Aligarh Journal of Statistics
Vol. 29 (2009), 39-50

ALTERNATIVE ESTIMATORS USING AUXILIARY INFORMATION IN SAMPLE SURVEYS

T.J. Rao

ABSTRACT

In many large scale sample surveys, usually, it is of interest to estimate parameters relating to several characteristics. Some of the study variables may be poorly correlated with the selection probabilities used, when the selection of units is done by a probability proportional to size (pps) sampling method. When such a situation of poor correlation exists, Amahia, Chaubey and Rao (ACR, 1989) proposed alternative estimators for the population total Y of the study variable y following Rao (1966) and Bansal and Singh (1985). It is established therein that the bias of the ACR estimator is smaller than the biases of the alternative estimators suggested. ACR also showed that their estimator has smaller variance compared to the conventional estimator under a usual super population model.

In this paper, our parameter of interest is mainly B, the bias of ACR estimator and our objective is to first discuss alternative estimators for estimating this bias when sampling is done with probability of selection of units proportional to the size measure. Next we shall deal with similar other situations where alternatives are to be sought, when the design used is the often adopted conventional Simple Random Sampling (SRS) design.

1. INTRODUCTION

It is shown in ACR (1989) that when y_i and p_i are poorly correlated one can consider a general estimator

$$\hat{Y}_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i^*}, \tag{1.1}$$

where

$$p_i^* = \frac{1-\rho}{N} + \rho\, p_i\,, \quad \rho = \operatorname{corr}(y, p)\,.$$

When $\rho = 0$, (1.1) reduces to the estimator

$$\hat{Y}_3 = \frac{N}{n}\sum_{i=1}^{n} y_i \tag{1.2}$$

considered by Rao (1966)

and when $\rho = 1$, this becomes the conventional design unbiased estimator

$$\hat{Y}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}. \tag{1.3}$$

Motivated by the suggestion that

$$\hat{Y}_4 = (1-\rho)\hat{Y}_3 + \rho\hat{Y}_1$$

will have a design bias smaller than $\hat{Y}_1$, ACR also considered $\hat{Y}_4$ rewritten as

$$\hat{Y}_4 = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p'_i}, \tag{1.4}$$

where

$$p'_i = \left[(1-\rho)N + \frac{\rho}{p_1}\right]^{-1}.$$

Notice that while p_i^* is the weighted arithmetic mean of $\frac{1}{N}$ and p_i, p'_i is nothing but the weighted harmonic mean of the same quantities.

ACR proved

Theorem 1.1: $B(\hat{Y}_3) > B(\hat{Y}_4) > B(\hat{Y}_2)$.

Furthermore, under a simple super population model assumed by Rao (1966), where

$$y_i = \mu + e_i, i + 1, 2, \ldots, N$$

$$\in(e_i \mid p_i) = 0, \quad \in(e_i^2 \mid p_i) = a, \quad a > 0$$

and

$$\in(e_i e_j \mid p_i, p_j) = 0, \tag{1.5}$$

where $\in$ denotes the average over all finite populations that can be drawn from the super population, ACR's and Rao's results show

Theorem 1.2: Both $\hat{Y}_2$ (and $\hat{Y}_4$) and $\hat{Y}_3$ have smaller expected variance than $\hat{Y}_1$.

Remark 1.1: No definite comparison between $\hat{Y}_3$ and $\hat{Y}_2$ (or $\hat{Y}_4$) is possible due to the presence of unknown super population parameters in the variance expressions.

Also, ACR considered the general super population model where

$$y_i = \beta\, p_i + e_i\,,\ \ i = 1, 2, \ldots, N$$

$$\in(e_i \mid p_i) = 0\,, \quad \in(e_i^2 \mid p_i) = a p_i^g\,,$$

and

$$\in(e_i e_j \mid p_i, p_j) = 0\,, \qquad a > 0\,,\ \ g \geq 0 \tag{1.6}$$

and obtained conditions under which $\hat{Y}_1$ is superior to $\hat{Y}_2$.

This happens when the parameter g is close to 2 where as in practice g can take any value in $[0, 2]$. This was also illustrated through numerical examples in ACR. Having noted that the proposed estimator $\hat{Y}_2$ is better than the conventional estimator $\hat{Y}_1$ in most of the situations and further that Bias of $\hat{Y}_2$ is the smallest, we now turn our attention to the parameter $B = B(\hat{Y}_2)$ in this paper and obtain some interesting results.

2. ESTIMATORS OF BIAS

It is shown in ACR that

$$B = B(\hat{Y}_2) = \sum_{i=1}^{N} Y_i \left(\frac{p_i}{p_i^*} - 1 \right) = \sum_{i=1}^{N} Z_i \text{ , say.}$$

If one is interested in obtaining the estimated MSE of $\hat{Y}_2$, it may be easier to derive the estimated variance and add to it the estimated squared bias. Alternatively if one is interested in design unbiasedness, one might make a bias-correction for $\hat{Y}_2$ by subtracting the estimated bias from $\hat{Y}_2$ to get (almost) unbiased estimates. Motivated by this, we shall look for alternative estimators for the Bias B. For the $PPSWR$ selection, as for the case of estimation of population total Y, we have estimators for bias given by

$$\hat{B}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{p_i} \tag{2.1}$$

$$\hat{B}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{p_i^*} \tag{2.2}$$

$$\hat{B}_3 = \frac{N}{n} \sum_{i=1}^{n} z_i \tag{2.3}$$

and

$$\hat{B}_4 = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{p_i'}\,. \tag{2.4}$$

Denoting the bias of $\hat{B}_i$ by γ_i, it is easy to see that

$$\gamma_4 = (1-\rho)\gamma_3 \text{ giving}$$

$$|\gamma_4| < |\gamma_3|.$$

Also

$$\gamma_3 - \gamma_2 = \frac{\rho(1-\rho)}{N}\sum_{i=1}^{N}\frac{Y_i\,(N\,P_i - 1)^2\,p_i}{p_i^*} > 0$$

which implies that $\gamma_3 > \gamma_2$ and

$$\gamma_4 - \gamma_2 = \frac{\rho(1-\rho)^2}{N}\sum_{i=1}^{N}\frac{Y_i\,(N\,P_i - 1)^3}{p_i^{*2}}$$

which may be positive or negative. However, notice that $B(\hat{Y}_4) > B(\hat{Y}_2)$, while $\gamma_4 - \gamma_2$ may be of any sign.

Remark 2.1: Following (1.4), one can consider another estimator

$$\hat{Y}_5 = (1-\rho)\hat{Y}_2 + \rho\hat{Y}_1$$

which has even still smaller bias than $\hat{Y}_2$. Next it is possible to extend this idea by constructing

$$\hat{Y}_6 = (1-\rho)\hat{Y}_5 + \rho\hat{Y}_1$$

$$= (1-\rho)^2\,\hat{Y}_5 + [1-(1-\rho)^2]\hat{Y}_1$$

which has $B(\hat{Y}_6) = (1-\rho)\,B(\hat{Y}_5) = (1-\rho)^2\,B(\hat{Y}_2)$.

After n steps, we have $\hat{Y}_n = (1-\rho)^n\,\hat{Y}_2 + [1-(1-\rho)^n]\hat{Y}_1$

which tends to the unbiased conventional estimator $\hat{Y}_1$. But the efficiencies will not improve.

Remark 2.2: Following exactly the same steps as for comparing the estimates, one can compare the variances of these estimated biases but the algebra is lengthy and the efficiency comparisons depend on unknown super population parameters.

3. ESTIMATORS OF POPULATION TOTAL UNDER SIMPLE RANDOM SAMPLING

In all the discussions so far, $PPSWR$ design was considered for the estimation of population total and when it is thought that the characteristic under study y and the probabilities of selection p are poorly correlated, several alternative estimators were suggested in the literature.

ACR estimator and related estimators following this were discussed in a series of papers by Bansal and Singh (1990), Mangat and Singh (1993), Rao (1987), Kumar and Agarwal (1997), Singh and Horn (1998), Chaubey and Tripathi (2004) and Arnab (2004) among others. All these results relate to the PPS case only. However, it is quite possible that a practical situation of the following type could occur.

Consider a situation where the design is the conventional simple random sampling with replacement $(SRSWR)$ design. Suppose first that our parameter of interest is the population total Y. We estimate this by the conventional unbiased estimator

$$\hat{Y}_1' = \frac{N}{n}\sum_{i=1}^{n} y_i \,. \tag{3.1}$$

After drawing the sample, if we have information that the population units are of highly varying sizes, one of the possible alternatives could be the use of the method of stratification after the selection of the sample. On the other hand, realizing that a PPS scheme would have been more appropriate, following Rao (1966) one can consider the biased estimator

$$\hat{Y}_3' = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i} \,, \tag{3.2}$$

where p_i's are available normed sizes. Also following ACR, we propose more general estimators

$$\hat{Y}_2' = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i^*} \tag{3.3}$$

$$\hat{Y}_4' = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{p_i'} \,. \tag{3.4}$$

We next have

$$B(\hat{Y}_3') = \sum_{i=1}^{N} Y_i \left(\frac{1}{Np_i} - 1 \right)$$

$$B(\hat{Y}_2') = \sum_{i=1}^{N} Y_i \left(\frac{1}{Np_i^*} - 1 \right)$$

and

$$B(\hat{Y}_2') - B(\hat{Y}_3') = \frac{1}{N}\sum_{i=1}^{N} Y_i \left(\frac{1}{p_i^*} - \frac{1}{p_i} \right)$$

$$= (1-\rho)\, N \,\mathrm{Cov}\left(p_i\,, \frac{Y_i}{p_i\,[(1-\rho) + N\rho\, p_i\,]} \right)$$

$$< 0\,,\ \text{since } \mathrm{Cov}\,(y,\, p) > 0\,.$$

Thus

$B(\hat{Y}_2') < B(\hat{Y}_3')$ as for the PPS case.

Furthermore, since

$\hat{Y}_4' = (1-\rho)\hat{Y}_1' + \rho\hat{Y}_3'$ with $0 < \rho < 1$,

$B(\hat{Y}_4') = \rho\, B(\hat{Y}_3') < B(\hat{Y}_3')\,.$

Also

$$B(\hat{Y}_4') - B(\hat{Y}_2') = \frac{1}{N}\sum_{i=1}^{N} Y_i \left(\frac{1}{p_i'} - \frac{1}{p_i^*} \right) > 0\,.$$

Thus we have

Theorem 3.1: $B(\hat{Y}_2') < B(\hat{Y}_4') < B(\hat{Y}_3')\,.$

Note that the order is the same as in Theorem 1.1 as can be expected.

Next we consider the general super population model (1.6) and find the Expectations $\in_1$, $\in_3$ and $\in_2$ under the model, of the expressions:

$$\frac{n}{N} V\,(\hat{Y}_1') = \sum_{i=1}^{N} Y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} Y_i \right)^2$$

$$\frac{n}{N} V\,(\hat{Y}_3') = \sum_{i=1}^{N} \frac{Y_i^2}{N^2 p_i^2} - \frac{1}{N}\left(\sum_{i=1}^{N} \frac{Y_i}{Np_i} \right)^2$$

and

$$\frac{n}{N} V\,(\hat{Y}_2') = \sum_{i=1}^{N} \frac{Y_i^2}{N^2 p_i^{*2}} - \frac{1}{N}\left(\sum_{i=1}^{N} \frac{Y_i}{Np_i^*} \right)^2 \quad \text{respectively.}$$

We then have

$$\in_1 = \sum_{i=1}^{N} (ap_i^g + \beta^2 p_i^2) - \frac{1}{N}\left\{ a\sum_{i=1}^{N} p_i^g + \beta^2 \left(\sum_{i=1}^{N} p_i \right)^2 \right\}$$

$$= a\left(\frac{N-1}{N}\right)\sum_{i=1}^{N} p_i^g + \beta^2 N \operatorname{Var}(p_i)$$

$$\in_2 = \sum_{i=1}^{N}\frac{(ap_i^g + \beta^2 p_i^2)}{N^2 p_i^{*2}} - \frac{1}{N}\left\{\sum_{i=1}^{N}\frac{ap_i^g}{N^2 p_i^{*2}} + \beta^2\left(\sum_{i=1}^{N}\frac{p_i}{Np_i^*}\right)^2\right\}$$

$$= \frac{a}{N^2}\left(\frac{N-1}{N}\right)\sum_{i=1}^{N}\frac{p_i^g}{p_i^{*2}} + \beta^2 N \operatorname{Var}\left(\frac{p_i}{Np_i^*}\right).$$

Therefore,

$$\in_1 - \in_2 = aC + \beta^2 D,$$

where

$$C = \frac{N-1}{N}\sum_{i=1}^{N} p_i^g\left(1 - \frac{1}{N^2 p_i^{*2}}\right)$$

and

$$D = N\operatorname{Var}(p_i) - N\operatorname{Var}\left(\frac{p_i}{Np_i^*}\right).$$

It is easy to verify that $D > 0$ (cf. ACR (1989)). Also

$$\frac{NC}{N-1} = \sum_{i=1}^{N} b_i c_i\,,$$

where

$$b_i = \frac{p_i^g (N\,p_i^* + 1)}{N^2 p_i^{*2}} \quad \text{and} \quad c_i = (N\,p_i^* - 1)\,.$$

Here $\sum_{i=1}^{N} c_i = 0$ and c_i increases with p_i.

Following Royall’s (1970) result, it can be established that $\sum_{i=1}^{N} b_i c_i > 0$ if b_i increases with p_i. A sufficient condition for this is $g > 2-\eta$, where $\eta \ge 0$. Thus we have, in most of the practical situations.

Theorem 3.2: The suggested estimator $\hat{Y}_2'$ is better than the conventional estimator $\hat{Y}_1'$.

Remark 3.1: It is easy to see that

$$\in_1 - \in_3 = \frac{a(N-1)}{N}\sum_{i=1}^{N} p_i^{g-2}\left(p_i^2 - \frac{1}{N^2}\right) + \beta^2 N\, Var(p_i)\,.$$

Here the first term is also positive if $g > 1 + \dfrac{1}{N\,p_i + 1}$ which is often the case.

This shows that $\hat{Y}_3'$ is also better than $\hat{Y}_1'$.

Remark 3.2: However, a definite choice between $\hat{Y}_2'$ and $\hat{Y}_3'$ is not possible. A comparison between $\in_2$ and $\in_3$ leads to the expression

$$\in_2 - \in_3 = \left(\frac{N-1}{N}\right)\frac{a}{N^2}\left[\sum_{i=1}^{N} p_i^g\left(\frac{1}{p_i^{*2}} - \frac{1}{p_i^2}\right)\right] + \text{a positive quantity}$$

and it can be derived that a sufficient condition for the first term to be positive is $g > 2 + \eta$, $\eta > 0$. However since in most of the practical situations g lies between 1 and 2, this cannot happen and thus we can expect that the first term could be negative and large enough to offset the second term and then $\in_2$ might be smaller than $\in_3$.

Next, we observe that for the SRS situation discussed above one could consider a simple ratio or a regression estimator. For instance, for the regression estimator

$$\hat{Y}_r' = N[\bar{\mathrm{y}} + b(\bar{\mathrm{X}} - \bar{x})]\,,$$

we have large sample variance given by

$$V(\hat{Y}_r') = (1-\rho^2)V(\hat{Y}_1'),\ \ 0 \le \rho \le 1 \text{ and under the model}$$

$$\in_r = \in V(\hat{Y}_r') = (1-\rho^2)\in_1.$$

Obviously, $\in_r < \in_1$.

In order to check how $\hat{Y}_r'$ compares with $\hat{Y}_3'$, let us consider

$$\in_r - \in_3 = a\left(\frac{N-1}{N}\right)\sum_{i=1}^{N} p_i^{g-2}\left((1-\rho^2)\,p_i^2 - \frac{1}{N^2}\right) + (1-\rho^2)\beta^2 N\, Var(p_i)\,.$$

For $\rho = 1$, $\in_r < \in_3$ as can be expected. Since the second term is always positive, we shall look at the first summation, say $f(g)$.

For $\rho = 0$,

$$f(g) = \sum_{i=1}^{N} p_i^g\left(1 + \frac{1}{Np_i}\right)\left(1 - \frac{1}{N\,p_i}\right)$$

$$=\sum_{i=1}^{N} b_i c_i \text{ , say}$$

where $c_i = Np_i - 1$ and $b_i = \dfrac{p_i^g (N\,p_i + 1)}{N^2 p_i^2}$,

$\sum c_i = 0$ and c_i increases with p_i .

By Royall's lemma, $\sum b_i c_i > 0$ if b_i also increases with p_i . To verify this we find that

$$f'(g) > 0 \text{ if } g > 1 + \frac{1}{N\,p_i + 1} .$$

This sufficient condition holds because in practice g lies between 1 and 2.

Thus for $\rho = 0$, in practice $\in_r > \in_3$ establishing the superiority of $\hat{Y}_3'$ over the regression estimator.

Combining the above results we can conclude that if ρ is not too large, then $\hat{Y}_3'$ dominates the regression estimator but for large values of ρ, the regression estimator $\hat{Y}_r'$ is better.

4. NUMERICAL ILLUSTRATIONS

We shall now consider the efficiency of various estimators considered in the previous section for 3 populations described below:

Population I: $n = 10$, $\rho = 0.488$

x	25	32	14	70	24	20	32	44	50	44
y	11	7	5	27	30	6	13	9	14	18

Population II: $n = 9$, $\rho = 0.9439$

x	1375	2065	1565	1363	1530	1328	1521	1474	1328
y	1780.294	2617.189	1776.566	1639.862	1859.913	1732.88	1811.252	1941.228	1744.531

Population III: $n = 12$, $\rho = 0.05$

x	41	34	54	39	49	45	41	33	37	41	47	39
y	36	47	41	47	49	45	32	37	40	41	37	48

Table 1: Expected variances of different estimators under a general super population model

g	$\in_1$	$\in_2$	$\in_3$	$\in_r$
Pop-I				
2	$0.1080a + 0.1995\beta^2$	$0.0862a + 0.005\beta^2$	$0.09a$	$0.0823a + 0.152\beta^2$
1.75	$0.1807a + 0.1995\beta^2$	$0.1503a + 0.005\beta^2$	$0.1652a$	$0.1377a + 0.152\beta^2$
1.5	$0.3054a + 0.1995\beta^2$	$0.265a + 0.005\beta^2$	$0.3071a$	$0.2327a + 0.152\beta^2$
Pop-II				
2	$0.1008a + 0.0023\beta^2$	$0.0986a + 0.000006\beta^2$	$0.0988a$	$0.0110a + 0.0003\beta^2$
1.75	$0.1733a + 0.0023\beta^2$	$0.0172a + 0.000006\beta^2$	$0.1715a$	$0.0189a + 0.0003\beta^2$
1.5	$0.2985a + 0.0023\beta^2$	$0.2975a + 0.000006\beta^2$	$0.2982a$	$0.0326a + 0.0003\beta^2$
Pop-III				
2	$0.0779a + 0.0017\beta^2$	$0.0776a + 0.0015\beta^2$	$0.0764a$	$0.0778a + 0.0017\beta^2$
1.75	$0.1440a + 0.0017\beta^2$	$0.1435a + 0.0015\beta^2$	$0.1426a$	$0.1436a + 0.0017\beta^2$
1.5	$0.2666a + 0.0017\beta^2$	$0.2658a + 0.0015\beta^2$	$0.2666a$	$0.2659a + 0.0017\beta^2$

It is seen from the above table that the proposed estimator $\hat{Y}_2'$, even though biased, is better than the conventional unbiased estimator $\hat{Y}_1'$ under the general super population model for various values of the parameter g which usually occur in practice. It is also noted that the alternative estimator $\hat{Y}_3'$ is also better than $\hat{Y}_1'$ and a choice between $\hat{Y}_2'$ and $\hat{Y}_3'$ depends on the model parameters a and β. It is also noted that the regression estimator does not fare well compared to $\hat{Y}_3'$ $(or\ \hat{Y}_2')$ for populations with small values of ρ.

5. SUMMARY AND CONCLUSIONS

Usually many large scale sample surveys are multi-subject enquiries and it is of interest to estimate parameters relating to several characteristics. When probability proportional to size (pps) sampling is used for selection of units, some of the study variables may be poorly correlated with the selection probabilities used for pps selection. Rao (1966) and Bansal and Singh (1985) suggested alternative biased estimators for the estimation of the population total Y of the study variate y. In Amahia, *et al.* (ACR, 1989), it is shown that the bias of the estimator suggested therein for Y has smaller bias than the biases of the Rao's, Bansal and Singh's and other alternative estimators. ACR also established that their estimator has smaller variance compared to the conventional estimator under a super population model.

In this paper, our parameter of interest is mainly B, the bias of ACR estimator. If one is interested in the estimated MSE of the estimator, it may be easier to derive the estimated variance and add to it the estimated squared bias. Furthermore, if design unbiasedness is demanded, one might make a bias correction for the estimator to get an (almost) unbiased estimator. Motivated by this, we first discussed alternative estimators for estimating this bias, when sampling is done by a pps method and compared these.

In all discussions so far, $ppswr$ design was considered for estimating the population total when it is thought that the characteristic under study y and the selection probabilities are poorly correlated. However, a simple practical situation may arise as follows, when the design used is the conventional Simple Random Sampling (SRS) design. Here, after selecting the sample by SRS, information may be available that the units are of varying sizes and a PPS design would have been more appropriate. Again, following the work of Rao and ACR, discussed in the earlier sections of the paper, we suggested certain alternative estimators and compared them. Towards the end, we have illustrated the results by numerical examples.

Acknowledgement

The author wishes to thank a referee for the useful suggestions which improved the presentation.

REFERENCES

Amahia, G.N., Chaubey, Y.P. and Rao, T.J. (1989): Efficiency of a new estimator in PPS sampling for multiple characteristics. *J. Statist. Plann. Inference*, **21**, 75-84.

Arnab, R. (2004): Optimum estimation of a finite population total in PPS sampling with replacement for multi-character surveys. *J. Indian Soc. Agricultural Statist*., **58**, 2, 231-243.

Bansal, M.L. and Singh, R. (1985): An alternative estimator for multiple characteristics in PPS sampling. *J. Statist. Plann. Inference*, **11**, 313-320.

Bansal, M.L. and Singh, R. (1989): An alternative estimator for multiple characteristics corresponding to Horvitz and Thompson estimator in probability proportional to size and without replacement sampling. *Statistica*, **XLIX**, 447-452.

Bansal, M.L. and Singh, R. (1990): An alternative estimator for multiple characteristics in RHC sampling scheme. *Comm. Statist. Theory Methods*, **19**, 1777-1784.

Kumar, P. and Agarwal, S.K. (1997): Alternative estimators for the population totals in multiple characteristic surveys. *Comm. Statist. Theory Methods*, **26**, 2527-2537.

Mangat, N.S. and Singh, R. (1992-93): Sampling with varying probabilities without replacement - a review. *Aligarh J. Statist*., **12 & 13**, 75-105.

Rao, J.N.K. (1966): Alternative estimators in *PPS* sampling for multiple characteristics. *Sankhy* $\bar{a}$ *, Ser. A*, **28**, 47-60.

Rao, T.J. (1987): On certain alternative estimators for multiple characteristics in varying probability sampling. *ISI Technical Report* No. 21/87.

Royall, R.M. (1970): On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.

Singh, S. and Horn, H. (1998): An alternative estimator for multi-character surveys. *Metrika*, **48**, 99-107.

Tripathi, T.P. and Chaubey, Y.P. (2004): Optimum *PPS* sampling design based on multivariate information for estimation of several totals. *Aligarh J. Statist*., **24**, 85-94.

Received : 18-01-2008
Revised : 01-09-2009

C.R. Rao AIMSCS, UoH Campus,
Hyderbad-500 046, India.
e-mail: tjrao@hotmail.com