ESTIMATION OF FINITE POPULATION VARIANCE

H. P. SINGH, L. N. UPADHYAYA* and U. D. NAMJOSHI

Department of Mathematics and Statistics, College of Agricultural Engineering, Jawaharlal Nehru Krishi Vishwa Vidyalaya, Jabalpur 482 004, India.

**Indian School of Mines, Dhanbad 826 004, India.*

ABSTRACT

This paper proposes two classes of estimators using supplimentary information on an auxiliary character in different situations, viz., (i) when $\bar{X}$, the population mean of auxiliary variable x is known, and (ii) when σ_x^2, the population variance of x is known, and analyses their properties.

INTRODUCTION

The use of supplimentary information has been dealt with at great length for improving estimators in sample surveys. It is a common practice to use auxiliary information on a character x in the estimation of finite population mean $\bar{Y}$ of a character y under study. It seems also reasonable that under suitable conditions the efficient estimation of σ_y^2, the variance of finite population of the character y is also possible. However, the problem considered is quite suitable for skewed populations. In the case of genetical, medical or biological studies, the estimation of variance assumes importance. In fact a number of estimators may be defined for population variance under different situations[1,2].

Let $U = (U_1, U_2, \ldots, U_N)$ denote the population of N units and let (y, x) be the variate defined on U taking values (y_i, x_i) on U_i $(i = 1, 2, \ldots, N)$. The problem is to estimate the population variance σ_y^2 of the study character y. The conventional unbiased estimator of σ_y^2 based on SRSWOR is given by

$$s_y^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2, \tag{1}$$

where $\bar{y}$ is the sample mean based on n observations.

To improve the conventional estimator s_y^2 information on an auxiliary variable x has been utilized[2] in different situations and proposed the following estimators:

(i) When the population mean $\bar{X}$ of auxiliary character x is known:

$$t_1 = s_y^2 \left(\frac{\bar{X}}{\bar{x}} \right)^W \tag{2}$$

$$t_2 = s_y^2 \frac{\bar{X}}{[\bar{X} + W(\bar{x} - \bar{X})]}, \tag{3}$$

where W being a suitably chosen constant and $\bar{x}$ is sample mean of x based on n observations.

(ii) When the population variance σ_x^2 of the auxiliary character x is known:

$$t_3 = s_y^2 (\sigma_x^2 / s_x^2)^\alpha$$

$$t_4 = s_y^2 \frac{\sigma_x^2}{[\sigma_x^2 + \alpha(s_x^2 - \sigma_x^2)]}, \tag{5}$$

where α is a suitably chosen constant.

For $\alpha = 1$, the estimator t_3 reduces to

$$t_5 = s_y^2 (\sigma_x^2 / s_x^2), \tag{6}$$

which is due to Isaki[3]. Das and Tripathi[2] also proposed two estimators when the coefficient of variation of auxiliary x is known and studied their properties to the first degree of approximation. We have not considered this case here.

In the present paper we have proposed two classes of estimators in the two different situations stated above. The properties of the proposed estimators have been discussed for exact sample size (ignoring finite population correction terms). To see the performance of our estimators over other estimators an empirical study is carried out.

CLASS OF ESTIMATORS AND ITS PROPERTIES

We have considered the following estimators for σ_y^2 in two situations:

(i) when $\bar{X}$ the population mean of x is known;

$$d_1 = W_1 s_y^2 - W_2 (\bar{x} - \bar{X}), \tag{7}$$

where W_1 and W_2 are suitably chosen constants to be determined such that MSE of d_1 is minimum.

(ii) When σ_x^2 the population variance of x is known;

$$d_2 = W_1^* s_y^2 - W_2^* (s_x^2 - \sigma_x^2), \tag{8}$$

where W_1^* and W_2^* are constants to be chosen suitably such that MSE of d_2 is least.

For $W_2 = W_2^* = 0$ both the estimators d_1 and d_2 reduce to[4]

$$d_3 = W^* s_y^2, \tag{9}$$

where W^* is a constant.

The exact biases and MSE's (ignoring the finite population correction term) of d_1 and d_2 are respectively, given by

$$B(d_1) = (W_1 - 1)\sigma_y^2 \tag{10}$$

$$B(d_2) = (W_1^* - 1)\sigma_y^2 \tag{11}$$

$$\mathrm{MSE}(D_1) = \frac{\sigma_y^4}{n}\left[W_1^2\{n + \beta_2^*(y)\} + W_2^2 \left(\frac{\bar{X}}{\sigma_y^2}\right) C_x^2 - 2W_1 W_2 \left(\frac{K\bar{X}}{\sigma_y^2}\right) - 2nW_1 + n\right], \tag{12}$$

$$\mathrm{MSE}(d_2) = \frac{\sigma_y^4}{n}\left[W_1^{*2}\{n + \beta_2^*(y)\} + W_2^{*2}\left(\frac{\sigma_x}{\sigma_y}\right)^4 \beta_2^*(x) - 2W_1^* W_2^* h^* \left(\frac{\sigma_x}{\sigma_y}\right)^2 - 2nW_2^* + n\right], \tag{13}$$

where

$\beta_2^*(y) = \{\beta_2(y) - 1\}$, $\beta_2^*(x) = \{\beta_2(x) - 1\}$,

$h^* = (h - 1)$, $\beta_2(y) = \mu_4(y)/\mu_2^2(y)$, $\beta_2(x) = \mu_4(x)/\mu_2^2(x)$, $h = \mu_{22}(y, x)/(\mu_2(x)\mu_2(y))$,

$K = \mu_{21}(y, x)/(\sigma_y^2 \bar{X})$,

$$\mu_2(z) = N^{-1} \sum_{i=1}^{N} (z_i - \bar{Z})^2,$$

$$\mu_4(z) = \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{Z})^4, \; z = y, x;$$

$$\mu_{21}(y, x) = N^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2 (x_i - \bar{X}),$$

$$\mu_{22}(y, x) = N^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2 (x_i - \bar{X})^2.$$

The MSE's of d_1 and d_2 are respectively minimized for

$$W_{10} = \frac{nC_x^2}{[(n + \beta_2^*(y))\, C_x^2 - K^2]};$$

$$W_{20} = -\frac{nC_y^2 R^2 \bar{X}}{[(n + \beta_2^*(y))\, C_x^2 - K^2]}; \tag{14}$$

and

$$\left.\begin{aligned} W_{10}^* &= \frac{n\beta_2^*(x)}{[(n + \beta_2^*(y))\beta_2^*(x) - h^{*2}]\sigma_x^2} \\ W_{20}^* &= -\frac{nh^*\{C_y^2 R^2 / C_x^2\}}{[(n + \beta_2^*(y))\beta_2^*(x) - h^{*2}]} \end{aligned}\right\} \tag{15}$$

where $R = (\bar{Y}/\bar{X})$, $C_y = \sigma_y/\bar{Y}$ and $C_x = \sigma_x/\bar{X}$.

Hence the resulting minimum MSE's of d_1 and d_2 are respectively, given by

$$\text{Min. MSE } (d_1) = \frac{\sigma_y^4[\beta_2^*(y) C_x^2 - K^2]}{[\{n + \beta_2^*(y)\}\, C_x^2 - K^2]}, \tag{16}$$

and

$$\text{Min. MSE } (d_2) = \frac{\sigma_y^4[\beta_2^*(y)\beta_2^*(x) - h^{*2}]}{[\{n + \beta_2^*(y)\}\,\beta_2^*(x) - h^{*2}]}. \tag{17}$$

Substituting the optimum values of weights in (7) and (8) we get the resulting biases of d_1 and d_2 as

$$B_0(d_1) = -\frac{\text{Min. MSE}(d_1)}{\sigma_y^2}, \tag{18}$$

and

$$B_0(d_2) = -\frac{\text{Min. MSE}(d_2)}{\sigma_y^2}. \tag{19}$$

In the case of bivariate normal population the optimum weight and minimum MSE's of d_1 and d_2 reduce to:

$$W_{10} = n/(n+2); \; W_{20} = 0 \tag{20}$$

and

$$\left.\begin{aligned} W_{10}^* &= \frac{n}{[n + 2(1 - \rho^4)]} \\ W_{20}^* &= \frac{-nC_y^2 R^2 \rho^2}{[n + 2(1 - \rho^4)]C_x^2} \end{aligned}\right\} \tag{21}$$

$$\text{Min. MSE}(d_1) = 2\sigma_y^4/(n+2), \tag{22}$$

$$\text{Min. MSE}(d_2) = \frac{2\sigma_y^4(1-\rho^4)}{[n+2(1-\rho^4)]}. \tag{23}$$

THEORETICAL COMPARISONS

For comparison we give the MSE's/minimum MSE's (ignoring the finite population correction terms) along with optimum weights of estimators s_y^2 and t_i; $i = 1$ to 5 in the following scheme:

Estimator	MSE's/ minimum MSE's	Optimum weight
s_y^2	$\frac{\sigma_y^4}{n}\beta_2^*(y)$	–
t_1, t_2	$\frac{\sigma_y^4}{nC_x^2}\{\beta_2^*(y)C_x^2-K^2\}$	$W_0 = K/C_x^2$
t_3, t_4	$\frac{\sigma_y^4}{n}\cdot\frac{\{\beta_2^*(x)\beta_2^*(y)-h^{*2}\}}{\beta_2^*(x)}$	$\alpha_0 = h^*/\beta_2^*(x)$
t_5	$\frac{\sigma_y^4}{n}[\beta_2^*(y)+\beta_2^*(x)-2h^*]$	–
d_3	$\sigma_y^4\beta_2^*(y)/\{n+\beta_2^*(y)\}$	$W_0^* = n/\{n+\beta_2^*(y)\}$

In the case of bivariate normal population the minimum MSE's of the above estimators are given in the following scheme:

Estimator	MSE's/ minimum MSE's	Optimum weight
s_y^2	$2\sigma_y^4/n$	–
t_1, t_2	$2\sigma_y^4/n$	$W_0 = 0$
t_3, t_4	$\frac{2\sigma_y^4}{n}(1-\rho^4)$	$\alpha_0 = \rho^2$
t_5	$\frac{2\sigma_y^4}{n}(1-\rho^4)$	–
d_3	$2\sigma_y^4/(n+2)$	$W_0^* = n/(n+2)$

We have from minimum MSE t_i's ($i = 1, 2$) and (16) that

$$\text{Min. MSE}(t_i) - \text{Min. MSE}(d_1) = \frac{\sigma_y^4[\beta_2^*(y)C_x^2-K^2]^2}{nC_x^2[(n+\beta_2^*(y))C_x^2-K^2]} > 0. \tag{24}$$

It follows that from (24) the estimator d_1 is more efficient than t_i's, $i = 1, 2$ proposed by Das and Tripathi[2].

Next, from minimum MSE of t_i's ($i = 3, 4$) and (17) we have

$$\underset{i=3,4}{\text{Min. MSE}(t_i)} - \text{Min. MSE}(d_2) = \frac{\sigma_y^4}{n\beta_2^*(x)}\cdot\frac{[\beta_2^*(y)\beta_2^*(x)-h^{*2}]^2}{[\{n+\beta_2^*(y)\}\beta_2^*(x)-h^{*2}]} > 0. \tag{25}$$

Further, from minimum MSE of t_i's ($i = 3, 4$) and MSE of t_5 we have

$$\text{MSE}(t_5) - \underset{i=3,4}{\text{min. MSE}(t_i)} = \frac{\sigma_y^4}{n}\cdot\frac{[\beta_2^*(x)-h^*]^2}{\beta_2^*(x)} > 0. \tag{26}$$

Hence it follows from (25) and (26) that

$$\text{Min. MSE}(d_2) \le \underset{i=3,4}{\text{min. MSE}(t_i)} \le \text{MSE}(t_5)$$

which establishes that the proposed estimator d_2 is more efficient than that of t_i's ($i = 3, 4$) and t_5 proposed by Das and Tripathi[2], and Isaki[3], respectively. Further from (16), (17) and min. MSE of d_3 we have

$$\text{Min. MSE}(d_3) - \text{min. MSE}(d_1) = \frac{n\sigma_y^4K^2}{\{n+\beta_2^*(y)\}[\{n+\beta_2^*(y)\}C_x^2-K^2]} > 0. \tag{27}$$

$$\text{Min. MSE}(d_3) - \text{min. MSE}(d_2) = \frac{n\sigma_y^4h^{*2}}{\{n+\beta_2^*(y)\}[\{n+\beta_2^*(y)\}\beta_2^*(x)-h^{*2}]} > 0. \tag{28}$$

It follows from (27) and (28) that both the estimators d_1 and d_2 proposed here are more efficient than d_3 considered by Singh *et al*[4].

It is interesting to note that in case of bivariate normal population the estimators t_1, t_2 and s_y^2 are equally efficient. Also there is no contribution of $\bar{X}$ in case of bivariate normal population as we see that min. MSE(d_1) in (22) and min. MSE(t_5) are equal but their min. MSE's are larger than that of proposed estimator d_2. Thus it is advisable that one should pick up the proposed estimator d_2 in case of bivariate normal population as it has smaller minimum MSE than other estimators.

EMPIRICAL STUDY

In order to study the performance of various estimators of σ_y^2 we have chosen a natural population data considered by Das[1]. This population consists of 278 villages towns/wards under Gajole police station of Malda district of West Bengal, India (in fact only those villages of towns/wards have been considered which are shown as inhabited and common to both census 1961 and census 1971 list). The variates considered are: x, the number of agricultural labourers for 1961; y, the number of agricultural labourers for 1971.

Data under consideration were taken from census 1961 and census 1971 West Bengal, District Census Hand Book Malda.

Values of required population parameters for the population are given below:

$$\bar{Y} = 39.0680, \quad C_y = 1.4451;$$
$$\bar{X} = 25.1110, \quad C_x = 1.6198;$$
$$\rho = 0.7213, \quad \beta_2(x) = 38.8898;$$
$$K = 5.5636, \quad \beta_2(y) = 25.8969;$$
$$\sigma_y^2 = 3187.30, \quad \sigma_x^2 = 1654.40;$$
$$n = 30, \quad h = 26.8142.$$

The relative efficiencies of the estimators considered here with respect to usual unbiased estimator s_y^2 for the above data are given in table 1.

It follows from table 1 that the estimators using knowledge on $\bar{X}$ are inferior to those estimators using knowledge on σ_x^2. It is also observed that the proposed estimators are more efficient than those estimators considered by Das and Tripathi[2], Isaki[3] and Singh *et al*[4] and the usual unbiased estimator s_y^2. The performance of the proposed estimator d_2 is better than others.

Table 1 *Per cent relative efficiency of various estimators of*

Estimator	Per cent relative efficiency	Optimum weight
s_y^2	100.00	–
t_1, t_2	190.06	$W_0 = 2.1205$
t_3, t_4	340.60	$\alpha_0 = 0.6813$
t_5	223.13	–
d_1	273.05	$\begin{cases} W_{10} = 0.6961 \\ W_{20} = -0.2689 \end{cases}$
d_2	423.60	$\begin{cases} W_{10}^* = 0.8041 \\ W_{20}^* = -1\,0554 \end{cases}$
d_3	183.00	$W_0^* = 0.5465$

ACKNOWLEDGEMENT

The authors are grateful to the learned referee for his valuable suggestions on an earlier draft of the paper.

8 December 1987; Revised 28 March 1988

1. Das, A. K., *Contributions to the theory of sampling strategies based on auxiliary information*, Ph.D. thesis, Indian Statistical Institute, Calcutta, India, 1982.
2. Das, A. K. and Tripathi, T. P., *Sankhya*, 1978, Sr. C., 139.
3. Isaki, C. T., *J. Am. Stat. Assoc.*, 1983, **78**, 117.
4. Singh, J., Pandey, B. N. and Hirano, K., *Ann. Inst. Stat. Math.*, 1973, **25**, 51.