

**A FAMILY OF ESTIMATORS OF POPULATION MEAN USING
INFORMATION ON AUXILIARY ATTRIBUTE**

H.S. Jhajj and M.K. Sharma

Department of Statistics
Punjabi University
Patiala-147 002, India
and

Lovleen Kumar Grover
Department of Mathematics
Guru Nanak Dev University
Amritsar-143 005, India

ABSTRACT

In practice, the information regarding the population proportion possessing certain attribute is easily available. So for estimating the population mean \bar{Y} of study variable y , a family of estimators of \bar{Y} has been proposed by using the known information of population proportion possessing an attribute (highly correlated with y). The expressions for the mean square error of the estimators of the proposed family and its minimum value have been obtained. It has been shown that the optimum estimator of the proposed family of estimators of \bar{Y} is always better than the mean per unit estimator. The results have also been extended for the case of the double sampling design. The results obtained have been illustrated numerically by taking some empirical populations considered in the literature.

1. INTRODUCTION

So far in the literature of survey sampling, the efficiencies of the estimators of unknown population mean of the study variable y have been increased by the use of known information on an auxiliary variable x which is highly correlated with variable y . But in several practical situations, instead of existence of auxiliary variables there exist some auxiliary attributes which are highly correlated with study variable y , such as

- (a) sex and height of the persons,
- (b) amount of milk produced and a particular breed of the cow,
- (c) amount of yield of wheat crop and a particular variety of wheat etc.

In such situations, taking the advantage of point biserial correlation between the study variable and the auxiliary attribute, the estimators of parameters of interest can be constructed by using prior knowledge of the parameters of auxiliary attribute. So by taking into consideration the point biserial correlation between a variable and an attribute, Naik and Gupta (1996) defined ratio, product and regression estimators of population

mean when the prior information of population proportion of units, possessing the same attribute is available.

In the present paper, a family of estimators of population mean \bar{Y} has been proposed by using the known information of population proportion P of units which possesses a particular attribute, highly correlated with study variable y . The expressions of the mean square error of the estimator of the proposed family and its minimum value have been obtained. The gain in efficiency of the optimum estimator of the proposed family over the mean per unit estimator has been obtained. It has also been shown that if the value of mean square error of the optimum estimator of \bar{Y} (belonging to the proposed family) is unknown then it can be estimated on the basis of same set of sampling observations simply by replacing unknown parameters involved with their corresponding conventional estimators. The results have been obtained under the double sampling design, in the case when population proportion P is unknown. The numerical illustrations have also been done by taking some empirical populations considered in the literature.

2. NOTATIONS AND EXPECTATIONS

Suppose there is a complete dichotomy in the population with respect to the presence or absence of an attribute, say \varkappa , and it is assumed that attribute \varkappa takes only the two values 0 and 1 according as

$$\begin{aligned} \varkappa_i &= 1, \text{ if } i^{\text{th}} \text{ unit of the population possesses attribute } \varkappa \\ &= 0, \text{ otherwise.} \end{aligned}$$

A simple random sample without replacement (SRSWOR) of size n is drawn from the population of size N . Let Y_i and \varkappa_i denote the observations on variable y and attribute \varkappa respectively at the i^{th} unit of the population ($i = 1, 2, \dots, N$). The corresponding small letters denote the values in the sample. We take the following:

$$\sum_{i=1}^N \varkappa_i = A \rightarrow \text{Total number of units in the population possessing attributes } \varkappa,$$

$$\sum_{i=1}^n \varkappa_i = a \rightarrow \text{Total number of sampling units possessing attributes } \varkappa,$$

$$\frac{A}{N} = P \rightarrow \text{Proportion of units in the population possessing attribute } \varkappa,$$

$$\frac{a}{n} = p \rightarrow \text{Proportion of units in the sample possessing attribute } \varkappa,$$

and also

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i, S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$S_{\varkappa}^2 = \frac{N}{N-1} P(1-P), S_{y\varkappa} = \frac{1}{N-1} \left[\sum_{i=1}^N Y_i \varkappa_i - NP\bar{Y} \right], \rho_{pb} = \frac{S_{y\varkappa}}{S_y S_{\varkappa}}$$

where ρ_{pb} denotes the point biserial correlation coefficient between study variable y and auxiliary attribute x (see Kendall and Stuart (1967) page-311).

Defining

$$\varepsilon = \frac{\bar{y}}{\bar{Y}} - 1, \quad \varphi = \frac{P}{P} - 1$$

So, we have the following expectations:

$$E(\varepsilon) = E(\varphi) = 0, \quad E(\varepsilon\varphi) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_{y\ast}}{\bar{Y}P}$$

$$E(\varepsilon^2) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_y^2}{\bar{Y}^2}, \quad E(\varphi^2) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_{\ast}^2}{P^2}$$

3. PROPOSED FAMILY OF ESTIMATORS OF \bar{Y} AND ITS MEAN SQUARE ERROR

Suppose that the information about proportion P of population units possessing attribute x , highly correlated with variable y , is known in advance. Using such known prior information, we propose a general family of estimators of \bar{Y} as

$$\tilde{t}_g = g_a(\bar{y}, \upsilon), \tag{3.1}$$

where $\upsilon = \frac{P}{P}$ and $g_a(\bar{y}, \upsilon)$ is a parametric function of \bar{y} and υ such that

$$g_a(\bar{Y}, 1) = \bar{Y}, \quad \forall \bar{Y} \tag{3.2}$$

and satisfying following regularity conditions:

- (i) Whatever be the sample chosen, the point (\bar{y}, υ) assume values in a bounded closed convex subset R_2 of the two-dimensional real space containing the point $(\bar{Y}, 1)$.
- (ii) The function $g_a(\bar{y}, \upsilon)$ is a continuous and bounded in R_2 .
- (iii) The first and second order partial derivatives of $g_a(\bar{y}, \upsilon)$ exist and are continuous as well as bounded in R_2 .

On expanding the function \tilde{t}_g in a second order Taylor's series about the point $(\bar{Y}, 1)$, we have

$$\begin{aligned} \tilde{t}_g &= \bar{Y} + (\bar{y} - \bar{Y})g_{1a} + (\nu - 1)g_{2a} \\ &+ \frac{1}{2} \left\{ (\bar{y} - \bar{Y})^2 g_{11a} + (\nu - 1)^2 g_{22a} + 2(\bar{y} - \bar{Y})(\nu - 1)g_{12a} \right\} \dots \end{aligned} \quad (3.3)$$

where g_{1a} and g_{2a} denote the respective first order partial derivatives of $g_a(\bar{y}, \nu)$ w.r.t. \bar{y} and ν at the point $(\bar{Y}, 1)$ and $g_{ija}(i, j = 1, 2)$ denote the corresponding second order partial derivatives of $g_a(\bar{y}, \nu)$ at the point (\bar{Y}^*, ν^*) ; $\bar{Y}^* = \bar{Y} + \theta(\bar{y} - \bar{Y})$,

$$\nu^* = 1 + \theta(\nu - 1) \text{ with } 0 < \theta < 1 \text{ and noting that } g_{1a} = 1.$$

On taking the expectation in (3.3) and noting that it exists, we have

$$E(\tilde{t}_g) = \bar{Y} + O(n^{-1}) \quad (3.4)$$

So the bias of an estimator belonging to the proposed class \tilde{t}_g is of order n^{-1} .

Theorem: Up to the terms of order n^{-1} , the mean square error of the estimators of the proposed family \tilde{t}_g is minimized for

$$g_{2a(opt)} = -P \frac{S_{y\bar{x}}}{S_{\bar{x}}^2} \quad (3.5)$$

and its minimum value is given by

$$Min.MSE(\tilde{t}_g) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 (1 - \rho_{pb}^2) \quad (3.6)$$

Remark 3.1: The expression (3.6) shows that optimum estimator of the proposed family \tilde{t}_g is always better than mean per unit estimator of \bar{Y} i.e. \bar{y} because

$$MSE(\bar{y}) - MinMSE(\tilde{t}_g) \geq 0 \quad (3.7)$$

Remark 3.2: The expression (3.6) obtained is exactly similar to the expression of mean square error of the ordinary linear regression estimator of \bar{Y} . But the only difference between these expressions is that in (3.6) there is point biserial correlation coefficient between the study variable y and the auxiliary attribute \bar{x} in place of ordinary correlation coefficient between the two variables y and x .

Remark 3.3: From (3.5), we see that the optimum values of the parameters involved in \tilde{t}_g depend upon the values of population parameters, which are assumed to be known for the efficient use of the proposed family \tilde{t}_g . But the values of these parameters are unknown, which can be obtained either from the pilot sample survey or from the past

experience. Following Srivastava and Jhajj (1983) 's approach, the estimators of the proposed family \tilde{t}_g will have the same minimum mean square error, up to the terms of order n^{-1} , if we replace the unknown values of the parameters involved in the optimum value of the parameter of \tilde{t}_g with their consistent estimators.

Remark 3.4: The proposed family \tilde{t}_g of estimators of \bar{Y} is very large. Any parametric function $g_a(\bar{y}, \nu)$ satisfying $g_a(\bar{Y}, 1) = \bar{Y}$, $\forall \bar{Y}$ and the same regularity conditions as given above, can generate an estimator of the proposed family \tilde{t}_g . For example, some simple estimators of \bar{Y} belonging to the proposed family \tilde{t}_g are obtained by taking the following simple functions:

$$g_a(\bar{y}, \nu) = \bar{y}\nu^\alpha, \quad g_a(\bar{y}, \nu) = \bar{y} + \alpha(\nu - 1), \quad g_a(\bar{y}, \nu) = \bar{y}e^{\alpha(\nu-1)} \text{ etc.}$$

The optimum value of α in all the above estimators is so obtained that satisfying (3.5) and the resulting estimators of \bar{Y} have the same minimum mean square error as given in (3.6). It is noted that the estimators of \bar{Y} , which are defined by Naik and Gupta (1996), are also particular members of the proposed family \tilde{t}_g .

4. AN ESTIMATE OF MINIMUM MEAN SQUARE ERROR OF \tilde{t}_g

In general, the estimate of $Min.MSE(\tilde{t}_g)$ can be obtained simply by replacing unknown population parameters with their corresponding conventional estimators in the expression (3.6). So we get the following estimate of $Min.MSE(\tilde{t}_g)$:

$$Est\{Min.MSE(\tilde{t}_g)\} = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 (1 - r_{pb}^2) \quad (4.1)$$

where s_y^2 denotes the sample variance of variable y which is an unbiased estimator of S_y^2 and r_{pb} denotes the sample point biserial correlation coefficient between the study variable y and the auxiliary attribute x .

From (4.1), we see that the efficiency of the optimum estimators of the proposed family \tilde{t}_g depend upon the value of the sample point biserial correlation coefficient r_{pb} . To verify that the value of the point biserial correlation coefficient is significant or insignificant, we test the null hypothesis $H_0 : \rho_{pb} = 0$ against the alternative hypothesis $H_a : \rho_{pb} \neq 0$.

The appropriate test statistic is

$$t = \frac{\sqrt{n-2} r_{pb}}{\sqrt{1-r_{pb}^2}} \quad (4.2)$$

where

$$r_{pb} = \frac{(\bar{y}_1 - \bar{y}_2) \sqrt{\frac{n}{n-1} p(1-p)}}{s_y} \quad (4.3)$$

and \bar{y}_1 and \bar{y}_2 denote the sample means of variable y corresponding to the sampling units possessing attribute \varkappa and not possessing attribute \varkappa respectively and p is proportion of sample possessing attribute \varkappa .

Under H_0 , the test statistic t as defined in (4.2) follows central t -distribution with $(n-2)$ degrees of freedom (as given in Kendall and Stuart (1967), page-312), because Tate (1954) proved that r_{pb} is asymptotically normally distributed with the mean ρ_{pb} and the variance

$$\text{var}(r_{pb}) \cong \frac{(1-\rho_{pb}^2)^2}{n} \left\{ 1 - \frac{3}{2} \rho_{pb}^2 + \frac{\rho_{pb}^2}{4p(1-p)} \right\}.$$

If the value of the point biserial correlation coefficient ρ_{pb} is found to be significant on the basis of above test then the optimum estimators of the proposed family \tilde{t}_g can be recommended to estimate \bar{Y} , which will be as much efficient as the linear regression estimator (defined by Naik and Gupta (1996)).

5. PROPOSED FAMILY OF ESTIMATORS OF \bar{Y} IN DOUBLE SAMPLING

When the value of P is unknown, we generally apply the double sampling design to obtain an efficient estimator of population mean \bar{Y} . Let p' denote the proportion of units possessing attribute \varkappa in the first phase sample of size n' ; p denote the proportion of units possessing attribute \varkappa in the second phase sample of size $n < n'$ and \bar{y} denote the mean of the study variable y in the second phase sample, under the double sampling design. In practice, the information of p' can be obtained with a little additional cost. So in such situations, we propose a family of estimators of \bar{Y} defined as

$$\tilde{t}_{gd} = g_{ad}(\bar{y}, v') \quad (5.1)$$

where

$$v' = \frac{P}{p'}$$

and $g_{ad}(\bar{y}, v')$ is a parametric function such that $g_{ad}(\bar{Y}, 1) = \bar{Y}$, $\forall \bar{Y}$ and satisfying certain regularity conditions similar to the regularity conditions as given in section 3.

When both the samples drawn in the double sampling are SRSWOR then following in the same way as in the section 3, the minimum mean square error of the proposed family \tilde{t}_{gd} , up to the terms of order n^{-1} , is given by

$$Min.MSE(\tilde{t}_{gd}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right) S_y^2 \rho_{pb}^2 \tag{5.2}$$

It is noted again that the expression obtained in (5.2) is an analogue to the expression of the mean square error of the ordinary linear regression estimator of \bar{Y} under the double sampling design.

6. NUMERICAL ILLUSTRATION

To have a rough idea about the gain in efficiency of the optimum estimator of \bar{Y} , say $\tilde{t}_{g(opt)}$, of the proposed family \tilde{t}_g over the mean per unit estimator \bar{y} , we take the two empirical populations considered in the literature. The source of population, variable y, attribute \varkappa , population size N, proportion P of the population units possessing attribute \varkappa and the population point biserial correlation coefficient ρ_{pb} are given in the table 6.1. The relative efficiency of the optimum estimator of the proposed family \tilde{t}_g w.r.t. the mean per unit estimator \bar{y} is shown in the table 6.2.

Table 6.1:
Description of populations

Sr. No.	Source of population	Variable y	Attribute \varkappa	N	P	ρ_{pb}	
1	Sukhatme and Sukhatme (1970) p-256 circles 1-89	i)	Number of villages in the circles	A circle consisting more than five villages	89	0.1235955	0.7662249
		ii)	Area (in acres) under wheat in the circles	A circle consisting more than five villages	89	0.1235955	0.6235605
2	Mukhopadhyay (2000) p-44 households 1-25	Household size	A household that availed an agricultural loan from a bank	25	0.400	-0.3873065	

Table 6.2:
Efficiencies of optimum estimators of proposed family \tilde{t}_g
w.r.t. mean per unit estimator \bar{y}

Population number		Efficiencies of estimators	
		\bar{y}	$\tilde{t}_{g(opt)}$
1	(i)	100	242.18974
	(ii)	100	163.61998
2		100	117.64793

From the Table 6.2, we can see that there is a significant gain in efficiency of the optimum estimator $\tilde{t}_{g(opt)}$ of the proposed family \tilde{t}_g over the mean per unit estimator \bar{y} . Here the increase in efficiency of the optimum estimator of the proposed family is due to readily available information regarding population proportion possessing an attribute, which is highly correlated to the study variable. So by having a little knowledge in the form of population proportion possessing an attribute, we can improve the efficiency to a great extent which is our ultimate aim.

REFERENCES

1. Kendall, M.G. and Stuart, A. (1967). *The advanced theory of statistics*, Vol. 2, Second Edition, Charles Griffin and Company limited, London.
2. Mukhopadhyay, P. (2000). *Theory and methods of survey sampling*. Prentice Hall of India, New Delhi, India.
3. Naik, V.D. and Gupta, P.C. (1996). A note on estimation of mean with known population proportion of an auxiliary character. *Jour. Ind. Soc. Agr. Stat.*, 48 (2), 151-158.
4. Srivastava, S.K. and Jhaji, H.S. (1983). A class of estimators of the population mean using multi-auxiliary information. *Cal. Stat. Assoc. Bull.*, 32, 47-56.
5. Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling theory of surveys with applications*. Asia Publishing House, New Delhi, India.
6. Tate, R.F. (1954). Correlation between a discrete and a continuous variable-Point biserial correlation. *Ann. Math. Statist.*, 25, 603.