J. Ind. Soc. Agril. Statist.
61(3), 2007 : 364-368

Estimating Population Mean Square through Predictive Approach when Auxiliary Character is Estimated

U.C. Sud, Prasenjit Pal and I.C. Sethi
Indian Agricultural Statistics Research Institute, New Delhi

(Received: April 2007)

SUMMARY

In the context of estimation of population mean square, Royall's (1970) prediction approach is extended to the case where information on auxiliary characters is not available. A double sampling procedure is proposed as an alternative under such a situation. The efficiency of the estimator based on this scheme of sampling is compared with the one where the information on an auxiliary character for all the units in the population is collected. Optimum values of the sample sizes are also obtained. Further, an empirical study is carried out to examine the situations wherein the double sampling plan is superior.

Key words : Population mean square, Model based estimation, Double sampling.

1. INTRODUCTION

Predictive approach (Royall 1970) for estimating the mean square of a finite population assumes knowledge about the auxiliary character (x, say) for the entire population. The approach fails where information is not available for all the population units. The approach is essentially model based. However, where a probability sample is selected, expectations taken over the sampling design as well as over the model are used, when studying the efficiency aspects of the design (Sarndal 1978). In case auxiliary information is not available the usual concept of two-phase sampling may be useful. Obviously, the approach of estimation will deviate from being completely model based.

Srivastava and Sud (1988) proposed a double sampling based approach. The present paper discusses a double sampling plan and the corresponding estimation procedure for estimating the population mean square in a finite population framework. The procedure is compared with the usual predictive estimator under the same cost constraint.

2. DOUBLE SAMPLING BASED PREDICTOR OF POPULATION MEAN SQUARE

Consider a finite population $\Omega = \left(U_1, U_2, ..., U_N\right)$ of 'N' identifiable units. Let 'y' be the variable under study taking value y_i on the i^th unit. Let the parameter of interest be

$$S_y^2 = \frac{1}{(N-1)}\sum_{i}^{N}\left(y_i - \bar{Y}\right)^2$$

where

$$\bar{Y} = \frac{1}{N}\sum_{i}^{N} Y_i$$

We consider the following model

$$y_i = x_i\beta + e_i \tag{2.1}$$

where

$$E_\psi(e_i) = 0$$

$$V_\psi(e_i) = \sigma^2 x_i$$

$$E_\psi\left(e_i e_j\right) = 0 \;\forall i \neq j$$

E_ψ refers to expectation under the model and x_i's are assumed unknown. Further, we assume that a sample of size n' is selected from a population of N units by simple random sampling without replacement (SRSWOR) for collecting information on the auxiliary character. Let a sample of size n be selected from n' to collect information on the character under study.

Since

$$S_y^2 = \frac{(n-1)}{(N-1)}s_y^2 + \frac{(N-n-1)}{(N-1)}S_r^2 + \frac{n}{N(N-1)}(N-n)(\bar{y}_s - \bar{y}_r)^2$$

where

$$S_r^2 = \left\{\frac{1}{N-n-1}\sum_{i=n+1}^{N} y_i^2 - \frac{N-n}{N-n-1}\bar{y}_r^2\right\}$$

$$s_y^2 = \frac{1}{(n-1)}\sum_{i}^{n}(y_i - \bar{y}_s)^2$$

$$\bar{y}_s = \frac{1}{n}\sum_{i}^{n} y_i\ ,\ \bar{y}_r = \frac{1}{(N-n)}\sum_{i}^{N-n} y_i$$

This implies

$$\frac{N-n-1}{N-1}E_\psi S_r^2$$

$$= \frac{1}{(N-1)}\left\{\sum_{i=n+1}^{N} x_i^2\beta^2 + \sigma^2\sum_{i=n+1}^{N} x_i - (N-n)\bar{x}_r^2\beta^2 - \sigma^2\bar{x}_r\right\}$$

$$\bar{x}_r = \frac{1}{(N-n)}\left(N\bar{X} - n\bar{x}_s\right)$$

$$\bar{X} = \frac{1}{N}\sum_{i}^{N} X_i$$

Under the model (2.1)

$$b = \frac{\bar{y}_s}{\bar{x}_s}$$

where b is an unbiased estimator of β, $\bar{x}_s = \frac{1}{n}\sum_{i}^{n} x_i$

Also, by definition

$$V_\psi(b) = E_\psi\left(b^2\right) - \beta^2$$

Thus, $\hat{\beta}^2 = b^2 - \hat{V}(b)$

and

$$E_\psi\frac{n(N-n)}{N(N-1)}(\bar{y}_s - \bar{y}_r)^2 = \frac{n(N-n)}{N(N-1)}\left[\bar{x}_s^2\beta^2 + \bar{x}_r^2\beta^2 - 2\bar{x}_s\bar{x}_r\beta^2 + \sigma^2\frac{\bar{x}_s}{n} + \sigma^2\frac{\bar{x}_r}{(N-n)}\right]$$

Thus

$$\hat{\bar{x}}_r = \frac{1}{(N-n)}\left(N\bar{x}' - n\bar{x}_s\right)$$

$$\bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i$$

Further

$$\bar{x}_r^2 = \frac{1}{(N-n)^2}\left[N^2\bar{X}^2 + n^2\bar{x}_s{}^2 - 2Nn\bar{X}\bar{x}_s\right]$$

An unbiased estimator of $\bar{X}^2$ on the basis of a sample of size n' is given by

$$\hat{\bar{X}}^2 = \bar{x}'^2 - \left(\frac{1}{n'} - \frac{1}{N}\right)s'^2$$

where

$$s'^2 = \frac{1}{(n'-1)}\sum_{i=1}^{n'}(x_i - \bar{x}')^2$$

Since

$$\sum_{i=n+1}^{N} x_i^2\beta^2 = \sum_{i=1}^{N} x_i^2\beta^2 - \sum_{i=1}^{n} x_i^2\beta^2$$

this implies that an unbiased estimator of $\sum_{i=n+1}^{N} x_i^2\beta^2$ is $\frac{N}{n'}\sum_{i=1}^{n'} x_i^2\hat{\beta}^2 - \sum_{i=1}^{n} x_i^2\hat{\beta}^2$

Thus, the double sampling based predictor is given by

$$\hat{T}'_Q = \frac{n-1}{N-1}s_y^2 + \frac{N}{(N-1)n'}\sum_{i=1}^{n'}(x_i - \bar{x}')^2\hat{\beta}^2 - \frac{1}{(N-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\sigma^2}{N}(N\bar{x}' - n\bar{x}) + \frac{N\hat{\beta}^2}{(N-1)}\left(\frac{1}{n'} - \frac{1}{N}\right)s'^2 + \sigma^2\bar{x}\frac{(N-n)}{N(N-1)} \quad (2.2)$$

Consider $E_p E_\psi\left(\hat{T}'_Q - S_y^2\right)$

where E_p pertains to expectation with respect to the sampling design.

Now

$$E_pE_\psi\left(\frac{N-n-1}{N-1}S_r^2+\frac{n(N-n)}{N(N-1)}(\bar{y}_s-\bar{y}_r)^2\right)$$

$$=\frac{N}{(N-1)}\frac{(n'-1)}{n'}\beta^2S^2-\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i-\bar{x}_s)^2\beta^2$$

$$+\sigma^2\bar{X}+\frac{N\beta^2}{(N-1)}\left(\frac{1}{n'}-\frac{1}{N}\right)S^2+\sigma^2\bar{x}_s\frac{(n-1)}{(N-1)}$$

and

$$E_pE_\psi\left[\frac{N}{(N-1)n'}\sum_{i=1}^{n'}(x_i-\bar{x}')^2\hat{\beta}^2\right.$$

$$-\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{\sigma^2}{N}(N\bar{x}'-n\bar{x})$$

$$\left.+\frac{N\hat{\beta}^2}{(N-1)}\left(\frac{1}{n'}-\frac{1}{N}\right)s'^2+\sigma^2\bar{x}\frac{(N-n)}{N(N-1)}\right]$$

$$=\frac{N}{(N-1)}\frac{(n'-1)}{n'}\beta^2S^2-\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i-\bar{x}_s)^2\beta^2$$

$$+\sigma^2\bar{X}+\frac{N\beta^2}{(N-1)}\left(\frac{1}{n'}-\frac{1}{N}\right)S^2+\sigma^2\bar{x}_s\frac{(n-1)}{(N-1)}$$

Thus

$$E_pE_\psi\left(\hat{T}_Q'-S_y^2\right)=0$$

indicating that $\hat{T}_Q'$ is design-model unbiased.

The variance of $\hat{T}_Q'$ i.e. $E_pE_\psi\left[\hat{T}_Q'-S_y^2\right]^2$ is given by

$$\approx\left\{\left(\frac{\mu_4-\sigma^4}{n'}\right)+\sigma^4\right.$$

$$\left.-\frac{2}{(N-1)}\sum_{i}^{N}(x_i-\bar{X})^2\frac{1}{(N-1)}\sum_{i}^{n}(x_i-\bar{x})^2\right\}\frac{4\sigma^2\beta^2}{n\bar{x}_s} \tag{2.3}$$

where

$$\mu_4=\frac{1}{N}\sum_{i}^{N}(Y_i-\bar{Y})^4$$

Therefore, $\dfrac{E_pE_\psi\left[\hat{T}_Q'-S_y^2\right]^2}{\sigma^4}$

$$\approx\left\{\left(\frac{\beta_2-1}{n'}\right)+1\right.$$

$$\left.-\frac{\frac{2}{(N-1)}\sum_{i=1}^{N}(x_i-\bar{X})^2\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^4}\right\}\frac{4\sigma^2\beta^2}{n\bar{x}_s} \tag{2.4}$$

where $\beta_2=\dfrac{\mu_4}{\sigma^4}$

Thus, in this case the optimal sampling plan is purposive i.e. a sample for which the auxiliary character values are largest.

3. EFFICIENCY OF THE DOUBLE SAMPLING BASED PREDICTOR UNDER A COST FUNCTION

It would be interesting to examine the efficiency of the predictor developed in Section 2, under the scheme where auxiliary information is not available and it is collected partially by a simple random sample of size n' and another sample of size n is drawn purposively to observe the character under study with the predictor based on the scheme where a sample is selected purposively from the population under the assumption that auxiliary information is not available and cost is incurred to collect this information on all the units in the population.

Now, $\dfrac{E_pE_\psi\left[\hat{T}_Q'-S_y^2\right]^2}{\sigma^4}$ can be written as

$$\left[\frac{A}{n'}+1-B\right]\frac{C}{n} \tag{3.1}$$

where

$$A=\beta_2-1$$

$$B=\frac{\frac{2}{(N-1)}\sum_{i=1}^{N}(X_i-\bar{X})^2\frac{1}{(N-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^4}$$

$$C=\frac{4\sigma^2\beta^2}{\bar{x}_s}$$

We consider the following cost function

$$C_0 = C_1 n + C_2 n'$$

where

C_0 = total cost of collecting information

C_1 = per unit cost of collecting information on character under study

C_2 = per unit cost of collecting information on auxiliary character

Consider the following function

$$\varphi = \left[\frac{A}{n'} + 1 - B\right]\frac{C}{n} - \lambda\left[C_0 - C_1 n - C_2 n'\right] \quad (3.2)$$

where λ is the Lagrangian multiplier.

Differentiating (3.2) with respect to n', n and λ and equating the resulting expressions to '0'we get optimum values of n' and n as

$$n'_{(opt)}$$

$$= \frac{-A_1 A^2 C + C_2^2 \pm \sqrt{\left(-A_1 A^2 C + C_2^2\right)^2 + 4\left(\left\{A_1 AC + (2-B)\right\} C_2 C_0\right)}}{2A_1 AC(2-B)}$$

$$n_{(opt)} = \frac{C_0 - C_2 n'_{(opt)}}{C_1} \quad .$$

where $A_1 = \dfrac{C_2^2}{A^2 C}$

For the case where cost is incurred in collecting auxiliary information on all the units in the population, the cost function can be written as

$$C_0 = C_1 n + C_2 N \quad (3.3)$$

The estimator relevant in this case is $\hat{T}_Q$. The variance of $\hat{T}_Q$ is given by

$$E_\psi\left(\hat{T}_Q - S_y^2\right)^2 = \approx 2\sigma^2 \frac{\beta^2 S_{r(x)}^4}{n\bar{x}_s}\left(\frac{\sigma^2}{n\bar{x}_s} + 2\beta^2\right)$$

It may be seen that the optimal sampling plan in this case is purposive i.e. a sample for which the auxiliary character values are largest.

Further, $$\frac{E_\psi\left(\hat{T}_Q - S_y^2\right)^2}{\sigma^4} \approx \frac{S_{r(x)}^4}{n\bar{x}_s}\left(\frac{1}{n\bar{x}_s} + 2\frac{\beta^2}{\sigma^2}\right)$$

The optimum value of n in this case is given by

$$n_{1(opt)} = \frac{C_0 - C_2 N}{C_1}$$

4. EMPIRICAL STUDY

For the purpose of empirical study, the following data set was considered. This data set, taken from Cochran (1977), pertained to the number of inhabitants (in 1000's) in each of a simple random sample of 48 cities drawn from the population of 196 large cities. This sample based data was taken as population for the purpose of empirical study.

The values of the parameters calculated on the basis of this data are given as

ρ	σ^2	β_2
0.35	10835.75	7.34

Using the above data set, optimum values of sample sizes as well as relative efficiency of the double sampling based predictor vis-à-vis a predictor which utilizes auxiliary information for all the units of the population were worked out. The results of analysis of data are given in Table 1.

Table 1 provides optimum values of sample sizes and relative efficiency of the double sampling based predictor vis-a-vis a predictor which makes use of auxiliary information for all the units of the population under the assumption that the auxiliary information is collected for all the units of the population for different values of cost ratio C_1/C_2 b. A close perusal of Table 1 reveals that the efficiency of double sampling procedure increases as the cost ratio C_1/C_2 increases. Thus, cheaper the cost of enumeration of auxiliary character vis-à-vis the character under study, greater are the efficiency gains of the double sampling based predictor. It may be seen from the table that there are infeasible values in some cases, particularly when $\dfrac{C_1}{C_2}$ is less than 6. This is attributable to poor correlation between 'y' and 'x'.

C_1/C_2	n	n′	$n_{1(opt)}$	RE
3	–	–	51	1.63
5	–	–	30	1.67
6	26	30	25	1.69
8	20	30	19	1.72
10	20	30	19	1.72

(– indicates the infeasible values of n′ i.e. $n' < n_{1(opt)}$)

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques.* (3rd Ed.), Wiley Eastern, New Delhi.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-389.

Sarndal, C.E. (1978). Design based and model-based inference in survey sampling. *Scand. J. Statist.*, **5**, 27-52.

Srivastava, A.K. and Sud, U.C. (1988). Estimating Population Total through Predictive Approach when Auxiliary Characters are Estimated. Wiley Eastern, 133-142.