Aligarh Journal of Statistics
Vol. 26 (2006), 57-64

A GENERALIZED ESTIMATOR OF POPULATION MEAN USING AUXILIARY INFORMATION IN GENERAL SAMPLING DESIGN

Vyas Dubey

ABSTRACT

A modified estimator of population mean using auxiliary information under general sampling design is proposed and its properties are studied under *PPS* sampling. It is seen that the proposed estimator is considerably more efficient than exiting estimators. Numerical illustration has also been included.

1. INTRODUCTION

In sampling strategies use of auxiliary information is known to increase considerably the precision of the estimator. Let $U=(U_1,U_2,\cdots,U_N)$ be a finite population of size N, y and x be study and auxiliary variables defined on U, $\overline{Y}$ and $\overline{X}$ be respectively population means of y and x. Then for estimating population mean $\overline{Y}$ Hansen *et al.* (1953) proposed a difference type estimator

$$\overline{y}_d=\overline{y}+\beta(\overline{X}-\overline{x}) \tag{1.1}$$

$\overline{y}$ and $\overline{x}$ are sample means of y and x based on a sample of size n drawn from U by simple random sampling, β is the population regression coefficient of y on x. Bedi and Hajela (1984), Jain (1987) and Rao (1991) studied a weighted estimator

$$\overline{y}_{w1}=\alpha_1\,\overline{y}+\alpha_2(\overline{X}-\overline{x}) \tag{1.2}$$

for $\alpha_2=\beta\alpha_1$, $\alpha_2=1-\alpha_1$ and $\alpha_1+\alpha_2\neq 1$ respectively. The estimator $\overline{y}_{w1}$ has greater precision over $\overline{y}_d$ for small sample sizes. Dubey and Singh (2001) proposed

$$\overline{y}_{w2}=\alpha_3\,\overline{y}+\alpha_4\,\overline{x}+(1-\alpha_3-\alpha_4)\,\overline{X} \tag{1.3}$$

with the restriction that the sum of coefficients of $\overline{y}$, $\overline{x}$ and $\overline{X}$ in $\overline{y}_d$ is unity. Dubey and Kant (2001) considered a more general estimator

$$\overline{y}_{w3}=\alpha_5\,\overline{y}+\alpha_6\,\overline{x}+(1-\alpha_5-\alpha_6)\,\alpha_7\,\overline{X} \tag{1.4}$$

Here α_i, $i=1,2,\cdots,7$ are constants. For specific choice of α_7, the estimator $\bar{y}_{w3}$ performs better than all the above estimators.

Dubey (2003) further extended the idea for general sampling scheme and studied its properties under *PPS* and stratified sampling. Let $\hat{\bar{Y}}$ and $\hat{\bar{X}}$ be unbiased estimators of $\bar{Y}$ and $\bar{X}$ under any sampling design. The estimator discussed by Dubey (2003) is

$$\hat{\bar{Y}}_w = \alpha_8 \hat{\bar{Y}} + \alpha_9 \hat{\bar{X}} + (1-\alpha_8-\alpha_9)\bar{X} \tag{1.5}$$

In section 2, we propose a more general estimator of $\bar{Y}$ analogous to (1.4) and study its properties.

2. PROPOSED ESTIMATOR AND ITS PROPERTIES

The proposed estimator of $\bar{Y}$ is

$$\hat{\bar{Y}}_g = \lambda_1 \hat{\bar{Y}} + \lambda_2 (\hat{\bar{X}} - \bar{X}) + (1-\lambda_1)\lambda_3 \bar{X} \tag{2.1}$$

where λ_i, $i=1,2,3$ are suitable constants. If $\lambda_3 = 0$, $\hat{\bar{Y}}_g$ reduces to

$$\hat{\bar{Y}}_{g1} = \lambda_1 \hat{\bar{Y}} + \lambda_2 (\hat{\bar{X}} - \bar{X}) \tag{2.2}$$

and for $\lambda_1 = 1$, $\hat{\bar{Y}}_g$ reduces to difference type estimator

$$\hat{\bar{Y}}_D = \hat{\bar{Y}} + \lambda_2 (\hat{\bar{X}} - \bar{X}) \tag{2.3}$$

which was considered by Sarndal *et al*. (1992). The suggested estimator has bias

$$B(\hat{\bar{Y}}_g) = (\lambda_1 - 1)\,(\bar{Y} - \lambda_3 \bar{X}) \tag{2.4}$$

and mean square error

$$M(\hat{\bar{Y}}_g) = \lambda_1^2 V(\hat{\bar{Y}}) + \lambda_2^2 V(\hat{\bar{X}}) + 2\lambda_1 \lambda_2 Cov(\hat{\bar{Y}}, \hat{\bar{X}}) + (\lambda_1 - 1)^2 (\bar{Y} - \lambda_3 \bar{X})^2 \tag{2.5}$$

The optimum values of λ_1 and λ_2, for which $M(\hat{\bar{Y}}_g)$ will be minimized, are given by

$$\lambda_{10} = \frac{1}{1+C_g} \tag{2.6}$$

$$\lambda_{20} = -\beta_g \lambda_{10} \tag{2.7}$$

where $C_g = \frac{V(\hat{\bar{Y}})(1-\rho_g^2)}{(\bar{Y}-\lambda_3\,\bar{X})^2}$, ρ_g and β_g are correlation and regression coefficients between $\hat{\bar{Y}}$ and $\hat{\bar{X}}$. For such choices of constants, $M(\hat{\bar{Y}}_g)$ reduces to

$$M_0(\hat{\bar{Y}}_g) = \frac{V(\hat{\bar{Y}})(1-\rho_g^2)}{(1+C_g)} \tag{2.8}$$

We note that as λ_3 tends to $R = \bar{Y}/\bar{X}$, $M_0(\hat{\bar{Y}}_g)$ tends to zero. In this case λ_{10} and λ_{20} also tend to zero, and the estimator $\hat{\bar{Y}}_g$ tends to $\bar{Y}$. But it is difficult to find the exact value of λ_3 as it depends upon $\bar{Y}$. However, in section 3, some values of λ_3 are suggested for which the proposed estimator records its superiority over existing estimators.

3. EFFICIENCY COMPARISIONS

The minimum *MSE* of $\hat{\bar{Y}}_D$, $\hat{\bar{Y}}_{g1}$ and $\hat{\bar{Y}}_w$ are respectively as follows

$$V_0(\hat{\bar{Y}}_D) = V(\hat{\bar{Y}})(1-\rho_g^2) \tag{3.1}$$

$$M_0(\hat{\bar{Y}}_{g1}) = \frac{V(\hat{\bar{Y}})\,(1-\rho^2)}{1+C_{1g}} \tag{3.2}$$

$$M_0(\hat{\bar{Y}}_w) = \frac{V(\hat{\bar{Y}})(1-\rho_g^2)}{1+C_{2g}} \tag{3.3}$$

where $C_{1g} = \frac{V(\hat{\bar{Y}})(1-\rho_g^2)}{\bar{Y}^2}$, $C_{2g} = \frac{V(\hat{\bar{Y}})(1-\rho_g^2)}{(\bar{Y}-\bar{X})^2}$

Comparing (2.8) with (3.1), (3.2) and (3.3) we have

$$M_0(\hat{\bar{Y}}_g) < V_0(\hat{\bar{Y}}_D) \text{ if } \lambda_3 \geq 0 \tag{3.4}$$

$$M_0(\hat{\bar{Y}}_g) < M_0(\hat{\bar{Y}}_{g1}) \quad \text{if} \quad 0 < \lambda_3 < 2R \tag{3.5}$$

$$M_0(\hat{\bar{Y}}_g) < M_0(\hat{\bar{Y}}_w) \quad \text{if} \quad 1 < \lambda_3 < 2R-1$$

$$\text{or} \quad 2R-1 < \lambda_3 < 1 \tag{3.6}$$

Thompson (1968) used guessed value of $\bar{Y}$ for estimating $\bar{Y}$. Let $\bar{Y}_{\min}$ be minimum value of $\bar{Y}$ which may be guessed from past data or repeated surveys. Let

$$R_{(1)} = \frac{\bar{Y}_{\min}}{\bar{X}}$$

Then replacing R by $R_{(1)}$ in (3.5) and (3.6), a value of λ_3 may easily be obtained for which the proposed estimator is more efficient than others.

4. CHOICE OF λ_{10} AND λ_{20}

Let $\hat{\beta}_g$ be a sample estimate of β_g. Sarndal *et al.* (1992) proposed regression type estimator under any sampling design as

$$\hat{\bar{Y}}_{glr} = \hat{\bar{Y}} + \hat{\beta}_g \, (\bar{X} - \hat{\bar{X}}) \tag{4.1}$$

which has MSE

$$M\,(\hat{\bar{Y}}_{glr}) = V\,(\hat{\bar{Y}})\,(1 - \rho_g^2) \tag{4.2}$$

upto the first order of approximation. Let $\hat{V}_g$ be a consistent estimate of $M\,(\hat{\bar{Y}}_{glr})$. Noting that

$$E\,(\hat{\bar{Y}}_{glr} - \lambda_3\,\bar{X})^2 = M\,(\hat{\bar{Y}}_{glr}) + (\bar{Y} - \lambda_3\,\bar{X})^2 \tag{4.3}$$

the estimates of λ_{10} and λ_{20} are given by

$$\hat{\lambda}_{10} = 1 - \frac{\hat{V}_g}{(\hat{\bar{Y}}_{glr} - \lambda_3\,\bar{X})^2} \tag{4.4}$$

$$\hat{\lambda}_{20} = -\hat{\beta}_g\,\hat{\lambda}_{10} \tag{4.5}$$

Therefore another estimator of $\bar{Y}$ is given by

$$\hat{\bar{Y}}_{glr}^{*} = \hat{\bar{Y}}_{glr} - \frac{\hat{V}_g}{(\hat{\bar{Y}}_{glr} - \lambda_3\,\bar{X})} \tag{4.6}$$

where λ_3 is taken such that the estimator $\hat{\bar{Y}}_{glr}^{*}$ exists. Let

$$\hat{V}_g = V_n + e_1$$

$$(\hat{\hat{Y}}_{gl}r-\lambda_3\,\overline{X})^2 = E(\hat{\hat{Y}}_{glr}-\lambda_3\,\overline{X})^2+e_2$$

where $V_n = V_g + O(n^{-2})$, $E(e_1)=0=E(e_2)$. Let $|e_2|<1$ so that terms with second and higher powers of e_2 may be neglected. We find that the estimator $\hat{\hat{Y}}^*_{glr}$ is as efficient as $\hat{\hat{Y}}_g$ upto first order of approximation.

5. SPECIAL CASE: *PPSWR* PROCEDURE

Whenever the units are taken with probability proportional to size with replacement ($PPSWR$) procedure, we have

$$\hat{\hat{Y}} = \bar{y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{N\,p_i}$$

$$\hat{\hat{X}} = \bar{x}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{N\,p_i}$$

where p_i is the probability of selecting $i-$th unit from the population

For $\lambda_1 = 1$ and $\lambda_2 = -\beta_{pps}$, $\hat{\hat{Y}}_g$ reduces to Tripathi (1969) type estimator

$$\bar{y}_{D\,(pps)} = \bar{y}_{pps} + \beta_{pps}\,(\overline{X}-\bar{x}_{pps}) \tag{5.1}$$

where

$$\beta_{pps} = \frac{\sigma^*_{yx}}{\sigma^{*2}_x}\ ,\ \sigma^{*2}_y = \sum_{i=1}^{N}(\frac{y_i}{N\,p_i}-\overline{Y})^2\,p_i$$

$$\sigma^{*2}_x = \sum_{i=1}^{N}(\frac{x_i}{N\,p_i}-\overline{X})^2\,p_i$$

$$\sigma^*_{yx} = \sum_{i=1}^{N}(\frac{y_i}{N\,p_i}-\overline{Y})(\frac{x_i}{N\,p_i}-\overline{X})$$

Furthermore $\hat{\hat{Y}}_g, \hat{\hat{Y}}_{g1}, \hat{\hat{Y}}_w$ reduce to

$$\bar{y}_{g\,(pps)} = \lambda_1\,\bar{y}_{pps} + \lambda_2\,(\bar{x}_{pps}-\overline{X}) + (1-\lambda_1)\,\lambda_3\,\overline{X} \tag{5.2}$$

$$\bar{y}_{g1(pps)} = \lambda_1\,\bar{y}_{pps} + \lambda_2\,(\bar{x}_{pps}-\overline{X}) \tag{5.3}$$

$$\bar{y}_{w(pps)} = \alpha_1\,\bar{y}_{pps} + \alpha_2\,\bar{x}_{pps} + (1-\alpha_1-\alpha_2)\,\overline{X} \tag{5.4}$$

Also (2.8), (3.1), (3.2) and (3.3) reduce to

$$M_0(\bar{y}_{g(pps)}) = \frac{\sigma_y^{*2}(1-\rho_{pps}^2)}{n + C_{1pps}} \tag{5.5}$$

$$V(\bar{y}_{D(pps)}) = \frac{\sigma_y^{*2}(1-\rho_{pps}^2)}{n} \tag{5.6}$$

$$M_0(\bar{y}_{g1(pps)}) = \frac{\sigma_y^{*2}(1-\rho_{pps}^2)}{n + C_{2pps}} \tag{5.7}$$

$$M_0(\bar{y}_{w(pps)}) = \frac{\sigma_y^{*2}(1-\rho_{pps}^2)}{n + C_{3pps}} \tag{5.8}$$

with $C_{1pps} = \dfrac{\sigma_y^{*2}(1-\rho_{pps}^2)}{(\bar{Y} - \lambda_3 \bar{X})^2}$, $C_{2pps} = \dfrac{\sigma_y^{*2}(1-\rho_{pps}^2)}{\bar{Y}^2}$,

$$C_{3pps} = \frac{\sigma_y^{*2}(1-\rho_{pps}^2)}{(\bar{Y} - \bar{X})^2}, \quad \rho_{pps} = \frac{\sigma_{yx}^*}{\sigma_y^* \sigma_x^*}.$$

Again $\hat{\bar{Y}}_{glr}^*$ is obtained as

$$\bar{y}_{lr(pps)}^* = \bar{y}_{lr(pps)} - \frac{s_y^{*2}(1-r^{*2})}{n(\bar{y}_{lr(pps)} - \lambda_3 \bar{X})} \tag{5.9}$$

where

$$\bar{y}_{lr(pps)} = \bar{y}_{pps} + \hat{\beta}_{pps}(\bar{X} - \bar{x}_{pps})$$

is regression estimator in *PPS* sampling considered by Tripathi (1969). And

$$s_y^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{N p_i} - \bar{y}_{pps}\right)^2, \quad s_x^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i}{N p_i} - \bar{x}_{pps}\right)^2,$$

$$s_{yx}^* = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i}{N p_i} - \bar{y}_{pps}\right)\left(\frac{x_i}{N p_i} - \bar{x}_{pps}\right), \quad r^* = \frac{s_{yx}^*}{s_y^* s_x^*}, \quad \hat{\beta}_p = \frac{s_{yx}^*}{s_x^{*2}}$$

The proposed estimator $\bar{y}_{lr(pps)}^*$ is more efficient than $\bar{y}_{lr(pps)}$ for any finite value of λ_3 such that $\bar{y}_{lr(pps)}^*$ exists.

6. NUMERICAL EXAMPLE

Let us consider the data from Sarndal *et al.* (1992) which relate to Gross National Product (z), import (y) and export (x) of 44 countries. Calculating probabilities (p_i) using variable z, we have

$$\overline{Y} = 485.23, \quad \overline{X} = 725.86, \quad \sigma_y^{*2} = 196372.94, \quad \sigma_x^{*2} = 263411.29, \quad \sigma_{yx}^{*} = 137460.80, \quad \rho_{pps} = 0.6044$$

Relative efficiencies (RE) of $\overline{y}_{g(pps)}$ with respect to (*w.r.t.*) $\overline{y}_{pps}$ defined by $V(\overline{y}_{pps})/V(\overline{y}_{g(pps)})$ for various values of λ_3 are given in the table1 below:

Table 1: Relative efficiencies

λ_3	RE $n=10$	RE $n=20$	RE $n=30$
0.0	1.66	1.62	1.60
0.1	1.69	1.63	1.61
0.2	1.75	1.66	1.63
0.3	1.85	1.71	1.67
0.4	2.09	1.83	1.75
0.5	2.89	2.23	2.01
0.6	9.52	5.55	4.22
0.7	39.12	20.35	14.09
0.8	3.73	2.65	2.29
0.9	2.27	1.92	1.81
1.0	1.91	1.75	1.69
1.1	1.78	1.68	1.64
1.2	1.70	1.64	1.61
1.3	1.67	1.62	1.60
1.4	1.65	1.61	1.59

RE of $\overline{y}_{D(pps)}$ w. r. t. $\overline{y}_{pps}$ is 1.58 for all sample sizes. Again, relative efficiencies of $\overline{y}_{g1(pps)}$ and $\overline{y}_{w(pps)}$ w. r. t. $\overline{y}_{pps}$ are given in table 1 for $\lambda_3 = 0$ and $\lambda_3 = 1$ respectively.

From table 1 it is clear that proposed estimator $\overline{y}_{w(pps)}$ is more efficient than $\overline{y}_{lr(pps)}$ if $0 < \lambda_3 < 1.4$. It has maximum efficiency for $\lambda_3 = 0.7$. It may further be noted that $\bar{y}_{w(pps)}$ has superiority over all the estimators for $0.3 < \lambda_3 < 1.0$.

Acknowledgement

Author is thankful to the referee for his suggestions leading to improvement of the paper.

REFERENCES

Bedi, P.K. and Hajela, D. (1984): An estimator for mean utilizing known coefficient of variation and auxiliary variable. *J. Statist. Res.*, **18**, 29-33.

Dubey, V. (2003): *On estimating population mean using auxiliary information.* Proceedings of the International Conference on 'Recent Statistical Techniques in Life Testing, Reliability, Sampling Theory and Quality Control, Dec. 29-31, 2003, BHU, Varanasi, India, 183-193.

Dubey, V. and Kant, S. (2001): *On improving regression estimator for estimating population mean*. National Conference on 'Mathematical and Applied Statistics, Nov. 5-7, 2001, Nagpur University, Nagpur, India.

Dubey, V. and Singh, S.K. (2001): An improved regression estimator for estimating population mean. *J. Indian Soc. Agricultural Statist.*, **52**, 179-183.

Hansen, M. H. and Hurwitz, W. N. and Madow, W. G. (1953): *Sample Survey Methods and Theory.* Vol. II, John Wiley, New York.

Jain, R.K. (1987): Properties of estimators in simple random sampling using auxiliary variable. *Metron*, **XLV**, 265-271.

Rao, T.J. (1991): On certain methods of improving ratio and regression estimators. *Comm. Statist. Theory Methods*, **20**, 3325-3340.

Sarndal, C.E., Swensson, B. and Wretman, J.H. (1992): *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Tripathi, T.P. (1969): A regression-type estimator in sampling with *PPS* and with replacement. *Aust. J. Statist.,* **11**, 140-148.

Thompson, J.R. (1968): Some Shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc.*, **63**, 113-123.

Received : 21-08-2004

Revised : 31-01-2006

School of Studies in Statistics
Pt. Ravishankar Shukla University
Raipur-492 010, India
e-mail: dubey_vyas@rediffmail.com