ON OPTIMAL AND RELATED STRATEGIES FOR SAMPLING ON TWO OCCASIONS WITH VARYING PROBABILITIES

ARIJIT CHAUDHURI
Indian Statistical Institute, Calcutta

(Received ; July, 1981)

SUMMARY

In estimating the 'current' total for a finite population total an 'optimal' sampling strategy is specified for a random permutation model with linear non-homogeneous estimators. Noting difficulties in implementing the optimal sampling scheme a pragmatically sensible alternative is suggested. In the latter, the 'current' 'unmatched' sample is taken from the complement of the 'matched' sub-sample of the 'initial' sample. The 'initial' and 'unmatched' samples are selected with unequal probabilities.

Introduction

The problem treated here is to find a suitable optimal estimator for the 'current' total of a finite population from data to be gathered on two consecutive ('current' and a 'previous' or 'initial') occasions. We consider selection schemes with varying probabilities with fixed sample-sizes and use an estimator within a class of non-homogeneous, linear 'model-design'—unbiased estimators and choose the 'random permutation model' to define our optimality criterion. We are able to identify a class of optimal sampling strategies involving a regression-type estimator based on an 'initial' sample, a subsequent 'unmatched' sample disjoint with it, both chosen with varying probabilities depending on available size-measures of sampling units and a 'matched' sub-sample from the initial

N.B. The work was (1) done when the author visited Prof. JNK Rao at Carleton University, Ottawa, Canada, at the latter's invitation in Summer, 1980 and (2) later presented in a conference at ISI, Delhi Centre in December, 1980 to mark the Sixtieth birthday of Prof. C. R. Rao.

sample chosen by SRSWOR method. Restricting within a sub-class of design-unbiased estimators a specific optimal sampling strategy is also derived. But noting the practical difficulties in implementing the sampling scheme involved we suggest an 'alternative' simpler sampling scheme retaining the earlier estimator. But we are unable to claim any optimality for the resulting strategy, rather we are able to find another 'fairly reasonable' sampling strategy (with both the selection scheme and the estimator altered) with a smaller average variance than that of this alternative. For the latter, the 'unmatched' component of the current sample is taken from the complement of the 'matched' portion of the 'initial' sample.

2. Formulation of the Problem, Notations and the Optimal Strategy

We suppose that a finite universe $U = (1, \ldots, i, \ldots, N)$ of N units is surveyed twice, y_j^* and x_j^* $(j = 1, \ldots, N)$ denoting respectively the true variate-values on the 'current' and a 'previous' occasion, y_j and x_j being the corresponding observable values subject to possible response errors. A fixed sample-size design P (using a generic symbol) is then employed with a scheme of selection with varying probabilities to yield an 'initial' sample S_1 of size n_1, a 'current' matched sub-sample S_2 of size n_2 from S_1 and a current 'unmatched' sample S_3 of size n_3 (with $n_1 = n_2 + n_3$, for simplicity) either from U or from S_1^c (complement of S_1) or from S_2^c (complement of S_2). Using survey data for the sample $S = (S_1, S_2, S_3)$ the problem is to estimate the true 'current' total $Y^* = \sum_1^N y_j^*$. By E_R, E_P, E_M respectively we shall denote the expectations with respect to response distributions, design and model (we assume here a random permutation model in a sense described below). Also, we will write $E_{PR} = E_P E_R$, $E_{MR} = E_M E_R$ and $E = E_P E_M E_R = E_M E_P E_R$. First we assume $E_R(y_j) = y_j^*$ and $E_R(x_j) = x_j^*$ $\forall\, j$. Also we assume that positive-valued size-measures W_j's $(j = 1, \ldots, N)$ are available for the units. Then, we postulate a model as follows:

Let $r = r_1, \ldots, r_j, \ldots, r_N)$, (with $r_j = y_j/w_j$) and $t = (t_1, \ldots, t_j, \ldots t_N)$, (with $t_j = x_j/w_j$) be respectively a randomly selected vector out of the N; vectors obtainable on permuting the co-ordinates in r and t; with respect to this random permutation the operation E_m is defined. Now we assume that the following is true:

$$r_j = \bar{R}^* + f_j,\ \bar{R}^* = \frac{1}{N} \Sigma\, r_j^*,\ E_{MR}(f_j) = 0,\ E_{MR}(f_j^2) = \delta_1,$$

$$E_{MR}(f_j f_k) = \delta_2\ (\forall\, j, k),\ E_M(\bar{R}^*) = \bar{R}^*$$

$$t_j = \overline{T}^* + h_j, \overline{T}^* = \frac{1}{N} \Sigma t_j^*, E_{MR}(h_j) = 0$$

$$E_{MR}(h_j^2) = \eta_1, E_{MR}(h_j h_k) = \eta_2 \ (\forall\, j, k), E_{MR}(f_j h_j) = \nu_1,$$

$$E_{MR}(f_j h_k) = \nu_2 (\forall\, j, k).$$

Under this model considered earlier by Rao and Bellhouse [3] we note that $\mu = E_M(\overline{y^*}) = \overline{R}^* \overline{W}$ and initially we seek a model-design unbiased estimator for μ. Let the present search be restricted within the sub-class of non-homogeneous linear 'model-design'—unbiased estimators of the following form:

$$e_b = e_b(s) = b_s + \sum_{s_2} b_{s_j}^{(1)} y_j + \sum_{s_3} b_{s_j}^{(2)} y_j$$

$$+ \sum_{s_2} b_{s_j}^{(3)} x_j + \sum_{\bar{s}_2} b_{s_j}^{(4)} x_j \tag{2.1}$$

(where $\bar{s}_2 = s_1 - s_2$) such that

$$E(e_b) = \mu \tag{2.2}$$

So, the coefficients are to satisfy the following conditions (they are independent of both x's and y's):

$$0 = E_p(b_s),\ E_p\left(\sum_{s_2} b_{s_j}^{(1)} W_j + \sum_{s_3} b_{s_j}^{(2)} W_j\right) = \overline{W},$$

$$E_p\left(\sum_{s_2} b_{s_j}^{(3)} W_j + \sum_{\bar{s}_2} b_{s_j}^{(4)} W_j\right) = 0$$

so that $E(e_b) = \mu$. Let e_d be a model-design unbiased estimator for 0, belonging to the same class as e_b so that we may write

$$e_d = d_s + \sum_{s_2} d_{s_j}^{(1)} y_j + \sum_{s_3} d_{s_j}^{(2)} y_j + \sum_{s_2} d_{s_j}^{(3)} x_j + \sum_{\bar{s}_2} d_{s_j}^{(4)} x_j$$

such that

$$0 = E_p(d_s) = E_p\left[\sum_{s_2} d_{s_j}^{(1)} W_j + \sum_{s_3} d_{s_j}^{(2)} W_j\right]$$

$$= E_p\left[\sum_{s_2} d_{s_j}^{(3)} W_i + \sum_{\bar{s}_2} d_{s_j}^{(4)} W_j\right] \tag{2.3}$$

We wish to choose the 'optimal' estimator e_b^* (say), for which we have

$$E(_b^* - \mu)^2 < E(e_b - \mu)^2 \tag{2.4}$$

with e_b, e^* satisfying (2.2) and the form (2.1). Applying C. R. Rao's

(1952) theorem the estimator e_b^* is the one for which we have

$E(e_b^* - \mu)\, e_d = 0$ for every estimator e_d of the form (2.3).

In order to find such an e_b^* (which may be called a UMV estimator) we confine the sampling designs to a sub-class (of P) of designs for which s_1 and s_2 are chosen as in P but s_3 is chosen necessarily from s_1^c. Then writing

$$\beta = \frac{\nu_1 - \nu_2}{\eta_1 - \eta_2}, \quad \beta' = \frac{\nu_1 - \nu_2}{\delta_1 - \delta_2}, \rho^2 = \beta\beta',$$

$$\phi = 1 - n_2/n_1, \quad \psi = \frac{1 - \phi}{1 - \phi^2 \rho^2}$$

we get the optimal estimator (which is design-unbiased) as

$$e_b^* = \bar{W}\,[\psi\, t + (1 - \psi)\, t_1], \text{ where}$$

$$t = \bar{r}_y(m) + \beta(\bar{r}_x(f) - \bar{r}_x(m)),\ t_1 = \bar{r}_y(n),$$

where

$$\bar{r}_y(m) = \frac{1}{n_2} \sum_{s_2} y_j/w_j,\ \bar{r}_x(f) = \frac{1}{n_1} \sum_{s_1} x_j/w_j,$$

$$\bar{r}_x(m) = \frac{1}{n_2} \sum_{s_2} x_j/w_j \text{ and } \bar{r}_y(u) = \frac{1}{n_3} \sum_{s_3} y_j/w_j\,.$$

To derive this briefly, we check, on putting $b_s = 0$, $b_{sj}^{(i)} = c_i/w_j$, $(i = 1, 2, 3, 4)$ with c_i's as constants to be determined, that

$$E(e_b e_d) = E_P E_{MR} \Big[\Big\{\bar{R}^*(c_1 n_2 + c_2 n_3) + \bar{T}^*(c_3 n_2 + c_4 n_3) + \Big(c_1 \sum_{s_2} f_j + c_2 \sum_{s_3} f_j\Big) + \Big(c_3 \sum_{s_2} h_j + c_4 \sum_{s_2} h_j\Big)\Big\}.$$

$$\Big\{d_s + \bar{R}^*\Big(\sum_{s_2} d_{s_j}^{(1)} w_j + \sum_{s_3} d_{s_j}^{(2)} w_j\Big) + \bar{T}^*\Big(\sum_{s_2} d_{s_j}^{(3)} w_j + \sum_{\bar{s}_2} d_{s_j}^{(4)} w_j\Big)$$

$$+ \Big(\sum_{s_2} d_{s_j}^{(1)} w_j f_j + \sum_{s_3} d_{s_j}^{(2)} w_j f_j\Big) + \Big(\sum_{s_2} d_{s_j}^{(3)} w_j h_j + \sum_{\bar{s}_2} d_{s_j}^{(4)} w_j h_j\Big)\Big\}\Big]$$

$$= E_P[\sum_{s_2} d_{s_j}^{(1)} w_j\{c_1(\delta_1 + (n_2 - 1)\,\delta_2) + c_2 n_3 \delta_2 + c_3\,(\nu_1 + (n_2 - 1)\,\nu_2$$

$$+ c_4 n_3 \nu_2\} + \sum_{s_3} d_{s_j}^{(2)} w_j\{c_1 n_2 + c_2(\delta_1 + (n_3 - 1)\,\delta_2) + c_3 n_2 \nu_2$$

$$+ c_4 n_3 \nu_2\} + \sum_{s_2} d_{s_j}^{(3)} w_j\{c_1(\nu_1 + (n_2 - 1)\,\nu_2) + c_2 n_3 \nu_2$$

$$+ c_3(n_2 - 1)\,\eta_2 + c_4 n_3 \eta_2\} + \sum_{s_3} d_{s_j}^{(4)} w_j\{c_1 n_2 \nu_2 + c_2 n_2 \nu_2 + c_3 n_2 \eta_2$$

$$+ c_4(n_1 + (n_3 - 1)\,\eta_2)\}]$$

In order to make it 0, and to make e_b design-unbiased for $\bar{y}$, we are to choose c_i's satisfying the four equations viz.,

$$c_1 - c_2 = -c_3\beta',\ c_3 - c_4 = -c_1\beta,\ c_1 n_2 + c_2 n_3 = 1,\ c_3 n_2 + c_4 n_3 = 0.$$

Recalling that $n_3 = n_1 - n_2$, s_2 is an SRSWOR and stipulating to choose S_1, S_3 with inclusion-probabilities proportional to w_j's unique solutions are obtained as

$$c_1 = \overline{W}\frac{\psi}{n_2},\ c_2 = \overline{W}\frac{(1-\psi)}{n_3},\ c_3 = \overline{W}\psi\beta\left(\frac{-n_3}{n_1 n_2}\right),\ c_4 = \overline{W}\frac{\psi\beta}{n_1}.$$

Hence the optimal estimator turns out as e_b^* as above.

It is easy to check that for each design P_1 we have a constant value for $E_{P_1}E_M E_R\,(e_b^* - \mu)^2$. So the next problem is to devise a scheme (and in fact just one scheme will do) for implementing an actual selection process corresponding to a design of the class P_1. In order to solve this let us consider a design P_2 within P_1 such that s_2 is an SRSWOR from s_1 and s_1 and s_3 are chosen in the manner described below.

Let $P_j = w_j/W, j = 1, \ldots, N$, (where $W = \sum_j w_j$), $s_1 = (i_1, \ldots, i_{n_1})$, $s_3 = (i_{n_1} + 1, \ldots, i_{n_1+n_3})$ and $(s_1, s_3) = (i_1, \ldots i_{n_1}, \ldots, i_{n_1+n_3})$ be an ordered sequence of distinct labels such that i_k stands for the unit selected on the kth draw from U, there being in all $n_1 + n_3$ draws such that s_1 consists of the outcomes of the first n_1 draws and s_3 of those of the last n_3 draws. Selection is made in $n_1 + n_3$ draws with the probability of selecting (s_1, s_3) as

$$p_{i_1}^{(1)} \times \frac{p_{i_2}^{(2)}}{1 - p_{i_1}^{(2)}} \times \frac{p_{i_3}^{(3)}}{1 - p_{i_1}^{(3)} - p_{i_2}^{(3)}} \times \ldots$$

$$\times \frac{p_{i_{n_1+n_3}}(n_1 + n_3)}{1 - p_{i_1}(n_1 + n_3) \ldots p_{i_{n_1+n_3-1}}(n_1 + n_3)}$$

(such that $1 \leqslant i_1 \neq \ldots \neq i_{n_1+n_3} \leqslant N$)

where $p_j(k)$'s are quantities such that

$$0 < p_{i_j}(k) < 1 \text{ for } 1 \leqslant i_j \leqslant N \ \forall j = 1, \ldots, N,$$

$$\sum_{i_j=1}^{N} p_{i_j}(k) = 1 \text{ for } k = 1, 2, \ldots, n_1 + n_3.$$

Writing $\delta_j(k)$ = the probability of selecting the j-th unit on the k-th draw for the scheme above, we require that the $p_j(k)$'s are so chosen that

$$\delta_j(k) \text{ equals } p_j \ \forall\ k = 1, \ldots n_1 + n_3 \text{ and } \forall\ k = 1, \ldots, N \ldots \quad (2.5)$$

Following Fellegi's [1] iterative method one may realize (2.5) at least approximately. So, our problem of specifying a strategy to yield an optimal (in the sense specified above) estimator is solved. Writing D_i $(i = 1, 2, 3)$, (say), to denote the designs corresponding to the sampling scheme for selecting s_1, s_2, s_3 for the sampling scheme (due to Fellegi given above and $\pi_j^{(i)}$, $(i = 1, 2, 3)$ for the corresponding first order inclusion-probabilities we have

$$\pi_j^{(1)} = \sum_{k=1}^{n_1} \delta_j(k) = n_1 \frac{W_j}{W}, \pi_j^{(2)} = n_2/n,$$

$$\pi_j^{(3)} = \sum_{k=n_1+1}^{n_1+n_3} \delta_j(k) = n_3 \frac{W_j}{W}, j = 1, \ldots, N.$$

Denoting the resulting design for choosing (s_1, s_3) by D and $s = (s_1, s_2, s_3)$ by P_2, we have

$$E_{P_2}(e_b^*) = E_D\,[E(e^* \mid s_1, s_3)]$$

$$= \frac{1}{N} E_D \left[\psi \sum_{j \in s_1} \frac{Y_j}{\pi_j^{(1)}} + (1-\psi) \sum_{j \in s_3} \frac{Y_j}{\pi_j^{(3)}} \right]$$

$$= \overline{Y} \text{ and } E_{P_2R}(e_b^*) = \overline{Y}^*,$$

If e_b is required to be not only model-design unbiased but also design-unbiased then whenever it is based on any member of P_1 it is easy to show that we have

$$E(e_b - \overline{Y}^*)^2 = E(e_b - \mu)^2 - E_{MR}(\overline{Y}^* - \mu)^2$$

$$\geqslant E(e_b^* - \mu)^2 - E_{MR}(\bar{y}^* - \mu)^2$$

$$= E_{P_1} E_{MR}(e_b^* - \mu)^2 - E_{MR}(\bar{y}^* - \mu)^2$$

$$= E_{P_2MR}(e_b^* - \mu)^2 - E_{MR}(\bar{y}^* - \mu)^2 = E_{P_2MR}(e_b^* - \bar{y}^*)^2.$$

So, we may claim that in the class e_b of non-homogeneous linear design-unbiased estimators for $\bar{y}^*$ the strategy (P_2, e_b^*) is optimal in the sense of yielding the minimum average mean square error.

3. A Couple of Related Strategies and their Uses

The strategy (P_2, e_b^*) is difficult to implement because Fellegi's scheme is not easy to execute accurately. So, we may consider employing an alternative strategy (P_3, e_b^*) where we retain the earlier estimator but employ a much simpler (and customary) design P_3 corresponding to a scheme for which s_1 is chosen with inclusion-probabilities $\pi_j^{(1)} = n_1 p_j$ and s_2 from s_1 with $\pi_j^{(2)} = n_2/n_1$ but s_3 is chosen not from s_1^c but from U with inclusion-probability $\bar{\pi}_j^{(3)} = n_3 p_j \; \forall \; j \varepsilon U$. Then, of course, $E_{P_3}(e_b^*) = \bar{Y}$. But $E_{P_3} E_{MR}$ $(e_b^* e_d)$ may not equal zero (uniformly). So, we are unable to claim any optimal property for (P_3, e_b^*). Rather, it is possible to employ another strategy $(P_4, \bar{e}_b)$, say, which fares better than $(P_3, \bar{e}_b^*)$ in the sense that $E_{P_4} \bar{e}_b = \bar{Y}$ but

$$E_{P_4} E_{MR} (\bar{e}_b - \bar{y}^*)^2 < E_{P_3} E_{MR} (e_b^* - \bar{y}^*)^2.$$

The design P_4 is such that s_1 and s_2 are chosen as in P_3 but s_3 is chosen from $s_2^c = U - s_2$ with inclusion-probabilities

$$\bar{\pi}_j(3) = \bar{\pi}_j(3)/Q(s_2) \text{ for } j \varepsilon s_2^c, \text{ where } Q(s_2) = \sum_{j \varepsilon s_2} {}_o p_j.$$

The estimator $\bar{e}_b$ is taken as

$$\bar{e}_b = \bar{W} [\psi t + (1 - \psi) t_1^*], \text{ with } \psi, t \text{ as in section 2,}$$

and

$$t_1^* = \frac{1}{W} \left[\frac{1}{n_3} \sum_{s_3} y_j / p_j^* + \sum_{s_3} y_j \right],$$

where

$$p_j^* = \frac{w_j}{W} \frac{1}{Q(s_2)} \cdot$$

It is easy to check that $E_{P_4} \bar{e}_b = \bar{y}$,

$$\text{Cov}_{P_4}(t, t^*) = 0 = \text{Cov}_{P_4}(t, t_1) = 0.$$

By V_P, Cov_P we denote variance and covariance with respect to a design P. One may now check that

$$E_{P_3} E_{MR} (e_b^* - \bar{y}^*)^2 - E_{P_4} E_{MR} (\bar{e}_b - \bar{y}^*)^2$$
$$= E_{MR} [E_{P_3} (e_b^* - \bar{y})^2 - E_{P_4} (\bar{e}_b - \bar{y})^2]$$
$$= E_{MR} [V_{P_3}(e^*) - V_{P_4}(\bar{e}_b)]$$

Now, $V_{P_3}(e_b^*) - V_{P_4}(\bar{e}_b) = \bar{W}^2 (1 - \psi)^2 [V_{P_3}(t_1) - V_{P_4}(t_1^*)]$

$$V_{P_3}(t_1) = \frac{1}{W^2} \sum_{j<k}\sum (\bar{\pi}_j^{(3)} \; (\bar{\pi}_k^{(3)} - \bar{\pi}_{jk}^{(3)}) \left(\frac{y_j}{\bar{\pi}_j^{(3)}} - \frac{y_k}{\bar{\pi}_k^{(3)}} \right)^2$$

$$E_{MR} V_{P_3}(t_1) = \frac{\delta_1 - \delta_2}{n_3^2} \left(n_3 - \sum_1^N \bar{\pi}_j^2 (3) \right) \cdot$$

Also, $E_{MR}V_P\,(t_1^*) = \dfrac{2(\delta_1 - \delta_2)}{n_3^2}\, E_1\Big[\sum\sum_{j<k\varepsilon s_2^c} (\overline{\overline{\pi}}_j^{(3)}\, \overline{\overline{\pi}}_k^{(3)} - \overline{\overline{\pi}}_{jk}^{(3)})\Big]\, Q^2(s_2)$

(E_1 denoting expectation with respect to the selection of s_1 for the design P_4) and writing $\overline{\pi}_{jk}^{(3)}, \overline{\overline{\pi}}_{jk}^{(3)}$'s for second order inclusion-probabilities for designs p_3 and P_4 respectively)

$$= \frac{\delta_1 - \delta_2}{n_3^2}\, E_1\Big[\, Q^2(s_2)\, n_3 - \sum_{j\varepsilon s_2^c} \overline{\overline{\pi}}_j^2(3)\,\Big]$$

$$= \frac{\delta_1 - \delta_2}{n_3^2}\Big[\, n_3 E_1\, Q^2(s_2) - E_1 \sum_{s_2^c} \overline{\pi}_j^2(3)\,\Big]$$

$$= \frac{\delta_1 - \delta_2}{n_3^2}\Big[\, n_3 E_1\, Q^2(s_2) - \sum_1^N \overline{\pi}_j^2(3) + E_1 \sum_{s_2} \overline{\pi}_j^2(3)\,\Big].$$

So, $E_{MR}(V_{P_3}(t_1) - V_{P_4}(t_1^*))$

$$= \frac{\delta_1 - \delta_2}{n_3^2}\Big[\, n_3\,\{1 - E_1 Q^2(s_2)\} - E_1 \sum_{s_2} \overline{\pi}_j^2(3)\,\Big]$$

$$> \frac{\delta_1 - \delta_2}{n_3^2}\Big[\, n_3 - E_1 \sum_{s_2} \overline{\pi}_j^{(3)} - n_3 E_1\, Q^2(s_2)\,\Big]$$

(since $\overline{\pi}_j^2(3) < \overline{\pi}_j^{(3)}$)

$$= \frac{\delta_1 - \delta_2}{n_3}\Big[\, E_1\Big(\, 1 - \sum_{s_2} p_j\Big) - E_1\, Q^2(s_2)\,\Big]$$

$$= \frac{\delta_1 - \delta_2}{n_3}\, E_1\,[Q(s_2)\,(1 - Q(s_2))] > 0,$$

since, $0 < Q(s_2) < 1 \;\forall\; s_2$.

So, $E_{P_3}E_{MR}(e_b^* - \overline{y}^*)^2 > E_{P_4}E_{MR}(\overline{e}_b - \overline{y}^*)^2$.

However, we are unable to claim any optimality property for $(P_4, \overline{e}_b)$, but it is easy to implement and it is more rational to use it rather than (P_3, e_b^*), because in the former we have a guarantee that the current sample (s_2, s_3) consists of distinct units only unlike in case of the latter.

If we restrict to the design-unbiased estimators above and use design P_4, then $(P_4, \overline{e}_b)$ can be further improved if one employs the strategy $(P_4, \widehat{e}_b)$, where $\widehat{e}_b = \overline{W}\theta\,[t + (1 - \theta)\, t_1^*]$ if one can choose θ so as to minimize $E_{P_4}E_{MR}(\widehat{e}_b - \overline{y}^*)^2$ for fixed values of $E_{MR}V_4(\mathrm{t})$ and $E_{MR}(V_4(t_1^*))$.

ACKNOWLEDGEMENT

The author is grateful to a referee for comments and to Prof. J. N. K. Rao who offered helpful suggestions and brought his attention to the applicability of Fellegi's scheme in deriving the optimality result reported above.

REFERENCES

[1] Fellegi, I. P. (1963) : Sampling with varying probabilities without replacement—rotating and non-rotating samples, *J. Amer. Statist. Assoc.*, **58**; 183-201.

[2] Rao, C. R. (1952) : Some theorems on minimum variance unbiased estimation, *Sankhya Ser. A*, **12**; 27-42.

[3] Rao, J. N. K. and Bellhouse, D. R. (1978) : Optimal estimation of a finite population mean under generalized random permutation models, *J. Statist. Plan. and Inf.*, **2**; 125-141.