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 $\phi_i$ denote the observations on variable y and $\phi$ respectively for the $i^{th}$ unit ($i=1,2,3,...N$). It is assumed that attribute $\phi$ takes only the two values 0 and 1 according as 

$\phi = 1$, if $i^{th}$ unit of the population possesses attribute $\phi$  
$\phi = 0$, 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 $\phi$ respectively, $p = \frac{A}{N}$ and $p = \frac{a}{n}$ denote the proportion of units in the population and sample, respectively, possessing attribute $\phi$.

Define,

$e_y = \frac{(\bar{y} - \bar{Y})}{\bar{Y}}$  
$e_\phi = \frac{(p - P)}{P}$,

$E(e_i) = 0, (i = y, \phi)$

$E\left(e_y^2\right) = fC_y^2$,  
$E\left(e_\phi^2\right) = fC_p^2$,  
$E\left(e_y e_\phi\right) = f\rho p b C_y C_p$.
Where
\[ f = \left( \frac{1}{n} - \frac{1}{N} \right) \quad C_y^2 = \frac{S_y^2}{\bar{Y}^2}, \quad C_p^2 = \frac{S_p^2}{p^2}, \]
and \( \rho_{pb} = \frac{S_{y\phi}}{S_y S_{\phi}} \) is the point biserial correlation coefficient.

Here,
\[ S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2, \quad S_{\phi}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\phi_i - P)^2 \]
and
\[ S_{y\phi} = \frac{1}{N-1} \left( \sum_{i=1}^{N} y_i \phi_i - NP\bar{Y} \right) \]

In order to have an estimate of the population mean \( \bar{Y} \) of the study variable \( y \), assuming the knowledge of the population proportion \( p \), Naik and Gupta (1996) defined following ratio and product estimators
\[ t_{NGR} = \bar{y} \left( \frac{P}{P} \right) \quad (1.1) \]
\[ t_{NGP} = \bar{y} \left( \frac{P}{P} \right) \quad (1.2) \]

The mean square error (MSE) of \( t_{NGR} \) and \( t_{NGP} \) up to the first order of approximation, respectively, are
\[ \text{MSE}(t_{NGR}) = f\bar{Y}^2 \left[ C_y^2 + C_p^2 - 2\rho_{pb} C_y C_p \right] \quad (1.3) \]
\[ \text{MSE}(t_{NGP}) = f\bar{Y}^2 \left[ C_y^2 + C_p^2 + 2\rho_{pb} C_y C_p \right] \quad (1.4) \]

2. Other estimators
Singh et al. (2008) suggested the following ratio estimator
\[ t_S = \frac{\bar{y} + b_{\phi} (P-p)}{(m_1 P + m_2)} (m_1 P + m_2) \quad (2.2) \]
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 \( \rho_{pb} \).

In Singh et al. (2008), MSE equation of these ratio-type estimators were given by
\[ \text{MSE}(t_S) = f \left[ R S_{\phi}^2 + S_y^2 \left( 1 - \rho_{pb}^2 \right) \right] \quad (2.3) \]
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
\[ \text{MSE}_{\text{min}}(t_{\text{Abd}}) = f \left[ S_y^2 \left( 1 - \rho_{pb}^2 \right) \right] \]  
(2.4)
The minimum MSE of \( t_{\text{Abd}} \) is equal to the MSE of regression estimator \( t_{\text{reg}} = \bar{y} + \hat{\beta}(P - p) \).
\[ \text{MSE}(t_{\text{reg}}) = f \left[ S_y^2 \left( 1 - \rho_{pb}^2 \right) \right]. \]

Shabbir and Gupta (2007) considered following estimator
\[ t_{SG} = \bar{y} \left[ d_1 + d_2 (P - p) \right] \left( \frac{P}{p} \right) \]  
(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
\[ \text{MSE} \left( t^{0}_{SG} \right) = \frac{fS_y^2 \left( 1 - \rho_{pb}^2 \right)}{1 + fC_y^2 \left( 1 - \rho_{pb}^2 \right)} \]
Unfortunately the expression obtained by Shabbir and Gupta (2007) is incorrect.
The corrected MSE of \( t_{SG} \) is given as-
\[ \text{MSE} \left( t_{SG} \right)_{\text{min}} = \left[ \bar{Y}^2 - \frac{\Delta_1 \Delta_5 + \Delta_3 \Delta_4 - 2 \Delta_2 \Delta_4 \Delta_5}{\Delta_1 \Delta_3 - \Delta_2^2} \right] \]  
(2.6)
where,
\[ \Delta_1 = \bar{Y}^2 + \bar{Y}^2 f \left( C_y^2 + 3C_x^2 - 4pC_yC_x \right), \quad \Delta_2 = \bar{X}Y f \left( 2C_x^2 - pC_yC_x \right), \]
\[ \Delta_3 = \bar{X}^2 fC_y^2, \quad \Delta_4 = \bar{Y}^2 f \left( C_x^2 - pC_yC_x \right) + \bar{Y}^2, \quad \Delta_5 = \bar{X}Y fC_x^2. \]

3. The proposed estimator

We define a family of ratio estimators of population mean \( \bar{Y} \) as
\[ t_{\alpha} = \alpha_1 \bar{y} + \alpha_2 \bar{y} \left( \frac{m_1P + m_2}{m_1P + m_2} \right) \]  
(3.1)
where \( m_1 \) and \( m_2 \) are same as defined in (2.2) and \( \alpha_1 \) and \( \alpha_2 \) are real constants to be determined such that the MSE of \( t_{\alpha} \) 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 $t_\alpha$ in terms of $e$’s we have

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

(3.2)

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

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

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

(3.3)

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

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

(3.4)

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

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

(3.5)

where

$$A_1 = 1 + f C_y^2,$$

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

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

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

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

Substituting these optimum values of $\alpha_1^*$ and $\alpha_2^*$ in (3.5), we get the minimum MSE of $t_\alpha$ as

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

(3.6)
4. Another estimator

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

\[ t_{SR} = \bar{y} \exp \left[ \frac{P - p}{P + p} \right] \quad (4.1) \]

\[ t_{SP} = \bar{y} \exp \left[ \frac{p - P}{p + P} \right] \quad (4.2) \]

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

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

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

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

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

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

\[ \text{Bias}(t_p) = (w_1 - 1)\bar{Y} + f \left[ (w_1 \bar{Y}A + w_2 PB)C_x^2 - w_1 \bar{Y}BpC_y C_x \right] \quad (4.6) \]

\[ \text{MSE}(t_p) = (w_1 - 1)^2 \bar{Y}^2 + w_1^2 (m_1 + 2m_3) + w_2^2 m_2 + 2w_1w_2 (-m_4 - m_5) - 2w_1m_3 + 2w_2m_5 \quad (4.7) \]

where,

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

\[ m_1 = \bar{Y}^2 f \left( C_y^2 + B^2C_x^2 - 2BpC_y C_x \right) \quad m_2 = \bar{X}^2 f \left( C_x^2 \right) \]

\[ m_3 = \bar{Y}^2 f \left( AC_x^2 - 2BpC_y C_x \right) \quad m_4 = \bar{Y}X f \left( -BC_x^2 + pC_y C_x \right) \]

\[ m_5 = \bar{X}Y f \left( -BC_x^2 \right) \]
Differentiating equation (4.7) with respect to $w_1$ and $w_2$ and then equating to zero we get

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

where

$$L_1 = (\bar{Y}^2 + m_1 + 2m_3) \quad L_2 = (-m_4 - m_3) \quad L_3 = m_2,$$

$$L_4 = (m_3 + \bar{Y}^2) \quad L_5 = (-m_5).$$

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

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

5. Efficiency comparison:

First, we compare the efficiency of proposed estimator $t_\alpha$ with usual estimator and than with regression estimator.

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

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

(5.1)

$$\text{MSE}(t_\alpha)_{\text{min}} \leq V(\bar{y})$$

$$\left[ 1 - \frac{A_2 + A_1A_4^2 - 2A_3A_4}{A_1A_2 - A_3^2} \right] \leq \bar{Y}^2f_1C_y^2$$

(5.2)

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

Similarly, it can be shown that

$$\text{MSE}(t_\alpha)_{\text{min}} \leq \text{MSE(\text{reg})} = \text{MSE}_{\text{min}}(t_{\text{Abd}})$$

If,

$$\bar{Y}^2 \left[ 1 - \frac{A_2 + A_1A_4^2 - 2A_3A_4}{A_1A_2 - A_3^2} \right] \leq \bar{Y}^2f_1C_y^2 \left( 1 - \rho_0^2 \right)$$

(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 then with regression estimator.

\[
\text{MSE}(t_p)_{\text{min}} \leq V(\bar{y})
\]

If,

\[
\left[ \bar{y}^2 - \frac{L_1 L_5^2 + L_3 L_4^2 - 2L_2 L_4 L_5}{L_1 L_3 - L_2^2} \right] \leq f_1 \bar{y}^2 C_y^2.
\]

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

Similarly it can be shown that

\[
\text{MSE}(t_p)_{\text{min}} \leq \text{MSE}(\text{reg}) = \text{MSE}_{\text{min}}(t_{\text{Abd}})
\]

If,

\[
\left[ \bar{y}^2 - \frac{L_1 L_5^2 + L_3 L_4^2 - 2L_2 L_4 L_5}{L_1 L_3 - L_2^2} \right] \leq \bar{y}^2 f_1 C_y^2 \left(1 - \rho^2_{\phi} \right)
\]

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 \( t_p \)

\[
\text{MSE}(t_p)_{\text{min}} \leq \text{MSE}(t_{\alpha})_{\text{min}}
\]

Or if,

\[
\left[ \bar{y}^2 - \frac{L_1 L_5^2 + L_3 L_4^2 - 2L_2 L_4 L_5}{L_1 L_3 - L_2^2} \right] \leq \bar{y}^2 \left[1 - \frac{A_2 + A_1 A_4^2 - 2A_3 A_4}{A_1 A_2 - A_3^2}\right]
\]

The conditions depends upon choice of \( \alpha \) and \( \beta \).

### 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
- \( \phi \) : 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.

<table>
<thead>
<tr>
<th>n</th>
<th>N</th>
<th>( \bar{Y} )</th>
<th>P</th>
<th>( \rho_{pb} )</th>
<th>Cy</th>
<th>C_p</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>89</td>
<td>1102</td>
<td>0.1236</td>
<td>0.643</td>
<td>0.65405</td>
<td>2.19012</td>
</tr>
</tbody>
</table>
Table 1: PRE of different estimators with respect to usual estimator

<table>
<thead>
<tr>
<th>Estimator</th>
<th>PRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}$</td>
<td>100</td>
</tr>
<tr>
<td>$t_{NGR}$</td>
<td>12.648</td>
</tr>
<tr>
<td>$t_{RER}$</td>
<td>60.603</td>
</tr>
<tr>
<td>$t_{s(opt)}$</td>
<td>170.488</td>
</tr>
<tr>
<td>$(t_{SG})_{min}$</td>
<td>172.120</td>
</tr>
<tr>
<td>$(t_{\alpha})_{min}$</td>
<td>173.132</td>
</tr>
<tr>
<td>$t_w$ $\alpha = 1, \beta = 0$</td>
<td>172.120</td>
</tr>
<tr>
<td>$\alpha = 0, \beta = 1$</td>
<td>187.804</td>
</tr>
<tr>
<td>$\alpha = 1, \beta = 1$</td>
<td>392.62</td>
</tr>
</tbody>
</table>

Conclusion

From Table 1, one can see that the proposed estimator $t_\alpha$ 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 $t_w$ depends upon choice of $\alpha$ and $\beta$. For $\alpha = 1$, $\beta = 1$, it attains maximum efficiency.

References


