SpringerPlus

RESEARCH Open Access

A note on a difference-type estimator for population mean under two-phase sampling design

Mursala Khan[1*] and Abdullah Yahia Al-Hossain[2]

*Correspondence: mursala.khan@yahoo.com
[1] Department of Mathematics, COMSATS Institute of Information Technology, Abbottabad 22060, Pakistan
Full list of author information is available at the end of the article

Abstract

In this manuscript, we have proposed a difference-type estimator for population mean under two-phase sampling scheme using two auxiliary variables. The properties and the mean square error of the proposed estimator are derived up to first order of approximation; we have also found some efficiency comparison conditions for the proposed estimator in comparison with the other existing estimators under which the proposed estimator performed better than the other relevant existing estimators. We show that the proposed estimator is more efficient than other available estimators under the two phase sampling scheme for this one example; however, further study is needed to establish the superiority of the proposed estimator for other populations.

Keywords: Study variable, Auxiliary variable, Bias, Mean squared-error, Two phase sampling, Exponential chain-type estimator, Efficiency

Background

In survey sampling, the use of the auxiliary information at the estimation stage is widely used in order to obtain improved designs and the precision of an estimator of the unknown population parameter. When the knowledge of the auxiliary variable is used at the estimation stage, the ratio, product and regression methods of estimation are widely employed in these situations.

The most important topic which is widely discussed in the various probability sampling schemes is the estimation of the population mean of the study variable. A large number of authors have paid their attention towards the formulation of new or modified estimators for the estimation of population mean, for the case, see Hansen and Hurwitz (1943), Sukhatme (1962), Srivastava (1970), Chand (1975), Cochran (1977), Kiregyera (1980, 1984), Srivastava et al. (1990), Bahl and Tuteja (1991), Singh et al. (2006, 2007, 2011), Singh and Choudhury (2012), Khare et al. (2013), Singh and Majhi (2014) and Khan (2015, 2016) etc.

Symbols and notations

Let us consider a finite population of size N of different units $U = \{U_1, U_2, U_3, \ldots, U_N\}$. Let y and x be the study and the auxiliary variable with corresponding values y_i and x_i respectively for the i-th unit $i = \{1, 2, 3, \ldots, N\}$ defined in a finite population U with

Springer Open

means $\bar{Y} = (1/N)\sum_{i=1}^{N} y_i$ and $\bar{X} = \left(1/N\right)\sum_{i=1}^{N} x_i$ of the study as well as auxiliary variable respectively.

Also let $S_y^2 = \left(1/N-1\right)\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2$ and $S_x^2 = \left(1/N-1\right)\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2$ be the population variances of the study and the auxiliary variable respectively and let C_y and C_x be the coefficient of variation of the study as well as auxiliary variable respectively, and ρ_{yx} be the correlation coefficient between x and y. Let y and x be the study and the auxiliary variable in the sample with corresponding values y_i and x_i respectively for the i-th unit $i = \{1, 2, 3\ldots, n\}$ in the sample with unbiased means $\bar{y} = \left(1/n\right)\sum_{i=1}^{n} y_i$ and $\bar{x} = \left(1/n\right)\sum_{i=1}^{n} x_i$ respectively.

Also let $\hat{S}_y^2 = \left(1/n-1\right)\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2$ and $\hat{S}_x^2 = \left(1/n-1\right)\sum_{i=1}^{n}(x_i-\bar{x})^2$ be the corresponding sample variances of the study as well as auxiliary variable respectively. Let $S_{yx} = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})}{N-1}$, $S_{yz} = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})(z_i-\bar{Z})}{N-1}$ and $S_{xz} = \frac{\sum_{i=1}^{N}(x_i-\bar{X})(z_i-\bar{Z})}{N-1}$ be the co-variances between their respective subscripts respectively. Similarly $b_{yx(n)} = \frac{\hat{S}_{yx}}{\hat{S}_x^2}$ is the corresponding sample regression coefficient of y on x based on a sample of size n. Also $C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$ and $C_z = \frac{S_z}{\bar{Z}}$ are the coefficients of variations of the study and auxiliary variables respectively.

Also $\theta = \left(\frac{1}{n}-\frac{1}{N}\right)$, $\theta_1 = \left(\frac{1}{n'}-\frac{1}{N}\right)$ and $\theta_2 = \left(\frac{1}{n}-\frac{1}{n'}\right)$.

Some existing estimators

Let us consider a finite population $U = \{U_1, U_2, U_3, \ldots, U_N\}$ of size N units. To estimate the population mean $\bar{Y}$ of the variable of interest say y taking values y_i, in the existence of two auxiliary variables say x and z taking values x_i and z_i for the ith unit U_i. We assume that there is a high correlation between y and x as compared to the correlation between y and z, (i.e. $\rho_{yx} > \rho_{yz} > 0$). When the population $\bar{X}$ of the auxiliary variable x is unknown, but information on the other cheaply auxiliary variable say z closely related to x but compared to x remotely to y, is available for all the units in a population. In such a situation we use a two phase sampling. In the two phase sampling scheme a large initial sample of size n' $(n' < N)$ is drawn from the population U by using simple random sample without replacement sampling (SRSWOR) scheme and measure x and z to estimate $\bar{X}$. In the second phase, we draw a sample (subsample) of size n from first phase sample of size n', i.e. $(n < n')$ by using (SRSWOR) or directly from the population U and observed the study variable y.

The variance of the usual simple estimator $t_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ up to first order of approximation is, given by

$$V(t_0) = \theta S_y^2 \tag{1}$$

The classical ratio and regression estimators in two-phase probability sampling and their mean square errors up to first order of approximation are, given by

$$t_1 = \frac{\bar{y}}{\bar{x}}\bar{x}' \tag{2}$$

$$MSE(t_1) = \bar{Y}^2\left[\theta C_y^2 + \theta_2\left(C_x^2 - 2\rho_{yx}C_yC_x\right)\right] \tag{3}$$

$$t_2 = \bar{y} + b_{yx(n)}\left(\bar{x}' - \bar{x}\right) \tag{4}$$

$$MSE(t_2) = S_y^2\left[\theta\left(1-\rho_{yx}^2\right)+\theta_1\rho_{yx}^2\right] \tag{5}$$

Chand (1975), suggested the following chain ratio-type estimator the suggested estimator is, given by

$$t_3 = \frac{\bar{y}}{\bar{x}}\frac{\bar{x}'}{\bar{z}'}\bar{Z} \tag{6}$$

The mean square error of the suggested estimator is, given as

$$MSE(t_3) = \bar{Y}^2\left[\theta C_y^2 + \theta_2\left(C_x^2 - 2\rho_{yx}C_yC_x\right) + \theta_1\left(C_z^2 - 2\rho_{yz}C_yC_z\right)\right] \tag{7}$$

Khare et al. (2013), proposed a generalized chain ratio in regression estimator for population mean, the recommended estimator is given by

$$t_4 = \bar{y} + b_{yx}\left\{\bar{x}'\left(\frac{\bar{Z}}{\bar{z}'}\right)^{\alpha} - \bar{x}\right\} \tag{8}$$

where α is the unknown constant, and the minimum mean square error at the optimum value of $\alpha = \frac{\rho_{yz}C_x}{\rho_{yx}c_z}$ is, given by

$$MSE(t_4) = \bar{Y}^2C_y^2\left[\theta - \theta_2\rho_{yx}^2 - \theta_1\rho_{yz}^2\right] \tag{9}$$

Recently Singh and Mahji (2014), suggested a chain-type exponential estimators for $\bar{Y}$ given by

$$t_5 = \frac{\bar{y}}{\bar{x}}\bar{x}'\exp\left(\frac{\bar{Z}-\bar{z}'}{\bar{Z}+\bar{z}'}\right) \tag{10}$$

$$t_6 = \bar{y} + b_{yx(n)}\left\{\bar{x}'\exp\left(\frac{\bar{Z}-\bar{z}'}{\bar{Z}+\bar{z}'}\right) - \bar{x}\right\} \tag{11}$$

$$t_7 = \bar{y}\exp\left(\frac{\bar{x}'-\bar{x}}{\bar{x}'+\bar{x}}\right)\frac{\bar{Z}}{\bar{z}'} \tag{12}$$

The mean square errors of the suggested estimators, up to first order of approximation are, given as follows

$$MSE(t_5) = \bar{Y}^2\left[\theta C_y^2 + \theta_2\left(C_x^2 - 2\rho_{yx}C_yC_x\right) + \frac{\theta_1}{4}\left(C_z^2 - 4\rho_{yz}C_yC_z\right)\right] \tag{13}$$

$$MSE(t_6) = \bar{Y}^2C_y^2\left[\theta_2\left(1-\rho_{yx}^2\right) + \theta_1\left(1 + \frac{\rho_{yx}^2}{4}\frac{C_z^2}{C_x^2} - \rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)\right] \tag{14}$$

$$MSE(t_7) = \bar{Y}^2\left[\theta C_y^2 + \frac{\theta_2}{4}\left(C_x^2 - 4\rho_{yx}C_yC_x\right) + \theta_1\left(C_z^2 - 2\rho_{yz}C_yC_z\right)\right] \tag{15}$$

The proposed estimator

On the lines of Khare et al. (2013), we propose a difference-type estimator for population mean under two-phase sampling scheme using two auxiliary variables; the suggested estimator is, given by

$$t_m = \bar{y} + k_1\left(\bar{x}'\frac{\bar{Z}}{\bar{z}'} - \bar{x}\right) + k_2\left(\bar{Z}\frac{\bar{x}'}{\bar{x}} - \bar{z}\right) \tag{16}$$

where k_1 and k_2 are the unknown constants,

To obtain the properties of the proposed estimator we define the following relative error terms and their expectations.

Let $e_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}$, $e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}$, $e'_1 = \frac{\bar{x}'-\bar{X}}{\bar{X}}$, $e_2 = \frac{\bar{z}-\bar{Z}}{\bar{Z}}$ and $e'_2 = \frac{\bar{z}-\bar{Z}}{\bar{Z}}$, such that

$$E(e_0) = E(e_i) = E\left(e'_i\right) = 0, \quad \text{for } i = 1,\, 2.$$

$$E\left(e_0^2\right) = \theta C_y^2, E\left(e_1^2\right) = \theta C_x^2, E\left(e_1'^2\right) = \theta_1 C_x^2, E\left(e_1e'_1\right) = \theta_1 C_x^2, E\left(e_2^2\right) = \theta C_z^2,$$

$$E\left(e_0e'_2\right) = \theta_1 C_{yz}, E(e_0e_1) = \theta C_{yx}, E\left(e_0e'_1\right) = \theta_1 C_{yx}, E(e_0e_2) = \theta C_{yz},$$

$$E\left(e_1e'_2\right) = E\left(e'_1e'_2\right) = \theta_1 C_{xz}, E(e_1e_2) = \theta C_{xz}, E\left(e_2'^2\right) = E\left(e_2e'_2\right) = \theta_1 C_z^2.$$

Rewriting (16), in terms of *e's*, we have

$$\begin{aligned} t_m = \Big[&\bar{Y}(1+e_0) + k_1\bar{X}\left(\left(1+e'_1\right)\left(1+e'_2\right)^{-1} - (1+e_1)\right) \\ &+ k_2\bar{Z}\left(\left(1+e'_1\right)(1+e_1)^{-1} - (1+e_2)\right)\Big] \end{aligned}$$

Expanding the right hand side of the above equation, and neglecting terms of *e's* having power greater than two, we have

$$t_m - \bar{Y} = \left[\bar{Y}e_0 - k_1\bar{X}\left(e_1 - e'_1 + e'_2 + e_2'^2 + e'_1e'_2\right) - k_2\bar{Z}\left(e_1 - e'_1 + e_2 - e_1^2 + e_1e'_1\right)\right] \tag{17}$$

On squaring and taking expectation on both sides of Eq. (17), and keeping terms up to second order, we have

$$\begin{aligned} MSE(t_m) = E\Big[&\bar{Y}^2e_0^2 + k_1^2\bar{X}^2\left(e_1^2 + e_1'^2 + e_2'^2 - 2e_1e'_1 + 2e_1e'_2 - 2e'_1e'_2\right) \\ &+ k_2^2\bar{Z}^2\left(e_1^2 + e_1'^2 + e_2^2 - 2e_1e'_1 + 2e_1e_2 - 2e'_1e_2\right) \\ &+ 2k_1k_2\bar{X}\bar{Z}\left(e_1^2 + e_1'^2 - 2e_1e'_1 + e_1e_2 - e'_1e_2 + e'_2e_1 - e_1'^2e'_1 + e'_1e_2\right) \\ &-2k_1\bar{Y}\bar{X}\left(e_0e_1 - e_0e'_1 + e_0e'_2\right) - 2k_2\bar{Y}\bar{Z}\left(e_0e_1 - e_0e'_1 + e_0e_2\right)\Big] \end{aligned}$$

Further simplifying, we get

$$\begin{aligned} MSE(t_m) = \Big[&\bar{Y}^2\theta C_y^2 + k_1^2\bar{X}^2\left(\theta_1C_z^2 + \theta_2C_x^2\right) + k_2^2\bar{Z}^2\left(\theta C_z^2 + \theta_2C_x^2 + 2\theta_2C_{xz}\right) \\ &+ 2k_1k_2\bar{X}\bar{Z}\left(\theta_2C_x^2 + \theta_1C_z^2 + \theta_2C_{xz}\right) \\ &- 2k_1\bar{Y}\bar{X}\left(\theta_2C_{yx} + \theta_1C_{yz}\right) - 2k_2\bar{Y}\bar{Z}\left(\theta_2C_{yx} + \theta C_{yz}\right)\Big] \end{aligned} \tag{18}$$

Now to find the minimum mean squared error of t_m, we differentiate Eq. (18) with respect to k_1 and k_2 respectively and putting it equal to zero, that is

$$\frac{\partial MSE(t_m)}{\partial k_1}=0 \quad \text{and} \quad \frac{\partial MSE(t_m)}{\partial k_2}=0$$

$$k_{1opt}=\frac{\bar{Y}(BC-DE)}{\bar{X}\left(AB-E^2\right)} \quad \text{and} \quad k_{2opt}=\frac{\bar{Y}(AD-CE)}{\bar{Z}\left(AB-E^2\right)}.$$

where $A=\theta_1 C_z^2+\theta_2 C_x^2$, $B=\theta C_z^2+\theta_2 C_x^2+2\theta_2 C_{xz}$, $C=\theta_2 C_{yx}+\theta_1 C_{yz}$, $D=\theta_2 C_{yx}+\theta C_{yz}$ and $E=\theta_1 C_z^2+\theta_2 C_x^2+\theta_2 C_{xz}$.

On substituting the optimum values of k_1 and k_2 in Eq. (18) we get the minimum mean square error (*MSE*) of the proposed estimator t_m up to order one is, given as

$$MSE(t_m)_{\text{min}}=\bar{Y}^2\left[\theta C_y^2-\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}\right] \tag{19}$$

Efficiency comparison

In this section, we have compare the propose estimator with the other existing estimators.

1. By Eqs. (1) and (19),

$$MSE(t_m)_{\text{min}}<MSE(t_0) \quad \text{if} \quad \left[\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}\right]>0.$$

2. By Eqs. (3) and (19)

$$MSE(t_m)_{\text{min}}<MSE(t_1) \quad \text{if} \quad \left[\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}+\theta_2\left(C_x^2-2\rho_{yx}C_yC_x\right)\right]>0.$$

3. By Eqs. (5) and (19),

$$MSE(t_m)_{\text{min}}<MSE(t_2) \quad \text{if} \quad \left[\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}-\theta_2 C_y^2\rho_{yx}^2\right]>0.$$

4. By Eqs. (7) and (19),

$$MSE(t_m)_{\text{min}}<MSE(t_3) \quad \text{if}$$
$$\left[\theta_2 C_x\left(C_x-2\rho_{yx}C_y\right)+\theta_1 C_z\left(C_z-2\rho_{yz}C_y\right)+\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}\right]>0.$$

5. By Eqs. (9) and (19),

$$MSE(t_m)_{\text{min}}<MSE(t_4) \quad \text{if} \quad \left[\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}-\left(\theta_2\rho_{yx}^2+\theta_1\rho_{yz}^2\right)C_y^2\right]>0.$$

6. By Eqs. (13) and (19),

$$MSE(t_m)_{\text{min}}<MSE(t_5) \quad \text{if}$$
$$\left[\theta_2\left(C_x^2-2C_{xy}\right)+\frac{\theta_1}{4}\left(C_z^2-4C_{yz}\right)+\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}\right]>0.$$

7. By Eqs. (14) and (19),

$MSE(t_m)_{\text{min}} < MSE(t_6)$ if

$$\left[\theta_1 C_y^2\left(\frac{\rho_{yx}^2}{4}\frac{C_z^2}{C_x^2} - \rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right) - \theta_2\rho_{yx}^2 C_y^2 + \frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)}\right] > 0.$$

8. By Eqs. (15) and (19),

$MSE(t_m)_{\text{min}} < MSE(t_7)$ if

$$\left[\frac{\left(AD^2+BC^2-2CDE\right)}{\left(AB-E^2\right)} + \frac{\theta_2}{4}\left(C_x^2 - 4C_{xy}\right) + \theta_1\left(C_z^2 - 2C_{yz}\right)\right] > 0.$$

Numerical comparison

To examine the performance of the proposed estimator with various existing estimators, we have considered a real data set from the literature the description of the population are, given by

Population Source, (Cochran 1977).

y: Number of placebo children;

x: Number of paralytic polio cases in the placebo group;

z: Number of paralytic polio cases in the not inoculated group.

$N = 34, n' = 15, n = 10, \bar{Y} = 4.92, \bar{X} = 2.59, \bar{Z} = 2.91$, $C_y^2 = 1.0248$, $C_x^2 = 1.5175$, $C_z^2 = 1.1492$, $C_{yx} = 0.9136$, $C_{yz} = 0.6978$, $\rho_{yx} = 0.7326$, $\rho_{yz} = 0.6430$, $\rho_{xz} = 0.6837$ (Table 1). We have use the following expression for Percentage Relative Efficiency (*PRE*)

$$PRE = \left[\frac{Var(t_0)}{MSE\left(t_j\right)\ or\ Var(t_j)}\right] * 100, \quad \text{for } j = 0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6, 7 \text{ and } m.$$

Table 1 The mean square errors (*MSE's*) and the Percent relative efficiencies (*PRE's*) of the estimators with respect to t_0

Population		
Estimator	***MSE's***	***PRE*** **(t_0,t_j)**
t_0	1.7525	100.00
t_1	1.5032	116.59
t_2	1.3073	134.06
t_3	1.2793	137.00
t4	0.9247	189.52
t_5	1.1312	154.92
t_6	1.0227	171.36
t_7	1.0982	159.58
t_m	0.8206	213.56

Conclusion

From the above table, we have observed that the proposed estimator has smaller mean square error and has higher percent relative efficiency than the other existing estimators. However, although the proposed estimator has the highest percent relative efficiency than other existing estimators for this one example, it could have lower relative efficiency for other populations. Further work is needed before it can be recommended for general use in practical surveys.

Authors' contributions

The authors contributed equally and significantly in writing this article. Both authors read and approved the final manuscript.

Author details

[1] Department of Mathematics, COMSATS Institute of Information Technology, Abbottabad 22060, Pakistan. [2] Department of Mathematics, Faculty of Science, Jazan University, Jazan 2097, Saudi Arabia.

Acknowledgements

The authors are very thankful to the editor and the anonymous learned referees for their valuable suggestions regarding the improvement of the paper.

Competing interests

The authors declare that they have no competing interests.

Received: 29 September 2015 Accepted: 18 May 2016
Published online: 14 June 2016

References

Bahl S, Tuteja RK (1991) Ratio and product type exponential estimator. Inf Optim Sci 12:159–163

Chand L (1975) Some ratio type estimator based on two or more auxiliary variables. Unpublished Ph.D. dissertation, Lowa State University, Ames, Lowa

Cochran WG (1977) Sampling techniques. Wiley, New-York

Hansen MH, Hurwitz WN (1943) On the theory of sampling from finite populations. Ann Math Stat 14:333–362

Khan M (2015) Improvement in estimating the finite population mean under maximum and minimum values in double sampling scheme. J Stat Appl Probab Lett 2(2):1–7

Khan M (2016) A ratio chain-type exponential estimator for finite population mean using double sampling. SpringerPlus 5:1–9

Khare BB, Srivastava U, Kumar K (2013) A generalized chain ratio in regression estimator for population mean using two auxiliary characters in sample survey. J Sci Res Banaras Hindu Univ Varanasi 57:147–153

Kiregyera B (1980) A chain ratio-type estimator in finite population mean in double sampling using two auxiliary variables. Metrika 27:217–223

Kiregyera B (1984) Regression-type estimator using two auxiliary variables and model of double sampling from finite populations. Metrika 31:215–223

Singh BK, Choudhury S (2012) Exponential chain ratio and product-type estimators for finite population mean under double sampling scheme. Glob J Sci Front Res Math Decis Sci 12(6):2249–4626

Singh GN, Majhi D (2014) Some chain-type exponential estimators of population mean in two-phase sampling. Stat Trans 15(2):221–230

Singh HP, Singh S, Kim JM (2006) General families of chain ratio type estimators of the population mean with known coefficient of variation of the second auxiliary variable in two phase sampling. J Korean Stat Soc 35(4):377–395

Singh R, Chuhan P, Swan N (2007) Families of estimators for estimating population mean using known correlation coefficient in two phase sampling. Stat Trans 8(1):89–96

Singh R, Chuhan P, Swan N, Smarandache F (2011) Improved exponential estimator for population variance using two auxiliary variables. Ital J Pure Appl Math 28:101–108

Srivastava SK (1970) A two phase estimator in sampling surveys. Austr J Stat 12:23–27

Srivastava SR, Khare BB, Srivastava SR (1990) A generalized chain ratio estimator for mean of finite population. J Indian Soc Agric Stat 42(1):108–117

Sukhatme BV (1962) Some ratio type estimators in two-phase sampling. J Am Stat Assoc 57:628–632