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*

Follow this and additional works at: [http://digitalcommons.wayne.edu/jmasm](http://digitalcommons.wayne.edu/jmasm)  
Part of the [Applied Statistics Commons](https://digitalcommons.wayne.edu/appliedstatistics), [Social and Behavioral Sciences Commons](https://digitalcommons.wayne.edu/sociologybehavioral), and the [Statistical Theory Commons](https://digitalcommons.wayne.edu/statistical-theory)

Recommended Citation  

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in *Journal of Modern Applied Statistical Methods* by an authorized editor of DigitalCommons@WayneState.
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 \( U = \{U_1, U_2, \ldots, U_N\} \). 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 \( x_{ih} \), respectively, for the \( h \)th unit. We define

\[
S_y^2 = \frac{N}{N-1} \sigma_y^2 \quad \text{with} \quad \sigma_y^2 = \frac{1}{N} \sum_{h=1}^{N} (y_h - \bar{Y})^2
\]

\[
S_{x_i}^2 = \frac{N}{N-1} \sigma_{x_i}^2 \quad \text{with} \quad \sigma_{x_i}^2 = \frac{1}{N} \sum_{h=1}^{N} (x_{ih} - \bar{X}_i)^2
\]

where

\[
\bar{Y} = \frac{1}{N} \sum_{h=1}^{N} y_h \quad \text{and} \quad \bar{X}_i = \frac{1}{N} \sum_{h=1}^{N} x_{ih}, i = 1, 2, \ldots, p
\]

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

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)
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

selection. In this sampling scheme, a first phase sample $S'$ ($S' \subset U$) 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 $S^2_{x_i} (i = 1, 2, \ldots, p)$. A second phase sample $S$ of size $n (n \leq n')$ 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. $S \subset S'$).

**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 $S^2_y$ whose two-phase sampling version may be defined as

$$t_1 = s^2_y(n) \frac{s^2_y(n')} {s^2_x(n)}$$

(1)

where

$$s^2_y(n) = \frac{1}{n-1} \sum_{h=1}^{n}(y_h - \bar{y})^2, \bar{y} = \frac{1}{n} \sum_{h=1}^{n} y_h$$

$$s^2_x(n) = \frac{1}{n-1} \sum_{h=1}^{n}(x_h - \bar{x})^2, \bar{x} = \frac{1}{n} \sum_{h=1}^{n} x_h$$

$$s^2_x(n') = \frac{1}{n'-1} \sum_{h=1}^{n'}(x_h - \bar{x}^{'})^2, \bar{x}' = \frac{1}{n'} \sum_{h=1}^{n'} x_h$$

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

$$t_2 = s^2_y(n) \exp \left[ \frac{s^2_y(n') - s^2_x(n)} {s^2_x(n') + s^2_x(n)} \right]$$

(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

\[
t_3 = s_y^2(n) \left( \frac{s_x^2(n')}{s_x^2(n)} \right) \frac{S_z^2}{s_z^2(n')}
\]

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

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

where

\[
s_x^2(n') = \frac{1}{n'-1} \sum_{h=1}^{n'} (z_h - \bar{z}')^2, \quad \bar{z}' = \frac{1}{n'} \sum_{h=1}^{n'} z_h
\]

\[
S_z^2 = \frac{1}{N-1} \sum_{h=1}^{N} (z_h - \bar{Z})^2, \quad \bar{Z} = \frac{1}{N} \sum_{h=1}^{N} z_h
\]

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 \( x_i \) \((i = 1, 2, \ldots, p)\) as

\[
T_{RK}(p) = K_1 s_y^2(n) + K_2 s_y^2(n) \sum_{i=1}^{p} \frac{s_x^2(n')}{s_x^2(n)} + K_3 s_y^2(n) \sum_{i=1}^{p} \frac{s_x^2(n)}{s_x^2(n')}
\]

where

\[
s_x^2(n) = \frac{1}{n-1} \sum_{h=1}^{n} (x_{ih} - \bar{x}_i)^2, \quad \bar{x}_i = \frac{1}{n} \sum_{h=1}^{n} x_{ih}, \quad i = 1, 2, \ldots, p
\]
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

\[ s^2_{x_i}(n') = \frac{1}{n'-1} \sum_{h=1}^{n'} (x_{i_h} - \bar{x}_i')^2, \bar{x}_i' = \frac{1}{n'} \sum_{h=1}^{n'} x_{i_h}, i = 1, 2, \ldots, p \]

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

\[ K_1 + pK_2 + pK_3 = 1 \]  \( (6) \)

Remark 1: The estimator \( T_{RK}(p) \) 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 \( T_{RK}(p) \) attains minimum.

Properties of the Estimator \( T_{RK}(p) \)

Noted from equation (5), the proposed estimator \( T_{RK}(p) \) is biased for \( S^2_y \).
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:

\[
\begin{align*}
  s^2_y(n) &= S^2_y(1+e_0), \\
  s^2_{x_i}(n) &= S^2_{x_i}(1+e_{i1}) \quad \text{for } i = 1, 2, \ldots, p \\
  s^2_{x_i}(n') &= S^2_{x_i}(1+e_{i2}) \\
  E(e_0) &= E(e_{i1}) = E(e_{i2}) = 0 \quad \forall i = 1, \ldots, p
\end{align*}
\]

Under the above transformations, the estimator \( T_{RK}(p) \) takes the following form:

\[
T_{RK}(p) = K_1 S^2_y (1+e_0) + K_2 S^2_y (1+e_0) \sum_{i=1}^{p} (1+e_{i2})(1+e_{i1})^{-1} \\
+ K_3 S^2_y (1+e_0) \sum_{i=1}^{p} (1+e_{i1})(1+e_{i2})^{-1}
\]  \( (7) \)
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

$$
\begin{align*}
E(e_0^2) &= f_1 C_0^2, E(e_{1i}^2) = f_1 C_i^2, E(e_{2i}^2) = f_2 C_i^2 \\
E(e_0^2_{1i}) &= f_1 \rho_{0i} C_0 C_i, E(e_0^2_{2i}) = f_2 \rho_{0i} C_0 C_i, E(e_{1i}^2) = f_2 C_i^2 \\
E(e_{1i} e_{ij}) &= f_1 \rho_{ij} C_i C_j, E(e_{2i} e_{2j}) = f_2 \rho_{ij} C_i C_j, E(e_{2i} e_{2j}) = f_2 \rho_{ij} C_i C_j
\end{align*}
$$

where, for integers $s, t \geq 0$,

$$
\begin{align*}
\mu(i)_{st} &= \frac{1}{N} \sum_{h=1}^{N} \left( (y_h - \bar{Y})^s (x_{ih} - \bar{X}_i)^t \right) \\
\lambda(i)_{st} &= \frac{\mu(i)_{st}}{\sqrt{\mu(i)_{20} \mu(i)_{02}}}, \\
C_0 &= \sqrt{\lambda(i)_{40} - 1}, C_i &= \sqrt{\lambda(i)_{04} - 1}, \rho_{0i} = \frac{(\lambda(i)_{22} - 1)}{\sqrt{(\lambda(i)_{40} - 1)(\lambda(i)_{04} - 1)}} \\
\mu(ij)_{st} &= \frac{1}{N} \sum_{h=1}^{N} \left( (x_{ih} - \bar{X}_i)^s (x_{ij} - \bar{X}_j)^t \right), \\
\lambda(ij)_{st} &= \frac{\mu(ij)_{st}}{\sqrt{\mu(ij)_{20} \mu(ij)_{02}}}, \quad \rho_{ij} = \frac{(\lambda(ij)_{22} - 1)}{\sqrt{(\lambda(ij)_{40} - 1)(\lambda(ij)_{04} - 1)}} \\
A_{0i} &= \rho_{0i} \frac{C_0}{C_i}, A_{ij} = \rho_{ij} \frac{C_i}{C_j}, \quad \forall i, j = 1, 2, \ldots, p,
\end{align*}
$$

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

\[ 163 \]
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

\[ B[T_{RK}(p)] = E[T_{RK}(p) - S_y^2] \]
\[ = K_2 f_3 S_y^2 \sum_{i=1}^{p} C_i^2 - (K_2 - K_3) f_3 S_y^2 \sum_{i=1}^{p} A_{0i} C_i^2 \]  

\[ M[T_{RK}(p)] = E[T_{RK}(p) - S_y^2]^2 \]
\[ = S_y^4 \left[ f_1 C_0^2 + \alpha^2 f_3 \left( \sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0j} C_j^2 \right) - 2 \alpha f_3 \sum_{i=1}^{p} A_{0i} C_i^2 \right] \]  

where \[ \alpha = K_2 - K_3 \]

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

\[ \alpha_{opt} = \frac{\left( \sum_{i=1}^{p} A_{0i} C_i^2 \right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0j} C_j^2} \]  

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

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

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

\[ \alpha_{opt} = (K_2)_{opt} - (K_3)_{opt} = \frac{\sum_{i=1}^{p} A_{0i} C_i^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_{0j} C_j^2} = R \text{(say)} \]
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

\[
B \left[ T_{RK}(p) \right] = 0
\]  

Thus from equation (9),

\[
K_2 \sum_{i=1}^{p} (1 - A_{bi}) C_i^2 + K_3 \sum_{i=1}^{p} A_{bi} C_i^2 = 0
\]  

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

\[
\begin{pmatrix}
1 & p & p \\
0 & 1 & -1 \\
0 & \sum_{i=1}^{p} (1 - A_{bi}) C_i^2 & \sum_{i=1}^{p} A_{bi} C_i^2
\end{pmatrix}
\begin{pmatrix}
K_1 \\
K_2 \\
K_3
\end{pmatrix} =
\begin{pmatrix}
1 \\
R \\
0
\end{pmatrix}
\]  

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

\[
(K_1)_{opt} = 1 - p \left[ \frac{\sum_{i=1}^{p} A_{bi} C_i^2 \sum_{i=1}^{p} (2A_{bi} - 1) C_i^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i=1}^{p} A_{bi} C_i^2} \right]
\]

\[
(K_2)_{opt} = \frac{\left( \sum_{i=1}^{p} A_{bi} C_i^2 \right)^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i=1}^{p} A_{bi} C_i^2}
\]

\[
(K_3)_{opt} = \frac{\sum_{i=1}^{p} A_{bi} C_i^2 \sum_{i=1}^{p} (A_{bi} - 1) C_i^2}{\sum_{i=1}^{p} C_i^2 + \sum_{i=1}^{p} A_{bi} C_i^2}
\]

From equation (18), substituting the values of \((K_1)_{opt}\), \((K_2)_{opt}\), and \((K_3)_{opt}\) in equation (5) yields the optimum unbiased estimator for \( S_y^2 \) as
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

\[ T_{RK}(p) = \left[ 1 - p \left\{ \frac{\sum_{i=1}^{p} A_{0i} C_i^2 \sum_{i=1}^{p} (2A_{0i} - 1) C_i^2}{\sum_{i=1}^{p} C_i^2 \left( \sum_{i=1}^{p} C_i^2 + \sum_{i=x,j=1}^{p} A_{ij} C_j^2 \right)} \right\} \right] s_y^2(n) \]

\[ + \frac{\left( \sum_{i=1}^{p} A_{0i} C_i^2 \right)^2}{\sum_{i=1}^{p} C_i^2 \left( \sum_{i=1}^{p} C_i^2 + \sum_{i=x,j=1}^{p} A_{ij} C_j^2 \right)} s_y^2(n) \sum_{i=1}^{p} s_{y_i}^2(n') \]

\[ + \frac{\sum_{i=1}^{p} A_{0i} C_i^2 \sum_{i=1}^{p} (A_{0i} - 1) C_i^2}{\sum_{i=1}^{p} C_i^2 \left( \sum_{i=1}^{p} C_i^2 + \sum_{i=x,j=1}^{p} A_{ij} C_j^2 \right)} s_y^2(n) \sum_{i=1}^{p} s_{y_i}^2(n) \]

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

\[ V[T_{RK}(p)]_{opt} = S_y^4 \left[ \frac{f_1 C_0^2 - f_3 \left( \sum_{i=1}^{p} A_{0i} C_i^2 \right)}{\sum_{i=1}^{p} C_i^2 + \sum_{i=x,j=1}^{p} A_{ij} C_j^2} \right] \]

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

\[ E\left(e_i^2\right) = f_1 C_i^2, \quad E\left(e_{ij}^2\right) = f_1 C_i^2, \quad E\left(e_{2i}^2\right) = f_2 C_i^2 \]

\[ E\left(e_0 e_{ij}\right) = f_1 \rho_{0i} C_0 C_i, \quad E\left(e_{ij} e_{1j}\right) = f_1 \rho_{ij} C_i C_j, \quad E\left(e_{2i} e_{2j}\right) = f_2 \rho_{ij} C_i C_j \]

\[ E\left(e_0 e_{2i}\right) = E\left(e_{ij} e_{2j}\right) = E\left(e_{ij} e_{2j}\right) = 0 \]

Proceeding as in Case I, the optimum unbiased estimator for \( S_y^2 \) is obtained as
MAJI ET AL

\[ T_{RK}(p) = \left[ 1 - \frac{pf_i}{(f_1 + f_2)^2} \right] \left[ \frac{\sum_{i=1}^{p} A_{0i} C_i^2 \sum_{i=1}^{p} (2f_i A_{0i} - f_3) C_i^2}{\sum_{i=1}^{p} C_i^2 \left( \sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_j C_j^2 \right)} \right] s_y^2(n) + \frac{f_1}{(f_1 + f_2)^2} \sum_{i=1}^{p} A_{0i} C_i^2 \left( \sum_{i=1}^{p} (f_2 + f_i A_{0i}) C_i^2 \right) s_y^2(n) \sum_{i=1}^{p} s_{x_i}^2(n') \]

(22)

with optimum variance up-to first order of approximations as

\[ V\left[T_{RK}(p)\right]_{\text{opt}} = S_y^4 \left[ f_i C_i^2 - \frac{f_1^2}{f_1 + f_2} \sum_{i=1}^{p} C_i^2 + \sum_{i \neq j=1}^{p} A_j C_j^2 \right] \]

(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 \( C_0, C_i, \rho_{0i}, \) and \( \rho_{ij} \) \( (i, j = 1, 2, \ldots, p) \). 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 \( \text{T}_{RK}(p) \) in its general form to carry out the numerical illustrations. Thus, for empirical investigations, consider \( \text{T}_{RK}(p) \) with \( p = 1 \) and 2, where the suggested estimator \( \text{T}_{RK}(p) \) is superior to \( t_1 \) and \( t_2 \) for \( \text{T}_{RK}(1) \) (i.e. \( p = 1 \)) and dominates \( t_3 \) and \( t_4 \) for \( p = 2 \). The performance of \( \text{T}_{RK}(1) \) 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 \( \text{T}_{RK}(p) \) (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

\[
\begin{align*}
M(t_1) &= S_y^4 \left[ f_1 C_0^2 + f_2 C_1^2 (1 - 2A_{01}) \right] \\
M(t_2) &= S_y^4 \left[ f_1 C_0^2 + \frac{1}{4} f_3 C_1^2 (1 - 4A_{01}) \right] \\
M(t_3) &= S_y^4 \left[ f_1 C_0^2 + f_3 C_1^2 (1 - 2A_{01}) + f_2 C_2^2 (1 - 2A_{02}) \right] \\
M(t_4) &= S_y^4 \left[ f_1 C_0^2 + f_3 C_1^2 (1 - 2A_{01}) + \theta f_2 C_2^2 (\theta - 2A_{02}) \right] \\
V[\text{T}_{RK} (1)] &= S_y^4 \left[ f_1 C_0^2 - f_3 A_{01}^2 C_1^2 \right] \\
V[\text{T}_{RK} (2)] &= S_y^4 \left[ f_1 C_0^2 - f_3 \frac{(A_{01} C_1^2 + A_{02} C_2^2)^2}{C_1^2 + C_2^2 + A_{12} C_2^2 + A_{21} C_1^2} \right]
\end{align*}
\]

where

\[
\theta = \frac{S_z^2}{S_z^2 + \beta_2(z)}
\]
**Case II**

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

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

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

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

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

\[
V \left[ T_{RK} (2) \right] = S_y^4 \left[ f_1 C_0^2 - \frac{f_1^2}{f_1 + f_2} \left( A_{01} C_1^2 + A_{02} C_2^2 \right) \right]
\]

with \( \theta \) as described above.

**Numerical Illustration using Known Natural Populations**

Six natural datasets were chosen to elucidate the efficacious performance of the proposed estimator \( T_{RK}(p) \) (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.
y: Food cost.
x: Size of the family.
z: Income.

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

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

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

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

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

$$\text{PRE} = \frac{\text{V}(s_y^2)}{\text{V}[T_{RK}(p)]_{opt}} \times 100$$

(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 $\sigma_y^2 = 100$, $\mu_y = 40$, $\sigma_x^2 = 225$, $\mu_x = 50$, $\sigma_z^2 = 25$, and $\mu_z = 30$ as

$$y_k = \mu_y + \sigma_y \left[ \rho_{xy}x'_k + \left(1 - \rho_{xy}^2\right) y'_k \right], \quad x_k = \mu_x + \sigma_x x'_k,$$

and

$$z_k = \mu_z + \sigma_z \left[ \rho_{xz}x'_k + \left(1 - \rho_{xz}^2\right) z'_k \right].$$

PREs of different estimators for fixed and varying values of $\rho_{xy}$ and $\rho_{xz}$ are presented in Tables 3 and 4, respectively.
## Table 2. PREs of different estimators

<table>
<thead>
<tr>
<th>Population</th>
<th>Case I</th>
<th>Percent Relative Efficiency</th>
<th>Case II</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pop. I</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n$</td>
<td>$t_1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>0.691</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.853</td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>0.931</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>0.341</td>
</tr>
<tr>
<td><strong>Pop. II</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>33</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Pop. III</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Note: "*" indicates no gain, i.e., PRE is less than 100
Table 2, continued.

<table>
<thead>
<tr>
<th>Population</th>
<th>Case I</th>
<th>Percent Relative Efficiency</th>
<th>Case II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop. IV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>n'</td>
<td>( t_1 )</td>
<td>( t_2 )</td>
</tr>
<tr>
<td>31</td>
<td>17</td>
<td>12</td>
<td>132.59</td>
</tr>
<tr>
<td>10</td>
<td>157.801</td>
<td>152.472</td>
<td>165.527</td>
</tr>
<tr>
<td>8</td>
<td>129.891</td>
<td>127.662</td>
<td>133.2</td>
</tr>
<tr>
<td>12</td>
<td>146.535</td>
<td>142.658</td>
<td>152.432</td>
</tr>
<tr>
<td>5</td>
<td>155.339</td>
<td>150.476</td>
<td>162.83</td>
</tr>
<tr>
<td>Pop. V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>n'</td>
<td>( t_1 )</td>
<td>( t_2 )</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>8</td>
<td>*</td>
</tr>
<tr>
<td>7</td>
<td>*</td>
<td>*</td>
<td>100.015</td>
</tr>
<tr>
<td>6</td>
<td>*</td>
<td>*</td>
<td>100.016</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>6</td>
<td>*</td>
<td>*</td>
<td>100.013</td>
</tr>
<tr>
<td>5</td>
<td>*</td>
<td>*</td>
<td>100.015</td>
</tr>
<tr>
<td>Pop. VI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>n'</td>
<td>( t_1 )</td>
<td>( t_2 )</td>
</tr>
<tr>
<td>34</td>
<td>25</td>
<td>12</td>
<td>130.044</td>
</tr>
<tr>
<td>10</td>
<td>132.338</td>
<td>145.463</td>
<td>149.733</td>
</tr>
<tr>
<td>8</td>
<td>134.343</td>
<td>148.581</td>
<td>153.251</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>123.927</td>
<td>132.792</td>
</tr>
<tr>
<td>6</td>
<td>126.494</td>
<td>136.592</td>
<td>139.801</td>
</tr>
</tbody>
</table>

Note: "*" indicates no gain, i.e., PRE is less than 100
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

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

<table>
<thead>
<tr>
<th>Artificial Population</th>
<th>Estimators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case I</td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$n'$</td>
</tr>
<tr>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>45</td>
<td>100</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Case II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>$n'$</td>
</tr>
<tr>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>45</td>
<td>100</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td>40</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

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

**Table 4.** PREs of Different estimators for varying values of $\rho_{xy}$ and $\rho_{xz}$

<table>
<thead>
<tr>
<th>$\rho_{xy}$</th>
<th>$\rho_{xz}$</th>
<th>$h$</th>
<th>$t_2$</th>
<th>$TR_k(1)$</th>
<th>$h$</th>
<th>$t_4$</th>
<th>$TR_k(2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>101.983</td>
<td>116.671</td>
<td>116.696</td>
<td>*</td>
<td>101.536</td>
<td>117.440</td>
</tr>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>126.096</td>
<td>121.277</td>
<td>127.007</td>
<td>109.818</td>
<td>115.747</td>
<td>118.407</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>115.223</td>
<td>117.736</td>
<td>119.551</td>
<td>*</td>
<td>*</td>
<td>109.180</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>*</td>
<td>119.547</td>
<td>119.551</td>
<td>*</td>
<td>*</td>
<td>111.733</td>
</tr>
<tr>
<td>0.5</td>
<td>0.8</td>
<td>*</td>
<td>*</td>
<td>100.349</td>
<td>*</td>
<td>*</td>
<td>100.390</td>
</tr>
<tr>
<td>0.6</td>
<td>*</td>
<td>102.123</td>
<td>103.171</td>
<td>*</td>
<td>*</td>
<td>101.590</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>*</td>
<td>*</td>
<td>100.159</td>
<td>*</td>
<td>*</td>
<td>100.227</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>*</td>
<td>*</td>
<td>102.017</td>
<td>*</td>
<td>*</td>
<td>100.300</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
<td>*</td>
<td>*</td>
<td>100.188</td>
<td>*</td>
<td>*</td>
<td>100.573</td>
</tr>
<tr>
<td>0.6</td>
<td>*</td>
<td>*</td>
<td>100.033</td>
<td>*</td>
<td>*</td>
<td>100.025</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>*</td>
<td>*</td>
<td>100.035</td>
<td>*</td>
<td>*</td>
<td>100.351</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>*</td>
<td>*</td>
<td>100.289</td>
<td>*</td>
<td>*</td>
<td>101.920</td>
<td></td>
</tr>
</tbody>
</table>

Note: * indicates no gain, i.e., PRE is less than 100
Table 4, continued.

<table>
<thead>
<tr>
<th>$\rho_{xy}$</th>
<th>$\rho_{xz}$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$T_{RK}(1)$</th>
<th>$t_3$</th>
<th>$t_4$</th>
<th>$T_{RK}(2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>*</td>
<td>119.103</td>
<td>132.247</td>
<td>*</td>
<td>*</td>
<td>133.885</td>
</tr>
<tr>
<td>0.6</td>
<td>106.535</td>
<td>156.666</td>
<td>156.841</td>
<td>*</td>
<td>103.806</td>
<td>136.045</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>*</td>
<td>136.901</td>
<td>138.644</td>
<td>*</td>
<td>*</td>
<td>116.728</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>*</td>
<td>118.359</td>
<td>138.644</td>
<td>*</td>
<td>*</td>
<td>121.799</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.8</td>
<td>*</td>
<td>*</td>
<td>100.596</td>
<td>*</td>
<td>*</td>
<td>100.666</td>
</tr>
<tr>
<td>0.6</td>
<td>*</td>
<td>105.528</td>
<td>*</td>
<td>*</td>
<td>102.740</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>*</td>
<td>*</td>
<td>100.272</td>
<td>*</td>
<td>*</td>
<td>100.387</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>*</td>
<td>*</td>
<td>103.488</td>
<td>*</td>
<td>*</td>
<td>100.511</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
<td>*</td>
<td>*</td>
<td>100.321</td>
<td>*</td>
<td>*</td>
<td>100.981</td>
</tr>
<tr>
<td>0.6</td>
<td>*</td>
<td>*</td>
<td>100.057</td>
<td>*</td>
<td>*</td>
<td>100.044</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>*</td>
<td>*</td>
<td>100.059</td>
<td>*</td>
<td>*</td>
<td>100.600</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>*</td>
<td>*</td>
<td>104.493</td>
<td>*</td>
<td>*</td>
<td>103.318</td>
<td></td>
</tr>
</tbody>
</table>

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 $\rho_{xy}$ and $\rho_{xz}$, 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.
VARIANCE ESTIMATION USING MULTI-AUXILIARY VARIABLES

References


