Jour. Ind. Soc. Ag. Statistics
48(2), 1996 : 151–158

A Note on Estimation of Mean with Known Population Proportion of an Auxiliary Character

V.D. Naik and P.C. Gupta
South Gujarat University, Surat

(Received : May, 1991)

SUMMARY

In this paper, when a character and an attribute are point biserialy correlated, feasibility of using the prior knowledge regarding proportion of units in the population has been investigated for estimating the population mean of the character of interest. Ratio, product and regression estimators have been defined for this case and the expressions, mostly of first order approximation, have been provided for their bias and MSE. Unbiased estimators corresponding to these expressions have also been given. Efficiency comparison among these estimators as well as with the sample proportion have been made. Using unbiased estimators, corresponding to various expressions of bias of these estimators, corresponding almost unbiased or unbiased estimators have also been defined using Hartley-Ross [1] technique. Deviating from the usual way of defining the regression estimator, it is defined in somewhat different manner.

Key words : Attribute, Biserial correlation, Ratio estimator, Product estimator, Regression estimator.

1. *Introduction*

Prior knowledge about population mean alongwith coefficient of variation of the population of an auxiliary variable is known to be very useful particularly when the ratio, product and regression estimators are used for estimation of population mean of a variable of interest. However, the fact that the known population proportion of an attribute also provides similar type of information has not drawn as much attention. In fact, such prior knowledge can also be very useful when a relation between the presence (or absence) of an attribute and the value of a variable, known as point biserial correlation, is observed. Relationship between height and sex in human being, milk production or fat content in milk and breed in cows etc. are some of the examples where such biserial correlation can be traced. Therefore, when such relationship is evident and the population proportion of the attribute involved is known, such information can also be exploited for estimating the mean by using ratio, product and regression estimator.

Let *Y* be the character of interest and *X* be the auxiliary attribute. Let there be complete dichotomy in the population with respect to the possession or otherwise of *X*. Let *Y* and *X* be the variables representing *Y* and *X* respectively such that corresponding to the ith unit of the population, $i = 1, 2, \ldots, N$ and N being the size of the population, they take values Y_i and X_i respectively. Further, let $X_i = 1$ if the ith unit of the population possesses *X* and $X_i = 0$, otherwise. Let a sample of size n be drawn from this population using a simple random sampling without replacement (SRSWOR) sampling procedure. Let $\bar{Z}$ and $\bar{z}$ be the population mean and sample mean respectively of variable Z; Z = Y and X. Let A and 'a' be the proportions of units in the population and sample respectively, which possess *X*. Let the value of A be assumed to be known. It is easy to check that $\bar{X} = A$ and $\bar{x} = a$. It may also be noted that when A is known Var (X) = A(1-A) is also known.

2. *The Ratio Estimator*

Analogous to the usual ratio estimator of $\bar{Y}$, defined for the case in which $\bar{X}$ is known, in the present case the ratio estimator of $\bar{Y}$ can also be defined as

$$e(\bar{Y})_r = \bar{y} - \frac{A}{a} \qquad \ldots(2.1)$$

Further, following Murthy (1967), as the sample mean and the sample proportion are unbiased estimators of population mean and population proportion respectively, the first order approximate expressions of bias and mean square error (MSE) of this estimator can be given as follows :

$$B(e(\bar{Y})_r)_I = \frac{\bar{Y}\,\mathrm{Var}(a)}{A^2} - \frac{\mathrm{Cov}(\bar{y}, a)}{A} \qquad \ldots(2.2)$$

$$= \theta\left\{\frac{\bar{Y}\,S_1^2}{A^2} - \frac{S_{1y}}{A}\right\} \qquad \ldots(2.3)$$

$$= \theta\left\{\frac{\bar{Y}\,S_1^2}{A^2} - \frac{N(\bar{Y}\,\bar{U} - \bar{Y}^2 A)}{(N-1)A}\right\} \qquad \ldots(2.4)$$

$$\mathrm{MSE}(e(\bar{Y})_r)_I = \left\{\mathrm{Var}(\bar{y}) + \frac{\bar{Y}^2\,\mathrm{Var}(\bar{y})}{A^2} - \frac{2\,\mathrm{Cov}(\bar{y}, a)}{A}\right\} \ldots(2.5)$$

$$= \theta \left\{ S_y^2 + \bar{Y}^2 \frac{S_1^2}{A^2} - \frac{2\bar{Y}\, S_{1y}}{A} \right\} \qquad ...(2.6)$$

$$= \theta \left\{ S_y^2 + \bar{Y}^2 \frac{S_1^2}{A^2} - \frac{2N\,(\bar{U} - \bar{Y}A)}{(N-1)\,A} \right\} \qquad ...(2.7)$$

where $\theta = (N-n)/(nN)$, $S_1^2 = N\,\mathrm{Var}\,(X)/(N-1) = NA\,(1-A)/(N-1)$

$S_{1y} = N\,\mathrm{Cov}\,(Y, X)/(N-1) = N\,(\bar{U} - \bar{Y}A)/(N-1)$ and

$\bar{U}$ is the population mean of the variable U such that with respect to the ith unit of the population its value is defined as follows.

$$U_i = \begin{cases} Y_i, \text{ if the ith unit possesses } X \\ 0, \text{ otherwise} \end{cases}$$

3. *Comparison of* e $(\bar{Y})_r$ with $\bar{y}$

From the preceding section, it follows that

$$\mathrm{Var}\,(\bar{y}) > \mathrm{MSE}\,(e\,(\bar{y})_r)_I$$

if $$\frac{\mathrm{Cov}\,(\bar{y}, a)}{\mathrm{Var}\,(a)} > \frac{\bar{Y}}{2A} \qquad ...(3.1)$$

i.e. $$\rho_{\bar{y}a} > \frac{CV\,(a)}{2CV\,(\bar{y})} \qquad ...(3.2)$$

where CV(a) and CV($\bar{y}$) are the coefficients of variation of a and $\bar{y}$ respectively and $\rho_{\bar{y}a}$ is the coefficient of correlation between $\bar{y}$ and a in their bivariate sampling distribution and is nothing but the coefficient of point biserial correlation between Y and X because $\rho_{\bar{y}a} = \rho_{YX}$.

Further, it can be seen that the inequality (3.1) is same as the inequality given below.

$$\frac{\mathrm{Cov}\,(Y, X)}{\mathrm{Var}\,(X)} > \frac{\bar{Y}}{2A} \qquad ...(3.3)$$

i.e. $$2\bar{U} - \bar{Y}\,(1 + A) > 0 \qquad ...(3.4)$$

Therefore it can be said that, under SRSWOR, upto first order of approximation, e $(\bar{Y})_r$ is more efficient than $\bar{y}$ if the point biserial correlation coefficient between the variable under study and the auxiliary attribute is positive and the inequality (3.4) holds.

For testing the condition (3.4) on the basis of sample, the corresponding unbiased estimator can be given by (3.5).

$$2\bar{u} - \bar{y}(1 + A) > 0 \qquad \text{...(3.5)}$$

where $\bar{u}$ is the sample mean of variable U.

4. *Unbiased estimators corresponding to the expressions of bias and MSE of* e $(\overline{Y})_r$

If s_y^2 and s_1^2 are the sample mean squares defined for Y and X respectively and if $s_{1y} = n(\bar{u} - \bar{y}a)/(n-1)$ then it can be seen that corresponding to the expressions of first order approximate bias of e $(\overline{Y})_r$ as given in (2.3) and (2.4), two different unbiased estimators can be obtained as follows :

$$(b(e(\overline{Y})_r)_l)_1 = \frac{\theta}{A^2}\{s_1^2 - As_{1y}\} \qquad \text{...(4.1)}$$

$$(b(e(\overline{Y})_r)_l)_2 = \frac{\theta}{A^2}\left\{s_1^2 - \frac{NA(\bar{u} - \bar{y}A)}{(N-1)}\right\} \qquad \text{...(4.2)}$$

Let

$$e(\overline{Y}^2) = \bar{y} - \frac{(N-n)s_y^2}{nN} \qquad \text{...(4.3)}$$

and

$$e(\overline{Y}\,\overline{U}) = \bar{y}\,\bar{u} - \frac{(N-n)s_{yu}}{nN} \qquad \text{...(4.4)}$$

where

$$s_{yu} = \frac{\sum\limits_{i=1}^{n} u_i^2 - n\bar{y}\bar{u}}{(n-1)} \qquad \text{...(4.5)}$$

are unbiased estimators of $\overline{Y}$ and $\overline{Y}\,\overline{U}$, it can easily be seen that the estimator given in (4.6) is an unbiased estimator of the first order approximate MSE of e $(\overline{Y})_r$ as given by (2.7).

$$\text{MSE}(e(\overline{Y})_r)_l = \theta\left\{s_y^2 + e(\overline{Y}^2)\frac{S_1^2}{A^2} - \frac{2(e(\overline{Y}\,\overline{U}) - e(\overline{Y}^2)A)}{A}\right\} \qquad \text{...(4.6)}$$

Moreover, using the unbiased estimators of expression of bias of e $(Y)_r$ almost unbiased ratio type estimators, which are obtained using Hartley-Ross technique, can be obtained as follows :

$$(e(\overline{Y})_{aur})_i = e(\overline{Y})_r - (b(e(\overline{Y})_r)_l)_i \qquad \text{...(4.7)}$$

It can easily be shown that the expression of first order approximate variance of the estimators defined in (4.7) are same as the expression of first order approximate MSE of $e\,(\overline{Y})_r$.

5. *Estimation with product estimator*

Following the discussions in the sections 2, 3 and 4, the definition of the product estimator and its corresponding results, under SRSWOR, are straight forward and are summerised below :

(a) The Estimator :

$$e\,(\overline{Y})_p = \bar{y}\,\frac{a}{A} \qquad ...(5.1)$$

(b) The Bias :

$$B\,(e\,(\overline{Y})_p) = \theta\,\frac{S_{1y}}{A} \qquad ...(5.2)$$

$$= \theta\,\frac{n\,(\overline{U} - \overline{Y}\,A)}{(N-1)\,A} \qquad ...(5.3)$$

(c) The first order approximate MSE :

$$MSE\,(e\,(\overline{Y})_p)_1 = \theta\left\{S_y^2 + \overline{Y}^2\,\frac{S_1^2}{A^2} + \frac{2\overline{Y}\,S_{1y}}{A}\right\} \qquad ...(5.4)$$

$$= \theta\left\{S_y^2 + \overline{Y^2}\,\frac{S_1^2}{A^2} + \frac{2N\,(\overline{U} - \overline{Y}\,A)}{(N-1)\,A}\right\} \qquad ...(5.5)$$

(d) Conditions under it is more efficient than $\bar{y}$:

$$\frac{Cov\,(\bar{y}, a)}{Var\,(a)} > -\frac{\overline{Y}}{2A} \qquad ...(5.6)$$

i.e.

$$\rho_{\bar{y}a} < -\frac{CV\,(a)}{2CV\,(Y)} \qquad ...(5.7)$$

Or alternatively $\quad 2\overline{U} + \overline{Y}\,(1-3A) \leq 0 \qquad ...(5.8)$

such that $\quad E\,[2\bar{u} + \bar{y}\,(1-3A)\,] = 2\overline{U} + \overline{Y}\,(1-3A)$

Thus the product estimator is more efficient than $\bar{y}$ if the biserial correlation between *Y* and *X* is negative and the condition (5.8) satisfied.

Note : It may be noted that a product estimator corresponding to a ratio estimator $e\,(\overline{Y})_r$ can also be constructed as $e\,(\overline{Y})_p = \bar{y}\,a'/A'$ with $a' = 1 - a$ and $A' = 1 - A$.

(e) Unbiased estimators of the first order approximate MSE, bias :

$$\mathrm{MSE}\,(e\,(\overline{Y})_p)_I = \theta\left\{s_y^2 + e\,(\overline{Y^2})\,\frac{S_1^2}{A^2} + \frac{2\,(e\,(\overline{Y}\,\overline{U}) - e\,(\overline{Y^2})\,A)}{A}\right\} \qquad ...(5.9)$$

$$(b\,(e(\overline{Y})_p))_1 = \theta\,\frac{s_{1y}}{A} \qquad ...(5.10)$$

$$(b\,(e\,(\overline{Y})_p))_2 = \theta\,\frac{N\,(\overline{u} - \overline{y}\,A)}{(N-1)\,A} \qquad ...(5.11)$$

(f) Unbiased product estimator having same efficiency as $e\,(\overline{Y})_p$:

$$(e\,(\overline{Y})_{up})_i = e(\overline{Y})_p - (b(e(\overline{Y})_p))_i \qquad ...(5.12)$$

6. *Estimation with regression estimator*

Like the cases of estimation with the ratio and the product estimators, the case of estimation with the regression estimator can also be considered. A regression estimator and corresponding properties of it, under SRSWOR, are presented below :

(g) The Estimator : $e\,(\overline{Y})_{lr} = \overline{y} + e\,(\beta)\,(A - a)$ ...(6.1)

where $e\,(\beta) = (\overline{u} - \overline{y}\,A)/(A\,(1 - A))$

(h) The Bias :

$$B\,(e\,(\overline{Y})_{lr}) = \frac{\mathrm{Cov}\,(\overline{y}, a)}{1 - A} - \frac{\mathrm{Cov}\,(\overline{u}, a)}{A\,(1 - A)} \qquad ...(6.2)$$

$$= \theta\left\{\frac{S_{1y}}{1 - A} - \frac{S_{1u}}{A}\right\} \qquad ...(6.3)$$

$$= \theta'\left\{\frac{\overline{U} - \overline{Y}\,A}{1 - A} - \frac{\overline{U}}{A}\right\} \qquad ...(6.4)$$

where $\theta' = (N - n)/(n\,(N - 1))$ and $S_{1u} = N\overline{U}\,(1 - A)/(n - 1)$

(i) The first order approximate MSE :

$$\mathrm{MSE}\,(e\,(\overline{Y})_{lr})_I = \theta\left\{S_y^2 - \frac{S_{1y}^2}{S_1^2}\right\} \qquad ...(6.5)$$

$$\mathrm{MSE}\,(e\,(\overline{Y})_{lr})_I = \theta\left\{S_y^2 - \frac{N\,(\overline{U}^2 + \overline{Y}^2\,A^2 - 2\overline{Y}^2\,A)}{(N-1)\,A\,(1 - A)}\right\} \qquad ...(6.6)$$

(j) Unbiased estimators of bias and first order approximate MSE :

$$b\,(e\,(\overline{Y})_{1r})_1 = \theta\left\{\frac{S_{1y}}{1-A} - \frac{S_{1u}}{A}\right\} \qquad ...(6.7)$$

$$b\,(e\,(\overline{Y})_{1r})_2 = \theta'\left\{\frac{\bar{u}-\bar{y}A}{1-A} - \frac{\bar{u}}{A}\right\} \qquad ...(6.8)$$

$$\text{MSE}\,(e\,(\overline{Y})_{1r})_1 = \theta\left\{s_y^2\,\frac{N\,(e\,(\overline{U^2}) + e\,(\overline{Y^2})\,A^2 - 2e\,(\overline{Y})\,A)}{(N-1)\,A\,(1-A)}\right\} ...(6.9)$$

(k) Unbiased estimators with same efficiency as $e\,(\overline{Y})_{1r}$:

$$(e\,(\overline{Y})_{u1r})_i = e\,(\overline{Y})_{1r} - b\,(e\,(\overline{Y})_{1r})_i,\ i = 1, 2 \qquad ...(6.10)$$

Note : So far as first order approximation is concerned, it can be easily verified that the regression estimators considered here are always more efficient than $y, e\,(\overline{Y})_r$, and $e\,(\overline{Y})_p$.

7. *Empirical Study*

For empirical study consider the problem of estimation of (I) Mean agricultural area and (II) Mean number of live stock in the population of villages. The population is consisted of 364 villages divided into two groups of small and large villages. It is known that there are 161 large villages and 203 small villages so that A = 0.4423 and A′ = 0.5577. For sake of computation of MSE/variance of the estimators under comparison, important parameters of the population for the two cases, summarized from Sukhatme *et al.* (1984), page no. 201, are as follows:

$$\text{Case - (I)}: \ \overline{Y} = 520.4,\ S_y^2 = 247571.14,\ \overline{U} = 384.5212$$

$$\text{Case - (II)}: \ \overline{Y} = 151.9,\ S_y^2 = 25091.934,\ \overline{U} = 110.6393$$

Table 1.

Estimator	[MSE/Variance]/θ	
	case - I	case - II
$\bar{y}$	247571.14	25091.934
$e(\overline{Y})_r = \bar{y}\,A/a$	225780.51	24336.633
$e(\overline{Y})_p = \bar{y}\,a'/A'$	174097.64	19705.347
$e(\overline{Y})_{1r}$	150725.34	17415.928

From the above table it can be observed that, up to the first order approximation, all the estimators using prior knowledge about the number (proportion) of large villages in the population are more efficient than $\bar{y}$ and the empirical results are in conformity with the theoretical results.

ACKNOWLEDGEMENT

The authors thank the referee for his valuable comments and suggestions which helped a lot in improving the presentation of the paper.

REFERENCES

[1] Hartley, H.O. and Ross, A., 1954. Unbiased ratio estimators. *Nature*, **174**, 270-271.

[2] Murthy, M.N., 1976. Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

[3] Sukhatme, P.V. ; Sukhatme, B.V.; Sukhatme, S. and Asok, C. 1984. Sampling Theory of Surveys with Applications. Indian Society of Agricultural Statistics, New Delhi, India.