Almost unbiased exponential estimator for the finite population mean

Rajesh Singh, Pankaj Chauhan, and Nirmala Sawan,
School of Statistics, DAVV, Indore (M.P.), India
(rsinghstat@yahoo.com)

Florentin Smarandache
Chair of Department of Mathematics, University of New Mexico, Gallup, USA
(smarand@unm.edu)

Abstract

In this paper we have proposed an almost unbiased ratio and product type exponential estimator for the finite population mean $\bar{Y}$. It has been shown that Bahl and Tuteja (1991) ratio and product type exponential estimators are particular members of the proposed estimator. Empirical study is carried to demonstrate the superiority of the proposed estimator.

Keywords: Auxiliary information, bias, mean-squared error, exponential estimator.

1. Introduction

It is well known that the use of auxiliary information in sample surveys results in substantial improvement in the precision of the estimators of the population mean. Ratio, product and difference methods of estimation are good examples in this context. Ratio method of estimation is quite effective when there is a high positive correlation between study and auxiliary variables. On other hand, if this correlation is negative (high), the product method of estimation can be employed effectively.

Consider a finite population with N units $(U_1, U_2, ..., U_N)$ for each of which the information is available on auxiliary variable x. Let a sample of size n be drawn with simple random sampling without replacement (SRSWOR) to estimate the population mean of character y under study. Let $(\bar{y},\bar{x})$ be the sample mean estimator of $(\bar{Y},\bar{X})$ the population means of y and x respectively.

In order to have a survey estimate of the population mean $\bar{Y}$ of the study character y, assuming the knowledge of the population mean $\bar{X}$ of the auxiliary character x, Bahl and Tuteja (1991) suggested ratio and product type exponential estimator

$$t_1 = \bar{y}\exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right) \tag{1.1}$$

$$t_2 = \bar{y}\exp\left(\frac{\bar{x}-\bar{X}}{\bar{x}+\bar{X}}\right) \tag{1.2}$$

Up to the first order of approximation, the bias and mean-squared error (MSE) of t_1 and t_2 are respectively given by

$$B(t_1) = \left(\frac{N-n}{nN}\right)\bar{Y}\frac{C_x^2}{2}\left(\frac{1}{2}-K\right) \tag{1.3}$$

$$MSE(t_1) = \left(\frac{N-n}{nN}\right)\bar{Y}^2\left[C_y^2 + C_x^2\left(\frac{1}{4}-K\right)\right] \tag{1.4}$$

$$B(t_2) = \left(\frac{N-n}{nN}\right)\bar{Y}\frac{C_x^2}{2}\left(\frac{1}{2}+K\right) \tag{1.5}$$

$$MSE(t_2) = \left(\frac{N-n}{nN}\right)\bar{Y}^2\left[C_y^2 + C_x^2\left(\frac{1}{4}+K\right)\right] \tag{1.6}$$

where $S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})^2$, $S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i-\bar{X})^2$, $C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$,

$K = \rho\left(\frac{C_y}{C_x}\right)$, $\rho = \frac{S_{yx}}{(S_y S_x)}$, $S_{yx} = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})$.

From (1.3) and (1.5), we see that the estimators t_1 and t_2 suggested by Bahl and Tuteja (1991) are biased estimator. In some applications bias is disadvantageous. Following Singh and Singh (1993) and Singh and Singh (2006) we have proposed almost unbiased estimators of $\overline{Y}$.

2. Almost unbiased estimator

Suppose $t_0 = \overline{y}$, $t_1 = \overline{y}\exp\left(\frac{\overline{X}-\overline{x}}{\overline{X}+\overline{x}}\right)$, $t_2 = \overline{y}\exp\left(\frac{\overline{x}-\overline{X}}{\overline{x}+\overline{X}}\right)$

such that $t_0,\ t_1,\ t_2 \in H$, where H denotes the set of all possible estimators for estimating the population mean $\overline{Y}$. By definition, the set H is a linear variety if

$$t_h = \sum_{i=0}^{2} h_i t_i \in H \tag{2.1}$$

for $\sum_{i=0}^{2} h_i = 1,\ h_i \in R$ (2.2)

where $h_i\,(i = 0,1,2)$ denotes the statistical constants and R denotes the set of real numbers.

To obtain the bias and MSE of t_h, we write

$\overline{y} = \overline{Y}(1+e_0),\ \overline{x} = \overline{X}(1+e_1)$.

such that

E (e_0)=E (e_1)=0.

$$E(e_0^2) = \left(\frac{N-n}{Nn}\right)C_y^2,\ E(e_1^2) = \left(\frac{N-n}{Nn}\right)C_x^2,\ E(e_0e_1) = \left(\frac{N-n}{Nn}\right)\rho C_y C_x .$$

Expressing t_h in terms of e's, we have

$$t_h = \overline{Y}(1+e_0)\left[h_0 + h_1\exp\left(\frac{-e_1}{2+e_1}\right) + h_2\exp\left(\frac{e_1}{2+e_1}\right)\right] \tag{2.3}$$

Expanding the right hand side of (2.3) and retaining terms up to second powers of e's, we have

$$t_h = \overline{Y}\left[1+e_0-\frac{e_1}{2}(h_1-h_2)+h_1\frac{e_1^2}{8}+h_2\frac{e_1^2}{8}-h_1\frac{e_0e_1}{2}+h_2\frac{e_0e_1}{2}\right] \tag{2.4}$$

Taking expectations of both sides of (2.4) and then subtracting $\overline{Y}$ from both sides, we get the bias of the estimator t_h, up to the first order of approximation as

$$B(t_h) = \left(\frac{N-n}{Nn}\right)\overline{Y}\frac{C_x^2}{2}\left[\frac{1}{4}(h_1+h_2)-K(h_1-h_2)\right] \tag{2.5}$$

From (2.4), we have

$$(t_h-\overline{Y}) \cong \overline{Y}\left[e_0-h\frac{e_1}{2}\right] \tag{2.6}$$

where $h=h_1-h_2$. (2.7)

Squaring both the sides of (2.7) and then taking expectations, we get MSE of the estimator t_h, up to the first order of approximation, as

$$MSE(t_h) = \left(\frac{N-n}{Nn}\right)\overline{Y}^2\left[C_y^2+C_x^2h\left(\frac{h}{4}-K\right)\right] \tag{2.8}$$

which is minimum when

$$h = 2K. \tag{2.9}$$

Putting this value of h = 2K in (2.1) we have optimum value of estimator as t_h(optimum).

Thus the minimum MSE of t_h is given by

$$\min.MSE(t_h) = \left(\frac{N-n}{Nn}\right)\overline{Y}^2C_y^2(1-\rho^2) \tag{2.10}$$

which is same as that of traditional linear regression estimator.

From (2.7) and (2.9), we have

$$h_1 - h_2 = h = 2K \, . \tag{2.11}$$

From (2.2) and (2.11), we have only two equations in three unknowns. It is not possible to find the unique values for h_i's, i=0,1,2. In order to get unique values of h_i's, we shall impose the linear restriction

$$\sum_{i=0}^{2} h_i B(t_i) = 0. \tag{2.12}$$

where $B(t_i)$ denotes the bias in the i^{th} estimator.

Equations (2.2), (2.11) and (2.12) can be written in the matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & B(t_1) & B(t_2) \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2K \\ 0 \end{bmatrix} \tag{2.13}$$

Using (2.13), we get the unique values of h_i's(i=0,1,2) as

$$\left.\begin{aligned} h_0 &= 1 - 4K^2 \\ h_1 &= K + 2K^2 \\ h_2 &= -K + 2K^2 \end{aligned}\right\} \tag{2.14}$$

Use of these h_i's (i=0,1,2) remove the bias up to terms of order $o(n^{-1})$ at (2.1).

3. Two phase sampling

When the population mean $\bar{X}$ of x is not known, it is often estimated from a preliminary large sample on which only the auxiliary characteristic is observed. The value of population mean $\bar{X}$ of the auxiliary character x is then replaced by this estimate. This technique is known as the double sampling or two-phase sampling.

The two-phase sampling happens to be a powerful and cost effective (economical) procedure for finding the reliable estimate in first phase sample for the unknown parameters of the auxiliary variable x and hence has eminent role to play in survey sampling, for instance, see; Hidiroglou and Sarndal (1998).

When $\bar{X}$ is unknown, it is sometimes estimated from a preliminary large sample of size n' on which only the characteristic x is measured. Then a second phase sample of size $n(n < n')$ is drawn on which both y and x characteristics are measured. Let $\bar{x} = \frac{1}{n'}\sum_{i=1}^{n'} x_i$ denote the sample mean of x based on first phase sample of size n'; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ be the sample means of y and x respectively based on second phase of size n.

In double (or two-phase) sampling, we suggest the following modified exponential ratio and product estimators for $\bar{Y}$, respectively, as

$$t_{1d} = \bar{y}\exp\left(\frac{\bar{x}' - \bar{x}}{\bar{x}' + \bar{x}}\right) \tag{3.1}$$

$$t_{2d} = \bar{y}\exp\left(\frac{\bar{x} - \bar{x}'}{\bar{x} + \bar{x}'}\right) \tag{3.2}$$

To obtain the bias and MSE of t_{1d} and t_{2d}, we write

$$\bar{y} = \bar{Y}(1 + e_0),\ \bar{x} = \bar{X}(1 + e_1),\ \bar{x}' = \bar{X}(1 + e_1')$$

such that

$$E(e_0) = E(e_1) = E(e_1') = 0$$

and

$$E(e_0^2) = f_1 C_y^2, \qquad E(e_1^2) = f_1 C_x^2, \qquad E(e_1'^2) = f_2 C_x^2,$$

$$E(e_0 e_1) = f_1 \rho C_y C_x,$$

$$E(e_0 e_1') = f_2 \rho C_y C_x ,$$

$$E(e_1 e_1') = f_2 C_x^2 .$$

where $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$, $f_2 = \left(\frac{1}{n'} - \frac{1}{N}\right)$.

Following standard procedure we obtain

$$B(t_{1d}) = \overline{Y} f_3 \left[\frac{C_x^2}{8} - \frac{1}{2}\rho C_y C_x\right] \quad (3.3)$$

$$B(t_{2d}) = \overline{Y} f_3 \left[\frac{C_x^2}{8} + \frac{1}{2}\rho C_y C_x\right] \quad (3.4)$$

$$MSE(t_{1d}) = \overline{Y}^2 \left[f_1 C_y^2 + f_3 \left(\frac{C_x^2}{4} - \rho C_x C_y\right)\right] \quad (3.5)$$

$$MSE(t_{2d}) = \overline{Y}^2 \left[f_1 C_y^2 + f_3 \left(\frac{C_x^2}{4} + \rho C_x C_y\right)\right] \quad (3.6)$$

where $f_3 = \left(\frac{1}{n} - \frac{1}{n'}\right)$.

From (3.3) and (3.4) we observe that the proposed estimators t_{1d} and t_{2d} are biased, which is a drawback of an estimator is some applications.

4. Almost unbiased two-phase estimator

Suppose $t_0 = \overline{y}$, t_{1d} and t_{2d} as defined in (3.1) and (3.2) such that $t_0, t_{1d}, t_{2d} \in W$, where W denotes the set of all possible estimators for estimating the population mean $\overline{Y}$. By definition, the set W is a linear variety if

$$t_W = \sum_{i=0}^{2} w_i t_i \in W . \quad (4.1)$$

for $\sum_{i=1}^{2} w_i = 1,\ w_i \in R$. (4.2)

where $w_i(i = 0,1,2)$ denotes thee statistical constants and R denotes the set of real numbers.

To obtain the bias and MSE of t_w, using notations of section 3 and expressing t_w in terms of e's, we have

$$t_w = \overline{Y}(1+e_0)\left[w_0 + w_1 \exp\left(\frac{e_1' - e_1}{2}\right) + w_2 \exp\left(\frac{e_1 - e_1'}{2}\right)\right] \tag{4.3}$$

$$t_w = \overline{Y}[1 + e_0 - \frac{w}{2}(e_1 - e_1') + \frac{w_1}{8}\left(e_1^2 + e_1'^2\right) + \frac{w_2}{8}\left(e_1^2 + e_1'^2\right) - \left(\frac{w_1}{4} + \frac{w_2}{4}\right)e_1 e_1'$$

$$+ \frac{w}{2}(e_0 e_1' - e_0 e_1)] \tag{4.4}$$

where $w = w_1 - w_2$. (4.5)

Taking expectations of both sides of (4.4) and then subtracting $\overline{Y}$ from both sides, we get the bias of the estimator t_w, up to the first order f approximation as

$$\text{Bias}(t_w) = \overline{Y} f_3\left[\left(\frac{w_1 + w_2}{8}\right)C_x^2 - \frac{w}{2}\rho C_y C_x\right] \tag{4.6}$$

From (4.4), we have

$$t_w \cong \overline{Y}\left[e_0 - \frac{w}{2}(e_1 - e_1')\right] \tag{4.7}$$

Squaring both sides of (4.7) and then taking expectation, we get MSE of the estimator t_w, up to the first order of approximation, as

$$\mathrm{MSE}(t_w) = \overline{Y}^2\left[f_1 C_y^2 + f_3 w C_x^2\left(\frac{w}{4}\right) - K\right] \qquad (4.8)$$

which is minimum when

$$w = 2K. \qquad (4.9)$$

Thus the minimum MSE of t_w is given by –

$$\min.\mathrm{MSE}(t_w) = \overline{Y}^2 C_y^2\left[f_1 - f_3\rho^2\right] \qquad (4.10)$$

which is same as that of two-phase linear regression estimator. From (4.5) and (4.9), we have

$$w_1 - w_2 = w = 2K \qquad (4.11)$$

From (4.2) and (4.11), we have only two equations in three unknowns. It is not possible to find the unique values for w_i's$(i = 0,1,2)$. In order to get unique values of h_i's, we shall impose the linear restriction

$$\sum_{i=0}^{2} w_i B(t_{id}) = 0 \qquad (4.12)$$

where $B(t_{id})$ denotes the bias in the i^{th} estimator.

Equations (4.2), (4.11) and (4.12) can be written in the matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & B(t_{1d}) & B(t_{2d}) \end{bmatrix}\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2K \\ 0 \end{bmatrix} \qquad (4.13)$$

Solving (4.13), we get the unique values of w_i's$(i = 0,1,2)$ as –

$$\left.\begin{aligned} w_0 &= 1 - 8K^2 \\ w_1 &= K + 4K^2 \\ w_2 &= -K + 4K^2 \end{aligned}\right\} \qquad (4.14)$$

Use of these w_i's$(i = 0,1,2)$ removes the bias up to terms of order $o(n^{-1})$ at (4.1).

5. Empirical study

The data for the empirical study are taken from two natural population data sets considered by Cochran (1977) and Rao (1983).

Population I: Cochran (1977)

$C_y = 1.4177$, $C_x = 1.4045$, $\rho = 0.887$.

Population II: Rao (1983)

$C_y = 0.426$, $C_x = 0.128$, $\rho = -0.7036$.

In table (5.1), the values of scalar h_i's (i = 0,1,2) are listed.

Table (5.1): Values of h_i's (i =0,1,2)

Scalars	Population	
	I	**II**
$\mathbf{h_0}$	-2.2065	-20.93
$\mathbf{h_1}$	2.4985	8.62
$\mathbf{h_2}$	0.7079	13.30

Using these values of h_i's (i = 0,1,2) given in the table 5.1, one can reduce the bias to the order o (n^{-1}) in the estimator t_h at (2.1).

In table 5.2, Percent relative efficiency (PRE) of $\bar{y}$, t_1, t_2 and t_h (in optimum case) are computed with respect to $\bar{y}$.

Table 5.2: PRE of different estimators of $\bar{Y}$ with respect to $\bar{y}$.

Estimators	PRE (., $\bar{y}$)	
	Population I	**Population II**
$\bar{y}$	100	100
t_1	272.75	32.55
t_2	47.07	126.81
t_h **(optimum)**	468.97	198.04

Table 5.2 clearly shows that the suggested estimator t_h in its optimum condition is better than usual unbiased estimator $\bar{y}$, Bahl and Tuteja (1991) estimators t_1 and t_2.

For the purpose of illustration for two-phase sampling, we consider following populations:

Population III: Murthy (1967)

y : Output
x : Number of workers
$C_y = 0.3542$, $C_x = 0.9484$, $\rho = 0.9150$, N = 80, $n' = 20$, n = 8.

Population IV: Steel and Torrie(1960)

$C_y = 0.4803$, $C_x = 0.7493$, $\rho = -0.4996$, N = 30, $n' = 12$, n = 4.

In table 5.3 the values of scalars w_i's$(i = 0,1,2)$ are listed.

Table 5.3: Values of w_i's$(i = 0,1,2)$

Scalars	Population I	Population II
w_0	0.659	0.2415
w_1	0.808	0.0713
w_2	0.125	0.6871

Using these values of w_i's$(i = 0,1,2)$ given in table 5.3 one can reduce the bias to the order $o(n^{-1})$ in the estimator t_w at 5.3.

In table 5.4 percent relative efficiency (PRE) of $\bar{y}$, t_{1d}, t_{2d} and t_w (in optimum case) are computed with respect to $\bar{y}$.

Table 5.4: PRE of different estimators of $\bar{Y}$ with respect to $\bar{y}$.

Estimators	PRE $(., \bar{y})$	
	Population I	**Population II**
$\bar{y}$	100	100
t_{1d}	128.07	74.68
t_{2d}	41.42	103.64
t_w	138.71	106.11

References

Bahl, S. and Tuteja, R.K. (1991): Ratio and product type exponential estimator. Information and optimization sciences, 12 (1), 159-163.

Cochran (1977):