Jour. Ind. Soc. Ag. Statistics
54(2), 2001 : 258–264

Predictive Estimation of Finite Population Mean in Two-Phase Sampling Using Two Auxiliary Variables

L.N. Sahoo and R.K. Sahoo
Utkal University, Bhubaneswar - 751004
(Received: March, 2001)

SUMMARY

Using predictive approach described by Basu [1], we focus attention on the creation of ratio-type estimators of a finite population mean in the presence of two auxiliary variables under a two-phase sampling procedure. We also report analytical as well as numerical studies to examine performance of the new estimators obtained.

Key words: Auxiliary variable, Predictive approach, Two-phase sampling.

1. Introduction

Let y and x denote study variable and auxiliary variable taking values y_i and x_i $(1 \le i \le N)$ respectively for the ith unit of a finite population U. When the two variables are strongly related but no information is available on the population mean $\overline{X}$ of x, we seek to estimate the population mean $\overline{Y}$ of y using a two-phase sampling mechanism. In many practical situations even if $\overline{X}$ is unknown, information on an additional auxiliary variable z, closely related to x, is readily available on all units of U such that z_i denotes its value on unit i and $\overline{Z}$ as the known population mean. For instance, if the elements of U are hospitals, and y_i, x_i and z_i are respectively the number of deaths, number of patients admitted and number of available beds relating to the ith hospital, then information on z_i's can be collected easily from the official records of the Health Department. Allowing simple random sampling (WOR) design in each phase our two-phase sampling scheme in this case will be as follows:

(a) The first phase sample s′ of size n′ (n′ < N) is drawn from U to observe x and z.

(b) The second phase sample s of size n (n < n′) is drawn from s′ to observe y only.

Let us define $\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$, $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$, $\bar{z} = \frac{1}{n}\sum_{i\in s} z_i$, $\bar{x}' = \frac{1}{n'}\sum_{i\in s'} x_i$ and

$$\bar{z}' = \frac{1}{n}\sum_{i\in s'} z_i$$

Basic contributions to the estimation of population mean in this area were given by Chand [2], Kiregyera [4,5] and subsequently studied by others. In many papers, the authors usually tried to develop estimators from a classical two-phase sampling estimators simply replacing $\bar{x}'$ by an improved estimator of $\bar{X}$ treating z as an auxiliary variable. But, in this work, the point of departure is different. The auxiliary variables are used to build up estimators under the predictive approach advocated by Basu [1].

2. *Predictive Approach and Estimators*

A typical attempt in this approach is to express $\bar{Y}$ as

$$\bar{Y} = \frac{1}{N}\sum_{i\in s} y_i + \frac{1}{N}\sum_{i\in r_2} y_i + \frac{1}{N}\sum_{i\in r_1} y_i \tag{2.1}$$

by decomposing U into three mutually exclusive domains $s, r_2 = \bar{s} \cap s'$ and $r_1 = U - s'$ of n, $(n' - n)$ and $(N - n')$ units respectively, where $\bar{s} = U - s$ denotes the collection of units in U which are not included in s. Writing

$(n'-n)\bar{Y}_2 = \sum_{i\in r_2} y_i$ and $(N-n')\bar{Y}_1 = \sum_{i\in r_1} y_i$ we have

$$\bar{Y} = f\bar{y} + (f'-f)\bar{Y}_2 + (1-f')\bar{Y}_1 \tag{2.2}$$

where $f = \frac{n}{N}$ and $f' = \frac{n'}{N}$. Since the first component of the right hand side of (2.2) is exactly known, the problem is therefore to predict the quantities $\bar{Y}_1$ and $\bar{Y}_2$ from the sample data. If T_1 and T_2 are their implied predictors, then an estimator of $\bar{Y}$ under our predictive approach is provided by the following equation :

$$\hat{\bar{Y}} = f\bar{y} + (f'-f)T_2 + (1-f')T_1 \tag{2.3}$$

If $s = s' = U$, we note that $\hat{\bar{Y}} = \bar{Y}$, the target of our prediction, i.e., if the whole population is surveyed. On the other hand, if no additional information on U is available, the simplest but obvious choice for both T_1 and T_2 would be $\bar{y}$ so that $\hat{\bar{Y}} = \bar{y}$, the expansion estimator of $\bar{Y}$.

Under classical approach, there are many alternative ways of selecting suitable predictors in terms of both auxiliary variables whereupon $\hat{\bar{Y}}$ will be determined by the predictive equation (2.3). But, our aim is just to study superiority of the method of estimation employed here over the earlier ones. Thus, for simplicity, assume that y, x and z are so related that, we have enough scope for considering ratio-type estimators as the predictors for our purpose.

Since information on x is available at the sample level, we predict y-values in the domain r_2 using x-values only by taking $T_2 = \bar{y}\frac{\bar{X}_2}{\bar{x}}$, where $(n'-n)\bar{X}_2 = \sum_{i \in r_2} x_i$. But, the auxiliary information on z is available at the population level. Accordingly, we make three different selections for T_1 and obtain the resulting estimators for $\bar{Y}$.

(i) Let $T_1 = \frac{\bar{y}}{\bar{z}'}\bar{Z}_1$, then $\hat{\bar{Y}}$ turns out to be $\bar{y}_{11} = f'(\bar{y}_{TR} - \bar{y}) + \bar{y}\frac{\bar{Z}}{\bar{z}'}$, where $(N-n')\bar{Z}_1 = \sum_{i \in r_1} z_i$ and $\bar{y}_{TR} = \bar{y}\frac{\bar{x}'}{\bar{x}}$ is the classical two-phase sampling ratio estimator.

(ii) Let $T_1 = \bar{y}\frac{\bar{x}'}{\bar{x}}\frac{\bar{Z}}{\bar{z}'} = \bar{y}_{RR}$ (say)

Then $\hat{\bar{Y}}$ reduces to an estimator $\bar{y}_{12} = f'\bar{y}_{TR} + (1-f')\bar{y}_{RR}$, which is an weighted combination of $\bar{y}_{TR}$ and $\bar{y}_{RR}$ with respective weights f' and $1-f'$. The estimator $\bar{y}_{RR}$ is the well known ratio-in-ratio estimator suggested by Chand [2].

(iii) Let $T_1 = \bar{y}\frac{\bar{x}'}{\bar{x}}\frac{\bar{Z}_1}{\bar{z}'}$. Then on simplification, $\hat{\bar{Y}}$ reduces to $\bar{y}_{RR}$, showing that the customary ratio-in-ratio estimator $\bar{y}_{RR}$ of $\bar{Y}$ is now known to be endowed with a predictive character under (2.3).

The above selections for T_1 and T_2 (although made independently) are not unique. Gathering information an x and z, one may opt for other ratio-type predictors and create an infinite number of estimators for $\bar{Y}$.

3. Asymptotic Efficiency of $\bar{y}_{11}$ and $\bar{y}_{12}$

Omitting details of the calculations (suppressed to save space) the mean square errors of $\bar{y}_{11}$ and $\bar{y}_{12}$ to a first order of approximation are obtained as follows:

$$M(\bar{y}_{11}) = \bar{Y}^2\left[\phi\left(C_y^2 - 2f'\rho_{yx}C_yC_x + f'^2C_x^2\right) + \phi'\left(2f'\rho_{yx}C_yC_x - f'^2C_x^2 + C_z^2 - 2\rho_{yz}C_yC_z\right)\right] \quad (3.1)$$

$$M(\bar{y}_{12}) = \bar{Y}^2\left[\phi\left(C_y^2 - 2\rho_{yx}C_yC_x + C_x^2\right) + \phi'\left(2\rho_{yx}C_yC_x - C_x^2 + (1-f')^2C_z^2 - 2(1-f')\rho_{yz}C_yC_z\right)\right] \quad (3.2)$$

where $\phi = \left(\frac{1}{n} - \frac{1}{N}\right), \phi' = \left(\frac{1}{n'} - \frac{1}{N}\right)$

C_y, C_x, C_z are the coefficient of variations and $\rho_{yx}, \rho_{yz}, \rho_{xz}$ are the correlation coefficients.

To the same order of approximation, the mean square error expressions for $\bar{y}_{TR}$ and $\bar{y}_{RR}$ (as available in the literature) are given by

$$M(\bar{y}_{TR}) = \bar{Y}^2\left[\phi\left(C_y^2 - 2\rho_{yx}C_yC_x + C_x^2\right) + \phi'\left(2\rho_{yx}C_yC_x - C_x^2\right)\right] \quad (3.3)$$

$$M(\bar{y}_{RR}) = \bar{Y}^2\left[\phi\left(C_y^2 - 2\rho_{yx}C_yC_x + C_x^2\right) + \phi'\left(2\rho_{yx}C_yC_x - C_x^2 + C_z^2 + 2\rho_{yz}C_yC_z\right)\right] \quad (3.4)$$

From (3.1), (3.3) and (3.4) it may be seen that $M(\bar{y}_{11}) < M(\bar{y}_{TR})$ if

$$D_{yx} < \frac{1+f'}{2} \text{ and } D_{yz} > \frac{1}{2} \quad (3.5)$$

and $M(\bar{y}_{11}) < M(\bar{y}_{RR})$ if

$$D_{yx} < \frac{1+f'}{2} \quad (3.6)$$

where $D_{yx} = \rho_{yx}\frac{C_y}{C_x}$ and $D_{yz} = \rho_{yz}\frac{C_y}{C_z}$. Thus, when $D_{yx} < \frac{1+f'}{2}$ and $D_{yz} > \frac{1}{2}$, $\bar{y}_{11}$ would be more efficient than both $\bar{y}_{TR}$ and $\bar{y}_{RR}$.

Since $0 \le f' \le 1$, we must have $D_{yx} < 1$ for $\bar{y}_{11}$ to be more precise than both $\bar{y}_{TR}$ and $\bar{y}_{RR}$ provided $D_{yz} > \frac{1}{2}$. But choice of a higher value of f' always leads to an undue increase in the cost of the survey. So, a good apriori knowledge on D_{yx} (if available) can be utilized to determine the value of f' according to the restriction $2D_{yx} - 1 < f'$. For example, if $D_{yx} = 0.6$ then we may choose $f' > 0.20$. But, when $D_{yx} < 0.5$, D_{yx} is always less than $\frac{1+f'}{2}$. Thus, we conclude that $\bar{y}_{11}$ may be more efficient than $\bar{y}$ and $\bar{y}_{TR}$ even if $\bar{y}_{TR}$ is less efficient than $\bar{y}$.

From (3.2), (3.3) and (3.4) it may also be seen that $M(\bar{y}_{12}) < M(\bar{y}_{TR})$

if
$$D_{yz} < \frac{1-f'}{2} \tag{3.7}$$

and $M(\bar{y}_{12}) < M(\bar{y}_{RR})$ if

$$D_{yz} < \frac{2-f'}{2} \tag{3.8}$$

Now we find that in situations where $M(\bar{y}_{RR}) < M(\bar{y}_{TR})$ i.e. when $D_{yz} < \frac{1}{2}$, $\bar{y}_{12}$ can also be more precise than $\bar{y}_{TR}$, and both $\bar{y}_{TR}$ and $\bar{y}_{RR}$, if

$$\frac{1-f'}{2} < D_{yz} < \frac{2-f'}{2} \tag{3.9}$$

Hence, for our purpose, we can select a suitable value of f' according to the restriction

$$1-2D_{yz} < f' < 2\left(1-D_{yz}\right) \tag{3.10}$$

if a good guess value of D_{yz} is available.

Finally from (3.1) and (3.2) we conclude that $M(\bar{y}_{12}) > M(\bar{y}_{11})$ if

$$D_{yx} < \frac{1+f'}{2} \text{ and } D_{yz} > \frac{2-f'}{2}$$

$$\Rightarrow \qquad f' > 2D_{yx} - 1 \text{ and } 2(1 - D_{yz}) \tag{3.11}$$

and $M(\bar{y}_{12}) < M(\bar{y}_{11})$ if

$$D_{yx} > \frac{1+f'}{2} \text{ and } D_{yz} < \frac{2-f'}{2}$$

$$\Rightarrow \qquad f' < 2D_{yx} - 1 \text{ and } 2(1 - D_{yz}) \tag{3.12}$$

4. Numerical Illustrations

To demonstrate the gain in precision of new estimators $\bar{y}_{11}$ and $\bar{y}_{12}$ over their competitors $\bar{y}_{TR}$ and $\bar{y}_{RR}$ numerically, we consider three sets of data. The first two sets of the data have been taken from Fisher [3], whereas the third set has been taken from Shukla [6].

Data Set I: Consisting of measurements on three variables namely sepal width (y), sepal length (x) and petal length (z) for 50 Iris flowers (versicolor) such that

$$\bar{Y} = 2.770, C_y^2 = 0.012566, C_x^2 = 0.007343, C_z^2 = 0.011924, \rho_{yx} = 0.5259$$

$$\rho_{yz} = 0.5605, \rho_{xz} = 0.7540$$

Data Set II: Consisting of measurements of petal width (y), sepal width (x) and petal length (z) on 50 Iris flowers (virginica) with

$$\bar{Y} = 2.026, C_y^2 = 0.018009, C_x^2 = 0.011524, C_z^2 = 0.009683, \rho_{yx} = 0.5377$$

$$\rho_{yz} = 0.3221, \rho_{xz} = 0.4010$$

Data Set III: Consisting of measurements on yield of fiber (y), height (x) and base diameter (z) for 50 jute plants (capsularies), such that

$$\bar{Y} = 2.5840, C_y^2 = 0.0866, C_x^2 = 0.1163, C_z^2 = 0.0170$$

$$\rho_{yx} = 0.4800, \rho_{yz} = 0.3700, \rho_{xz} = 0.7300$$

The combinations of n' and n for data sets I, II, III are respectively taken as (20, 10), (18, 8) and (15, 8) and the relative efficiencies of different estimators with respect to the expansion estimators $\bar{y}$ are displayed in Table 4.1.

Findings of Table 4.1 show that $\bar{y}_{12}$ attains the maximum precision amongst all for the first two data sets. $\bar{y}_{11}$ is better than $\bar{y}$, $\bar{y}_{TR}$ and $\bar{y}_{RR}$ in data set I,

Table 4.1. Relative efficiency of different estimators w.r.t. $\bar{y}$ (in %)

Data set	$\bar{y}$	$\bar{y}_{TR}$	$\bar{y}_{RR}$	$\bar{y}_{11}$	$\bar{y}_{12}$
I	100	116	124	126	134
II	100	117	114	116	120
III	100	89	94	133	113

whereas its performance compared to $\bar{y}_{TR}$ is poor in data set II as the conditions in (3.5) are not fulfilled. For the third data set $\bar{y}_{11}$ is the most efficient followed by $\bar{y}_{12}$. But in this case both $\bar{y}_{TR}$ and $\bar{y}_{RR}$ are less efficient than $\bar{y}$. However, the empirical findings of the study essentially show that the estimators $\bar{y}_{11}$ and $\bar{y}_{12}$ may be superior to $\bar{y}$ even if $\bar{y}_{TR}$ and $\bar{y}_{RR}$ are inferior to $\bar{y}$.

REFERENCES

[1] Basu, D. (1971). *An Essay on the Logical Foundations of Survey Sampling*, Part I. Foundations of Statistical Inference, V.P. Godambe and D.A. Sprott (eds.), Holt, Rinehart and Winston, Toronto, Canada, 203-242.

[2] Chand, L. (1975). *Some ratio-type estimators based on two or more auxiliary variables*. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

[3] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**, 179-188.

[4] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika*, **27**, 217-223.

[5] Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika*, **31**, 215-226.

[6] Shukla, G.K. (1966). An alternative multivariate ratio estimate for finite population. *Calcutta Stat. Assoc. Bull.*, **15**, 127-134.