DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 10

5-2017

Effective Estimation Strategy of Finite Population Variance Using Multi-Auxiliary Variables in Double Sampling

Reba Maji
Sarojini Naidu College for Women, Kolkata, India, rebamaji09@gmail.com

G. N. Singh
Indian School of Mines, Dhanbad, India, gnsingh_ism@yahoo.com

Arnab Bandyopadhyay
Asansol Engineering College, Asansol, India, arnabbandyopadhyay4@gmail.com

Recommended Citation

Maji, R., Singh, G. N., & Bandyopadhyay, A. (2017). Effective estimation strategy of finite population variance using multi-auxiliary variables in double sampling. Journal of Modern Applied Statistical Methods, 16(1), 158-178. doi: 10.22237/jmasm/1493597340

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 158-178.
doi: 10.22237/jmasm/1493597340

ISSN 1538 − 9472

Effective Estimation Strategy of Finite Population Variance Using Multi-Auxiliary Variables in Double Sampling

Reba Maji
Sarojini Naidu College for Women
Kolkata, India

G. N. Singh
Indian School of Mines
Dhanbad, India

Arnab Bandyopadhyay
Asansol Engineering College
Asansol, India

Estimation of population variance in two-phase (double) sampling is considered using information on multiple auxiliary variables. An unbiased estimator is proposed and its properties are studied under two different structures. The superiority of the suggested estimator over some contemporary estimators of population variance was established through empirical studies from a natural and an artificially generated dataset.

Keywords: Double sampling, study variable, auxiliary variable, chain-type, regression, bias, variance, efficiency

Introduction

Auxiliary information plays a role in the planning, selection, and estimation stages of a sample survey. Sometimes information on several auxiliary variables may be readily available. For instance, to study the case of public health and welfare of a state or a country, the number of beds in different hospitals, doctors, and supporting staffs may be known, as well as the amount of funds available for medicine. When such information is lacking, it may be possible to obtain a large preliminary sample in which the auxiliary variable is measured, which is the premise of two-phase sampling, also known as double sampling. It is a powerful and cost-effective technique for obtaining reliable estimates in the first phase sample for the unknown parameters of the auxiliary variables.

Variation is an inherent phenomenon of nature. The use of auxiliary information in the estimation of population variance was considered by Das and Tripathi (1978), and extended by Isaki (1983), R. K. Singh (1983), Srivastava and

Reba Maji is an Assistant Professor of Mathematics. Email them at: rebamaji09@gmail.com. G. N. Singh is in the Department of Applied Mathematics. Email them at: gnsingh_ism@yahoo.com. Arnab Bandyopadhyay is in the Department of Mathematics. Email them at: arnabbandyopadhyay4@gmail.com.

Jhaji (1980), Upadhyaya and Singh (1983), Tripathi, Singh, and Upadhyaya (1988), Prasad and Singh (1990, 1992), S. Singh and Joarder (1998), R. Singh, Chauhan, Sawan, and Smarandache (2011), and Tailor and Sharma (2012), among others. However, most of these estimators of population variance are biased and based on the assumptions that the population mean or variance of the auxiliary variables are known, which may become a serious drawback in estimating population parameters in sample surveys.

Motivated with the above arguments, the objective of the present work is to propose an efficient and unbiased estimator of the population variance. The properties of the proposed estimator have been studied under two different structures of double sampling and results are supported with suitable simulation studies carried over six real datasets and an artificially generated data set.

Formulation of the Proposed Estimator

Consider a finite population $\mathrm{U} = \left(U_1, U_2, \ldots, U_N\right)$. Let y be the character under study and x_i, $i = 1, 2, \ldots, p$, be p (non-negative integer constant) auxiliary variables, taking values y_h and x_{i_h}, respectively, for the h^{th} unit. We define

$$S_y^2 = \frac{N}{N-1}\sigma_y^2 \text{ with } \sigma_y^2 = \frac{1}{N}\sum_{h=1}^{N}\left(y_h - \bar{Y}\right)^2$$

$$S_{x_i}^2 = \frac{N}{N-1}\sigma_{x_i}^2 \text{ with } \sigma_{x_i}^2 = \frac{1}{N}\sum_{h=1}^{N}\left(x_{i_h} - \bar{X}_i\right)^2$$

where

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{N} y_h \text{ and } \bar{X}_i = \frac{1}{N}\sum_{h=1}^{N} x_{i_h},\ i = 1, 2, \ldots, p$$

are the population means of y and x_i, respectively. For large N, $S_y^2 \cong \sigma_y^2$ and $S_{x_i}^2 \cong \sigma_{x_i}^2$ $\forall i = 1, 2, \ldots, p$.

Estimate the population variance S_y^2 of y when the population variances $S_{x_i}^2$ of x_i ($i = 1, 2, \ldots, p$) are unknown. When the variables y and the x_i are closely related but no information is available on the population variances $S_{x_i}^2$ of x_i, we seek to estimate S_y^2 from a sample S, obtained through a two-phase (or double) selection. In this sampling scheme, a first phase sample S' (S' ⊂ U) of size *n*' is drawn by a simple random sampling without replacement (SRSWOR) scheme from the entire population U and the auxiliary variables x_i are observed to furnish the estimates of $S_{x_i}^2$ ($i = 1, 2, \ldots, p$). A second phase sample S of size n ($n \le n'$) is drawn according to one of the following rules by the method of SRSWOR to observe the study variable *y*:

Case I: The second phase sample is drawn as a subsample of the first phase sample (i.e. S ⊂ S').

Case II: The second phase sample is drawn independently of the first phase sample.

Using one auxiliary variable *x*, Isaki (1983) suggested a ratio estimator for S_y^2 whose two-phase sampling version may be defined as

$$t_1 = s_y^2(n)\frac{s_x^2(n')}{s_x^2(n)} \tag{1}$$

where

$$s_y^2(n) = \frac{1}{n-1}\sum_{h=1}^{n}(y_h - \bar{y})^2,\ \bar{y} = \frac{1}{n}\sum_{h=1}^{n} y_h$$
$$s_x^2(n) = \frac{1}{n-1}\sum_{h=1}^{n}(x_h - \bar{x})^2,\ \bar{x} = \frac{1}{n}\sum_{h=1}^{n} x_h$$
$$s_x^2(n') = \frac{1}{n'-1}\sum_{h=1}^{n'}(x_h - \bar{x}')^2,\ \bar{x}' = \frac{1}{n'}\sum_{h=1}^{n'} x_h$$

The two-phase sampling version of the exponential estimator for S_y^2 proposed by R. Singh et al. (2011) is

$$t_2 = s_y^2(n)\exp\left[\frac{s_x^2(n') - s_x^2(n)}{s_x^2(n') + s_x^2(n)}\right] \tag{2}$$

Additional auxiliary variables which are highly correlated to the study variable *y* can be used to enhance the precision of the estimator. Motivated by

Chand (1975), consider a chain ratio-type estimator using information on two auxiliary variables *x* and *z* for estimating S_y^2 as

$$t_3 = s_y^2\left(n\right)\frac{s_x^2\left(n'\right)}{s_x^2\left(n\right)}\frac{S_z^2}{s_z^2\left(n'\right)} \tag{3}$$

A modified chain ratio-type estimator for S_y^2 suggested by H. P. Singh, Mathur, and Chandra (2009) is

$$t_4 = s_y^2\left(n\right)\frac{s_x^2\left(n'\right)}{s_x^2\left(n\right)}\left\{\frac{S_z^2+\beta_2\left(z\right)}{s_z^2\left(n'\right)+\beta_2\left(z\right)}\right\} \tag{4}$$

where

$$s_z^2\left(n'\right) = \frac{1}{n'-1}\sum_{h=1}^{n'}\left(z_h - \bar{z}'\right)^2,\ \bar{z}' = \frac{1}{n'}\sum_{h=1}^{n'} z_h$$
$$S_z^2 = \frac{1}{N-1}\sum_{h=1}^{N}\left(z_h - \bar{Z}\right)^2, \bar{Z} = \frac{1}{N}\sum_{h=1}^{N} z_h$$

and $\beta_2(z)$ is the known population coefficient of kurtosis of the variable *z*. There may be several auxiliary information, which if efficiently utilized can improve the precision of the estimates.

Motivated by the above, consider an unbiased estimator for the population variance S_y^2 of the study variable *y* using *p* (non-negative integer constant) auxiliary variables x_i (i = 1, 2,…, p) as

$$\mathrm{T_{RK}}\left(p\right) = K_1 s_y^2\left(n\right) + K_2 s_y^2\left(n\right)\sum_{i=1}^{p}\frac{s_{x_i}^2\left(n'\right)}{s_{x_i}^2\left(n\right)} + K_3 s_y^2\left(n\right)\sum_{i=1}^{p}\frac{s_{x_i}^2\left(n\right)}{s_{x_i}^2\left(n'\right)} \tag{5}$$

where

$$s_{x_i}^2\left(n\right) = \frac{1}{n-1}\sum_{h=1}^{n}\left(x_{i_h} - \bar{x}_i\right)^2, \bar{x}_i = \frac{1}{n}\sum_{h=1}^{n} x_{i_h}, i = 1, 2, \ldots, p$$

$$s_{x_i}^2(n') = \frac{1}{n'-1}\sum_{h=1}^{n'}\left(x_{i_h} - \overline{x}_i'\right)^2,\ \overline{x}_i' = \frac{1}{n'}\sum_{h=1}^{n'} x_{i_h},\ i = 1, 2, \ldots, p$$

and the K_i (i = 1, 2, 3) are real scalars suitably chosen such that

$$K_1 + pK_2 + pK_3 = 1 \tag{6}$$

Remark 1: The estimator T$_{\text{RK}}$(p) is proposed under the following conditions:

i. The sum ($K_1 + pK_2 + pK_3$) is one.
ii. The weights of the linear form are chose such that the approximate bias is zero.
iii. The approximate variance of T$_{\text{RK}}$(p) attains minimum.

Properties of the Estimator T$_{\text{RK}}$(*p*)

Noted from equation (5), the proposed estimator T$_{\text{RK}}$(p) is biased for S_y^2. Following Remark 1, it may be made unbiased for up to the first order of approximations. The variance V(.) up to the first order of approximations are derived under large sample approximations using the following transformations:

$$\begin{aligned}
&s_y^2(n) = S_y^2(1+e_0),\\
&\left.\begin{aligned} s_{x_i}^2(n) &= S_{x_i}^2(1+e_{1i})\\ s_{x_i}^2(n') &= S_{x_i}^2(1+e_{2i})\end{aligned}\right\}\text{for } i = 1, 2, \ldots, p\\
&\mathrm{E}(e_0) = \mathrm{E}(e_{1i}) = \mathrm{E}(e_{2i}) = 0 \quad \forall i = 1, \ldots, p
\end{aligned}$$

Under the above transformations, the estimator T$_{\text{RK}}$(p) takes the following form:

$$\begin{aligned}
\mathrm{T}_{\text{RK}}(p) = K_1S_y^2(1+e_0) + K_2S_y^2(1+e_0)\sum_{i=1}^{p}(1+e_{2i})(1+e_{1i})^{-1}\\
+ K_3S_y^2(1+e_0)\sum_{i=1}^{p}(1+e_{1i})(1+e_{2i})^{-1}
\end{aligned} \tag{7}$$

Hence, the bias and mean square error of the estimator T_{RK}(*p*) must be derived separately for Cases I and II of the two-phase sampling structure.

Case I

The second phase sample S is drawn as a subsample of the first phase sample S'. In this case, the expected values of the sample statistics are

$$\left.\begin{aligned}
&\mathrm{E}\left(e_0^2\right) = f_1 C_0^2,\ \mathrm{E}\left(e_{1i}^2\right) = f_1 C_i^2,\ \mathrm{E}\left(e_{2i}^2\right) = f_2 C_i^2 \\
&\mathrm{E}\left(e_0 e_{1i}\right) = f_1 \rho_{0i} C_0 C_i,\ \mathrm{E}\left(e_0 e_{2i}\right) = f_2 \rho_{0i} C_0 C_i,\ \mathrm{E}\left(e_{1i} e_{2i}\right) = f_2 C_i^2 \\
&\mathrm{E}\left(e_{1i} e_{1j}\right) = f_1 \rho_{ij} C_i C_j,\ \mathrm{E}\left(e_{2i} e_{2j}\right) = f_2 \rho_{ij} C_i C_j,\ \mathrm{E}\left(e_{1i} e_{2j}\right) = f_2 \rho_{ij} C_i C_j
\end{aligned}\right\} \qquad (8)$$

where, for integers *s*, *t* ≥ 0,

$$\mu(i)_{st} = \frac{1}{N}\sum_{h=1}^{N}\left\{\left(y_h - \bar{Y}\right)^s \left(x_{i_h} - \bar{X}_i\right)^t\right\}, \lambda(i)_{st} = \frac{\mu(i)_{st}}{\sqrt{\mu(i)_{20}^{s}\,\mu(i)_{02}^{t}}},$$

$$C_0 = \sqrt{\left(\lambda(i)_{40} - 1\right)}, C_i = \sqrt{\left(\lambda(i)_{04} - 1\right)}\ , \rho_{0i} = \frac{\left(\lambda(i)_{22} - 1\right)}{\sqrt{\left(\lambda(i)_{40} - 1\right)\left(\lambda(i)_{04} - 1\right)}}$$

$$\mu(ij)_{st} = \frac{1}{N}\sum_{h=1}^{N}\left\{\left(x_{i_h} - \bar{X}_i\right)^s \left(x_{j_h} - \bar{X}_j\right)^t\right\},$$

$$\lambda(ij)_{st} = \frac{\mu(ij)_{st}}{\sqrt{\mu(ij)_{20}^{s}\,\mu(ij)_{02}^{t}}}, \rho_{ij} = \frac{\left(\lambda(ij)_{22} - 1\right)}{\sqrt{\left(\lambda(ij)_{40} - 1\right)\left(\lambda(ij)_{04} - 1\right)}},$$

$$A_{0i} = \rho_{0i}\frac{C_0}{C_i},\ A_{ij} = \rho_{ij}\frac{C_i}{C_j}, \forall i, j = 1, 2, \ldots, p,$$

$$\text{and } f_1 = \frac{1}{n} - \frac{1}{N},\ f_2 = \frac{1}{n'} - \frac{1}{N},\ f_3 = \frac{1}{n} - \frac{1}{n'}$$

Expanding the right-hand side of equation (7) in terms of the *e* and using the results from equation (8), the expression of bias and mean square error of the estimator T_{RK}(*p*) using large sample approximations is

$$\begin{aligned}\mathrm{B}\left[\mathrm{T_{RK}}\left(p\right)\right] &= \mathrm{E}\left[\mathrm{T_{RK}}\left(p\right)-S_y^2\right] \\ &= K_2 f_3 S_y^2 \sum_{i=1}^{p} C_i^2 - \left(K_2 - K_3\right) f_3 S_y^2 \sum_{i=1}^{p} A_{0i} C_i^2\end{aligned} \tag{9}$$

$$\begin{aligned}\mathrm{M}\left[\mathrm{T_{RK}}\left(p\right)\right] &= \mathrm{E}\left[\mathrm{T_{RK}}\left(p\right)-S_y^2\right]^2 \\ &= S_y^4 \left[f_1 C_0^2 + \alpha^2 f_3 \left\{ \sum_{i=1}^{p} C_i^2 + \sum_{i\neq j=1}^{p} A_{ij} C_j^2 \right\} - 2\alpha f_3 \sum_{i=1}^{p} A_{0i} C_i^2 \right]\end{aligned} \tag{10}$$

where

$$\alpha = K_2 - K_3 \tag{11}$$

Minimization of the mean square error in equation (10) with respect to *α* yields its optimum value as

$$\alpha_{\text{opt}} = \frac{\left(\sum_{i=1}^{p} A_{0i} C_i^2\right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i\neq j=1}^{p} A_{0i} C_i^2} \tag{12}$$

Substituting the optimum value of *α* in equation (10) we obtain the minimum mean square error of T_{RK}(*p*) as

$$\text{Min.}\mathrm{M}\left[\mathrm{T_{RK}}\left(p\right)\right] = S_y^4 \left[f_1 C_0^2 - f_3 \frac{\left(\sum_{i=1}^{p} A_{0i} C_i^2\right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i\neq j=1}^{p} A_{0i} C_i^2} \right] \tag{13}$$

Further, from equations (11) and (12),

$$\alpha_{\text{opt}} = \left(K_2\right)_{\text{opt}} - \left(K_3\right)_{\text{opt}} = \frac{\sum_{i=1}^{p} A_{0i} C_i^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i\neq j=1}^{p} A_{0i} C_i^2} = R\left(\text{say}\right) \tag{14}$$

From equations (6) and (14), note that only two equations in three unknowns are not sufficient to find the unique values of the K_i (i = 1, 2, 3). In order to get unique values of the K_i, impose a linear restriction as

$$\mathrm{B}\left[\mathrm{T}_{\mathrm{RK}}\left(p\right)\right]=0 \tag{15}$$

Thus from equation (9),

$$K_2\sum_{i=1}^{p}\left(1-A_{0i}\right)C_i^2+K_3\sum_{i=1}^{p}A_{0i}C_i^2=0 \tag{16}$$

Equations (6), (14), and (16) can be written in matrix form as

$$\begin{pmatrix} 1 & p & p \\ 0 & 1 & -1 \\ 0 & \sum_{i=1}^{p}\left(1-A_{0i}\right)C_i^2 & \sum_{i=1}^{p}A_{0i}C_i^2 \end{pmatrix}\times\begin{pmatrix} K_1 \\ K_2 \\ K_3 \end{pmatrix}=\begin{pmatrix} 1 \\ R \\ 0 \end{pmatrix} \tag{17}$$

Solving (17), we get the unique values of the K_i as

$$\begin{aligned}
\left(K_1\right)_{\mathrm{opt}} &= 1-p\left[\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(2A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right] \\
\left(K_2\right)_{\mathrm{opt}} &= \frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)} \\
\left(K_3\right)_{\mathrm{opt}} &= \frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}
\end{aligned} \tag{18}$$

From equation (18), substituting the values of $(K_1)_{\mathrm{opt}}$, $(K_2)_{\mathrm{opt}}$, and $(K_3)_{\mathrm{opt}}$ in equation (5) yields the optimum unbiased estimator for S_y^2 as

$$
\begin{aligned}
\mathrm{T}_{\mathrm{RK}}\left(p\right)=&\left[1-p\left\{\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(2A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right\}\right]s_y^2\left(n\right)\\
&+\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2\left(n\right)\sum_{i=1}^{p}\frac{s_{x_i}^2\left(n'\right)}{s_{x_i}^2\left(n\right)}\\
&+\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}\left(A_{0i}-1\right)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2\left(n\right)\sum_{i=1}^{p}\frac{s_{x_i}^2\left(n\right)}{s_{x_i}^2\left(n'\right)}
\end{aligned}
\qquad (19)
$$

whose optimum variance up to the first degree of approximations is given by

$$
\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}\left(p\right)\right]_{\mathrm{opt}}=S_y^4\left[f_1C_0^2-f_3\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2}\right] \qquad (20)
$$

Case II

When the second-phase sample S is drawn independently of the first-phase sample S'. In this case, the following expected values of the sample statistics are

$$
\left.
\begin{aligned}
&\mathrm{E}\left(e_0^2\right)=f_1C_0^2,\ \mathrm{E}\left(e_{1i}^2\right)=f_1C_i^2,\ \mathrm{E}\left(e_{2i}^2\right)=f_2C_i^2\\
&\mathrm{E}\left(e_0e_{1i}\right)=f_1\rho_{0i}C_0C_i,\ \mathrm{E}\left(e_{1i}e_{1j}\right)=f_1\rho_{ij}C_iC_j,\ \mathrm{E}\left(e_{2i}e_{2j}\right)=f_2\rho_{ij}C_iC_j,\\
&\mathrm{E}\left(e_0e_{2i}\right)=\mathrm{E}\left(e_{1i}e_{2i}\right)=\mathrm{E}\left(e_{1i}e_{2j}\right)=0
\end{aligned}
\right\} \qquad (21)
$$

Proceeding as in Case I, the optimum unbiased estimator for S_y^2 is obtained as

$$\begin{aligned}\mathrm{T}_{\text{RK}}(p)=&\left[1-\frac{pf_1}{(f_1+f_2)^2}\left\{\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(2f_1A_{0i}-f_3)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}\right\}\right]s_y^2(n)\\&+\frac{f_1}{(f_1+f_2)^2}\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(f_2+f_1A_{0i})C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n')}{s_{x_i}^2(n)}\\&+\left(\frac{f_1}{f_1+f_2}\right)^2\frac{\sum_{i=1}^{p}A_{0i}C_i^2\sum_{i=1}^{p}(A_{0i}-1)C_i^2}{\sum_{i=1}^{p}C_i^2\left(\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2\right)}s_y^2(n)\sum_{i=1}^{p}\frac{s_{x_i}^2(n)}{s_{x_i}^2(n')}\end{aligned}\tag{22}$$

with optimum variance up-to first order of approximations as

$$\mathrm{V}\left[\mathrm{T}_{\text{RK}}(p)\right]_{\text{opt}}=S_y^4\left[f_1C_0^2-\frac{f_1^2}{f_1+f_2}\frac{\left(\sum_{i=1}^{p}A_{0i}C_i^2\right)^2}{\sum_{i=1}^{p}C_i^2+\sum_{i\neq j=1}^{p}A_{ij}C_j^2}\right]\tag{23}$$

Remark 2: It is to be noted from equation (18) that the unique value of the scalars K_i ($i = 1, 2, 3$) involved in estimator depend on unknown population parameters C_0, C_i, ρ_{0i}, and ρ_{ij} ($i, j = 1, 2, \ldots, p$). Thus, to make the estimator practicable, one has to use the guessed or estimated values of these unknown population parameters. Guessed values of population parameters can be obtained either from past data or experience gathered over time; see Murthy (1967), Reddy (1978), and Tracy, Singh, and Singh (1996). If the guessed values are not known then it is advisable to use their respective sample estimates as suggested by Upadhyaya and Singh (1999), H. P. Singh, Chandra, Joarder, and Singh (2007), and Gupta and Shabbir (2008). The minimum variance of the proposed class of estimators remains the same up to the first order of approximations, even if population parameters are replaced by their respective sample estimates.

Empirical Investigations

As p, the number of auxiliary variables, is a non-negative integer, therefore it is not practically possible to deal with the suggested estimator $\mathrm{T_{RK}}(p)$ in its general form to carry out the numerical illustrations. Thus, for empirical investigations, consider $\mathrm{T_{RK}}(p)$ with $p = 1$ and 2, where the suggested estimator $\mathrm{T_{RK}}(p)$ is superior to t_1 and t_2 for $\mathrm{T_{RK}}(1)$ (i.e. $p = 1$) and dominates t_3 and t_4 for $p = 2$. The performance of $\mathrm{T_{RK}}(1)$ is examined under two different cases of double sampling. The MSEs of the estimators t_1, t_2, t_3, and t_4 and the variance of $\mathrm{T_{RK}}(p)$ (for $p = 1, 2$) up to first order of approximations under both the Cases I and II of two-phase sampling set up are presented below.

Case I

$$\mathrm{M}(t_1) = S_y^4\left[f_1C_0^2 + f_3C_1^2\left(1-2A_{01}\right)\right]$$

$$\mathrm{M}(t_2) = S_y^4\left[f_1C_0^2 + \frac{1}{4} f_3C_1^2\left(1-4A_{01}\right)\right]$$

$$\mathrm{M}(t_3) = S_y^4\left[f_1C_0^2 + f_3C_1^2\left(1-2A_{01}\right) + f_2C_2^2\left(1-2A_{02}\right)\right]$$

$$\mathrm{M}(t_4) = S_y^4\left[f_1C_0^2 + f_3C_1^2\left(1-2A_{01}\right) + \theta f_2C_2^2\left(\theta-2A_{02}\right)\right]$$

$$\mathrm{V}\left[\mathrm{T_{RK}}\left(1\right)\right] = S_y^4\left[f_1C_0^2 - f_3A_{01}^2C_1^2\right]$$

$$\mathrm{V}\left[\mathrm{T_{RK}}\left(2\right)\right] = S_y^4\left[f_1C_0^2 - f_3\frac{\left(A_{01}C_1^2 + A_{02}C_2^2\right)^2}{C_1^2 + C_2^2 + A_{12}C_2^2 + A_{21}C_1^2}\right]$$

where

$$\theta = \frac{S_z^2}{S_z^2 + \beta_2\left(z\right)}$$

Case II

$$\mathrm{M}(t_1) = S_y^4\left[f_1 C_0^2 + (f_1 + f_2)C_1^2\left(1 - 2\frac{f_1}{f_1 + f_2}A_{01}\right)\right]$$

$$\mathrm{M}(t_2) = S_y^4\left[f_1 C_0^2 + \frac{1}{4}(f_1 + f_2)C_1^2\left(1 - 4\frac{f_1}{f_1 + f_2}A_{01}\right)\right]$$

$$\mathrm{M}(t_3) = S_y^4\left[f_1 C_0^2 + (f_1 + f_2)C_1^2\left(1 - 2\frac{f_1}{f_1 + f_2}A_{01}\right) + f_2 C_2^2\left(1 - 2A_{12}\right)\right]$$

$$\mathrm{M}(t_4) = S_y^4\left[f_1 C_0^2 + (f_1 + f_2)C_1^2\left(1 - 2\frac{f_1}{f_1 + f_2}A_{01}\right) + \theta f_2 C_2^2\left(\theta - 2A_{12}\right)\right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}\left(1\right)\right] = S_y^4\left[f_1 C_0^2 - \frac{f_1^2}{f_1 + f_2}A_{01}^2 C_1^2\right]$$

$$\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}\left(2\right)\right] = S_y^4\left[f_1 C_0^2 - \frac{f_1^2}{f_1 + f_2}\frac{\left(A_{01}C_1^2 + A_{02}C_2^2\right)^2}{C_1^2 + C_2^2 + A_{12}C_2^2 + A_{21}C_1^2}\right]$$

with θ as described above.

Numerical Illustration using Known Natural Populations

Six natural datasets were chosen to elucidate the efficacious performance of the proposed estimator $\mathrm{T}_{\mathrm{RK}}(p)$ (for $p = 1, 2$) over the estimators stated above. The source of the variables y, x, and z and the values of the various parameters are given below.

Population I: Source: Murthy (1967, p. 288).

y: Output.
x: Fixed capital.
z: Number of workers.

Population II: Source: Cochran (1977, p. 182).

y: Food cost.
x: Size of the family.
z: Income.

Population III: Source: Anderson (1958).

y: Head length of second son.
x: Head length of first son.
z: Head breadth of first son.

Population IV: Source: Wang and Chen (2012, p. 39).

y: Volume.
x: Diameter.
z: Height.

Population V: Source: Dobson (1990, p. 192).

y: Survival time.
x: White blood cell count.
z: White blood cell count at page number 74.

Population VI: Source: Sukhatme and Sukhatme (1970, p. 185).

y: Area (acres) under wheat in 1937.
x: Area (acres) under wheat in 1936.
z: Total cultivated area (acres) in 1931.

Table 1. Parametric values of different populations

Population	N	θ	C_0	C_1	C_2	ρ_{01}	ρ_{02}	ρ_{12}
I	80	0.999996	1.1255	1.6065	1.3662	0.7319	0.7940	0.9716
II	33	0.981200	1.0104	1.1780	1.0691	0.1341	0.4630	0.3905
III	25	0.953485	1.3512	1.4295	1.2853	0.5057	0.5683	0.4213
IV	31	0.943500	1.2634	1.2018	1.1962	0.7448	0.0547	0.3256
V	17	0.152800	0.8351	1.4049	1.0818	-0.0144	0.4468	0.5790
VI	34	1.000000	1.5959	1.5105	1.3200	0.6251	0.8007	0.5342

The values of various parameters obtained from above populations are presented in Table 1.

To obtain a tangible idea about the performance of the proposed estimator T_{RK}(p) (for p = 1, 2), the percent relative efficiencies (PREs) of T_{RK}(p) (for p = 1, 2) and other estimators were computed with respect to the sample variance $s_y^2(n)$, the natural estimator for S_y^2, for both the cases of two-phase sampling set up. The results are demonstrated in Tables 2 and 3.

The PRE of an estimator T_{RK}(p) with respect to sample variance estimator s_y^2 is defined as

$$\mathrm{PRE} = \frac{\mathrm{V}\left(s_y^2\right)}{\mathrm{V}\left[\mathrm{T}_{\mathrm{RK}}\left(p\right)\right]_{\mathrm{opt}}} \times 100 \tag{24}$$

Numerical Example using Artificially Generated Population

Three sets of independent random numbers were generated of size N (N = 100), x_k', y_k', and z_k' (k = 1, 2, 3,…, N) from a standard normal distribution via R. Motivated by the artificial data set generation techniques adopted by S. Singh and Deo (2003) and S. Singh, Joarder, and Tracy (2001), the following transformed variables of U were generated with the values of $\sigma_y^2 = 100$, μ_y = 40, $\sigma_x^2 = 225$, μ_x = 50, $\sigma_z^2 = 25$, and μ_z = 30 as

$$y_k = \mu_y + \sigma_y \left[\rho_{xy} x_k' + \left(\sqrt{1-\rho_{xy}^2}\right) y_k' \right], \; x_k = \mu_x + \sigma_y x_k',$$

$$\text{and } z_k = \mu_z + \sigma_z \left[\rho_{xz} x_k' + \left(\sqrt{1-\rho_{xz}^2}\right) z_k' \right]$$

PREs of different estimators for fixed and varying values of ρ_{xy} and ρ_{xz} are presented in Tables 3 and 4, respectively.

Table 2. PREs of different estimators

Population			**Percent Relative Efficiency**											
Pop. I			**Case I**						**Case II**					
N	***n'***	***n***	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
80	65	45	103.796	160.387	160.447	120.674	120.675	170.217	*	162.396	170.389	100.937	100.937	182.593
		40	104.167	170.012	170.085	116.933	116.933	182.212	*	171.764	177.066	101.913	101.913	191.046
		30	104.691	185.605	185.703	112.068	112.068	202.155	*	186.854	188.867	103.313	103.313	206.274
	50	35	102.853	139.961	139.996	131.523	131.523	145.523	*	142.380	157.539	*	*	166.643
		25	103.931	163.758	163.823	119.287	119.287	174.391	*	165.682	172.679	101.290	101.290	185.479
		20	104.341	174.910	174.991	115.265	115.265	188.407	*	176.515	180.635	102.376	102.376	195.613
Pop. II			**Case I**						**Case II**					
N	***n'***	***n***	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
33	25	12	*	*	101.492	*	*	111.007	*	*	101.545	*	*	111.432
		10	*	*	101.574	*	*	111.665	*	*	101.605	*	*	111.923
		8	*	*	101.642	*	*	112.224	*	*	101.66	*	*	112.369
	15	8	*	*	101.121	*	*	108.079	*	*	101.317	*	*	109.611
		6	*	*	101.337	*	*	109.768	*	*	101.441	*	*	110.595
		4	*	*	101.525	*	*	111.267	*	*	101.568	*	*	111.622
Pop. III			**Case I**						**Case II**					
N	***n'***	***n***	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
25	20	12	*	124.425	124.489	100.282	101.028	144.897	*	123.551	126.228	*	*	148.651
		10	*	127.01	127.083	*	*	150.529	*	126.351	128.074	*	*	152.734
		7	*	129.934	130.017	*	*	157.146	*	129.531	130.39	*	*	158.008
	15	8	*	121.231	121.286	102.2	103.257	138.205	*	120.107	124.172	*	*	144.22
		6	*	125.23	125.297	*	100.499	146.629	*	124.422	126.78	*	*	149.87
		4	*	128.665	128.743	*	*	154.24	*	128.149	129.352	*	*	155.623

Note: “*” indicates no gain, i.e., PRE is less than 100

Table 2, continued.

Population			Percent Relative Efficiency											
Pop. IV			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
31	17	12	132.59	130.114	136.282	*	*	113.144	104.343	157.393	157.471	*	*	118.941
		10	145.21	141.475	150.88	100.041	103.811	117.253	118.68	164.905	166.253	103.718	106.548	121.049
		8	157.601	152.472	165.527	116.059	119.744	120.88	133.876	171.675	175.826	119.654	122.393	123.179
	12	8	129.891	127.662	133.2	*	*	112.203	110.422	160.73	161.166	*	*	119.847
		6	146.535	142.658	152.432	101.654	105.424	117.659	120.25	165.656	167.229	105.331	108.158	121.274
		5	155.339	150.476	162.83	112.972	116.688	120.245	131.013	170.48	174.001	116.597	119.363	122.786

Pop. V			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
17	12	8	*	*	100.013	*	*	102.832	*	*	104.438	*	*	100.098
		7	*	*	100.015	*	*	103.197	*	*	104.721	*	*	100.104
		6	*	*	100.016	*	*	103.498	*	*	104.981	*	*	100.11
	10	7	*	*	100.011	*	*	102.281	*	*	104.067	*	*	100.09
		6	*	*	100.013	*	*	102.779	*	*	104.399	*	*	100.097
		5	*	*	100.015	*	*	103.197	*	*	104.721	*	*	100.104

Pop. VI			Case I						Case II					
N	*n'*	*n*	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
34	25	12	130.044	141.943	145.778	155.47	155.47	209.612	103.4	143.687	143.736	108.245	108.245	202.714
		10	132.338	145.463	149.733	151.613	151.613	223.762	112.559	147.867	148.505	116.48	116.48	219.25
		8	134.343	148.581	153.251	148.495	148.495	237.318	120.683	151.21	152.888	123.628	123.628	235.875
	15	7	123.927	132.792	135.581	167.614	167.614	177.625	*	138.968	139.116	*	*	188.047
		6	126.494	136.592	139.801	162.149	162.149	190.144	104.639	144.281	144.371	109.37	109.37	204.829
		4	131.394	144.008	148.096	153.161	153.161	217.773	115.766	149.225	150.218	119.319	119.319	225.573

Note: “*” indicates no gain, i.e., PRE is less than 100

Table 3. PREs of different estimators under artificially generated populations for ρ_{xy} = 0.7 and ρ_{xz} = 0.5

Artificial Population			Estimators						
			Case I						
N	***n'***	***n***	$s_y^2(n)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
100	80	55	100	*	108.8652	109.8181	*	*	108.0435
		45	100	*	110.2873	111.4091	*	*	109.3225
		40	100	*	110.8303	112.0177	*	*	109.8100
	70	50	100	*	107.1820	107.9408	*	*	106.5256
		40	100	*	109.1417	110.1271	*	*	108.2923
		30	100	*	110.5859	111.7437	*	*	109.5906
			Case II						
N	***n'***	***n***	$s_y^2(n)$	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
100	80	55	100	*	105.3662	110.9395	*	*	108.9443
		45	100	*	107.8582	111.9666	*	*	109.7676
		40	100	*	108.8234	112.4033	*	*	110.1168
	70	50	100	*	102.4825	109.9041	*	*	108.1116
		40	100	*	105.8466	111.1271	*	*	109.0948
		30	100	*	108.3879	112.2033	*	*	109.9569

Note: “*” indicates no gain, i.e., PRE is less than 100

Table 4. PREs of Different estimators for varying values of ρ_{xy} and ρ_{xz}

		Case I Estimators					
ρ_{xy}	ρ_{xz}	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
0.8	0.8	101.983	116.671	116.696	*	101.536	117.440
	0.6	126.096	121.277	127.007	109.818	115.747	118.407
	0.4	115.223	117.736	119.551	*	*	109.180
	0.2	*	119.547	119.551	*	*	111.733
0.5	0.8	*	*	100.349	*	*	100.390
	0.6	*	102.123	103.171	*	*	101.590
	0.4	*	*	100.159	*	*	100.227
	0.2	*	*	102.017	*	*	100.300
0.2	0.8	*	*	100.188	*	*	100.573
	0.6	*	*	100.033	*	*	100.025
	0.4	*	*	100.035	*	*	100.351
	0.2	*	*	100.289	*	*	101.920

Note: “*” indicates no gain, i.e., PRE is less than 100

Table 4, continued.

		Case II Estimators					
ρ_{xy}	ρ_{xz}	t_1	t_2	$T_{RK}(1)$	t_3	t_4	$T_{RK}(2)$
0.8	0.8	*	119.103	132.247	*	*	133.885
	0.6	106.535	156.666	156.841	*	101.806	136.045
	0.4	*	136.901	138.644	*	*	116.728
	0.2	*	118.359	138.644	*	*	121.799
0.5	0.8	*	*	100.596	*	*	100.666
	0.6	*	*	105.528	*	*	102.740
	0.4	*	*	100.272	*	*	100.387
	0.2	*	*	103.488	*	*	100.511
0.2	0.8	*	*	100.321	*	*	100.981
	0.6	*	*	100.057	*	*	100.044
	0.4	*	*	100.059	*	*	100.600
	0.2	*	*	100.493	*	*	103.318

Note: “*” indicates no gain, i.e., PRE is less than 100

Conclusion

For natural population datasets, Table 2 exhibits that, under different structures of two-phase sampling set up, our suggested estimator $T_{RK}(p)$ (for $p = 1$ and 2) is superior to the existing one under its respective optimality condition and also preferable in general situations. For fixed n' (first-phase sample size), the PRE of the proposed estimator is increasing with decreasing values of n (second-phase sample size), i.e. the smaller the second phase sample, the more efficiency in $T_{RK}(p)$ will be achieved, which reduces the cost of the survey.

For the artificially generated data set, the results compiled in Table 3 indicate the proposed methodology yielded impressive gains in efficiency over the existing methods, and same behavior in efficiency of $T_{RK}(p)$ was reflected, indicating the proposed methodology is cost-effective.

It can also be observed from Table 4 that if several populations are generated artificially for various combinations of values of ρ_{xy} and ρ_{xz}, our proposed methodology is always preferable over the existing one. The proposition of the estimator in the present study is justified as it unifies several desirable results including unbiased and efficient estimation strategy, and may be recommended for practical applications.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York, NY: John Wiley & Sons.

Chand, L. (1975). *Some ratio type estimators based on two or more auxiliary variables* (Unpublished doctoral dissertation). Iowa State University, Ames, IA. Retrieved from: http://lib.dr.iastate.edu/rtd/5190/

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: Wiley.

Das, A. K., & Tripathi, T. P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhyā, Series C, 40*(2), 139-148.

Dobson, A. J. (1990). *An introduction to generalized linear models* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9781420057683

Gupta, S., & Shabbir, J. (2008). On improvement in estimating the population mean in simple random sampling. *Journal of Applied Statistics, 35*(5), 559-566. doi: 10.1080/02664760701835839

Isaki, C. T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association, 78*(381), 117-123. doi: 10.2307/2287117

Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta, India: Statistical Publishing Society.

Prasad, B., & Singh, H. P. (1990). Some improved ratio type estimators of finite population variance in sample surveys. *Communications in Statistics – Theory and Methods, 19*(3), 1127-1139. doi: 10.1080/03610929008830251

Prasad, B., & Singh, H. P. (1992). Unbiased estimators of finite population variance using auxiliary information in sample surveys. *Communications in Statistics – Theory and Methods, 21*(5), 1367-1376. doi: 10.1080/03610929208830852

Reddy, V. N. (1978). A study on the use of prior knowledge on certain population parameters in estimation. *Sankhyā, Series C, 40*(1), 29-37.

Singh, H. P., Chandra, P., Joarder, A. H., & Singh, S. (2007). Family of estimators of mean, ratio and product of a finite population using random nonresponse. *TEST, 16*(3), 565-597. doi: 10.1007/s11749-006-0020-z

Singh, H. P., Mathur, N., & Chandra, P. (2009). A chain type estimator for population variance using two auxiliary variables in two-phase sampling. *Statistics in Transition New Series, 10*(1), 75-84.

Singh, R., Chauhan, P., Sawan, N., & Smarandache, F. (2011). Improved exponential estimator for population variance using two auxiliary variables. *Italian Journal of Pure and Applied Mathematics, 28-2011*, 101-108.

Singh, R. K. (1983). Estimation of finite population variance using ratio and product method of estimation. *Biometrical Journal, 25*(2), 193-200.

Singh, S., & Deo, B. (2003). Imputation by power transformation. *Statistical Papers, 44*(4), 555-579. doi: 10.1007/bf02926010

Singh, S., & Joarder, A. H. (1998). Estimation of finite population variance using random non-response in survey sampling. *Metrika, 47*(1), 241-249. doi: 10.1007/bf02742876

Singh, S., Joarder, A. H., & Tracy, D. S. (2001). Median estimation using double sampling. *Australian & New Zealand Journal of Statistics, 43*(1), 33-46. doi: 10.1111/1467-842x.00153

Srivastava, S. K., & Jhaji, H. S. (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā, Series C, 42*(12), 87-96.

Sukhatme, P. V., & Sukhatme, B. V. (1970). *Sampling theory of surveys with application*. Ames, IA: Iowa State University Press.

Tailor, R., & Sharma, B. (2012). Modified estimators of population variance in presence of auxiliary information. *Statistics in Transition New Series, 13*(1), 37-46.

Tracy, D. S., Singh, H. P., & Singh, R. (1996). An alternative to the ratio-cum-product estimator in sample surveys. *Journal of Statistical Planning and Inference, 53*(3), 375-387. doi: 10.1016/0378-3758(95)00136-0

Tripathi, T. P., Singh, H. P., & Upadhyaya, L. N. (1988). A generalized method of estimation in double sampling. *Journal of the Indian Statistical Association, 26*, 91-101.

Upadhyaya, L. N., & Singh, H. P. (1983). Use of auxiliary information in the estimation of population variance. *Mathematical Forum, 6*(2), 33-36.

Upadhyaya, L. N., & Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal, 41*(5), 627-636. doi: 10.1002/(SICI)1521-4036(199909)41:5<627::AID-BIMJ627>3.0.CO;2-W

Wang, Y. W., & Chen, H.-J. (2012). Use of percentiles and Z-scores in anthropometry. In V. R. Preedy (Ed.), *Hand book of anthropometry: Physical* *measures of human form in health and disease* (pp. 29-48). New York, NY: Springer. doi: 10.1007/978-1-4419-1788-1_2