Available online at www.isas.org.in
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 63(2) 2009 133-137

A Note on Alternative Estimators for Multi-Character Surveys

Raghunath Arnab[1*], Sarjinder Singh[2] and P.A.E. Serumaga-Zake[3]
[1]*University of Botswana, Botswana*
[2]*The University of Texas at Brownsville and Southmost.*
[3]*Department of Statistics, North-West University, Mafikeng Campus*

(Received: June 2008, Revised: March 2009, Accepted: April 2009)

SUMMARY

The problems of estimating the population total in multi-character surveys in varying probability sampling schemes when the measure of size is not well-related to the study variables, have been considered by Rao (1966), Scott and Smith (1969) and Arnab (2001). In the present note, their results are extended for a wider class of superpopulation models and sampling designs.

Keywords: Auxiliary information, Model design unbiased estimator, Multi-character surveys, Optimal estimator, PPSWR sampling, Superpopulation model.

1. INTRODUCTION

In large-scale surveys, we generally estimate population parameters like totals, means and variances for more than one character at a time. In such a survey if a sample is selected by a varying probability sampling scheme using an auxiliary variable x as a measure of size, then the resulting sampling design may yield efficient estimators for those characters which are well-related to the auxiliary variable but may not provide efficient estimators for the characters which are poorly related to the auxiliary variable. Rao (1966) first addressed the requirement for the adjustments of the conventional estimators in such a multicharacter survey and provided with some alternative estimators for estimation of a finite population total under various sampling schemes when the correlation between the study and auxiliary variable is very low. The alternative estimators, proposed by Rao (1966), fare better than the conventional estimators under the following superpopulation model:

Model $M1$: $E_{M1}(y_i) = \mu$, $V_{M1}(y_i) = \sigma^2$ and $C_{M1}(y_i, y_j) = 0$ for $i \neq j$ (1)

where, μ, $\sigma^2 (> 0)$ are unknown model parameters and E_{M1}, V_{M1} and C_{M1} denote respectively the expectation, variance and covariance with respect to the model $M1$. Following Rao (1966), Scott and Smith (1969), Bansal and Singh (1985), Kumar and Agarwal (1997), Mangat and Singh (1992-93) and Singh and Horn (1998), among others also suggested some alternative estimators under the PPSWR sampling scheme. Arnab (2001) extended Rao's (1966) results for an arbitrary varying probability sampling scheme and showed that Rao's (1966) results could be derived from his results as special cases. For the sake of clarity, let us describe Rao (1966), Arnab (2001) and Scott and Smith (1969) results relevant to our present discussion as follows.

1.1 Estimators due to Rao (1966), Arnab (2001) and Scott and Smith (1969)

Let $U = \{1, ..., i, ..., N\}$ be a finite population of N units and y_i (x_i) be the value of the study (auxiliary)

*Corresponding author : Raghunath Arnab
E-mail address : arnabr@mopipi.ub.bw

variable for the i^{th} unit of the population and $Y(X)$ be their total. Here x_i's are assumed to be known and positive for every $i \in U$. Let a sample s of size n be selected from U by a varying probability sampling scheme using x_i as a measure of size for the i^{th} unit. Rao (1966), Arnab (2001) and Scott and Smith (1969) alternative estimators are given below.

1.1.1 Rao's (1966) estimators

The conventional estimators for a finite population total Y under PPSWR, πps and Rao-Hartley-Cochran (1963, RHC) sampling schemes are respectively given by

$$t_{pps} = \frac{1}{n}\sum_{i \in s} n_i(s)\frac{y_i}{p_i} \qquad (2)$$

$$t_{hte}(\pi ps) = \sum_{i \in s}\frac{y_i}{\pi_i} = \sum_{i \in s}\frac{y_i}{np_i} \qquad (3)$$

and

$$t_{rhc} = \sum_{i \in s} y_i \frac{P_i}{p_i} \qquad (4)$$

where $p_i = x_i/X$, $n_i(s)$ = frequency of the i^{th} unit in s, π_i = inclusion probability for the i^{th} unit and P_i = sum of p_j's for the group containing the $i\ (\in s)^{th}$ unit for selection of sample under RHC sampling scheme.

Rao (1966) showed that the alternative estimators

$$t_{pps}(1) = \frac{1}{n}\sum_{i \in s} n_i(s) y_i = N\,\bar{y}_n t_0(s) = N\sum_{i \in s} y_i/n \quad = N\,\bar{y}_s$$

and $t_{rhc}(1) = N\sum_{i \in s} y_i P_i$ are unbiased for Y under model $M1$ and more efficient than the corresponding conventional estimators t_{pps}, $t_{hte}\ (\pi ps)$ and t_{rhc}.

The Murthy's (1957) estimator for PPSWOR sampling scheme is given by

$$\overset{*}{t}_{mur} = \frac{1}{p(s)}\sum_{i \in s} y_i\, p(s \mid i) \qquad (5)$$

where $p(s)$ and $p(s|i)$ denote respectively the probability of selection of an unordered sample s based on PPSWOR sampling scheme and the conditional probability of selection s given that the unit i was chosen on the first draw. Rao (1966) proposed an alternative estimator of $\overset{*}{t}_{mur}(2)$ (which is $\overset{*}{t}_{mur}$ with n = 2) as

$$t_{mur}(2) = \frac{N}{2 - p_i - p_j}\{(1-p_j)y_i + (1-p_i)y_j\} \qquad (6)$$

The estimator t_{mur} (2) is inconsistent but unbiased under model M_1. Rao (1966) did not prove theoretically whether or not the proposed alterntive estimator t_{mur} (2) is superior to the conventional estimator $\overset{*}{t}_{mur}(2)$. However, he showed empirically the superiority of $t_{mur}(2)$ over $\overset{*}{t}_{mur}(2)$.

1.1.2 Arnab's (2001) estimators

Let P_n be the class of fixed effective size n sampling design and C be the class of linear homogeneous unbiased estimators for Y consisting of estimators of the form

$$t(s) = \sum_{i \in s} b_{si}\, y_i \qquad (7)$$

where b_{si}'s are constants free from y_i's satisfying the unbiasedness condition

$$\sum_{s \supset i} b_{si}\, p(s) = 1 \quad \forall\, i \in U \qquad (8)$$

Arnab (2001) showed that the alternative estimators $t_0(s) = N\,\bar{y}_s$ fares better than any estimator belonging to C in the sense that

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(t(s)\right) \;\; \forall\; p \in P_n, t(s) \in C \qquad (9)$$

From equation (9), we can establish the following inequalities

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(\sum_{i \in s}\frac{y_i}{\pi_i}\right),\; E_{M1}V_p(\overset{*}{t}_{mur})$$

and also

$$E_{M1}V_p\left(t_0(s)\right) \le E_{M1}V_p\left(t_{rhc}(1)\right) \le E_{M1}V_p(t_{rhc})$$

1.1.3 Scott and Smith's (1969) estimators

Scott and Smith considered a class C^* of linear homogeneous model design unbiased estimators of the population total Y based on a sampling design $p \in P_n$ of n distinct units. The class C^* consists of estimators of the form

$$t(s) = \sum_{i \in s} b_{si} y_i$$

satisfying the model-design unbiasedness condition

$$\sum_s p(s) \sum_{i \in s} b_{si} y_i = N \qquad (10)$$

Scott and Smith (1969) proved that

$$E_{M1}(MSE\ (t_0(s))) = E_{M1}\ E_p\ (t_0(s) - Y)^2$$

$$\leq E_{M1}(MSE(t(s))) = E_{M1} E_p (t(s) - Y)^2$$

Here E_p denotes expectation with respect to design p.

2. PROPOSED ESTIMATOR UNDER MODEL M2

In this present note we have showed that Rao (1966) and Arnab (2001)'s results can be extended further for a wider superpopulation model given below.

Model $M2$: $E_{M2}(y_i) = \mu$, $V_{M2}(y_i) = \sigma_i^2 = \sigma^2 v(x_i)$

and $C_{M2}(y_i, y_j) = 0$ for $i \neq j$

where $v(x_i)$ is a function of x_i only. Various forms of the variance function $v(x_i)$ specially $v(x_i) = x_i^g$ with $g \geq 0$ are referred to by Cassel *et al.* (1971), and Chaudhuri and Stenger (1992) among others. We have also extended the Scott and Smith's (1969) result by showing that their result is valid also for the wider classes of sampling designs $P_n^*(\supset P_n)$ consisting of n units which may not necessarily be distinct.

Theorem 1. $E_{M2} V_p \left(t_{pps}(1)\right) \leq E_{M2} V_p (t_{pps})$

Proof. $$E_{M2} V_p(t_{pps}) = E_{M2} V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{y_i}{p_i} \right)$$

$$= E_{M2} \left(\frac{1}{n} \left(\sum_{i=1}^N \frac{y_i^2}{p_i} - Y^2 \right) \right)$$

$$= \frac{1}{n} \left(\sum_{i=1}^N \sigma_i^2 \left(\frac{1}{p_i} - 1 \right) \right)$$

$$+ \mu^2 V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{1}{p_i} \right) \qquad (11)$$

$$E_{M2} V_p[t_{pps}(1)] = E_{M2} \left(\frac{1}{n} \left(\sum_{i=1}^N \frac{z_i^2}{p_i} - Z^2 \right) \right)$$

where $z_i = y_i p_i$ and $z = \sum_{i=1}^N z_i$

$$E_{M2} V_p[t_{pps}(1)] = \frac{1}{n} \sum_{i=1}^N \sigma_i^2 p_i (1 - p_i) \qquad (12)$$

From (11) and (12), we get

$$E_{M2} V_p(t_{pps}) - E_{M2} V_p(t_{pps}(1))$$

$$= \sum_{i=1}^N \sigma_i^2 \frac{(1+p_i)}{p_i} (1-p_i)^2 + \mu^2 V_p \left(\frac{1}{n} \sum_{i \in s} n_i(s) \frac{1}{p_i} \right) \geq 0$$

Theorem 2. $E_{M2} V_p(t_0(s)) \leq E_{M2} V_p(t(s))\ \ \forall\ t(s) \in C$, $p \in P_n$, if σ_i^2 is a decreasing function of π_i.

Proof. $V_p(t(s)) = E_p(t(s))^2 - Y^2$

$$= \sum_{i=1}^N y_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right) + \sum_{i \neq}^N \sum_{j=1}^N y_i y_j \left(\sum_{s \supset i, j} b_{si} b_{sj} p(s) - 1 \right)$$

and

$$E_{M2} V_p(t(s)) = \sum_{i=1}^N \sigma_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right) + \mu^2 V_p \left(\sum_{i \in s} b_{si} \right)$$

$$\geq \sum_{i=1}^N \sigma_i^2 \left(\sum_{s \supset i} b_{si}^2 p(s) - 1 \right)$$

$$\geq \sum_{i=1}^N \sigma_i^2 \left(\frac{1}{\pi_i} - 1 \right) \qquad (13)$$

$\left(\sum_{s \supset i} b_{si}^2 p(s) \geq \frac{\left(\sum_{s \supset i} b_{si} p(s) \right)^2}{\sum_{s \supset i} p(s)} = \frac{1}{\pi_i} \right.$ follows from the unbiasedness condition (8))

$$V_p(t_0(s)) = \frac{N^2}{n^2} E_p \left(\left(\sum_{i \in s} y_i \right)^2 - \left(\sum_{i=1}^N y_i \pi_i \right)^2 \right)$$

$$= \frac{N^2}{n^2} \left(\sum_{i=1}^N \pi_i (1 - \pi_i) y_i^2 - \sum_{1 \neq}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) y_i y_j \right) \qquad (14)$$

Equation (14) yields

$$E_{M2}V_p(t_0(s)) = \frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2$$

$$+\frac{N^2}{n^2}\mu^2\left(\sum_{i=1}^{N}\pi_i(1-\pi_i)-\sum_{i\neq}^{N}\sum_{j=1}^{N}(\pi_i\pi_j-\pi_{ij})\right)$$

$$= \frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2 \qquad (15)$$

(noting $\sum_{i=1}^{N}\pi_i = n$ and $\sum_{i\neq}^{N}\sum_{j=1}^{N}\pi_{ij} = n(n-1)$)

Finally from (13) and (15), we get

$$E_{M2}V_P(t(s)) - E_{M2}V_p(t_0(s))$$

$$\geq \sum_{i=1}^{N}\sigma_i^2\left(\frac{1}{\pi_i}-1\right)-\frac{N^2}{n^2}\sum_{i=1}^{N}\pi_i(1-\pi_i)\sigma_i^2$$

$$=\frac{1}{N}\sum_{i=1}^{N}q_i\left(\pi_i-\frac{n}{N}\right)$$

$$= Cov(q_i, \pi_i)$$

where

$$q_i = -\frac{\sigma_i^2}{n^2}\left(\frac{1}{\pi_i}-1\right)(n+N\pi_i)$$

$$= -\frac{\sigma_i^2}{n^2}\left(n\left(\frac{1}{\pi_i}-1\right)+N(1-\pi_i)\right) \qquad (16)$$

Now if σ_i^2 is a decreasing function of π_i, then q_i will be an increasing function of π_i since $n\left(\frac{1}{\pi_i}-1\right)+N(1-\pi_i)$ is a decreasing function of π_i. In this situation $Cov(q_i, \pi_i)$ becomes positive.

Corollary 1. For an IPPS sampling design where $\pi_i = np_i$ and for the model with $M2$, $\sigma_i^2 = \sigma^2 x_i^g$, σ_i^2/π_i becomes a decreasing function of π_i if $g \leq 1$. In this case $E_{M2}V_p(t_0(s)) \leq E_{M2}V_p(t(s))$.

In particular if $\sigma_i^2 = \sigma^2$, Theorem 2 reduces to inequality (9).

Theorem 3. For a sampling design $p \in P_n^*$ and $t(s) \in C^*$

$$E_{M1}(MSE(t(s)) = E_{M1}E_p(t(s) - Y)^2 \geq \sigma^2 N\left(\frac{N}{\gamma}-1\right)$$

$$= E_{M1}E_{p_0}\left(t_0(s)-Y\right)^2$$

where $\gamma = E_p(\gamma_s)$ = expected effective sample size $=\sum_s \gamma_s p(s)$ and p_0 is a fixed effective size sampling design with Prob$\{\gamma_s = \gamma\} = 1$.

Proof. $E_{M1}(MSE(t(s)))$

$$= E_{M1}E_P(t(s) - Y)^2$$

$$= E_pE_{M1}(t(s) - Y)^2$$

$$= \sigma^2\sum_s p(s)\left(\sum_{i\in s}{b_{si}}^2 + N - 2\sum_{i\in s}b_{si}\right)$$

$$+ \mu^2\sum_s p(s)\left(\sum_{i\in s}b_{si} - N\right)^2 \qquad (17)$$

and

$$\sum_s p(s)\left(\sum_{i\in s}{b_{si}}^2 + N - 2\sum_{i\in s}b_{si}\right)$$

$$= \sum_s p(s)\sum_{i=1}^{N}I_{si}(b_{si}-1)^2 + N - \gamma \qquad (18)$$

Further the model-design unbiased condition (10) yields

$$\sum_s p(s)\sum_{i=1}^{N}I_{si}(b_{si}-1)^2 = \sum_s\sum_{i=1}^{N}I_{si}\,p(s)(b_{si}-1)^2$$

$$\geq \frac{\left(\sum_s\sum_{i=1}^{N}I_{si}\,p(s)(b_{si}-1)\right)^2}{\sum_s\sum_{i=1}^{N}I_{si}\,p(s)}$$

$$= \frac{(N-\gamma)^2}{\gamma} \qquad (19)$$

Finally using (17), (18) and (19), we get

$$E_{M1}(MSE(t(s))) \geq N\left(\frac{N}{\gamma}-1\right)\sigma^2 \qquad (20)$$

Equality in equation (20) holds for a sampling strategy based on fixed effective size sampling design p_0 satisfying Prob$\{\gamma_s = \gamma\} = 1$ and an estimator *t(s)* with $b_{si} = N/\gamma_s = N/\gamma$.

Remark 1: Scott and Smith's (1969) assertions of non-existence of the lower bound given in (20) for a with replacement sampling design is clearly incorrect.

ACKNOWLEDGEMENTS

The authors are grateful to the referee and editor for their valuable suggestions that led to an improvement of the paper.

REFERENCES

Arnab, R. (2001). Estimation of a finite population total in varying probability sampling for multi-character surveys. *Metrika,* **54**, 159-177.

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

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). *Foundation of Inference in Survey Sampling.* John Wiley and Sons, New York.

Chaudhuri, A. and Stenger, H. (1992). *Survey Sampling Theory and Methods.* Marcel Deckker, Inc.

Hansen, M.H., Hurwitz, W.N. (1943). On theory of sampling from finite populations. *Ann. Math. Statist.,* **14**, 433-362.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc*., **47**, 663-685.

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. Stat.,* **12&13**, 75-105.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18**, 379-390.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1963). A simple procedure of unequal probability without replacement. *J. Roy. Statist. Soc.,* **B24**, 482-491.

Rao, J.N.K. (1966). Alternative estimator for PPS sampling for multiple characteristics. *Sankhya,* **A28**, 47-60.

Scott, A. and Smith, T.M.F. (1969). A note on estimating secondary characteristics in multivariate surveys. *Sankhya,* **A31**, 497-498.

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