An Improved Regression Estimator for Estimating Population Mean

Vyas Dubey and S.K. Singh
Pt. Ravishankar Shukla University, Raipur 492010
(Received : May, 1994)

SUMMARY

The paper deals with a new estimator of population mean. Under certain conditions, the proposed estimator is more efficient than usual difference estimator and the modified estimators proposed by Jain [3] and Rao [6]. Further a regression type estimator has been developed after estimating the constants involved in the proposed estimator. This estimator is more efficient than usual regression estimator for bivariate symmetric populations. At last, numerical illustrations have also been made in order to highlight the important results.

Key words : Bias, Mean square error (MSE), Simple random sampling without replacement (SRSWOR) procedure, Finite population correction factor, Coefficient of variation (CV), Relative efficiency.

1. Introduction

Hansen, Hurwitz and Madow [2] have suggested a difference method of estimation for estimating population mean $\overline{Y}$ of a certain characteristic y. Let x be auxiliary characteristic highly correlated with y. $\bar{y}$ and $\bar{x}$ be sample means of y and x respectively; and $\overline{X}$ be known population mean of x. The difference estimator is defined as

$$\bar{y}_d = \bar{y} + K\left(\overline{X} - \bar{x}\right) \tag{1.1}$$

where K is a suitable constant. Modifications over this estimator have been done by Bedi and Hajala [1], Jain [3] and Rao [6]. These estimators are respectively as follows :

$$\bar{y}_{BH} = W\left[\bar{y} + \beta(\overline{X} - \bar{x})\right] \tag{1.2}$$

$$\bar{y}_J = \alpha_1 \bar{y} + \alpha_2\left(\overline{X} - \bar{x}\right), \quad \alpha_1 + \alpha_2 = 1 \tag{1.3}$$

and $$\bar{y}_{RG} = K_1 \bar{y} + K_2 \left(\bar{X} - \bar{x}\right) \qquad (1.4)$$

$W, \alpha_1, \alpha_2, K_1$ and K_2 are known constants. β is the regression coefficient of y on x. Gupta and Kothwala [5] have suggested an estimator similar to $\bar{y}_{BH}$ in case multi-auxiliary information is available. For small samples, $\bar{y}_{RG}$ is generally more efficient than others. But Rao [6] has shown in his study that for large samples $\bar{y}_d, \bar{y}_{BH}$, and $\bar{y}_{RG}$ are approximately equally efficient.

The purpose of this paper is to develop a new estimator which is more efficient than above estimators in large samples. Its efficiency has been compared in section 3.

2. *Proposed Estimator*

The proposed estimator is

$$T = w_1 \bar{y} + w_2 \bar{x} + (1 - w_1 - w_2) \bar{X} \qquad (2.1)$$

where w_1 and w_2 are suitably chosen constants. If a sample of size n is taken from the population of size N by SRSWOR procedure so that finite population correction factor can be ignored, the corresponding bias and MSE of T are given by

$$B(T) = (w_1 - 1)\left(\bar{Y} - \bar{X}\right) \qquad (2.2)$$

$$M(T) = \left[\frac{\left(w_1^2 S_{02} + w_2^2 S_{20} + 2 w_1 w_2 S_{11}\right)}{n}\right] + (w_1 - 1)^2 \left(\bar{Y} - \bar{X}\right)^2 \qquad (2.3)$$

where $$S_{ij} = E\left(x - \bar{X}\right)^i \left(y - \bar{Y}\right)^j \quad i, j = 0, 1, 2$$

The best values of w_1 and w_2 for which M (T) will be minimum are given by

$$w_1^* = \left(1 + \frac{Q}{n}\right)^{-1}, \; Q = \frac{S_{02}\left(1 - \rho^2\right)}{\left(\bar{Y} - \bar{X}\right)^2} \qquad (2.4)$$

and $$w_2^* = -\beta w_1^* \qquad (2.5)$$

respectively. In this case M(T) reduces to

$$M_0(T) = \frac{S_{02}\left(1 - \rho^2\right)}{(n + Q)} \qquad (2.6)$$

3. Efficiency Comparisons

The minimum MSE attained by the estimators $\bar{y}_d$, $\bar{y}_{BH}$, $\bar{y}_J$ and $\bar{y}_{RG}$, are as under

$$M_0(\bar{y}_d) = \frac{S_{02}\left(1-\rho^2\right)}{n} \tag{3.1}$$

$$M_0(\bar{y}_{BH}) = \frac{S_{02}\left(1-\rho^2\right)}{(n+C)} = M_0(\bar{y}_{RG}), C = \frac{S_{02}\left(1-\rho^2\right)}{\overline{Y}^2} \tag{3.2}$$

$$M_0(\bar{y}_J) = \frac{\left[S_{02}\left(n\overline{Y}^2\right)+S_{20}\left(1-\rho^2\right)\right]}{\left[n\left\{n\overline{Y}^2+(S_{02}+S_{20}+2S_{11})\right\}\right]} \tag{3.3}$$

It is easy to verify that $M_0(T)$ is always smaller than $M_0(\bar{y}_d)$. For others we have

$$M_0(T) < M_0(\bar{y}_{BH}) \text{ if } 0 < \overline{X} < 2\overline{Y} \tag{3.4}$$

and $M_0(T) < M_0(\bar{y}_J)$ if $n\left[n\rho^2\overline{Y}^2 - 2\left(1-\rho^2\right)S_{11}\right]D^2 + \left(1-\rho^2\right)S_{02}S_{20} > 0$

(3.5)

$D = \left(\overline{Y}-\overline{X}\right)$. The condition (3.5) is always true for negatively correlated characteristics. A sufficient condition under which it will be true for positively correlated characteristics is

$$\rho^2 > \frac{2S_{11}}{2S_{11}+n\overline{Y}^2} \tag{3.6}$$

Thus for highly positively correlated variables, T will be more efficient than $\bar{y}_J$.

4. The Case where Parameters are Unknown

A good guess of w_1^* and w_2^* may be available from past data or repeated surveys. But it is not always possible, so we propose a modified regression estimator as

$$t = \bar{y}_{1r} + \frac{q}{n}\left(\bar{y}_{1r} - \overline{X}\right) \tag{4.1}$$

where

$$q = \frac{s_{02}\left(1-r^2\right)}{(\bar{y}-\bar{x})^2}, \quad s_{ij} = \sum_{1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{(n-1)}, (i, j = 0, 1, 2)$$

and

$$\bar{y}_{lr} = \bar{y} + \hat{\beta}\left(\bar{X} - \bar{x}\right), \quad \hat{\beta} = \frac{s_{11}}{s_{20}} \tag{4.2}$$

is usual regression estimator, r is sample correlation coefficient between y and x. If we assume that y and x follow bivariate normal distribution then following Sukhatme *et al.* [7] and writing

$$(\bar{y}-\bar{x})^2 = \left(\bar{Y}-\bar{X}\right)^2 (1+e)^2 \tag{4.3}$$

$E(e) = 0$ with $e < 1$ so that its higher order terms may be neglected, we find that t is generally biased having value

$$B(t) = \frac{\left(\bar{Y}-\bar{X}\right)Q}{n} \text{ up to order } n^{-1} \tag{4.4}$$

Its MSE up to order n^{-2} is obtained as

$$M(t) = \frac{M(\bar{y}_{lr}) - \left(\bar{Y}-\bar{X}\right)^2 Q^2}{n^2} \tag{4.5}$$

where

$$M\left(\bar{y}_{lr}\right) = \frac{S_{02}}{n}\left(1 + \frac{1}{(n-3)}\right) \text{ (See Tikkiwal [8])} \tag{4.6}$$

It is clear from (4.4) that t is more efficient than regression estimator.

5. Numerical Illustration

Let us consider the data from Jessen [4] where a sample was taken in two successive years of the same 100 randomly selected Iowa farms from a population of more than 2000 farms to determine net cash income each year. By denoting net cash income in the first and second year by x and y respectively, summarized data is given below

$$\left(\frac{\bar{Y}}{\bar{X}}\right) = 1.20, CV(y) = CV(x) = 1.80 \text{ and } \rho = 0.78$$

Table 1 shows the efficiencies (%) of above estimators with respect to $\bar{y}$.

Table 1

Estimator	Efficiency
$\bar{y}_d$	255.3626
$\bar{y}_{RG}$	258.6026
$\bar{y}_J$	109.9215
T	372.0025
$\bar{y}_{lr}$	252.7568
t	461.3204

It is evident from Table 1 that proposed estimators T and t are more efficient than existing ones.

REFERENCES

[1] Bedi, P.K. and Hajela, D. (1984). An estimator for population mean utilizing known coefficient of variation and auxiliary variable. *J. Stat. Res.*, **18**, 29-33.

[2] Hensen, M.N., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*, John Wiley and Sons, New York.

[3] Jain, R.K. (1987). Properties of estimators in simple random sampling using auxiliary variable. *Metron*, **45 (1-2)**, 265-271.

[4] Jessen, R.J. (1978). *Statistical Survey Techniques*. John Wiley & Sons, New York.

[5] Kothwala, N.H. and Gupta, P.C. (1989). Estimation of population mean with knowledge of coefficient of variation with p auxiliary variables. *Metron*, **47 (1-4)**, 107-119.

[6] Rao, T.J. (1991). On certain methods of improving ratio and regression estimators. *Comm. Statist.–Theory and Methods*, **20(10)**, 3325-3340.

[7] Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. ISAS Publication, New Delhi, (INDIA).

[8] Tikkiwal, B.D. (1985). Some aspects of elements of theory of sample surveys. Monograph on "*All India Advanced Institute in Sample Surveys and Inference*" organised in October -November 1985, University of Jaipur, Rajasthan.