On Estimation of Population Mean Using Information on Auxiliary Attribute

Rajesh Singh
Department of Statistics
BHU, Varanasi (U.P.), India
rsinghstat@yahoo.com

Abstract

We consider the problem of estimating the finite population mean when some information on auxiliary attribute is available. We obtain the mean square error (MSE) equation for the proposed estimators. It has been shown that the proposed estimator is better than Naik and Gupta (1996), Singh et al. (2008), Abd-Elfattah (2010) estimators. The results have been illustrated numerically by taking some empirical population considered in the literature.

Keywords: SRSWOR, Attribute, Point bi-serial correlation, MSE, Efficiency.

1. Introduction

In survey sampling the use of auxiliary information can increase the precision of an estimator when study variable y is highly correlated with the auxiliary variable x. But in several practical situations, instead of existence of auxiliary variables there exists some auxiliary attributes, which are highly correlated with study variable y, such as (i) use of drugs and gender (ii) amount of milk produced and a particular breed of cow.

Consider a sample of size n drawn by simple random sampling without replacement (SRSWOR) from a population of size N. Let y_i and ϕ_i denote the observations on variable y and ϕ respectively for the i^{th} unit (i=1,2,3....N.). It is assumed that attribute ϕ takes only the two values 0 and 1 according as

ϕ = 1, if i^{th} unit of the population possesses attribute ϕ
= 0, if otherwise.

Let $A=\sum_{i=1}^{N}\phi_i$ and $a=\sum_{i=1}^{n}\phi_i$ denote the total number of units in the population and sample possessing attribute ϕ respectively, $p=\frac{A}{N}$ and $p=\frac{a}{n}$ denote the proportion of units in the population and sample, respectively, possessing attribute ϕ.

Define,

$$e_y=\frac{(\bar{y}-\bar{Y})}{\bar{Y}} \quad e_\phi=\frac{(p-P)}{P},$$
$$E(e_i)=0,(i=y,\phi)$$
$$E(e_y^2)=fC_y^2,\quad E(e_\phi^2)=fC_p^2,\quad E(e_y e_\phi)=f\rho_{pb}C_yC_{p.}$$

Where

$$f=\left(\frac{1}{n}-\frac{1}{N}\right)\quad C_y^2=\frac{S_y^2}{\overline{Y}^2},\qquad C_p^2=\frac{S_p^2}{P^2},$$

and $\rho_{pb}=\frac{S_{y\phi}}{S_y S_\phi}$ is the point biserial correlation coefficient.

Here,

$$S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2,\qquad S_\phi^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(\phi_i-P\right)^2\qquad \text{and}$$

$$S_{y\phi}=\frac{1}{N-1}\left(\sum_{i=1}^{N}y_i\phi_i-NP\overline{Y}\right)$$

In order to have an estimate of the population mean $\overline{Y}$ of the study variable y, assuming the knowledge of the population proportion p, Naik and Gupta (1996) defined following ratio and product estimators

$$t_{NGR}=\overline{y}\left(\frac{P}{p}\right)\tag{1.1}$$

$$t_{NGP}=\overline{y}\left(\frac{p}{P}\right)\tag{1.2}$$

The mean square error (MSE) of t_{NGR} and t_{NGP} up to the first order of approximation, respectively, are

$$MSE\left(t_{NGR}\right)=f\overline{Y}^2\left[C_y^2+C_p^2-2\rho_{pb}C_yC_p\right]\tag{1.3}$$

$$MSE\left(t_{NGP}\right)=f\overline{Y}^2\left[C_y^2+C_p^2+2\rho_{pb}C_yC_p\right]\tag{1.4}$$

2. Other estimators

Singh et al. (2008) suggested the following ratio estimator

$$t_S=\frac{\overline{y}+b_\phi(P-p)}{\left(m_1p+m_2\right)}\left(m_1P+m_2\right)\tag{2.2}$$

where m_1 ($\neq 0$) and m_2 are either real numbers or the functions of the parameters of the attribute such as $C_{p,}\ \beta_2(\phi)$ and ρ_{pb}.

In Singh et al. (2008), MSE equation of these ratio- type estimators were given by

$$MSE\left(t_S\right)=f\left[RS_\phi^2+S_y^2\left(1-\rho_{pb}^2\right)\right]\tag{2.3}$$

where R depends on the choice of the parameters.

Abd-Elfattah et al. (2010) proposed some ratio type estimators. The minimum MSE attained in Abd-Elfattah et al. (2010) was

$$MSE_{min}(t_{Abd}) = f\left[S_y^2\left(1-\rho_{pb}^2\right)\right] \tag{2.4}$$

The minimum MSE of t_{Abd} is equal to the MSE of regression estimator $t_{reg} = \bar{y} + \hat{\beta}(P - p)$.

$(MSE(t_{reg}) = f\left[S_y^2\left(1-\rho_{pb}^2\right)\right])$.

Shabbir and Gupta (2007) considered following estimator

$$t_{SG} = \bar{y}\left[d_1 + d_2(p - P)\right]\left(\frac{P}{p}\right) \tag{2.5}$$

where d_1 and d_2 are constants and whose sum is not necessarily equal to one.

The optimum MSE reported by Shabbir and Gupta (2007) of t_{SG} is

$$MSE\left(t_{SG}^0\right) = \frac{fS_y^2\left(1-\rho_{pb}^2\right)}{1+fC_y^2\left(1-\rho_{pb}^2\right)}$$

Unfortunately the expression obtained by Shabbir and Gupta (2007) is incorrect.
The corrected MSE of t_{SG} is given as-

$$MSE\left(t_{SG}\right)_{min} = \left[\bar{Y}^2 - \frac{\Delta_1\Delta_5^2 + \Delta_3\Delta_4^2 - 2\Delta_2\Delta_4\Delta_5}{\Delta_1\Delta_3 - \Delta_2^2}\right] \tag{2.6}$$

where,

$\Delta_1 = \bar{Y}^2 + \bar{Y}^2 f\left(C_y^2 + 3C_x^2 - 4\rho C_y C_x\right)$, $\quad \Delta_2 = \bar{X}\bar{Y}f\left(2C_x^2 - \rho C_y C_x\right)$,

$\Delta_3 = \bar{X}^2 f C_x^2$, $\quad \Delta_4 = \bar{Y}^2 f\left(C_x^2 - \rho C_y C_x\right) + \bar{Y}^2$, $\quad \Delta_5 = \bar{X}\bar{Y} f C_x^2$.

3. The proposed estimator

We define a family of ratio estimators of population mean $\bar{Y}$ as

$$t_\alpha = \alpha_1\bar{y} + \alpha_2\bar{y}\left(\frac{m_1P + m_2}{m_1p + m_2}\right)^\alpha \tag{3.1}$$

where m_1 and m_2 are same as defined in (2.2) and α_1 and α_2 are real constants to be determined such that the MSE of t_α is minimum.

Remark 1: Here we would like to mention that the choice of the estimator depends on the availability and values of the various parameter(s) used (for choice of the parameters m_1 and m_2 refer to Singh et al. (2008) and Singh and Kumar (2009)).

Expressing t_α in terms of e's we have

$$t_\alpha = \overline{Y}\left[\alpha_1\left(1+e_y\right)+\alpha_2\left(1+\theta e_\phi\right)^{-\alpha}\right] \tag{3.2}$$

where $\theta = \dfrac{aP}{ap+b}$.

Expanding the right hand side of (3.2) and retaining terms up to second power of e's, we have

$$t_\alpha = \overline{Y}\left[\alpha_1\left(1+e_y\right)+\alpha_2\left\{1-\alpha\theta e_\phi + \frac{\alpha(\alpha+1)}{2}\theta^2 e_\phi^2 + e_y - \alpha\theta e_y e_\phi\right\}\right] \tag{3.3}$$

Subtracting $\overline{Y}$ from both side of (3.3) and then taking expectations, we get the bias of the estimator t_α up to the first order of approximation, as

$$B(t_\alpha) = \overline{Y}\left[(\alpha_1+\alpha_2-1)+\alpha_2 f\left\{\frac{\alpha(\alpha+1)}{2}\theta^2 C_p^2 - \alpha\theta\rho_\phi C_y C_p\right\}\right] \tag{3.4}$$

Subtracting $\overline{Y}$ from both side of (3.3), squaring and then taking expectations, we get MSE of the estimator t_α up to the first order of approximation, as

$$MSE(t_\alpha) = \overline{Y}^2\left[\alpha_1^2 A_1+\alpha_2^2 A_2 + 2\alpha_1\alpha_2 A_3 - 2\alpha_2 A_4 - 2\alpha_1 + 1\right] \tag{3.5}$$

where

$$A_1 = 1+fC_y^2,$$

$$A_2 = 1+f\left\{C_y^2 + \alpha^2\theta^2 C_p^2 - 4\alpha\theta\rho_\phi C_{yCp} + \alpha(\alpha+1)\theta^2 C_p^2\right\},$$

$$A_3 = 1+f\left\{\frac{\alpha(\alpha+1)}{2}\theta^2 C_p^2 + C_y^2 - 2\alpha\theta\rho_\phi C_y C_p\right\},$$

$$A_4 = 1+f\left\{\frac{\alpha(\alpha+1)}{2}\theta^2 C_p^2 - \alpha\theta\rho_\phi C_y C_p\right\},$$

$$\alpha_1^* = \frac{A_2 - A_3 A_4}{A_1 A_2 - A_3^2} \quad \text{and} \quad \alpha_2^* = \frac{A_1 A_4 - A_3}{A_1 A_2 - A_3^2}.$$

Substituting these optimum values of α_1^* and α_2^* in (3.5), we get the minimum MSE of t_α as-

$$MSE(t_\alpha)_{min} = \overline{Y}^2\left[1-\frac{A_2 + A_1 A_4^2 - 2A_3 A_4}{A_1 A_2 - A_3^2}\right] \tag{3.6}$$

4. Another estimator

Singh et al. (2007) suggested exponential ratio type and exponential product type estimators, respectively, as

$$t_{SR} = \bar{y}\exp\left[\frac{P-p}{P+p}\right] \tag{4.1}$$

$$t_{SP} = \bar{y}\exp\left[\frac{p-P}{p+P}\right] \tag{4.2}$$

MSE expressions for the estimators t_{SR} and t_{SP} are given, respectively, as

$$MSE(t_{SR}) = f\bar{Y}^2\left[C_y^2 + \frac{C_p^2}{4} - \rho_\phi C_y C_p\right] \tag{4.3}$$

$$MSE(t_{SP}) = f\bar{Y}^2\left[C_y^2 + \frac{C_p^2}{4} + \rho_\phi C_y C_p\right] \tag{4.4}$$

Using (3.1) and Singh et al. (2007) estimator, we define another family of estimators for population mean $\bar{Y}$ as

$$t_w = \{w_1\bar{y} + w_2(P-p)\}\left\{\frac{aP+b}{ap+b}\right\}^{\alpha}\exp\left\{\frac{(aP+b)-(ap+b)}{(aP+b)+(ap+b)}\right\}^{\beta} \tag{4.5}$$

where w_1 and w_2 are constants and whose sum is not necessarily equal to one.

The Bias and MSE expressions of t_w are respectively, given by

$$Bias(t_p) = (w_1-1)\bar{Y} + f\left[(w_1\bar{Y}A + w_2PB)C_x^2 - w_1\bar{Y}B\rho C_y C_x\right] \tag{4.6}$$

$$\begin{aligned} MSE(t_p) = (w_1-1)^2\bar{Y}^2 + w_1^2(m_1+2m_3) + w_2^2 m_2 \\ + 2w_1w_2(-m_4-m_5) - 2w_1m_3 + 2w_2m_5 \end{aligned} \tag{4.7}$$

where,

$$A = \frac{\theta^2}{8}\left[4\alpha(\alpha+1) + \beta(\beta+2) + 4\alpha\beta\right], \qquad B = \left(\alpha + \frac{\beta}{2}\right)\theta,$$

$$m_1 = \bar{Y}^2 f\left(C_y^2 + B^2C_x^2 - 2B\rho C_y C_x\right), \qquad m_2 = \bar{X}^2 f\left(C_x^2\right),$$

$$m_3 = \bar{Y}^2 f\left(AC_x^2 - 2B\rho C_y C_x\right), \qquad m_4 = \bar{Y}\bar{X} f\left(-BC_x^2 + \rho C_y C_x\right),$$

$$m_5 = \bar{X}\bar{Y} f\left(-BC_x^2\right).$$

Differentiating equation (4.7) with respect to w_1 and w_2 and than equating to zero we get

$$w_1^* = \frac{L_3L_4 - L_2L_5}{\left(L_1L_3 - L_2^2\right)^2} \quad \text{and} \quad w_2^* = \frac{L_1L_5 - L_2L_4}{\left(L_1L_3 - L_2^2\right)^2}$$

where

$$L_1 = \left(\overline{Y}^2 + m_1 + 2m_3\right), \qquad L_2 = \left(-m_4 - m_5\right), \qquad L_3 = m_2,$$

$$L_4 = \left(m_3 + \overline{Y}^2\right), \qquad L_5 = \left(-m_5\right).$$

Substituting these optimum values of w_1^* and w_2^* in (4.7), we get the minimum MSE of t_p as-

$$MSE\left(t_p\right)_{min} = \left[\overline{Y}^2 - \frac{L_1L_5^2 + L_3L_4^2 - 2L_2L_4L_5}{L_1L_3 - L_2^2}\right]$$

5. Efficiency comparison:

First, we compare the efficiency of proposed estimator t_α with usual estimator and than with regression estimator.

The variance of the usual estimator $\bar{y}$ is given by

$$V(\bar{y}) = fC_y^2 \tag{5.1}$$

$$MSE\left(t_\alpha\right)_{min} \le V\left(\bar{y}\right)$$

$$\left[1 - \frac{A_2 + A_1A_4^2 - 2A_3A_4}{A_1A_2 - A_3^2}\right] \le \overline{Y}^2 f_1 C_y^2 \tag{5.2}$$

On solving, we observe that above condition always holds true. Therefore, proposed estimator t_α under optimum condition performs better than usual estimator.

Similarly, it can be shown that

$$MSE\left(t_\alpha\right)_{min} \le MSE\left(reg\right) = MSE_{min}\left(t_{Abd}\right)$$

If,

$$\overline{Y}^2\left[1 - \frac{A_2 + A_1A_4^2 - 2A_3A_4}{A_1A_2 - A_3^2}\right] \le \overline{Y}^2 f_1 C_y^2\left(1 - \rho_\phi^2\right). \tag{5.3}$$

This is also true for all values of $\alpha(-1,0,1)$.

Next, we compare the efficiency of proposed estimator t_p with usual estimator and than with regression estimator.

$$MSE(t_p)_{min} \le V(\bar{y})$$

If,

$$\left[\bar{Y}^2 - \frac{L_1L_5^2 + L_3L_4^2 - 2L_2L_4L_5}{L_1L_3 - L_2^2}\right] \le f_1\bar{Y}^2C_y^2. \tag{5.4}$$

On simplification, we observe that above condition is always true. Therefore proposed estimator (t_w)min performs better than usual estimator in all situations.

Similarly it can be shown that

$$MSE(t_p)_{min} \le MSE(reg) = MSE_{min}(t_{Abd})$$

If,

$$\left[\bar{Y}^2 - \frac{L_1L_5^2 + L_3L_4^2 - 2L_2L_4L_5}{L_1L_3 - L_2^2}\right] \le \bar{Y}^2 f_1C_y^2\left(1 - \rho_\phi^2\right) \tag{5.5}$$

This is also true for all values of $\alpha(-1,0,1.)$ and $\beta(-1,0,1.)$

Finally we have compared the efficiency of proposed estimator t_w with the estimator t_p

$$MSE(t_p)_{min} \le MSE(t_\alpha)_{min}$$

Or if,

$$\left[\bar{Y}^2 - \frac{L_1L_5^2 + L_3L_4^2 - 2L_2L_4L_5}{L_1L_3 - L_2^2}\right] \le \bar{Y}^2\left[1 - \frac{A_2 + A_1A_4^2 - 2A_3A_4}{A_1A_2 - A_3^2}\right] \tag{5.6}$$

The conditions depends upon choice of α and β.

6. Empirical study

We have used the data given in Sukhatme and Sukhatme ((1970) p. 256). Where,

Y : Number of villages in the circle and

ϕ : represent A circle consisting more than five villages.

The following Table shows percent relative efficiencies (PRE's) of different estimator's with respect to usual estimator.

n	N	$\bar{Y}$	P	ρ_{pb}	Cy	C_p
23	89	1102	0.1236	0.643	0.65405	2.19012

Table 1: PRE of different estimators with respect to usual estimator

Estimator	PRE
$\bar{y}$	100
t_{NGR}	12.648
t_{RER}	60.603
$t_{s(opt)}$	170.488
$(t_{SG})_{min}$	172.120
$(t_{\alpha})_{min}$.	173.132
t_w $\alpha = 1, \beta = 0$	172.120
$\alpha = 0, \beta = 1$	187.804
$\alpha = 1, \beta = 1$	**392.62**

Conclusion

From Table 1, one can see that the proposed estimator tα under optimum condition performs better than the Shabbir and Gupta (2007) estimator, Singh et al. (2008) estimator and usual estimator. Also, the performance of the second proposed estimator tw depends upon choice of α and β. For α = 1, β = 1, it attains maximum efficiency.

References

1. Abd-Elfattah, A.M. El-Sherpieny, E.A. Mohamed, S.M. Abdou, O.F. (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. Applied Mathematics Computation, doi:10.1016/j.amc.2009.12.041.
2. Naik, V.D., Gupta, P.C. (1996). A note on estimation of mean with known population of an auxiliary character. Journal of Indian Society of Agricultural Statistics 48(2) 151–158.
3. Shabbir, J., Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. Pakistan Journal of Statistics 23 (1) 1–9.
4. Singh, R. Chauhan, P. Sawan, N. Smarandache, F. (2007). Ratio-product type exponential estimator for estimating finite population mean using information on auxiliary attribute. Renaissance High Press.

5. Singh, R. Chauhan, P. Sawan, N. Smarandache, F. (2008). Ratio estimators in simple random sampling using information on auxiliary attribute. Pakistan journal of statistics and operations research 4 (1) 47–53

6. Singh, R., Kumar, M. and Smarandache, F. (2008). Almost unbiased estimator for estimating population mean using known value of some population parameter(s). Pakistan journal of statistics and operations research 4(2) pp63-76.

7. Singh, R. and Kumar, M. (2009). A note on transformations on auxiliary variable in survey sampling. Model Assisted Statistics and Applications (Accepted).

8. Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling theory of surveys with applications. Iowa State University Press, Ames, U.S.A.