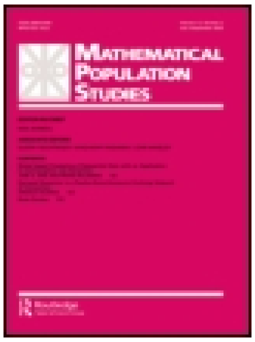

Routledge
Taylor & Francis Group

Mathematical Population Studies
An International Journal of Mathematical Demography

ISSN: 0889-8480 (Print) 1547-724X (Online) Journal homepage: https://www.tandfonline.com/loi/gmps20

Improved chain-ratio type estimator for population total in double sampling

Saurav Guha & Hukum Chandra

To cite this article: Saurav Guha & Hukum Chandra (2019): Improved chain-ratio type estimator for population total in double sampling, Mathematical Population Studies, DOI: 10.1080/08898480.2019.1626635

To link to this article: https://doi.org/10.1080/08898480.2019.1626635

Published online: 19 Jun 2019.

MATHEMATICAL POPULATION STUDIES
https://doi.org/10.1080/08898480.2019.1626635

Improved chain-ratio type estimator for population total in double sampling

Saurav Guha and Hukum Chandra

Indian Agricultural Statistics Research Institute, New Delhi, India

ABSTRACT

Chain-ratio estimators are often used to improve the efficiency of the estimation of the population total or the mean using two auxiliary variables, available in two different phases. An improved chain-ratio estimator for the population total based on double sampling is proposed when auxiliary information is available for the first variable and not available for the second variable. The bias and the mean square error of this estimator are obtained for a large sample. Empirical evaluations using both model-based and design-based simulations show that the proposed estimator performs better than the ratio, the regression, and the difference estimators.

KEYWORDS

chain-ratio estimator; two-phase sampling; auxiliary variables; bias; mean square error; efficiency

1. Introduction

Survey statisticians may use auxiliary information to improve the precision of estimators of population parameters, notably means and totals. The ratio, the regression, and the difference estimators are used when available auxiliary information is used for estimation. The ratio estimator is used when the character y and the auxiliary variable x are positively correlated to one another and the product estimator is used when y and x are correlated negatively to one another. When the information on the auxiliary variable is not available, a large preliminary sample is necessary to record the auxiliary character and a subsample is drawn from this preliminary sample to record the character y. This method of first selecting a large preliminary sample and then subsampling from that large sample is known as "two-phase sampling or double sampling" (Neyman, 1938). The double sampling technique is appropriate when the auxiliary variables are easy to obtain (Hidiroglou and Särndal, 1998; Fuller, 2000; Hidiroglou, 2001). Consider a first-phase sample comprising n' units drawn from a finite population comprising N units, in order to record the auxiliary variable x. From this first-phase sample of n'units, a second-phase sample comprising n units is drawn to estimate the character y. The ratio estimator for the population mean in double sampling is given by Sukhatme (1962) as

CONTACT Hukum Chandra hchandra12@gmail.com Library Avenue, Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India

$$\bar{y}_{\mathrm{Rd}} = \bar{y}\frac{\bar{x}'}{\bar{x}}, \tag{1}$$

where $\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$ is the sample mean of y for the second-phase sample, $\bar{x}' = \frac{1}{n'}\sum_{k=1}^{n'} x_k$ is the sample mean of the auxiliary variable x for the first-phase sample, and $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ is the sample mean of x for the second-phase sample.

We assume that the two auxiliary variables x_1 and x_2 are available at different phases of sampling, that is, x_1 is completely known while x_2 is available at the first phase and $\mathrm{cor}(x_1, x_2) > 0.50$, with $\mathrm{cor}(x_1, y) < \mathrm{cor}(x_2, y)$. Then, the population mean of x_2 is estimated more accurately than $\bar{x}_2'$ alone by

$$\hat{\bar{X}}_{2,\mathrm{Rd}} = \frac{\bar{x}_2'}{\bar{x}_1'}\bar{X}_1, \tag{2}$$

if

$$\rho_{x_1 x_2} > \frac{1}{2}\frac{C_{x_1}}{C_{x_2}}. \tag{3}$$

Here

$$C_{x_i} = \frac{S_{x_i}}{\bar{X}_i};\ S_{x_i}^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(x_{ik} - \bar{X}_i\right)^2,\ i = 1, 2, \tag{4}$$

and $\bar{X}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik}$ is the population mean of the auxiliary variable x_i, $i = 1, 2$. Chand (1975) introduced chain-ratio estimators and Kiregyera (1980) a chain-ratio estimator with two auxiliary variables with a regression in the first phase to estimate the auxiliary variable. Bahl and Tuteja (1991) developed the ratio and the product exponential estimators and compared with the ratio estimator. Singh and Tailor (2003) proposed an exponential ratio estimator by incorporating the correlation coefficient and Kadilar and Cingi (2004) a ratio estimator in simple random sampling. Singh and Choudhury (2012) developed an exponential chain-ratio and product estimator for correlated auxiliary variables, and Vishwakarma and Gangele (2014) a class of chain-ratio product estimators using two auxiliary variables in double sampling. We improve the chain-ratio estimator under simple random sampling with a weighted combination of the ratio estimator (Cochran, 1977) and the chain-ratio estimator (Chand, 1975). The weight is such that the mean square error of the proposed estimator is smaller than that of the ratio estimator (Cochran, 1977) or that of the chain-ratio estimator (Chand, 1975). We evaluate the estimator on the basis of four data sets.

2. Theory

Consider a finite population $\Omega = (1, 2, \ldots, k, \ldots, N)$. Complete information on the auxiliary variable x_1 is available. A large preliminary sample $s' \in \Omega$ of size n' is drawn from Ω by simple random sampling without replacement to collect information on x_2. Subsequently, a second-phase sample $s \subset s'$ of size n is drawn from s' by simple random sampling without replacement to record the study variable y. For every $k \in s$, the kth unit pertaining to y is denoted by y_k. The kth unit for both auxiliary variables are x_{1k} and x_{2k}. The population total $\sum_{k \in \Omega} x_{1k}$ of x_1 and the values of x_{1k} and x_{2k} are known for every $k \in s'$. The sample means of the auxiliary variables for the first-phase sample are $\bar{x}'_1 = \frac{1}{n'} \sum\limits_{k=1}^{n'} x_{1k}$ and $\bar{x}'_2 = \frac{1}{n'} \sum\limits_{k=1}^{n'} x_{2k}$; and for the second-phase sample $\bar{x}_1 = \frac{1}{n} \sum\limits_{k=1}^{n} x_{1k}$, $\bar{x}_2 = \frac{1}{n} \sum\limits_{k=1}^{n} x_{2k}$, and $\bar{y} = \frac{1}{n} \sum\limits_{k=1}^{n} y_k$ for the variables corresponding to the population means $\bar{X}_1 = \frac{1}{N} \sum\limits_{k=1}^{N} x_{1k}$, $\bar{X}_2 = \frac{1}{N} \sum\limits_{k=1}^{N} x_{2k}$, and $\bar{Y} = \frac{1}{N} \sum\limits_{k=1}^{N} y_k$. The coefficients of variation of x_1, x_2, and y are $C_{x_1} = \frac{S_{x_1}}{\bar{X}_1}$, $C_{x_2} = \frac{S_{x_2}}{\bar{X}_2}$, and $C_y = \frac{S_y}{\bar{Y}}$, where $S^2_{x_i} = \frac{1}{N-1} \sum\limits_{k=1}^{N} \left(x_{ik} - \bar{X}_i\right)^2$, $i = 1,\ 2$ and $S^2_y = \frac{1}{N-1} \sum\limits_{k=1}^{N} \left(y_k - \bar{Y}\right)^2$.

The population correlations are ρ_{yx_1}, ρ_{yx_2}, and $\rho_{x_1 x_2}$ between the subscripted variables. We define

$$\varepsilon_y = \frac{\bar{y} - \bar{Y}}{\bar{Y}},\ \varepsilon'_{x_1} = \frac{\bar{x}'_1 - \bar{X}_1}{\bar{X}_1},\ \varepsilon_{x_1} = \frac{\bar{x}_1 - \bar{X}_1}{\bar{X}_1},\ \varepsilon'_{x_2} = \frac{\bar{x}'_2 - \bar{X}_2}{\bar{X}_2},\ \text{and } \varepsilon_{x_2} = \frac{\bar{x}_2 - \bar{X}_2}{\bar{X}_2} \tag{5}$$

such that

$$\begin{aligned}
&E(\varepsilon_y) = E(\varepsilon'_{x_1}) = \mathrm{E}(\varepsilon_{x_1}) = \mathrm{E}(\varepsilon'_{x_2}) = \mathrm{E}(\varepsilon_{x_2}) = 0\\
&\mathrm{E}(\varepsilon_j^2) = \delta_1 S_j^2 \text{ for } j = y,\ x_1,\ x_2\\
&\mathrm{E}(\varepsilon_j'^2) = \mathrm{E}(\varepsilon_j \varepsilon'_j) = \delta_2 S_j^2,\ \text{for} j = x_1,\ x_2\\
&\mathrm{E}(\varepsilon_j \varepsilon'_{j'}) = \mathrm{E}(\varepsilon'_j \varepsilon'_{j'}) = \mathrm{E}(\varepsilon'_j \varepsilon_{j'}) = \delta_2 S_{jj'} \text{ for } j = x_1,\ j' = x_2\\
&\mathrm{E}(\varepsilon_y \varepsilon_{x_2}) = \delta_1 S_{x_2 y},\ \mathrm{E}(\varepsilon_y \varepsilon'_{x_1}) = \delta_2 S_{x_1 y},\ \mathrm{E}(\varepsilon_y \varepsilon'_{x_2}) = \delta_2 S_{x_2 y},\ \mathrm{E}(\varepsilon_{x_1} \varepsilon_{x_2}) = \delta_1 S_{x_1 x_2}.
\end{aligned} \tag{6}$$

Here,

$$S_{x_1 y} = \rho_{yx_1} S_y S_{x_1},\ S_{x_2 y} = \rho_{yx_2} S_y S_{x_2},\ S_{x_1 x_2} = \rho_{x_1 x_2} S_{x_1} S_{x_2}, \tag{7}$$

$$\delta_1 = \left(\frac{1}{n} - \frac{1}{N}\right), \ \delta_2 = \left(\frac{1}{n'} - \frac{1}{N}\right), \text{ and } \delta_3 = \left(\frac{1}{n} - \frac{1}{n'}\right). \tag{8}$$

Estimators for the population total are

(1) the sample mean estimator (Cochran, 1977)

$$\hat{Y}_{\mathrm{m}} = N\bar{y}, \tag{9}$$

with variance

$$\mathrm{Var}(\hat{Y}_{\mathrm{m}}) = N^2\bar{Y}^2\delta_1 C_y^2; \tag{10}$$

(2) the ratio estimator in double sampling (Sukhatme, 1962)

$$\hat{Y}_{\mathrm{RD}} = N\bar{y}\frac{\bar{x}'_2}{\bar{x}_2}, \tag{11}$$

whose bias, to the first-order approximation, is

$$\mathrm{Bias}\,(\hat{Y}_{\mathrm{RD}}) = N\bar{Y}\delta_3\left(C_{x_2}^2 - \rho_{yx_2}C_yC_{x_2}\right), \tag{12}$$

and whose first-order approximation of the mean square error is

$$\mathrm{MSE}\,(\hat{Y}_{\mathrm{RD}}) = N^2\bar{Y}^2\left(\delta_1 C_y^2 + \delta_3\left(C_{x_2}^2 - 2\rho_{yx_2}C_yC_{x_2}\right)\right); \tag{13}$$

(3) the chain-ratio estimator (Chand, 1975)

$$\hat{Y}_{\mathrm{C}} = N\bar{y}\frac{\bar{x}'_2}{\bar{x}_2}\frac{\bar{X}_1}{\bar{x}'_1}, \tag{14}$$

whose bias, to the first-order approximation, is

$$\mathrm{Bias}\,(\hat{Y}_{\mathrm{C}}) = N\bar{Y}\left(\delta_2\rho_{yx_2}C_yC_{x_2} - \delta_1\rho_{yx_2}C_yC_{x_2} - \delta_2\rho_{yx_1}C_yC_{x_1} - \delta_2C_{x_2}^2\right), \tag{15}$$

and first-order approximation of the mean square error of $\hat{Y}_{\mathrm{C}}$ is

$$\mathrm{MSE}\,(\hat{Y}_{\mathrm{C}}) = N^2\bar{Y}^2\left(\delta_1 C_y^2 + \delta_3\left(C_{x_2}^2 - 2\rho_{yx_2}C_yC_{x_2}\right) + \delta_2\left(C_{x_1}^2 - 2\rho_{yx_1}C_yC_{x_1}\right)\right); \tag{16}$$

(4) the chain exponential estimator (Kiregyera, 1980)

$$\hat{Y}_{\mathrm{CE}} = N\frac{\bar{y}}{\bar{x}_2}(\bar{x}'_2 + b_{x_2x_1}(\bar{X}_1 - \bar{x}'_1)), \tag{17}$$

whose bias, to the first-order approximation, is

$$\mathrm{Bias}\,(\hat{Y}_{\mathrm{CE}}) = N\bar{Y}\left(\delta_2\rho_{x_2x_1}C_{x_2}\left(C_{x_2} - \rho_{yx_1}C_yC_{x_1}\right) - \delta_3\rho_{yx_2}C_yC_{x_2} - \delta_2C^2_{x_2}\right), \tag{18}$$

and first-order approximation of the mean square error of $\hat{Y}_{\mathrm{CE}}$ is

$$\mathrm{MSE}\,(\hat{Y}_{\mathrm{CE}}) = N^2\bar{Y}^2\left(\delta_1C^2_y + \delta_3\left(C^2_{x_2} - 2\rho_{yx_2}C_yC_{x_2}\right) + \delta_2\rho_{x_2x_1}C_{x_2}\left(\rho_{x_2x_1}C_{x_2} - 2\rho_{yx_1}C_y\right)\right); \tag{19}$$

(5) the multivariate ratio estimator in double sampling with two auxiliary variables (Sukhatme, 1962)

$$\hat{Y}_{\mathrm{MRD}} = N\bar{y}\left(w\frac{\bar{x}'_1}{\bar{x}_1} + (1-w)\frac{\bar{x}'_2}{\bar{x}_2}\right), \tag{20}$$

where w is obtained by minimizing the mean square error of $\hat{Y}_{\mathrm{MRD}}$:

$$w = \frac{\mathrm{MSE}\left(\bar{y}\frac{\bar{x}'_2}{\bar{x}_2}\right) - \mathrm{E}\left(\bar{y}\frac{\bar{x}'_1}{\bar{x}_1} - \bar{Y}\right)\left(\bar{y}\frac{\bar{x}'_2}{\bar{x}_2} - \bar{Y}\right)}{\mathrm{MSE}\left(\bar{y}\frac{\bar{x}'_1}{\bar{x}_1}\right) + \mathrm{MSE}\left(\bar{y}\frac{\bar{x}'_2}{\bar{x}_2}\right) - 2\mathrm{E}\left(\bar{y}\frac{\bar{x}'_1}{\bar{x}_1} - \bar{Y}\right)\left(\bar{y}\frac{\bar{x}'_2}{\bar{x}_2} - \bar{Y}\right)}, \tag{21}$$

and whose minimum mean square error of $\hat{Y}_{\mathrm{MRD}}$ (at optimum value of w) is

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{MRD}}\right) = N^2\bar{Y}^2\left(\delta_1C^2_y - \frac{\delta_3\Delta}{C^2_{x_1} + C^2_{x_2} - 2\rho_{x_1x_2}C_{x_1}C_{x_2}}\right), \tag{22}$$

where

$$\begin{aligned}\Delta = {} & C^2_y\left(\rho_{yx_1}C_{x_1} - \rho_{yx_2}C_{x_2}\right)^2 - C^2_{x_1}C^2_{x_2}\left(1 - \rho^2_{x_1x_2}\right) \\ & + 2C_yC_{x_1}C_{x_2}\left(\left(\rho_{yx_2}C_{x_1} + \rho_{yx_1}C_{x_2}\right) - \rho_{x_1x_2}\left(\rho_{yx_1}C_{x_1} + \rho_{yx_2}C_{x_2}\right)\right);\end{aligned}$$

(6) the exponential-chain-ratio estimator (Singh and Choudhury, 2012)

$$\hat{Y}_{\mathrm{CS}} = N\bar{y}\exp\left(\frac{(\bar{x}'_2/\bar{x}'_1)\bar{X}_1 - \bar{x}_2}{(\bar{x}'_2/\bar{x}'_1)\bar{X}_1 + \bar{x}_2}\right), \tag{23}$$

whose bias, to the first-order approximation, is

$$\mathrm{Bias}\left(\hat{Y}_{\mathrm{CS}}\right) = N\bar{Y}\left(\frac{3}{8}\left(\delta_3C^2_{x_2} + \delta_2C^2_{x_1}\right) - \frac{1}{2}\left(\delta_3\rho_{yx_2}C_yC_{x_2} + \delta_2\rho_{yx_1}C_yC_{x_1}\right)\right), \tag{24}$$

and whose first-order approximation of the mean square error of $\hat{Y}_{\mathrm{CS}}$ is

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{CS}}\right) = N^2\bar{Y}^2\left(\delta_1 C_y^2 + \frac{1}{4}\left(\delta_3 C_{x_2}^2 + \delta_2 C_{x_1}^2\right) - \left(\delta_3 \rho_{yx_2} C_y C_{x_2} + \delta_2 \rho_{yx_1} C_y C_{x_1}\right)\right); \quad (25)$$

(7) the chain exponential estimator in double sampling (Vishwakarma and Gangele, 2014)

$$\hat{Y}_{\mathrm{VS}} = N\bar{y}\exp\left(\frac{\bar{x}'_2\left(\alpha\bar{x}'_1 + \beta\right)^{-1}\left(\alpha\bar{X}_1 + \beta\right) - \bar{x}_2}{\bar{x}'_2\left(\alpha\bar{x}'_1 + \beta\right)^{-1}\left(\alpha\bar{X}_1 + \beta\right) + \bar{x}_2}\right), \quad (26)$$

where α and β are real numbers used as parameters in $\hat{Y}_{\mathrm{VS}}$, and whose minimum mean square error is

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{VS}}\right) = N^2\bar{Y}^2\left(\delta_1 C_y^2 + \delta_3\frac{C_{x_2}^2 - 4\rho_{yx_2} C_y C_{x_2}}{4} - \delta_2 C_y^2 \rho_{yx_1}^2\right). \quad (27)$$

The sample mean $\hat{Y}_{\mathrm{m}}$ in Eq. (9) is the basic estimator to estimate the population total and it uses no auxiliary variable. The precision of an estimator increases by including auxiliary variables (Cochran, 1977). The ratio estimator in double sampling $\hat{Y}_{\mathrm{RD}}$ in Eq. (11) uses a single auxiliary variable, which is positively correlated with the study variable y (Sukhatme, 1962). The multivariate ratio estimator in double sampling $\hat{Y}_{\mathrm{MRD}}$ in Eq. (20) with two auxiliary variables $x_1$and x_2, suggested by Sukhatme (1962), requires both that auxiliary variables are available at the first phase of the sampling. This estimator is defined as a linear combination of two separate ratio estimators and does not use the correlation between $x_1$and x_2. Two auxiliary variables x_1 and x_2 may be available at different phases of the sampling, for example, when x_1 is completely known and x_2 is already available at the first phase. The chain-ratio estimator $\hat{Y}_{\mathrm{C}}$ in Eq. (14) uses two auxiliary variables and the correlation between them (Chand, 1975), but the regression line between the study variable y and the auxiliary variables x_1 and x_2 passes through the origin and the auxiliary variable x_2 is estimated with the ratio estimator. Kiregyera (1980) described the chain exponential estimator $\hat{Y}_{\mathrm{CE}}$ in Eq. (17) avoiding the constraint that the regression line passes through the origin and he used the regression method to estimate x_2. The exponential-chain-ratio estimator $\hat{Y}_{\mathrm{CS}}$ in Eq. (23) of Singh and Choudhury (2012) also involves two auxiliary variables x_1 and x_2, but, this estimator of x_2 is based on the ratio estimator. The chain exponential estimator in double sampling $\hat{Y}_{\mathrm{VS}}$ in Eq. (26) is a generalized class of the exponential ratio estimator and several estimators are obtained as special cases of this estimator (Vishwakarma and Gangele, 2014), but this estimator does not minimize the mean square error.

3. The weighted chain-ratio estimator

In the case of a single auxiliary variable, an estimator performs satisfactorily when the auxiliary variable is highly correlated with the variable y under study, but in the case of two auxiliary variables, the performance of the estimator depends on how these auxiliary variables are correlated with y. They must also be positively correlated with each other. We propose the weighted chain-ratio estimator for the population total in simple random sampling without replacement as

$$\hat{Y}_S = N\bar{y}\left(\alpha\frac{\bar{X}_1}{\bar{x}_1} + (1-\alpha)\frac{\bar{x}'_2}{\bar{x}_2}\frac{\bar{X}_1}{\bar{x}'_1}\right), \tag{28}$$

where the value of the weight α $(0<\alpha<1)$ is determined by minimizing the mean square error of $\hat{Y}_S$. The estimator in Eq. (28) is a weighted form of the ratio estimator (Cochran, 1977) and Chand's (1975) estimator. If ρ_{yx_1} is greater than ρ_{yx_2}, then the first term $\alpha\frac{\bar{X}_1}{\bar{x}_1}$ dominates in the estimator given in Eq. (28); if it is lower, the second term $(1-\alpha)\frac{\bar{x}'_2}{\bar{x}_2}\frac{\bar{X}_1}{\bar{x}'_1}$ dominates. The estimator in Eq. (28), whose mean square error is smaller than each component $\bar{y}\frac{\bar{X}_1}{\bar{x}_1}$ and $\bar{y}\frac{\bar{x}'_2}{\bar{x}_2}\frac{\bar{X}_1}{\bar{x}'_1}$, takes the advantages of both the ratio estimator (Cochran, 1977) and Chand's (1975) estimator. It is expected to perform satisfactorily in estimating finite population totals. The estimator in Eq. (28) is

$$\hat{Y}_S = N\bar{Y}(1+\varepsilon_y)\left(\alpha(1+\varepsilon_{x_1})^{-1} + (1-\alpha)(1+\varepsilon'_{x_2})(1+\varepsilon_{x_2})^{-1}(1+\varepsilon'_{x_1})^{-1}\right). \tag{29}$$

The bias of $\hat{Y}_S$, to the first-order approximation, is

$$\begin{aligned}\text{Bias}\,(\hat{Y}_S) &= E(\hat{Y}_S - N\bar{Y})\\ &= N\bar{Y}\left((\alpha-1)\left(\delta_2 C^2_{x_2} + \delta_3\rho_{yx_2}C_yC_{x_2}\right) - \alpha\delta_3\rho_{yx_1}C_yC_{x_1} - \delta_2\rho_{yx_1}C_yC_{x_1}\right).\end{aligned} \tag{30}$$

The mean square error of $\hat{Y}_S$, to the first-order approximation, is

$$\text{MSE}\,(\hat{Y}_S) = E(\hat{Y}_S - N\bar{Y})^2 = N^2\bar{Y}^2(\alpha^2 D_1 + (1-\alpha)^2 D_2), \tag{31}$$

where

$$D_1 = \delta_1\left(C^2_y + C^2_{x_1} - 2\rho_{yx_1}C_yC_{x_1}\right), \text{and} \tag{32}$$

$$D_2 = \delta_1 C^2_y + \delta_3\left(C^2_{x_2} - 2\rho_{yx_2}C_yC_{x_2}\right) + \delta_2\left(C^2_{x_1} - 2\rho_{yx_1}C_yC_{x_1}\right). \tag{33}$$

Solving $\partial \mathrm{MSE}(\hat{Y}_S)/\partial\alpha = 0$ leads to the optimum value of the weight α as

$$\alpha_{\mathrm{opt}} = \frac{D_2}{D_1 + D_2}. \tag{34}$$

Replacing the value of α_{opt} in the mean square error of $\hat{Y}_S$ given in Eq. (31), the minimum mean square error of the proposed estimator $\hat{Y}_S$ is

$$\mathrm{MSE}(\hat{Y}_S)_{\min} = N^2\bar{Y}^2\left(\frac{D_1D_2}{D_1 + D_2}\right). \tag{35}$$

4. Comparison of the mean square error of different estimators of the population total

We compare the mean square error of the proposed estimator $\hat{Y}_S$ of Eq. (35) with the estimators mentioned in section 2.

(i) From Eq. (10) and (35),

$$\mathrm{MSE}\,(\hat{Y}_S)_{\min} < \mathrm{MSE}\,(\hat{Y}_m)\ \text{if}\ \frac{A_0^2 - A_1(A_2 + A_3)}{2A_0 + A_1 + A_2 + A_3} > 0, \tag{36}$$

where

$$A_0 = \delta_1 C_y^2,\ A_1 = \delta_1\left(C_{x_1}^2 - 2\rho_{yx_1}C_yC_{x_1}\right), \tag{37}$$

$$A_2 = \delta_2\left(C_{x_1}^2 - 2\rho_{yx_1}C_yC_{x_1}\right)\ \text{and}\ A_3 = \delta_3\left(C_{x_2}^2 - 2\rho_{yx_2}C_yC_{x_2}\right). \tag{38}$$

(ii) From Eq. (13) and (35),

$$\mathrm{MSE}\,(\hat{Y}_S)_{\min} < \mathrm{MSE}\,(\hat{Y}_{\mathrm{RD}})\ \text{if}\ \frac{A_0^2 + A_3(2A_0 + A_3) + A_2(A_3 - A_1)}{B_1} > 0, \tag{39}$$

with

$$B_1 = 2A_0 + A_1 + A_2 + A_3. \tag{40}$$

(iii) From Eq. (16) and (35),

$$\mathrm{MSE}\,(\hat{Y}_S)_{\min} < \mathrm{MSE}\,(\hat{Y}_C)\ \text{if}\ \frac{(A_0 + A_2 + A_3)^2}{B_1} > 0. \tag{41}$$

(iv) From Eq. (19) and (35),

$\mathrm{MSE}\,(\hat{Y}_S)_{\min} < \mathrm{MSE}\,(\hat{Y}_{\mathrm{CE}})$ if

$$\frac{A_0(A_0+2A_3+(2+A_1+A_2+A_3)A_4)+A_3^2+A_2(A_3-A_1)}{B_1}>0, \tag{42}$$

with

$$A_4=\delta_2\rho_{x_1x_2}C_{x_2}\left(\rho_{x_1x_2}C_{x_2}-2\rho_{yx_1}C_y\right). \tag{43}$$

(v) From Eq. (22) and (35),

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{S}}\right)_{\min}<\mathrm{MSE}\left(\hat{Y}_{\mathrm{MRD}}\right)\ \text{ if }\ \frac{B_2}{B_1}-\delta_3\frac{\Delta}{C_{x_1}^2+C_{x_2}^2-2\rho_{x_1x_2}C_{x_1}C_{x_2}}>0, \tag{44}$$

with

$$B_2=A_0^2-A_1(A_2+A_3). \tag{45}$$

(vi) From Eq. (25) and (35),

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{S}}\right)_{\min}<\mathrm{MSE}\left(\hat{Y}_{\mathrm{CS}}\right)\ \text{ if }\ \frac{B_2}{B_1}+\frac{1}{4}(A_2+A_3)-\frac{3}{4}\left(\delta_3\rho_{yx_2}C_yC_{x_2}+\delta_2\rho_{yx_1}C_yC_{x_1}\right)>0. \tag{46}$$

(vii) From Eq. (27) and (35),

$$\mathrm{MSE}\left(\hat{Y}_{\mathrm{S}}\right)_{\min}<\mathrm{MSE}\left(\hat{Y}_{\mathrm{VS}}\right)\ \text{ if }\ \frac{B_2}{B_1}+\delta_3\frac{C_{x_2}^2-4\rho_{yx_2}C_yC_{x_2}}{4}-\delta_2C_y^2\rho_{yx_1}^2>0. \tag{47}$$

These analytical expressions do not yield the performance of the estimator in Eq. (28) in terms of mean square error compared to the other estimators. That is why a simulation is necessary.

5. Comparison by simulation

First we simulate, second we use empirical datasets. We compute the percentage of relative efficiency of the estimators compared with the sample mean estimator

$$\mathrm{PRE}=\frac{\mathrm{MSE}\left(\hat{Y}_{\mathrm{m}}\right)}{\mathrm{MSE}\left(\hat{Y}_{\mathrm{i}}\right)}\times 100, \tag{48}$$

where $\mathrm{MSE}\left(\hat{Y}_{\mathrm{i}}\right)$ denotes the mean square error of an estimator $\hat{Y}_i$.

5.1. *Model-based simulation*

We generate the unknown auxiliary variable x_2 using the model

$$x_{2k}=1.5\,x_{1k}+v_k\ ,\ \ k=1,\ldots N \text{ where } x_{1k}\sim\chi^2(5) \text{ and } v_k\sim N\left(0,\sigma_v^2\right). \tag{49}$$

Then, we generate a finite population of size $N = 5000$ using the model

$$y_k = x_{1k} + x_{2k} + \varepsilon_k\,,\ \ k = 1,\ \ldots N, \tag{50}$$

where $\varepsilon_k \sim N\left(0, \sigma_\varepsilon^2\right)$. σ_υ and σ_ε take three different values each to generate nine different data-sets: $\sigma_\upsilon \in \{4,\ 6,\ 8\}$ and $\sigma_\varepsilon \in \{5,\ 10,\ 15\}$. Table 1 presents the nine population datasets with different correlations between y, x_1, and x_2. For each corresponding population, we draw a first-phase sample s' of size 500 units and a subsample s of size 100 units from s' by simple random sampling without replacement, and we estimate the population total. We draw 2500 random samples from the population and for each sample we calculate the percentage of relative efficiency, presented in Table 2.

First, the percentage of relative efficiency increases with the correlations between y, x_1, and x_2. As the case 1 ($\sigma_\upsilon = 4$ and $\sigma_\varepsilon \in \{5,\ 10,\ 15\}$) has the highest correlation between x_1 and x_2, the gain in efficiency is the highest one in this case. For a fixed correlation between x_1 and x_2, the percentage of relative efficiency increases with the difference between the correlations $\rho(y,\ x_1)$ and $\rho(y,\ x_2)$. In case 2 ($\sigma_\upsilon = 6$ and $\sigma_\varepsilon \in \{5,\ 10,\ 15\}$), $\rho(x_1, x_2) = 0.61$. The gain in efficiency $\rho(y,\ x_1)$-$\rho(y,\ x_2)$= 0.14 for $\sigma_\upsilon = 6$ and $\sigma_\varepsilon = 5$ is higher than for $\sigma_\upsilon = 6$ and $\sigma_\varepsilon = 10$ ($\rho(y,\ x_1)$-$\rho(y,\ x_2)$= 0.10) and for $\sigma_\upsilon = 6$ and $\sigma_\varepsilon = 15$ ($\rho(y,\ x_1)$-$\rho(y,\ x_2)$= 0.08). Likewise for cases 1 and 3. Second, the proposed estimator $\hat{Y}_S$ outperforms all the existing estimators in terms of percentage of relative efficiency for any level of correlation between y, x_1, and x_2. As expected, the relative gain for $\hat{Y}_S$ is highest for case 1.

5.2. *Design-based simulation*

1. Population 1, taken from Singh and Chaudhary (1986)

y = area planted (in acres) with wheat in 1974; x_1= area planted (in acres) with wheat in 1973; x_2= area planted (in acre) with wheat in 1971.

Table 1. Parameters of the simulation.

Situation	Correlation level	Standard deviation of error υ σ_υ	Standard deviation of error ε σ_ε	Correlation between y and x_1 $\rho(y,\ x_1)$	Correlation between y and x_2 $\rho(y,\ x_2)$	Correlation between x_1 and x_2 $\rho(x_1,\ x_2)$
1	High	4	5	0.78	0.84	0.76
	Medium	4	10	0.60	0.64	0.76
	Low	4	15	0.46	0.49	0.76
2	High	4	5	0.71	0.85	0.61
	Medium	6	10	0.57	0.67	0.61
	Low	6	15	0.45	0.53	0.61
3	High	8	5	0.64	0.88	0.50
	Medium	8	10	0.53	0.72	0.50
	Low	8	15	0.43	0.57	0.50

Table 2. Percentage of relative efficiency of estimators of the population total in the model-based simulation.

		Estimators							
Situation	Correlation level	Sample mean $\hat{Y}_{\mathrm{m}}$	Ratio $\hat{Y}_{\mathrm{RD}}$	Multivariate ratio $\hat{Y}_{\mathrm{MRD}}$	Singh and Choudhury $\hat{Y}_{\mathrm{CS}}$	Vishwakarma and Gangele $\hat{Y}_{\mathrm{VS}}$	Chand $\hat{Y}_{\mathrm{C}}$	Kiregiyera $\hat{Y}_{\mathrm{CE}}$	Weighted $\hat{Y}_{\mathrm{S}}$
1	High	100	182.7	65.7	268.8	283.7	227.6	226.8	613.8
	Medium	100	131.8	75.0	158.1	160.8	143.7	143.4	464.2
	Low	100	114.4	81.9	128.1	129.1	119.5	119.3	369.5
2	High	100	150.6	77.5	296.5	311.7	174.3	173.6	562.6
	Medium	100	119.2	82.0	170.8	173.6	127.8	127.5	430.7
	Low	100	107.0	86.2	134.3	135.3	111.1	110.9	352.2
3	High	100	126.0	92.8	326.4	341.1	138.9	138.3	505.9
	Medium	100	105.9	91.5	186.6	189.4	111.7	111.3	395.4
	Low	100	97.9	92.0	142.3	143.3	101.0	100.8	332.2

Statistics are

$N = 34,\ n' = 20,\ n = 5,\ \bar{Y} = 856.4$ acres, $\bar{X}_1 = 199.4$ acres, $\bar{X}_2 = 208.8$ acres, $C_y = 0.86,\ C_{x_1} = 0.75,\ C_{x_2} = 0.72,\ \rho_{yx_1} = 0.45,\ \rho_{yx_2} = 0.45,\ \rho_{x_1x_2} = 0.98.$ (51)

2. Population 2, taken from Cochran (1977)

y = total number of children taking a placebo; x_1= total number of paralytic polio cases in the group taking placebos; x_2= total number of paralytic polio cases in the group not taking placebos.

Statistics are

$N = 34,\ n' = 15,\ n = 7,\ \bar{Y} = 4.9$ children, $\bar{X}_1 = 2.9$ children, $\bar{X}_2 = 2.5$ children, $C_y = 1.01,\ C_{x_1} = 1.05,\ C_{x_2} = 1.23,\ \rho_{yx_1} = 0.64,\ \rho_{yx_2} = 0.73,\ \rho_{x_1x_2} = 0.68.$ (52)

3. Population 3, taken from Ahmed (1997)

y = total number of literate persons; x_1= total population size; x_2= total number of cultivators.

Statistics are

$N = 376,\ n' = 100,\ n = 20,\ \bar{Y} = 316.6$ literate persons, $\bar{X}_1$= 1075.3 persons, $\bar{X}_2 = 141.1$ cultivators, $C_y = 0.77,\ C_{x_1} = 0.77,\ C_{x_2} = 0.84,\ \rho_{yx_1} = 0.90,\ \rho_{yx_2} = 0.91,\ \rho_{x_1x_2} = 0.86.$ (53)

4. Population 4, taken from Abu-Dayyeh et al. (2003)

y = total number of cultivators; x_1= total number of households in a village; x_2= area of the village (in acres).

Statistics are

$N = 332,\ n' = 70,\ n = 15,\ \bar{Y} = 1093.1$ cultivators, $\bar{X}_1$= 143.3 households, $\bar{X}_2 = 181.5$ acres, $C_y = 0.76,\ C_{x_1} = 0.76,\ C_{x_2} = 0.77,\ \rho_{yx_1} = 0.86,\ \rho_{yx_2} = 0.97,\ \rho_{x_1x_2} = 0.84.$ (54)

Table 3 presents the results obtained from design-based simulations. The percentage of relative efficiency of the proposed estimator in Eq. (28) is higher than for the other estimators, and this for the four populations. We also performed a sensitivity analysis of the proposed estimator by taking three different samples for the first and the second phases from populations 3 and 4. We combine first-phase samples with $n' = 80,\ 100,\ 120$, and second-phase samples with $n = 10,\ 20,\ 30$. We perform no sensitivity analysis for populations 1 and 2 which are too small. Table 4 shows that the percentage of relative efficiency of the

Table 3. Percentage of relative efficiency of estimators of the population total in empirical datasets.

	Estimators							
Population data	Sample mean $\hat{Y}_{\text{m}}$	Ratio $\hat{Y}_{\text{RD}}$	Multivariate ratio $\hat{Y}_{\text{MRD}}$	Singh and Choudhury $\hat{Y}_{\text{CS}}$	Vishwakarma and Gangele $\hat{Y}_{\text{VS}}$	Chand $\hat{Y}_{\text{C}}$	Kiregiyera $\hat{Y}_{\text{CE}}$	Weighted $\hat{Y}_{\text{S}}$
1	100	104.8	62.3	125.2	125.2	105.1	105.7	221.4
2	100	125.4	60.6	192.6	194.5	140.2	148.6	265.7
3	100	304.9	54.9	324.1	353.7	497.3	500.3	1030.2
4	100	452.9	57.9	340.5	369.6	1070.9	1109.6	1436.8

Table 4. Percentage of relative efficiency of estimators of the population total in population 3 and 4 for different choices of first- and second-phase sample sizes.

Sample size		Estimators							
First-phase	Second-phase	Sample mean $\hat{Y}_m$	Ratio $\hat{Y}_{RD}$	Multivariate ratio $\hat{Y}_{MRD}$	Singh and Choudhury $\hat{Y}_{CS}$	Vishwakarma and Gangele $\hat{Y}_{VS}$	Chand $\hat{Y}_C$	Kiregiyera $\hat{Y}_{CE}$	Weighted $\hat{Y}_S$
Population 3									
80	10	100	350.9	53.3	326.2	345.1	494.3	496.2	1027.1
	20	100	270.3	56.5	322.1	362.5	500.4	504.4	1033.2
	30	100	217.5	60.2	317.9	382.9	506.9	513.3	1039.7
100	10	100	378.0	52.7	327.2	341.2	492.9	494.3	1025.7
	20	100	304.9	54.9	324.1	353.7	497.4	500.3	1030.2
	30	100	253.2	57.4	320.9	367.9	502.2	506.8	1035.0
120	10	100	398.5	52.2	327.9	338.6	492.0	493.1	1024.8
	20	100	333.4	53.9	325.5	348.0	495.4	497.7	1028.2
	30	100	284.3	55.8	323.0	358.6	499.1	502.6	1031.9
Population 4									
80	10	100	654.3	55.6	350.6	366.9	1213.7	1242.5	1570.9
	20	100	399.1	58.6	336.3	368.8	950.6	987.7	1307.8
	30	100	281.8	62.2	322.4	370.9	772.2	810.4	1129.3
100	10	100	777.5	54.9	354.3	366.4	1303.0	1327.4	1660.2
	20	100	498.5	57.0	343.5	367.8	1069.1	1103.3	1426.2
	30	100	360.5	59.5	332.6	369.3	897.3	935.1	1254.5
120	10	100	889.0	54.4	356.8	366.1	1370.2	1390.6	1727.4
	20	100	597.7	56.0	348.4	367.2	1165.9	1196.8	1523.1
	30	100	442.9	57.8	339.9	368.3	1006.0	1041.9	1363.1

proposed estimator performs better for all choices of sample sizes than the other estimators in terms of relative efficiency.

6. Conclusion

We have expressed an efficient chain-ratio estimator for the population total using two auxiliary variables. This estimator is a weighted combination of the ratio estimator and Chand's (1975) chain-ratio estimator. The weights are obtained by minimizing the mean square error of this estimator. This estimator is more efficient than the ratio estimator and the chain-ratio estimator. With empirical data, our estimator in Eq. (28) has higher efficiency than other estimators.

Acknowledgments

The authors acknowledge the helpful comments of two referees.

References

Abu-Dayyeh, W. A., Ahmed, M. S., Ahmed, R. A., et al. (2003). Some estimators of a finite population mean using auxiliary information. *Applied Mathematics and Computation*, *139* (2–3): 287–298. doi:10.1016/S0096-3003(02)00180-7

Ahmed, M. S. (1997). The general class of chain estimators for ratio of two means using double sampling. *Communication in Statistics Theory and Methods*, *26*(9): 2249–2254. doi:10.1080/03610929708832044

Bahl, S. and Tuteja, R. K. (1991). Ratio and product type exponential estimator. *Journal of Information and Optimization Sciences*, *12*(1): 159–164. doi:10.1080/02522667.1991.10699058

Chand, L. (1975). Some ratio type estimators based on two or more auxiliary variables (Ph. D. thesis). Iowa State University.

Cochran, W. G. (1977). *Sampling Techniques* (3rd). New York: John Wiley & Sons.

Fuller, W. A. (2000). Two-phase sampling. *In Proceedings of the Survey Methods Section: SSC Annual Meeting*. Ottawa, Canada: Statistical Society of Canada, 23–30.

Hidiroglou, M. A. (2001). Double sampling. *Survey Methodology*, *27*(2): 143–154.

Hidiroglou, M. A. and Särndal, C. E. (1998). Use of auxiliary information for two phase sampling. *Survey Methodology*, *24*(1): 11–20.

Kadilar, C. and Cingi, H. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation*, *151*(3): 893–902. doi:10.1016/S0096-3003(03)00803-8

Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, *27*(1): 217–223. doi:10.1007/BF01893599

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*(201): 101–116. doi:10.1080/01621459.1938.10503378

Singh, B. K. and Choudhury, S. (2012). Exponential chain ratio and product type estimators for finite population mean under double sampling scheme. *Global Journal of Science Frontier Research*, *12*(6): 13–24.

Singh, D. and Chaudhary, F. S. (1986). *Theory and Analysis of Sample Survey Designs*. New York: John Wiley and Sons.

Singh, H. P. and Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean. *Statistics in Transition*, *6*(4): 555–560.

Sukhatme, B. V. (1962). Generalized hartley ross unbiased ratio type estimators. *Nature*, *196*: 1238. doi:10.1038/1961238a0

Vishwakarma, G. K. and Gangele, R. K. (2014). A class of chain ratio-type exponential estimators in double sampling using two auxiliary variates. *Applied Mathematics and Computation*, *227*: 171–175. doi:10.1016/j.amc.2013.11.027