Pak. J. Statist.
1997 Vol. 13(3) pp 267 - 272

ASYMPTOTICALLY OPTIMAL DOUBLE SAMPLING STRATEGIES

Arijit Chaudhuri and Arun Kumar Adhikary
Applied Statistics Unit,
Indian Statistical Institute, Calcutta, India

(*Date Received: December, 1996 Date Accepted: August, 1997*)

Abstract

Postulating a simple regression model, asymptotic optimality is demonstrated for a class of sampling strategies in two phases to estimate a survey population total.

Key Words

Double sampling; optimality; survey population.

1. INTRODUCTION

Rao and Bellhouse (1978) gave model-based optimal double sampling strategies to estimate a survey population total under appropriate super-population modelling. But these strategies involve unknowable model parameters. Chaudhuri and Adhikary (1983, 1985) gave alternatives free of model parameters but with drastically over-simplified modelling. Mukerjee and Chaudhuri (1990) resorted to asymptotic analysis to derive generalized regression estimators allowing flexibility to double sampling designs recommended by their predecessors. Chaudhuri and Roy (1994) derived asymptotically optimal strategies on deriving Godambe-Joshi (1965) and Godambe-Thompson (1977) type lower bounds attained by 'unknown parameter-based' as well as by 'parameter-free' estimators. In this note we simplify Chaudhuri and Roy's (1994) model and derive simplified strategies with estimators of both of the above two kinds attaining sharper bounds. In addition, we observe Chaudhuri and Adhikary's (1983, 1985) simple two-phase strategies to constitute a sub-class of the above asymptotically optimal strategies. Chaudhuri and Roy (1994), to be abbreviated as CR (1994), is our key reference and we avoid repeating some of the discussions therein to save space. Särndal and Swensson (1987) gave several results on double sampling with varying probabilities. Also, two recent texts by Särndal, Swensson and Wretman (1992) and Chaudhuri and Stenger (1992) cover many relevant topics. The present work is a supplement to them.

2. THE SAMPLING STRATEGIES, MODEL AND OPTIMALITY

Let $U = (1, \ldots, i, \ldots, N)$ denote a survey population of size N. Let s_1 denote a first phase sample of distinct units n_1 in number chosen with probability $p_1(s_1)$ from U. Let s_2 be a sub-sample of s_1 consisting of distinct units, $n_2 (< n_1)$ in number chosen with the conditional probability $p_2(s_2|s_1)$. The over-all "double sample" thus chosen in two phases is $s = (s_1, s_2)$ having the selection-probability $p(s) = p_1(s_1)p_2(s_2|s_1)$.

Let y_i, x_i, w_i denote respectively values of variables y, x, w for $i \in U$ and let for every j, in U,

$$P_j = \sum_{s_1 \ni j} p_1(s_1) > 0,\ R_j(s_1) = \sum_{s_2 \ni j} p_2(s_2|s_1) > 0$$

$$Q_j = \sum_{s_1 \ni j} \sum_{s_2 \ni j} p_2(s_2|s_1)p_1(s_1) > 0.$$

The survey data to be gathered may be denoted by $d = (s, y_i, x_j | i \in s_2;\ j \in s_1)$. The values w_i will be supposed to be known and positive with a total W and may be used to specify the designs p_1, p_2 and p. Our problem is to estimate the total Y of y_i for $i \in U$ using an estimator t based on d for which its value is $t(d)$. Using notations and definitions given by CR (1994), to be mostly persisted with here, t is required to satisfy the condition of being 'Asymptotically design unbiased' (ADU) for Y in Brewer's (1979) sense and thus be subject to

$$\lim E_p(t - Y) = 0 \tag{2.1}$$

—— the notation $\lim E_p$ stands for 'limiting design expectation' as conceived by Brewer (1979) and applied by CR (1994).

CR (1994) postulated the super-population model, denoted by $\underline{M}(\theta)$ which essentially stipulates the following. Writing $\underline{Z} = (z_1, \ldots, z_i, \ldots, z_N)$, where $z_i = y_i, x_i, w_i, i \in U, E_1(V_1), E_2(V_2)$ the operators for expectation (variance) over distributions of $\underline{X}$ given $\underline{W}$ and of $\underline{Y}$ given $\underline{X}$, $\underline{W}$, let

$$E_1(x_i|\underline{W}) = \beta_1 w_i, E_2(y_i|\underline{X}, \underline{W}) = \theta x_i + \beta_2 w_i,$$

$$V_1(x_i|\underline{W}) = \sigma_{1i}^2,\ V_2(y_i|\underline{X}, \underline{W}) = \sigma_{2i}^2, E_1(\sigma_{2i}^2|\underline{W}) = \Psi_i^2.$$

Further, w_i's are 'non-stochastic', y_i's are 'independent' conditionally on $\underline{X}$ and x_i's are 'independent'.

CR (1994) have recommended double sampling strategies with certain desirable properties under this model with θ, β_1 and β_2 unknown. Our purpose here is to show that a simplification with a higher efficiency is available if θ is known. Assuming θ to be known we shall take θ as unity with no loss of generality because one may replace x_i above by $x_i^* = \theta\, x_i$ for i in U when θ is known but different from unity. Though Särndal, Swensson and Wretman (1992) in their Chapter 9 have discussed numerous details about double sampling methods they have not presented any asymptotically optimal strategies as in CR (1994) nor as the ones here to follow.

In this note with these preliminaries, we shall restrict to the special case $\underline{M}(1)$, say, of $\underline{M}(\theta)$ taking $\theta = 1$ above. This will lead to (1) simplifications of strategies as well as (2) sharpening in the efficiency levels discussed below. By E_m we shall denote over-all 'model' expectation.

Our findings are enunciated in the following theorems and remarks.

Theorem 1.

$$E_m \lim E_p(t - Y)^2 \geq \sum_1^N \Psi_j^2 \left(\frac{1}{P_j^2} \sum_{s_1 \ni j} \frac{p_1(s_1)}{R_j(s_1)} - 1 \right)$$

$$+ \sum_1^N \sigma_{1j}^2 \left(\frac{1}{P_j} - 1 \right) = E_m \lim E_p(t_0 - Y)^2, \text{ where}$$

$$t_0 = t_0(d) = \sum_{j \in s_2} \{(y_j - x_j - \beta_2 w_j)/P_j R_j(s_1)\}$$

$$+ \sum_{j \in s_1} (x_j - \beta_1 w_j)/P_j + (\beta_1 + \beta_2)W.$$

Proof.

Omitted as it is an obvious special case of Theorem 1 in CR (1994) with $\theta = 1$.

Theorem 2.

$$\frac{1}{P_j^2} \sum_{s_1 \ni j} \frac{p_1(s_1)}{R_j(s_1)} \geq \frac{1}{Q_j}. \tag{2.2}$$

Proof. Follows, on applying Cauchy inequality on

$$\left\{ \sum_{s_1 \ni j} p_1(s_1) R_j(s_1) \right\} \left\{ \sum_{s_1 \ni j} \frac{p_1(s_1)}{R_j(s_1)} \right\}.$$

Remark I.

The inequality (2.2) reduces to equality if

$$R_j(s_1) = \frac{Q_j}{P_j} \text{ for every } s_1 \text{ with } j \in s_1. \tag{2.3}$$

A design p for which (2.3) is satisfied will be denoted by p_0. For equal probability sampling in both the phases, (2.3) is satisfied. Chaudhuri and Adhikary (1983) showed the existence of 'unequal probability' sampling designs also satisfying (2.3). The class of sampling strategies (p_0, t_0) is then 'asymptotically optimal' by virtue of

Theorem 3.

$$E_m \lim E_p(t-Y)^2 \geq \sum_1^N \Psi_j^2 \left(\frac{1}{Q_j}-1\right)$$

$$+\sum_1^N \sigma_{1j}^2 \left(\frac{1}{P_j}-1\right) = E_m \lim E_{p_0}(t_0-Y)^2.$$

Proof. Follows from Theorems 1,2 and condition (2.3) on p.

Remark II.

We may observe that when based on p_0, t_0 becomes

$$t_0 = \sum_{j\in s_2} (y_j - x_j - \beta_2 w_j)/Q_j + \sum_{j\in s_1} (x_j - \beta_1 w_j)/P_j + (\beta_1+\beta_2)W.$$

Yet t_0 is not usable. So, as in CR (1994) we proceed to replace unknown model parameters in t_0 by their estimators and derive as follows 'a generalized regression' type estimator t_0^* to be used in practice in lieu of t_0 itself. Let $u_i = y_i - x_i$; $s_{ab} = \sum_{i\in s_2} a_i' b_i'$ where a_i', b_i' stand for $x_i, w_i, u_i, i \in U$. Noting $E_m(u_i) = \beta_2 w_i$, we take b_1, b_2 as estimators of β_1, β_2 given by $b_1 = \frac{s_{wx}}{s_{ww}}$, $b_2 = \frac{s_{wu}}{s_{ww}}$. Clearly, $E_m(b_j) = \beta_j, j = 1,2$. Our proposed generalized regression type estimator for Y is

$$t_0^* = \sum_{j\in s_2} (y_j - x_j - b_2 w_j)/Q_j + \sum_{j\in s_1} (x_j - b_1 w_j)/P_j + (b_1+b_2)W,$$

which satisfies, as may easily be checked,

$$\lim E_p(t_0^* - Y) = 0.$$

Next, we have

Theorem 4.

$$\begin{aligned} E_m \lim E_{p_0}(t_0^*-Y)^2 &= \sum_1^N \Psi_j^2 \left(\frac{1}{Q_j}-1\right) + \sum_1^N \sigma_{1j}^2 \left(\frac{1}{P_j}-1\right) \\ &= E_m \lim E_{p_0}(t_0-Y)^2. \end{aligned}$$

Proof. Easy and hence omitted; one may see p.360 in CR (1994).

Remark III.

An alternative way to get t_0 free of β_1, β_2 is to impose the following restrictions on the class of designs p_0:

$$P_j = \frac{n_1 w_j}{W} \text{ for } j \in s_1 \text{ and } Q_j = \frac{n_2 w_j}{W} \text{ for } j \in s_2. \tag{2.4}$$

For the resulting sub-class, say, p_1 of designs within the class of designs p_0 subject to (2.4) we note that t_0 reduces to

$$t_1 = \sum_{j \in s_2} (y_i - x_j)/Q_j + \sum_{j \in s_1} x_j/P_j. \tag{2.5}$$

This sub-class (p_1, t_1) of strategies within (p_0, t_0) was earlier recommended by Chaudhuri and Adhikary (1983, 1985). Of course we have

Theorem 5.

$$E_m \lim E_{p_1}(t_1 - Y)^2 = \sum_1^N \Psi_j^2 \left(\frac{1}{Q_j} - 1\right) + \sum_1^N \sigma_{1j}^2 \left(\frac{1}{P_j} - 1\right)$$

provided P_j, Q_j are subject to (2.4).

Remark IV.

In spite of Theorem 5 justifying the use of (p_1, t_1) we recommend the use, in practice, of (p_0, t_0^*) in preference to (p_1, t_1) in case $\underline{M}(1)$ seems tenable, because the restriction (2.4) may curtail the efficiency level.

Remark V. We omit a formula for an estimator of the mean square error of t_0^* derivable analogously to the one given in Section 4 in CR (1994).

ACKNOWLEDGEMENT

The authors deeply appreciate the suggestions from a referee incorporated here to achieve an improvement upon an earlier draft.

REFERENCES

(1) Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Jour. Amer. Statist. Assoc*, 74, 911-915.

(2) Chaudhuri, A. and Adhikary, A.K. (1983). On optimality of double sampling strategies with varying probabilities. *Jour. Statist. Plan. Inference*, 8, 257-265.

(3) Chaudhuri, A., and Adhikary, A.K. (1985). Some results on admissibility and uniform admissibility in double sampling. *Jour. Statist. Plan. Inference*, 12, 199-202.

(4) Chaudhuri, A. and Roy, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.

(5) Chaudhuri, A. and Stenger, H. (1992). *Survey Sampling: Theory and Methods.* Marcel Dekker, New York.

(6) Godambe, V.P. and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations. *Ann. Math. Statist*, 36, 1707-1722.

(7) Godambe V.P. and Thompson, M.E. (1977). Robust near optimal estimation in survey practice. *Bull. Int. Statist. Inst*, 47, 3, 129-146.

(8) Mukerjee, R. and Chaudhuri, A. (1990). Asymptotic optimality of double sampling plans employing generalized regression estimators. *Jour. Statist. Plan. Inference*, 26, 173-183.

(9) Rao, J.N.K. and Bellhouse, D.R. (1978). Optimal estimation of a finite population mean under generalized random permutation models. *Jour. Statist. Plan. Inference*, 2, 125-141.

(10) Särndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse . *Int. Stat. Rev.* 55 (3), 279-294.

(11) Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag, New York.