IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 11, Issue 3 Ver. I (May - Jun. 2015), PP 78-81
www.iosrjournals.org

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 knownas 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 subsamplefrom 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 mean square denoted by($\bar{Y}$, S_y^2) for the y-variable and ($\bar{X}$, S_x^2)for the x-variable,the respective measurements on the y and the x variables for the jth unit being denoted by y_j and x_j ,j=1,2,…..,N.Let n' and n 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 $\bar{x}'$and $\bar{x}$be the means of auxiliary variable x based on n' and n units,respectively,and $\bar{y}$ 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

$$\bar{y}_{rd}=\frac{\bar{y}}{\bar{x}}\bar{x}'. \qquad (1.1)$$

Agrawal and Jain [2] have shown that$\bar{y}_{rd}$ is predictive in character.
For this purpose,they have split the population total Y in the following form:

$$Y=\sum_{j\in s_2} y_j+\sum_{j\in s_1\bar{s}_2} y_j + \sum_{j\in \bar{s}_1} y_j, \qquad (1.2)$$

where$s_1$ands_2 denote the first phaseand the second phase samples, respectively,$\bar{s}_1$and $\bar{s}_2$ being their respective compliments. The first component of right sideof (1.2) being exactly known,each y_j in the segments $s_1\bar{s}_2$ and$\bar{s}_1$,in keeping with the sampling situation at hand,is predicted by meansof$(\bar{y}/\bar{x})x_j$ and $(\bar{y}/\bar{x})\bar{x}'$,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 y_j in $s_1\bar{s}_2$ and $\bar{s}_1$.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 $s_1\bar{s}_2$ and the completely non-surveyed (unobserved) segment$\bar{s}_1$.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 $s_1\bar{s}_2$and $\bar{s}_1$,it is clear from (1.2) that the population total Y can be estimated if each y_j in these segments is appropriately predicted.Since the auxiliary information is fully available in the segment $s_1\bar{s}_2$ as per the procedure of two-phase sampling,an apparently broad-based sensible predictor (employing two potential predictors) of y_j in $s_1\bar{s}_2$ that we propose is

$$\hat{y}_j=\alpha\frac{\bar{y}}{\bar{x}}x_j + (1-\alpha)\bar{y}, \qquad j \in s_1\bar{s}_2 (2.1)$$

whereα 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$\bar{y}$is the mean for the segment s_2,the quantity$(\bar{y}/\bar{x})x_j$ is the usual predictor for y_j (j∈ $s_1\bar{s}_2$),see Agrawal and Jain [1].As regards the non-surveyed segment$\bar{s}_1$,a plausible weighted predictor would then be

$\hat{y}_j = \alpha\frac{\bar{y}}{\bar{x}}\bar{x}' + (1-\alpha)\bar{y}$, j∈ $\bar{s}_1$(2.2)

whichrepresentstheweighted mean of the potential predictors$(\bar{y}/\bar{x})\bar{x}'$and $\bar{y}$ for each y_j,j∈ $\bar{s}_1$.

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

$\bar{y}_{\alpha d} = \alpha\frac{\bar{y}}{\bar{x}}\bar{x}' + (1-\alpha)\bar{y}$. (2.3)

Note that, $\bar{y}_{\alpha d}$ reduces to well-known estimators in two-phase sampling via specific values of α,e.g.,

(a) $\bar{y}_{rd}$ (theusual ratio estimator in two-phasesamplinggiven by(1.1))

If $\alpha = 1$;

$(b)\bar{y}_{ld}$ (the usual regression estimator in twophase sampling) $if\ \alpha = b\bar{x}/\bar{y}$,where b is the sample regression coefficient.

It is evident that even the predictors$\hat{y}_j$ given in (2.1) and (2.2) in respect of $s_1\bar{s}_2$and $\bar{s}_1$,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 definedby

$\bar{y}_{HRd} = \bar{r}\bar{x}' + \frac{n(n'-1)}{n'(n-1)}(\bar{y} - \bar{r}\bar{x})$ (2.4)

$\bar{y}_{Td} = \bar{y}_{rd}[1-(\frac{1}{n}-\frac{1}{n'})(\frac{s_x^2}{\bar{x}^2}-\frac{s_{xy}}{\bar{x}\bar{y}})]$(2.5)

and$\bar{y}_{Bd} = \bar{y}_{rd}\ [1+(\frac{1}{n}-\frac{1}{n'})\frac{s_{xy}}{\bar{x}\bar{y}}]/[1+(\frac{1}{n}-\frac{1}{n'})\frac{s_x^2}{\bar{x}^2}]$, (2.6)

where$\bar{r}=\frac{1}{n}\sum_{j=1}^{n}\frac{y_j}{x_j}$,$s_x^2$ and s_{xy} are,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 α 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 $\bar{y}_{\alpha d}$, taking α as a pre-assigned weight,is obtained as

M($\bar{y}_{\alpha d}$) = $(\frac{1}{n}-\frac{1}{N})S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(\alpha^2R^2S_x^2 - 2\alpha R\rho S_yS_x\right)$,(3.1)

whereρ 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 errors,to the first degree of approximation,of $\bar{y}_{rd}$ and $\bar{y}_{HRd}$ given in (1.1) and (2.3),respectively,are known to be

M($\bar{y}_{rd}$) = $(\frac{1}{n}-\frac{1}{N})S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(R^2S_x^2 - 2R\rho S_yS_x\right)$ (3.2)

andM($\bar{y}_{HRd}$) = $(\frac{1}{n}-\frac{1}{N})S_y^2 + \left(\frac{1}{n}-\frac{1}{n'}\right)\left(\bar{R}^2S_x^2 - 2\bar{R}\rho S_yS_x\right)$, (3.3)

where$\bar{R} = \frac{1}{N}\sum_{j=1}^{N}\frac{y_j}{x_j}$,see Sukhatme et al. ([4],pp.212-213). Using (3.1) and (3.2),a condition for better performance of$\bar{y}_{\alpha d}$ relative to $\bar{y}_{rd}$,namely

$$(\alpha^2-1)RS_x - 2(\alpha-1)\rho S_y \le 0$$

leads to

$$\rho \ge \left(\frac{1+\alpha}{2}\right)\frac{C_x}{C_y}\text{if}\alpha \ge 1;$$

otherwise

$$\rho \le \left(\frac{1+\alpha}{2}\right)\frac{C_x}{C_y}\text{if}\alpha \le 1;$$

which, in turn,yield the following equivalent conditions on the range of α:

1≤ $\alpha \le 2\Delta - 1$ if Δ≥ 1(3.4)otherwise, $2\Delta - 1 \le \alpha \le 1$ if Δ≤ 1, (3.5)

forwhich $\bar{y}_{\alpha d}$ is to be preferred to $\bar{y}_{rd}$ where Δ= $\rho C_y/C_x$ and C_y 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 α can invariably be chosen with a view to rendering $\bar{y}_{\alpha d}$ more efficient than $\bar{y}_{rd}$.Since $\bar{y}_{rd}$ is a widely used estimator,it would be worthwhile to note that the condition Δ≥ 1 always points to the y-variablity being higher than the x-variability,while the 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, $\bar{y}_{\alpha d}$ 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 α for $\bar{y}_{\alpha d}$ 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 $\varphi = \bar{R}/R$.It may be noted that

$$\varphi \geq 1 => \bar{R} \geq R => \rho_{zx} \leq 0$$

and $\varphi \leq 1 => \bar{R} \leq R => \rho_{zx} \geq 0,$

where ρ_{zx} 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 α can unexceptionably be made so that $\bar{y}_{\alpha d}$ fares better than $\bar{y}_{HRd}$.

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

$$M(\bar{y}_{\alpha d}) = (\frac{1}{n} - \frac{1}{N})S_y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\rho^2 S_y^2$$

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 $\bar{y}_{\alpha d}$, it is quite natural to examine its performance vis-à-vis the usual sample mean $\bar{y}$ having the variance

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})S_y^2.$$

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

Table 3.1 Choice of estimator for various values of α

Competing Estimators	Estimator to be used	Choice of α
$\bar{y}_{\alpha d}$ vs $\bar{y}_{rd}$ or $\bar{y}_{Td}$ or $\bar{y}_{Bd}$	$\bar{y}_{\alpha d}$	$1 \leq \alpha \leq 2\Delta - 1$ if $\Delta \geq 1$ $2\Delta - 1 \leq \alpha \leq 1$ if $\Delta \leq 1$
$\bar{y}_{\alpha d}$ vs $\bar{y}_{HRd}$	$\bar{y}_{\alpha d}$	$\varphi \leq \alpha \leq 2\Delta - \varphi$ if $\Delta \geq \varphi$ $2\Delta - \varphi \leq \alpha \leq \varphi$ if $\Delta \leq \varphi$
$\bar{y}_{\alpha d}$ vs $\bar{y}_{ld}$	$\bar{y}_{\alpha d}$	$\alpha = \Delta$
$\bar{y}_{\alpha d}$ vs $\bar{y}$	$\bar{y}_{\alpha d}$	$\alpha \leq \Delta$

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

As regards the choice of α equal to Δ,it can be said that the population coefficients of variation C_y and C_x and the correlation coefficient ρ 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 Δ may not be a problem,see Ray and Sahay[3].

To conclude the foregoing discussion,it can be said that the composite estimator $\bar{y}_{\alpha d}$,employing a suitable choice of α,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 α

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

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

where δ symbolises proportional deviation in Δ. As a result of use of α in stead of Δ,there will be proportional increase in mean square error measured by

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

which,for large N,can be worked out as

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

and the same can then yield

$\boldsymbol{P_I} \leq \boldsymbol{\delta^2}$ **if** $\boldsymbol{\rho^2} < \frac{\boldsymbol{n'}}{\boldsymbol{2(n'-n)}}$, (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 α would be less than the square of proportional deviation δ in optimum α. In other words, if δ 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

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

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

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

V. Numerical Illustration

We now illustrate the performance of the composite estimator $\bar{y}_{\alpha d}$ 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 the basis of a sample survey,the following quantities are obtained:

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

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

Table 5.1 Estimated relative efficiency of the competing estimators w.r.t.$\bar{y}$

Estimator	Estimated Relative Efficiency w.r.t. $\bar{y}$
$\bar{y}$	1.00
$\bar{y}_{rd}$ or $\bar{y}_{Td}$ or $\bar{y}_{Bd}$	1.19
$\bar{y}_{HRd}$	1.18
$\bar{y}_{\alpha d}$ ($with\ \alpha = \Delta$= 0.60)	1.53

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

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

Table 5.2 Impact of variation inΔ on the relative efficiency of $\bar{y}_{\alpha d}$

$\alpha = \hat{\Delta}$ (guessed Δ)	Estimated Loss in Efficiency of $(\bar{y}_{\alpha d})_{\alpha=\Delta}$
0.45	0.0331
0.55	0.0037
0.65	0.0037
0.75	0.0031

Table 5.2 makes it abundantly clear that even if Δ is subject to the error to the extent of 25%, the superiority of $(\bar{y}_{\alpha d})_{\alpha=\Delta}$ 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 samplingexcels 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

[1]. M.C.Agrawal and N.Jain.Predictive estimation in doublesampling procedures.American Statistician.42.1988.184-186.
[2]. M.C.Agrawal and N. Jain. A new predictive product estimator.Biometrika.76. 1989. 822-823.
[3]. S.K. Rayand ASahai. Efficient families of ratio and ratio-type estimators.Biometrika.67. 1980.211-215.
[4]. P. V. Sukhatme,B .V.Sukhatme,S.Sukhatme and C.Asok.Sampling theory of surveys with applications. (ISAS, New Delhi, India and Iowa State University Press, Iowa, U.S.A. 1980)