An Efficient Predictive Approach to Estimation in Two-phase Sampling

K. B. Panda

Reader, Department of Statistics, Utkal University, Bhubaneswar, Odisha, India

Abstract: Agrawal and Jain [1] employed a predictive framework to examine the predictive character of ratio, ratio-type and regression estimators in two-phase sampling. In this paper, an efficient predictive estimator, which is the fountainhead of a family of widely used estimators in two-phase sampling, is proposed. The newly proposed estimator has been shown to excel its competing estimators provided a weighting factor is appropriately chosen. In the absence of knowledge of the optimum weighting factor, performance-sensitivity of the proposed estimator has been carried out.

Keywords: Efficient predictive estimator in two-phase sampling; performance-sensitivity; ratio, ratio-type and regression estimators in two-phase sampling;

I. Introduction

In the ratio method of estimation, we, with a view to obtaining more efficient estimators of the population mean of the survey variable y, a known and closely related auxiliary variable x. However, when the population mean of x is not available, we invoke the technique known as two-phase sampling or double sampling. This technique essentially consists in selecting a large sample in the first phase for collecting information on x, followed by a selection of a subsample from the first-phase sample in the second phase for measuring y.

Consider a population of N units arbitrarily labelled 1, 2, …., N having mean and square denoted by (, ) for the y-variable and (, ) for the x-variable, the respective measurements on the y and the x variables for the jth unit being denoted by and , for j = 1, 2, …., N. Let n1 and n2 be the sample sizes in the first and the second phases, respectively, drawn according to the method of simple random sampling without replacement. Further, let and be the means of auxiliary variable x based on and units, respectively, and be the mean of the survey variable y based on n units. Then, the usual ratio-type estimator in two-phase sampling is given by

\[
\bar{Y}_{rd} = \frac{X'}{X} \cdot (1.1)
\]

Agrawal and Jain [2] have shown that is predictive in character. For this purpose, they have split the population total Y in the following form:

\[
Y = \sum_{j \in \xi} y_j + \sum_{j \in \xi^c} y_j + \sum_{j \in \eta} y_j + \sum_{j \in \eta^c} y_j \quad (1.2)
\]

where and denote the first phase and the second phase samples, respectively, and being their respective complements. The first component of right side of (1.2) being exactly known, each in the segments and in keeping with the sampling situation at hand, is predicted by means of ( ) and ( ), respectively. Although this approach adopted by Agrawal and Jain is quite justifiable and intuitively appealing, there is need to generalize the same as regards the prediction of each in and . In a practical situation, it would be ideal to utilize, for prediction purposes, the available information on the main and the auxiliary variables to form suitably weighted predictors for the x-observed segment and the completely non-surveyed (unobserved) segments. It is in the light of this background that we, in the following section, come up with an efficient predictive estimator in two-phase sampling.

II. An Efficient Predictive Estimator in Two-phase Sampling

Since no information on y has been collected in respect of the segments and , it is clear from (1.2) that the population total Y can be estimated if each is in these segments is appropriately predicted. Since the auxiliary information is fully available in the segment and as per the procedure of two-phase sampling, an apparently broad-based sensible predictor (employing two potential predictors) of in is proposed as

\[
\hat{Y}_j = \alpha \frac{X_j}{\bar{X}} + (1 - \alpha) \bar{Y}, \quad j \in s_1 \xi_2 \quad (2.1)
\]

where is a weight which might be preassigned or might depend on quantities estimated from the sample. In this context, it would be apt to point out that, while is the mean for the segment , the quantity is the usual
An Efficient Predictive Approach to Estimation in Two-phase Sampling

predictor for \( y_j \) (\( j \in s_1, s_2 \)), see Agrawal and Jain [1]. As regards the non-surveyed segment \( s_1 \), a plausible weighted predictor would then be

\[
y_j = a \frac{\bar{y}}{\bar{x}} + (1-a)y_j \in s_1(2.2)
\]

which represents the weighted mean of the potential predictors \((\bar{y}/\bar{x})\bar{x}^i\) and \( \bar{y} \) for each \( y_j \) \( j \in s_1 \).

Now, to estimate the population mean \( \bar{Y} \), we follow up the predictive decomposition of \( Y \) as given in (1.2) and employing the predictors given in (2.1) and (2.2), the proposed estimator is

\[
y_{rd} = a \frac{\bar{y}}{\bar{x}} + (1-a)y_j. \quad (2.3)
\]

Note that, \( y_{rd} \) reduces to well-known estimators in two-phase sampling via specific values of \( a \), e.g.,

(a) \( y_{rd} \) (the usual regression estimator in two-phase sampling) if \( a = b \bar{x}/\bar{y} \), where \( b \) is the sample regression coefficient.

It is evident that even the predictors \( y_j \) given in (2.1) and (2.2) in respect of \( s_1, s_2 \), and \( s_1, s_2 \), respectively, reduce to the known forms, cf. Agrawal and Jain [1].

We refer to Sukhatme et al. ([4], p.213) for a discussion of the other estimators employed in two-phase sampling, namely, the Hartley-Ross, Tin’s and Beale’s estimators defined by

\[
y_{HRd} = \bar{Y} - \frac{n(n-1)}{n-n-1} (\bar{y} - \bar{x}) \quad (2.4)
\]

\[
y_{rd} = \frac{1}{n} [1 - \frac{1}{n} S_2^2 / S_{xy}^2] \quad (2.5)
\]

and \( \bar{y}_{BD} = \frac{1}{n} [1 + \frac{1}{n} S_2^2 / S_{xy}^2] \quad (2.6) \)

where \( \bar{y} = \sum_{j=1}^{n} y_j \) and \( S_2^2 \) and \( S_{xy} \) represent, respectively, the sample mean square of \( x \) and the sample covariance between \( x \) and \( y \). The estimators given in (2.4), (2.5) and (2.6) are obtainable from (2.3) choosing a suitable \( a \) in each case.

The results based on predictive approach that is developed here can also apply to one-phase sampling when \( n = n \) in relation to the customary ratio and regression methods of estimation.

III. Performance of the Proposed Estimator vis-à-vis the Competing Estimators in Two-phase Sampling

The mean square error, to the first degree of approximation, of the composite estimator \( y_{ad} \), taking \( a \) as a pre-assigned weight, is obtained as

\[
M(y_{ad}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_2^2 + \left( \frac{1}{n} - \frac{1}{n} \right) \left( a^2 R^2 S_2^2 - 2a R p S_x S_y \right). \quad (3.1)
\]

where \( p \) is the correlation coefficient between \( x \) and \( y \) and \( R = \bar{Y} / \bar{X} \), the other notations having the same meaning as given in section 1.

The mean square error, to the first degree of approximation, of \( y_{rd} \) and \( \bar{y}_{HRd} \) given in (1.1) and (2.3) respectively, are known to be

\[
M(y_{rd}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_2^2 + \left( \frac{1}{n} - \frac{1}{n} \right) \left( R^2 S_2^2 - 2R p S_x S_y \right) \quad (3.2)
\]

and \( M(\bar{y}_{HRd}) = \left( \frac{1}{n} - \frac{1}{N} \right) S_2^2 + \left( \frac{1}{n} - \frac{1}{n} \right) \left( R^2 S_2^2 - 2R p S_x S_y \right) \quad (3.3) \)

where \( R = \frac{1}{N} \sum_{j=1}^{n} \frac{y_j}{x_j} \). See Sukhatme et al. ([4], pp.212-213). Using (3.1) and (3.2), a condition for better performance of \( y_{ad} \) relative to \( y_{rd} \) namely

\[
(\alpha^2 - 1) R S_x - 2(\alpha - 1) p S_x \leq 0
\]

leads to

\[
\rho \geq \left( \frac{1+\alpha}{2} \right) \frac{C_p}{C_x} \quad \text{if} \ a \geq 1;
\]

otherwise

\[
\rho \leq \left( \frac{1+\alpha}{2} \right) \frac{C_p}{C_x} \quad \text{if} \ a < 1;
\]

which, in turn, yield the following equivalent conditions on the range of \( a \):

\[
1 \leq a \leq 2a - 1 \quad \text{if} \ A \geq 1 \quad (3.4) \text{otherwise,} \quad 2a - 1 \leq a \leq 1 \quad \text{if} \ A \leq 1, \quad (3.5)
\]

For which \( y_{ad} \) is to be preferred to \( y_{rd} \) where \( A = p C_p/C_x \) and \( C_p \) and \( C_x \) are the coefficients of variation of \( y \) and \( x \), respectively. It is thus clear from (3.4) and (3.5) that a suitable value of \( \alpha \) can invariably be chosen with a view to rendering \( y_{ad} \) more efficient than \( y_{rd} \). Since \( y_{ad} \) is a widely used estimator, it would be worthwhile to note that the condition \( A \geq 1 \) always points to the \( y \)-variability being higher than the \( x \)-variability, while the

DOI: 10.9790/5728-11317881 www.iosrjournals.org 79 | Page
condition $\Delta \leq 1$ would often point to the reverse case. As a matter of fact, we are faced with the condition $\Delta \geq 1$ in a large variety of practical situations.

In this context, it would be apt to consider two well-known ratio-type estimators in two-phase sampling given in (2.3) and (2.4), namely, Tin’s and Beale’s estimators $\bar{y}_{Td}$ and $\bar{y}_{Bd}$ which have the same approximate mean square error as that of $\bar{y}_{rd}$ given in (3.2), see Sukhatme et al. ([4], p. 213) and hence, $\bar{y}_{ad}$ would fare better than $\bar{y}_{Td}$ and $\bar{y}_{Bd}$ under the same conditions as given in (3.4) and (3.5).

Analogously, employing (3.1) and (3.3), the conditions on $\alpha$ for $\bar{y}_{ad}$ to perform better than $\bar{y}_{Hrd}$ can be expressed as

$\varphi \leq \alpha \leq 2 \Delta - \varphi$ if $\Delta \geq \varphi$ (3.6)

or $2 \Delta - \varphi \leq \alpha \leq \varphi$ if $\Delta \leq \varphi$, (3.7)

where $\varphi = R/R$. It may be noted that

$\varphi \geq 1 \Rightarrow \bar{R} \geq R \Rightarrow \rho_{xx} \leq 0$

and $\varphi \leq 1 \Rightarrow \bar{R} \leq R \Rightarrow \rho_{xx} \geq 0$, where $\rho_{xx}$ is the correlation coefficient between $z = y/x$ and $x$. Thus, a choice, in accordance with (3.6) or (3.7), of a suitable value of $\alpha$ can unexceptionably be made so that $\bar{y}_{ad}$ fares better than $\bar{y}_{Hrd}$.

Now, a comparison of $\bar{y}_{ad}$ with the usual regression estimator $\bar{y}_{id}$ in two-phase sampling whose mean square error, to the first degree of approximation, is given by

$$M(\bar{y}_{ad}) = \left(\frac{1}{n} - \frac{1}{n}\right)\delta^2 + \left(\frac{1}{n} - \frac{1}{n}\right)\rho^2 \delta^2$$

shows that the former will be as efficient as the latter when $\alpha = \Delta$.

In the context of our foregoing appraisal of the proposed estimator $\bar{y}_{ad}$, it is quite natural to examine its performance vis-à-vis the usual sample mean $\bar{y}$ having the variance

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{n}\right)\delta^2.$$

The results obtained in this section are now concisely presented in Table 3.1.

<table>
<thead>
<tr>
<th>Competing Estimators</th>
<th>Estimator to be used</th>
<th>Choice of $\alpha$</th>
</tr>
</thead>
</table>
| $\bar{y}_{ad}$ vs $\bar{y}_{Td}$ or $\bar{y}_{Bd}$ | $\bar{y}_{ad}$ | $1 \leq \alpha \leq 2 \Delta - 1$ if $\Delta \geq 1$
| | | $2 \Delta - 1 \leq \alpha \leq 1$ if $\Delta \leq 1$
| $\bar{y}_{ad}$ vs $\bar{y}_{Hrd}$ | $\bar{y}_{ad}$ | $\varphi \leq \alpha \leq 2 \Delta - \varphi$ if $\Delta \geq \varphi$
| | | $2 \Delta - \varphi \leq \alpha \leq \varphi$ if $\Delta \leq \varphi$
| $\bar{y}_{ad}$ vs $\bar{y}_{id}$ | $\bar{y}_{ad}$ | $\alpha = \Delta$
| $\bar{y}_{ad}$ vs $\bar{y}$ | $\bar{y}_{ad}$ | $\alpha \leq \Delta$

As evidenced from the above table, a common single value of $\alpha$ that renders $\bar{y}_{ad}$ the best among the competing estimators considered by us is $\Delta = \alpha C_y/C_x$ which, in fact, yields the minimum value of the mean square error of $\bar{y}_{ad}$ given in (3.1).

As regards the choice of $\alpha$ equal to $\Delta$, it can be said that the population coefficients of variation $C_y$ and $C_x$ and the correlation coefficient $\rho$ may often be more or less known on the basis of past data, experience, a pilot survey or otherwise and hence some prior information on $\Delta$ may not be a problem, see Ray and Sahay [3].

To conclude the foregoing discussion, it can be said that the composite estimator $\bar{y}_{ad}$, employing a suitable choice of $\alpha$, can invariably be invoked with a view to scoring over the well-known estimators in two-phase sampling.

### IV. Performance-Sensitivity due to Lack of Optimality of $\alpha$

We now appraise performance-sensitivity of $\bar{y}_{ad}$ when optimum $\alpha$, viz., $\Delta$, is not available, meaning thereby that we examine the performance of the estimator $\bar{y}_{ad}$ if the optimum $\alpha$ (i.e., $\Delta$) is not employed and instead we use a weight $\alpha$, which embodies a certain error in $\Delta$, defined as

$$\alpha = (1 + \delta)\Delta,$$

where $\delta$ symbolises proportional deviation in $\Delta$. As a result of use of $\alpha$ in stead of $\Delta$, there will be a proportional increase in mean square error measured by

$$P_I = \frac{M(\bar{y}_{ad}) - M(\bar{y}_{ad})_{\alpha=\Delta}}{M(\bar{y}_{ad})_{\alpha=\Delta}}.$$

which, for large $N$, can be worked out as

$$P_I = \left(\frac{1}{n} - \frac{1}{n}\right)\delta^2 \rho^2 / \left(\frac{1}{n} - \frac{1}{n}\right)\rho^2 + \frac{\rho^2}{n}$$

and the same can then yield
$P_I \leq \delta^2 \bar{p}^2 < \frac{n}{2(n-n)}$, \quad (4.1)

which will always hold if $n^* \leq 2n$. From (4.1), it is clear that, if $n^* \leq 2n$, the proportional increase in mean square error ($P_I$) resulting from lack of optimality of $\alpha$ would be less than the square of proportional deviation $\delta$ in optimum $\alpha$. In other words, if $\delta$ is of the order of 10% or 20%, then $P_I$ will not exceed 1% or 4% as the case may be.

However, we can obtain $P_I$ as

$$P_I = \delta^2 \left\{ \frac{v(\bar{y}) - M(\bar{y}_{ad})_{\alpha=\Delta}}{M(\bar{y}_{ad})_{\alpha=\Delta}} \right\},$$

from which it can be interpreted that $P_I$ is $\delta^2$ times the gain in efficiency of $(\bar{y}_{ad})_{\alpha=\Delta}$ relative to $\bar{y}$.

From the above results, we can conclude that, unless $\delta$ is quite large, the inflation in variance of $\bar{y}_{ad}$ resulting from the use of non-optimum $\alpha$ will not be significant. Note that $P_I$ is symmetric with respect to deviations from $\Delta$.

V. Numerical Illustration

We now illustrate the performance of the composite estimator $\bar{y}_{ad}$ vis-à-vis some well-known estimators in two-phase sampling.

For a certain population, it is a priori known that $\Delta = 0.60$. On the basis of a sample survey, the following quantities are obtained:

N=117, $\bar{n}=40$, n=17, $\bar{R} = \bar{y}/\bar{x} = 0.99$, $\bar{r} = \frac{1}{n} \sum_{j=1}^{n} y_j/x_j = 1.00$, $s_2^2 = 287.85$, $s^2 = 458.56$ and $\bar{\rho} = 0.72$.

For the above example, the estimated relative efficiency of each of the estimators $\bar{y}_{rd}$ (or $\bar{y}_{rd}$ or $\bar{y}_{bd}$), $\bar{y}_{HRd}$ and $\bar{y}_{ad}$ with respect to $\bar{y}$ is presented in Table 5.1 given below.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Estimated Relative Efficiency w.r.t. $\bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}$</td>
<td>1.00</td>
</tr>
<tr>
<td>$\bar{y}<em>{rd}$ or $\bar{y}</em>{rd}$ or $\bar{y}_{bd}$</td>
<td>1.19</td>
</tr>
<tr>
<td>$\bar{y}_{HRd}$</td>
<td>1.18</td>
</tr>
<tr>
<td>$\bar{y}_{ad}$ (with $\alpha = \Delta = 0.60$)</td>
<td>1.53</td>
</tr>
</tbody>
</table>

The above table demonstrates that, in the context of two-phase sampling, appreciable gain in efficiency can be achieved through the use of $\bar{y}_{ad}$.

In the light of our findings of section 4, we examine the impact of variation in $\Delta (=0.60)$ on the relative efficiency of $\bar{y}_{ad}$. For this purpose, we have prepared the following table:

<table>
<thead>
<tr>
<th>$\alpha = \Delta$ (guessed $\Delta$)</th>
<th>Estimated Loss in Efficiency of $(\bar{y}<em>{ad})</em>{\alpha=\Delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.45</td>
<td>0.0371</td>
</tr>
<tr>
<td>0.55</td>
<td>0.0037</td>
</tr>
<tr>
<td>0.65</td>
<td>0.0037</td>
</tr>
<tr>
<td>0.75</td>
<td>0.0031</td>
</tr>
</tbody>
</table>

Table 5.2 makes it abundantly clear that even if $\Delta$ is subject to the error to the extent of 25%, the superiority of $(\bar{y}_{ad})_{\alpha=\Delta}$ remains considerably intact in the sense that the estimated loss in efficiency is around 3% or less.

VI. Conclusion

Besides being predictive in character, the newly proposed estimator in two-phase sampling excels its competing estimators from the standpoint of efficiency if the weighting factor is optimally determined. In case there is a problem in the determination of optimum weighting factor, one can go ahead with a guessed value since the variation between the true value and the guessed value results in a negligible loss in efficiency.

Acknowledgement

The author is greatly thankful to Dr. M.C. Agrawal, former Professor of Statistics, University of Delhi for his guidance in the preparation of this paper.

References