) - (f - 8)V]k^2 + [(f - 10) - (5f - 23)V]k$

$+ [(4f + 24) + (4f - 22)V] = 0$ (33)

Remark 7

Reddy (1978) has shown that quantity $V = \rho\frac{C_Y}{C_X}$ is stable over moderate length time period and could be initially known or guessed by past data. Therefore, pair (*f*, *V*) be treated as known and Equation 33 generates

maximum of four roots (some may imaginary) on which optimum level of m.s.e. will be attained.

Remark 8

The Equation 33 has only unknown k in the power of order four, other V and f by Reddy (1978) are known in advance. We can solve Equation 33 for obtaining optimal estimator in the suggested class.

COMPARISON OF THE ESTIMATORS

(1) Let $D_1 = M\left[\left(\bar{y}_{FTCI}\right)_1\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_2\right]_{\min}$

$= \left(M_1 - M_2\rho^2\right)S_Y^2 - \left(M_1 - M\rho^2\right)S_Y^2$

$= \left(M - M_2\right)\rho^2 S_Y^2$ (34)

$\left(\bar{y}_{FTCI}\right)_2$ is better than $\left(\bar{y}_{FTCI}\right)_1$ if

$$D_1 > 0 \Rightarrow r < \frac{n}{(2-f)}$$

If population *N* is large, *f* is small then working rule is; $D_1 > 0$ if $r < \left(\frac{n}{2}\right)$ which usually happens and high chance for $D_1 > 0$.

(2) Let $D_2 = M\left[\left(\bar{y}_{FTCI}\right)_1\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_3\right]_{\min}$

$$D_2 > 0 \Rightarrow M\rho^2 S_Y^2 > 0 \quad (35)$$

Which is always positive. So, $\left(\bar{y}_{FTCI}\right)_3$ is always better than $\left(\bar{y}_{FTCI}\right)_1$ until *n* = *r*

(3) Let $D_3 = M\left[\left(\bar{y}_{FTCI}\right)_2\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_3\right]_{\min}$

$$D_3 > 0 \quad \Rightarrow \quad N > n \quad (36)$$

Which is always true. So, $\left(\bar{y}_{FTCI}\right)_3$ is always better than $\left(\bar{y}_{FTCI}\right)_2$ until *n* = *N*. But $\left(\bar{y}_{FTCI}\right)_2$ is better than $\left(\bar{y}_{FTCI}\right)_1$ therefore, estimation strategy $\left(\bar{y}_{FTCI}\right)_3$ is most preferable over other two.

(4) $D_4 = M\left[\bar{y}_{COMP}\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_1\right]_{\min}$

$$D_4 > 0 \Rightarrow \quad r > \frac{n}{2-f} \Rightarrow r > \frac{n}{2} \text{ for } 0 < f < 1 \quad (37)$$

This generally does not happen until huge non-response occurs.

(5) $D_5 = M\left[\bar{y}_{COMP}\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_2\right]_{\min}$

$$= \left(M_1 - M\rho^2\right)S_Y^2 - \left(M_1 - M\rho^2\right)S_Y^2 = 0 \quad (38)$$

Therefore, both are equally efficient.

(6) $D_6 = M\left[\bar{y}_{COMP}\right]_{\min} - M\left[\left(\bar{y}_{FTCI}\right)_3\right]_{\min}$

$$D_6 > 0 \quad \Rightarrow M_2\rho^2 S_Y^2 > 0 \quad (39)$$

which is always true.

It seems $\left(\bar{y}_{FTCI}\right)_1$ and $\left(\bar{y}_{FTCI}\right)_3$ are better offer over compromised imputation procedure and $\left(\bar{y}_{FTCI}\right)_3$ is best.

Remark 9

One can think of comparing the proposed with other existing like Hot-deck, Cold-deck, Nearest neighbour, and Regression imputation methods. But all these are based on replacing by single unit to single missing unit of sample. The proposed is on replacing single unit by an average of random sub-group units of auxiliary variate. The comparison environment in proposed case is different.

Remark 10

It may be interesting to derive the Horvitz-Thompson estimator based on imputation data and compare with the proposed. It is an open problem to extend the content further.

EMPIRICAL STUDY

The attached Appendix A has generated artificial population of size *N* = 200 containing values of main variable *Y* and auxiliary variable *X*. Parameter of these are given as follows:

$\bar{Y}$ = 42.485; $\bar{X}$ = 18.515; S_Y^2 = 199.0598; S_X^2 = 48.5375; ρ = 0.8652; C_X = 0.3763; C_Y = 0.3321.

Using random sample SRSWOR of size, *n* = 30; *r* = 22;

Table 2. Bias and optimum m.s.e. at $k = k_i$ (i =1,2).

Estimator	Bias (.)	M(.)
$\bar{y}_{COMP}$	0.0150 (at α_{opt} = 0.2365)	6.2504 (at α_{opt} = 0.2365)
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_1}$	0.0079	3.8254
$\left[\left(\bar{y}_{FTCI}\right)_1\right]_{k_2}$	0.0109	3.8254
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_1}$	0.0034	6.2504
$\left[\left(\bar{y}_{FTCI}\right)_2\right]_{k_2}$	0.0046	6.2504
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_1}$	0.0113	2.0225
$\left[\left(\bar{y}_{FTCI}\right)_3\right]_{k_2}$	0.0156	2.0225

f = 0.15, α = 0.2365. Solving optimum condition $\theta = -V$ (Equation 33), the equation of power four in k provides only two real values k_1 = 0.8350; k_2 =4.1043. Rest other two roots appear imaginary (Table 2).

ALMOST UNBIASED *FTCI* ESTIMATOR:

If $B\left[\left(\bar{y}_{FTCI}\right)_1\right] = 0$

$\Rightarrow$ $-\bar{Y}\theta M_2\left(\alpha_2 C_X^2 - \rho C_Y C_X\right) = 0$ Using Equation 22

When $\theta = 0 \Rightarrow (1-k)\alpha = 0$ (40)

Case (i)

$(1-k) = 0 \quad \Rightarrow \quad k = 1 = k_1$

Case (ii)

$\alpha = 0 \quad \Rightarrow \quad \alpha_1 - \alpha_2 = 0$

$\Rightarrow \quad \frac{fB}{A + fB + C} - \frac{C}{A + fB + C} = 0$

$\Rightarrow \quad k = k_2^{'} = 4$ [since (k - 4) = 0]

$k = k_3^{'} = \frac{1}{2}\left[(5+f) + \sqrt{f^2 + 6f + 1}\right]$

and $k = k_4^{'} = \frac{1}{2}\left[(5+f) - \sqrt{f^2 + 6f + 1}\right]$

Case (iii)

When $\left(\alpha_2 C_X^2 - \rho C_Y C_X\right) = 0 \quad \Rightarrow \alpha_2 = \rho\frac{C_Y}{C_X} = V$

$\Rightarrow$ [$AV + fBV + (V-1)C$]= 0 (41)

Equation 41 is a cubic equation in k, solvable for fix f and guess value V, and one can obtain three values $k = k_1^{''}$; $k = k_2^{''}$ and $k = k_3^{''}$ to make the proposed estimator imputed unbiased to the first order of approximation.

At seven values $k_1^{'}, k_2^{'}, k_3^{'}, k_4^{'}, k_1^{''}, k_2^{''}$ and $k_3^{''}$, bias is almost zero or completely zero. Using data set in Appendix A, the k-values are compared in Table 3. The best choice is that having the lowest m.s.e.

DISCUSSION

The estimator $\bar{y}_{COMP}$ of Singh and Horn (2000) bears bias 0.0150, m.s.e. 6.2504 at $\alpha_{opt} = 0.2365$. But, bias is not at minimal level. Proposed estimators have shown two choices k_1, k_2 on a data set for known (f, V) at which m.s.e. is minimum and there are open options to choose that optimum k-value having the lowest bias. These two choices may extend to the maximum four on some data sets. So, FTCI estimators have an upper hand over Compromised estimator of Singh and Horn (2000) in

Table 3. Almost unbiased FTCI.

k-Values	$\left[\left(\bar{y}_{FTCI}\right)_1\right]$		$\left[\left(\bar{y}_{FTCI}\right)_2\right]$		$\left[\left(\bar{y}_{FTCI}\right)_3\right]$	
	***Bias* (.)**	***M* (.)**	***Bias* (.)**	***M* (.)**	***Bias* (.)**	***M* (.)**
$k_1^{'}$ = 1.0000	0	8.0424	0	8.0424	0	8.0424
$k_2^{'}$ = 4.0000	0	8.0424	0	8.0424	0	8.0424
$k_3^{'}$ = 3.2682	0	8.0424	0	8.0424	0	8.0424
$k_4^{'}$ = 1.8819	0	8.0424	0	8.0424	0	8.0424
$k_1^{''}$ = 1.5743	-0.00004	13.1265	-0.00004	10.1586	-0.00006	15.4300
$k_2^{''}$ = 2.3355	-0.0004	56.7904	-0.0001	28.8276	-0.0006	77.7691
$k_3^{''}$ = 8.8231	-0.0005	315.6901	-0.0002	139.5245	-0.0009	447.2762

terms of efficiency comparison. The $\left(\bar{y}_{FTCI}\right)_2$ is equally efficient to $\bar{y}_{COMP}$ in terms of m.s.e. but has unique option of lower bias. Moreover, estimator $\left(\bar{y}_{FTCI}\right)_3$ is more efficient than all others used herein. Another observation is that, FTCI could be made almost unbiased also over seven different choices of *k*-values, the best is that having lowest m.s.e. Thus, FTCI has wider prospects of superiority over Singh and Horn (2000). One more interesting feature is that FTCI contains only one parameter *k*, same as used by Singh and Horn (2000), but mixed up in mathematical structure in such a way so as to generate more efficient properties.

ACKNOWLEDGEMENT

Authors are thankful to the referee for the critical comments and useful suggestions which has improved the quality of the manuscript.

REFERENCES

Ahmed MS, Al-Titi O, Al-Rawi Z, Abu-Dayyeh W (2006). Estimation of a population mean using different imputation methods. Stat. Transit. 7(6):1247-1264.

Cochran WG (2005). Sampling Techniques, John Wiley and Sons, New York.

Heitjan DF, Basu S (1996). Distinguishing 'Missing at random' and 'missing completely at random'. Am. Stat. 50:207-213.

Liu L, Tu Y, Li Y, Zou G (2005, 2006). Imputation for missing data and variance estimation when auxiliary information is incomplete. Model Assist. Stat. Appl. pp. 83-94.

Rao JNK, Sitter RR (1995). Variance estimation under two-phase sampling with application to imputation for missing data. Biometrica 82:453-460.

Reddy VN (1978). A study on the use of prior knowledge on certain population parameters in estimation. Sankhya C40:29-37.

Rueda M, González S (2008). A new ratio-type imputation with random disturbance. Appl. Math. Lett. 21(9):978-982.

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

Shukla D, Thakur NS (2008). Estimation of mean with imputation of missing data using factor type estimator. Stat. Transit. 9(1):33-48.

Singh GN, Kumari P, Jong-Min K, Singh S (2010). Estimation of population mean using imputation techniques in sample survey. J. Korean Stat. Soc. 39(1):67-74.

Singh S (2009). A new method of imputation in survey sampling. Statistics 43(5):499-511.

Singh S, Deo B (2003). Imputing with power transformation. Stat. Pap. 44:555-579.

Singh S, Horn S (2000). Compromised imputation in survey sampling. Metrika 51:266-276.

Singh S, Horn S, Tracy DS (2001). Hybrid of calibration and imputation: Estimation of mean in survey sampling. Statistics 61:27-41.

Appendix A. Artificial 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