Jour. Ind. Soc. Ag. Statistics
43 *(3), 1991 : 282-290*

On Sampling Over Two Occasions Using Varying Probabilities

Raghunath Arnab*
Indian Statistical Institute, Calcutta
(Received: March, 1995)

Summary

In estimating the finite population total on a current occasion, three strategies are proposed and studied involving sample selection with varying probabilities on an earlier occasion, stratified sub-sampling in three different ways from the initial sample and current sampling independently from the entire population and suitable combination of the survey data and available values of an auxiliary variable. Some of these strategies are found better than comparable strategies available.
Keywords : Sampling over two occasions, SRSWOR, PPSWOR, Rao- Hartley-Cochran Strategies.

Introduction

Following the works of Raj [5], Ghangurde and Rao [4], Chotai [2], Chaudhri and Arnab [1] we consider and compare performances of three strategies for estimating the finite population (of size N) total of a variable on a current occasion (y) using the values of it on a previous occasion (x) available from an initial sample, a current sub-sample from that and an independent sample currently drawn from the entire population. Specifically, on the first occasion a PPSWR sample S_1 of size n_1 is taken from the entire population U using known size-measures z_i's (>0 for i = 1, . . ., N). Utilizing the ascertained x-values for them, on the basis of certain criteria, the n_1 sample units are assigned to L strata. Let y_{hj} , x_{hj} and z_{hj} be the value of jth unit of hth (=1, . . ., L) stratum for the characters y, x and z respectively. Typically, a random number

* Present address: The National University of Lesotho, P.O. Roma 180, Lesoths, Africa.

$n_{1h}\left(0 < n_{1h} \le n_1, \sum_h n_{1h} = n_1\right)$ of these n_1 units will constitute the h th stratum say, S_{1h}. On the second occasion, independent sub-samples S_{2h}'s (say), of sizes $m_h = \gamma_h \; n_{1h}$ (with γ_h pre-assigned, $0 < \gamma_h < 1$) are chosen independently from respective S_{1h}'s (h = 1, . . ., L), each following suitable schemes utilizing known z_i's and ascertained x_i's. We will write $m = \sum_{h=1}^{L} m_h$ $u = n_2 - m$ with n_2 (<N) chosen as a positive integer, large if necessary, such that u may not be negative. Here n_2 is pre- assigned but u and m are random. A sample of size u is then drawn from U again by PPSWR method using z_i's. In the next section we describe 3 procedures for drawing the stratified sub-sample which lead to three distinct strategies for estimating the y-total, say $Y = \sum_{i}^{N} y_i$. The study however, is purely theoretical.

2. *The Proposed Strategies And Related Results:*

For the strategies denoted respectively as 1, 2 and 3, S_{2h}'s are selected from S_{1h}'s respectively by (i) SRSWOR method, (ii) PPSWR method with normed size measures taken as $q_{hi} = (x_{hi}/z_{hi})/\sum_{S_{1h}} (x_{hi}/z_{hi})$ expecting high correlation between y and x and (iii) Rao-Hartley-Cochran [7] (RHC in brief) method with normed size measures again as q_{hi}. Let us write $Y_h = \sum_j Y_{hj}$, $X_h = \sum_j x_{hj}$, $Z_h = \sum_j z_{hj}$, $Y = \sum_h Y_h$, $X = \sum_h X_h$, $Z = \sum_h Z_h$, $p_{hj} = z_{hj}/Z$, $P_h = Z_h/Z$, $w_h = n_{1h}/n_1$, $V(y \mid z) = Z \sum_h \sum_j \frac{y_{hj}^2}{z_{hj}} - Y^2$

$$T_1(h) = \sum_{S_{2h}} \frac{y_{hj} - c_{h1} x_{hj}}{m_h \, p_{hj}} + c_{h1} \sum_{S_{1h}} \frac{x_{hj}}{n_{1h} \, p_{hj}}$$

$$T_2(h) = \sum_{S_{2h}} \frac{y_{hj} - c_{h2} x_{hj}}{n_{1h} \, m_h \, p_{hj} \, q_{hj}} + c_{h2} \sum_{S_{1h}} \frac{z_{hj}}{n_{1h} \, p_{hj}}$$

$$T_3(h) = \sum_{S_{2h}} \frac{(y_{hj} - c_{h3} z_{hj})\, Q_{hj}}{n_{1h} \, p_{hj} \, q_{hj}} + c_{h3} \sum_{S_{1h}} \frac{z_{hj}}{n_{1h} \, p_{hj}}$$

Q_{hj} = sum of q_{hk} values for the group containing jth unit of hth stratum that was formed in selecting S_{2h} by RHC scheme of sampling, c_{hi}'s are constants minimizing variances of $T_i(h)$, $T_i = \sum_h w_h \; T_i(h)$ for i = 1, 2, 3 and $T = \sum_{S_{2u}} \frac{y_i}{up_i}$

The proposed estimators for Y based on strategy i (= 1, 2, 3) is

$$t_i = \varphi_i T_i + (1 - \varphi_i T \tag{1}$$

where, φ_i is a constant to be chosen to minimize $V(t_i)$, the variance of t_i.

Theorem 1 : $E(T_1) = Y$ and $V(T_1) = \dfrac{\left[\sum_h \left(\frac{1}{\gamma_h} - 1\right) V_h \frac{(d_1|z)}{P_h} + V(y|z)\right]}{n_1}$

with $V_h(d_1|z) = \sum_j \frac{(y_{hj} - c_{h1} x_{hj})^2}{p_{hj}} - (Y_h - c_{h1} X_h)^2$

Proof : Let $E_1(V_1)$ be the expectation (variance) over $\mathbf{n}_1$ $(n_{11}, \ldots, n_{1L})$ and $E_2(V_2)$, $E_3(V_3)$ be the conditional expectations (variance) over S_{1h}'s for fixed $\mathbf{n}_1$ and over S_{2h}'s for fixed S_{1h} and $\mathbf{n}_1$ respectively. The covariance operators V_{ij} $(i \neq j = 1, 2, 3)$ are similarly defined.

$$E(T_1) = E_1 E_2 E_3 (T_1) = E_1 E_2 \sum_h w_h \frac{\sum_{S_{1h}} y_{hj}/p_{hj}}{n_{1h}}$$

$$= E_1 \sum_h w_h Y_h/P_h = Y$$

and $V(T_1) = E_1 V_{23}(T_1) + V_1 E_{23}(T_1) = E_1 \sum_h w_h^2 V_{23} T_1(h) + V_1\left(\sum_h w_h Y_h/P_h\right)$

$$= E_1 \sum_h w_h^2 \left[\frac{\left(\sum_j y_{hj}^2 P_h/p_{hj} - Y_h^2\right)}{(n_{1h} P_h)} + (1/m_h - 1/n_{1h})\right.$$

$$\left.\left\{\sum_j (y_{hj} - c_{h1} x_{hj})^2 P_h/p_{hj} - (Y_h - c_{h1} X_h)^2\right\}\right] + V\left(\sum_h w_h Y_h/P_h\right)$$

$$= \frac{\left[\sum_h \left(\frac{1}{\gamma_h} - 1\right) V_h(d_1|z)/P_h + V(y|z)\right]}{n_h}$$

[Since $V(w_h) = \frac{P_h(1-P_h)}{n_1}$, $\text{cov}(w_h\, w_k) = \frac{-P_h P_k}{n_1}$ for $h \neq k = 1, \ldots, L$]

Now following the proof of the theorem 1 one may prove the following theorems:

Theorem 2: $E(T_2) = Y$ and $V(T_2) = \frac{\sum_h \{1 - 1/(n_1 P_h)\} \frac{V_h(d_2|x)}{P_h \gamma_h} + V(y|z)}{n_1}$

with $V_h(d_2|x) = \sum_j (y_{hj} - c_{h2} z_{hj})^2 \frac{X_h}{x_{hj}} - (Y_h - c_{h2} z_h)^2$

Theorem 3: $E(T_3) = Y$ and $V(T_3) = \frac{\sum_h (1/\gamma_h - 1) \frac{V_h(d_3|x)}{P_h} + V(y|z)}{n_1}$

with $V_h(d_3|x) = \sum_j (y_{hj} - c_{h3} z_{hj})^2 \frac{X_h}{x_{hj}} - (Y_h - c_{h3} z_h)^2$

Let us denote by $V_h(r|t) = \left(\sum_j t_{hj}\right) \sum_j \left(\frac{r_{hj}^2}{t_{hj}}\right) - \left(\sum_j t_{hj}\right)^2$: r, t = x, y, z.

$$\delta_h(r, s|t) = \frac{\sum t_{hj} \sum \frac{(r_{hj}\, s_{hj})}{t_{hj}} - \left(\sum_j t_{hj}\right)\left(\sum_j s_{hj}\right)}{\left[V_h(r|t)\, V_h(s|t)\right]^{1/2}}$$

$$\theta_h = \left\{1 - \delta_h^2(y, x|z)\right\} \frac{V_h(y|z)}{V(y|z)} \;;$$

$$\theta'_h = \left\{1 - \frac{1}{(n_1 P_h)}\right\} \left\{1 - \delta_h^2(y, z|x)\right\} \frac{V_h(y|x)}{V(y|z)} \;;$$

$$\theta''_h = \left\{1 - \delta_h^2(y, z|x)\right\} \frac{V_h(y|x)}{V(y|z)} \;;$$

$$A_1 = \sum_h \theta_h \frac{(1/\gamma_h - 1)}{P_h} \quad ; \qquad A_2 = \sum_h \theta'_h (P_h \gamma_h) \quad ;$$

$$A_3 = \sum_h (1/\gamma_h - 1)\, \theta''_h / P_h \quad ;$$

The optimum values of c_{hi}'s (to be written $c^{\bullet}_{hi}$) obtained by minimizing $V(T_i)$'s with respect to c_{hi}'s ($i = 1, 2, 3$) and corresponding values of $V(T_i)$'s to be written as $V_o(T_i)$'s are

$$c^{\bullet}_{h1} = \delta_h (y, x \mid z) \sqrt{V_h(y \mid z)/V_h(x \mid z)} \; ; \; V_o(T_1) = (1 + A_1)\, V(y \mid z)/n_1$$

$$c^{\bullet}_{h2} = \delta_h (y, z \mid x) \sqrt{V_h(y \mid x)/V_h(z \mid x)} \; ; \; V_o(T_2) = (1 + A_2)\, V(y \mid z)/n_1$$

$$c^{\bullet}_{h3} = \delta_h (y, z \mid x) \sqrt{V_h(y \mid x)/V_h(z \mid x)} \; ; \; V_o(T_3) = (1 + A_3)\, V(y \mid z)/n_1$$

Let E_u, V_u denote expectation and variance operators for variation over u and $E(T \mid u)$, $V(T \mid u)$ denote conditional expectation and conditional variance for fixed u. Then

$$V(T) = \frac{E_u\left(\sum_i \frac{y_i^2}{p_i} - Y^2\right)}{u} \simeq \frac{\left\{\frac{1 - V(u)}{(E(u))^2}\right\} V(y \mid z)}{E(u)}$$

$$= \frac{V(y \mid z)}{\xi^2 E(u)}, \; [\text{where } 1/\xi^2 = 1 - \text{ (coefficient of variation of u)}^2]$$

The optimum values of Φ_i's to be denoted by Φ_{io} and corresponding values of $V(t_i)$ written as $V_o(t_i)$ (say) come out as

$$\Phi_{io} = \left\{1 + (1 + A_i)\xi^2 \mu\right\}^{-1}, \; V_o(t_i) = \left\{1 + (1/A_i) + \xi^2 \mu\right\}^{-1} V(y \mid z)/n_1$$

writing $\mu = E(u)/n_1$

The optimum values of γ'_hs for given n_2 denoted by $\gamma^{\bullet}_{ih}$ obtained by minimizing $V_o(t_i)$ with respect to γ_h come out as :

$$\gamma^{\bullet}_{1h} = \frac{1 - xi \sum \sqrt{\theta_h}}{1 - \sum \theta_h / P_h} \; , \; \gamma^{\bullet}_{2h} = \xi\left(1 - 2\sum \sqrt{\theta''_h}\right) \sqrt{\theta''_h} \,/ P_h$$

$$\gamma_{3h}^{*} = \frac{(1-\sqrt{\theta_h''})\sqrt{\theta_h''}}{\xi P_h\left(1-\sum_h \frac{\theta_h''}{P_h}\right)}$$

Let us denote by c, c_o and $\alpha_h(\alpha_h')$ the total cost, overhead cost and cost per unit for the hth stratum on second (first) occasion respectively. Then for the cost function of the form $c = c_o + \sum \alpha_h u_h + \sum \alpha_h' m_h$ considered by Rao [6] with u_h as the un-matched sample size in hth strutam we have optimum values of γ_{ih}^{*} for the given expected cost $c^* = E(c) = c_o + n_1 \sum_h \alpha_h P_h + \left(n_2 - n_1 \sum_h \gamma_h P_h\right) \sum \alpha_h P_h$ for the three strategies as follows:

$$\gamma_{1h}^{*} = \sqrt{\theta_h}\,(\sqrt{\Sigma\,\alpha_h P_h} - \xi\Sigma\sqrt{\alpha_h' P_h}\,)\left\{\xi P_h\sqrt{\alpha_h'}(1-\Sigma\theta_h/P_h)\right\}^{-1}$$

$$\gamma_{2h}^{*} = \sqrt{\theta_h'}\,(\sqrt{\Sigma\,\alpha_h P_h} - \sqrt{\Sigma\,\alpha_h'\theta_h'}\,)\,(\xi P_h\sqrt{\alpha_h'}\,)^{-1}$$

$$\gamma_{3h}^{*} = \sqrt{\theta_h''}\,(\sqrt{\Sigma\,\alpha_h P_h} - \sqrt{\alpha_h' P_h}\,)\left\{\xi P_h\sqrt{\alpha_h'}(1-\Sigma\theta_h''/P_h)\right\}^{-1}$$

If the total sample size for the second occasion n_2 in kept fixed and the proportional allocation is used, the optimum γ_{ih}^{*}'s written as γ_i^{*} and the value of $V_o(t_i)$ denoted as $V_{min}(t_i)$ come out as

$$\gamma_1^{*} = \frac{\sqrt{B_1}}{1+\sqrt{B_1}}\;;\quad V_{min}(t_1) = \frac{(1+\sqrt{B_1})V(y\,|\,z)}{2n_1}$$

$$\gamma_2^{*} = \frac{\sqrt{B_2}}{1+\sqrt{B_2}}\;;\quad V_{min}(t_2) = \left\{n_2 + n_1(1-\sqrt{B_2})^2\right\}^{-1}$$

$$\gamma_3^{*} = \frac{\sqrt{B_3}}{1+\sqrt{B_3}}\;;\quad V_{min}(t_3) = \left\{n_2 - n_1 + \frac{2n_1}{(1+\sqrt{B_3})}\right\}^{-1}$$

where, $B_1 = \frac{\sum \theta_h}{P_h}$; $B_2 = \frac{\sum \theta_h'}{P_h}$; $B_3 = \frac{\sum \theta_h''}{P_h}$

In particular when $n_1 = n_2 = n$ we have the minimum variances as:

$$V_{min}(t_1) = (1+\sqrt{B_1}\,)\,V(y\,|\,z)(2n)^{-1}\;;$$

$$V_{min}(t_2) = \left[n\left\{1 + (1-\sqrt{B_2})^2\right\}\right]^{-1} V(y|z)$$

$$V_{min}(t_3) = (1+\sqrt{B_3})\, V(y|z)(2n)^{-1}$$

3. *Relative Efficiencies of the Proposed Strategies :*

To compare the efficiencies of the proposed strategies we note that if $V_o(T_i) \lesseqgtr V_o(T_j)$ for a fixed set of γ_h's ($h=1, \ldots, L$) we have $V_o(t_i) \lesseqgtr V_o(t_j)$ for $i \neq j = 1, 2, 3$. Thus comparing $V_o(T_1)$ with $V_o(T_2)$ we note that the strategy 1 is superior or inferior to strategy 2 according as $\gamma_h >$ or $< 1 - \theta'_h/\theta_h \ \forall\ h = 1, \ldots, L$. Similarly strategy 1 is superior or inferior to strategy 3 according as $\theta_h <$ or $> \theta''_h \ \forall\ h = 1, \ldots, L$. Strategy 3 is superior to strategy 2 since we have assumed $m_h \geq 1 \ \forall\ h$ and hence $Em_h = n_1 P_h \gamma_h \geq 1 \ \forall\ h = 1, \ldots, L$. Raj [5], Chaudhuri and Arnab [1], Ghangurde and Rao [4], Chotai [2], considered the strategies (to be denoted respectively as 0, 4, 5, 6) of sampling over two occasions with $n_1 = n_2 = n$. The expressions for the variances for their estimators of Y denoted by $V(t_o)$, $V(t_4)$, $V(t_5)$, $V(t_6)$ respectively are

$$V(t_o) = \left[1 + \left\{2(1-\rho)\right\}^{1/2}\right] V(y|z)\,(2n)^{-1}$$

$$V(t_4) = \left[1 + \left\{1 - \delta^2(y, x|z)\right\}^{1/2}\right] V(y|z)\,(2n)^{-1}$$

$$V(t_5) = N\left[\frac{1 - n\left\{2(1-\rho)(1+\beta n/N)\right\}^{1/2}}{N}\right] V(y|z)$$

$$V(t_6) = N\left[1 - \frac{n}{N} + 2\left\{1 - \delta(y, x|z)\right\}^{1/2}\right] \frac{V(y|z)}{(2n)(N-1)}$$

where
$$\rho = \sum_{i=1}^{N} \frac{(y_i - \overline{Y})(x_i - \overline{x})}{\left\{\sum_1^N (y_i - \overline{Y})^2 \sum_1^N (x_i - \overline{x})^2\right\}^{1/2}}$$

$$\beta = N\left\{1 - \delta(y, x|z)\right\} \sum (y_i - \overline{Y})^2 \left\{(1-\rho)V(y|z)\right\}^{-1} - 1$$

$$\overline{Y} = \sum_1^N \frac{Y_i}{N}, \qquad \overline{x} = \sum_1^N \frac{x_i}{N}$$

The expression for $V(t_5)$ was obtained by Chotai [2]. Chaudhury and Arnab modified Raj's estimator and it is of the form $t_4 = \varphi T_4 + (1-\varphi) T$

$$\text{Where } T_4 = \sum_{s_m} \frac{y_i}{mp_i} - \delta(x, y \mid z) \left\{ \sum_{s_m} \frac{x_i}{mp_i} - \sum_{s_1} \frac{y_i}{np_i} \right\} \left\{ \frac{V(y \mid z)}{V(x \mid z)} \right\}^{1/2}$$

and $\sum_{s_m}$ denotes the sum over the matched sample.

Now if we put $c_{h1} = \delta(y,x \mid z)\ \forall\ h$ and $\gamma_h = \frac{m}{n}$, $n_1 = n_2 = n$ we have

$$V(T_1) = \left(\frac{1}{m} - \frac{1}{n}\right) \sum_h \left[\sum_j \frac{\{y_{hj} - \delta(y, x \mid z)\, x_{hj}\}^2}{P_{hj}} - \frac{\{Y_h - \delta(y, x \mid z)\, X_h\}^2}{P_h} \right] + \frac{V(y \mid z)}{n}$$

$$\leq \left(\frac{1}{m} - \frac{1}{n}\right) \sum_h \left[\sum_j \frac{\{y_{hj} - \delta(y, x \mid z)\, x_{hj}\}^2}{P_{hj}} - \{Y - \delta(x, y \mid z)\, X\}^2 \right] + \frac{V(y \mid z)}{n} = V(T_4)$$

Hence the proposed strategy 1 is better than Chaudhuri and Arnab's [1] strategy. It was already shown by Chaudhuri and Arnab that their strategy is better than Raj's [5] strategy. Hence strategy 1 is better than Raj's strategy as well. Thus we can always improve on Raj's and Chaudhuri and Arnab's strategy by (i) stratifying the initial sample S_1 (ii) taking matched samples from respective strata by proportional allocation and (iii) using the estimator t_1 on taking $c_{h1} = \delta(x, y \mid z)$ in $T_1(h)$ for $h = 1, \ldots, L$.

Taking some numerical data on the yields of barley and maize in two successive years treated as x and y and the area under the crops as z we checked that for several combinations of the parameters involved, the strategy 3 fares the best and strategy 1 better than the strategies 4 - 6. Details are easy to check and hence omitted to save space.

ACKNOWLEDGEMENT

The author is grateful to the referee and Dr. Arijit Chaudhuri for their fruitful suggestions.

REFERENCES

[1] Chaudhuri, Arijit and Arnab, R., 1979. On estimating of a finite population sampled over two occasions with varying probabilities. *Aust. Jour. Statist.* **21**, 162-165.

[2] Chotai, J., 1974. A note on Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhya* **36**, Ser. C, 173-180.

[3] Cochran, W.G., 1963. Sampling Techniques. New-York Wiley.

[4] Ghangurde, P.D. and Rao, J.N.K., 1969. Some results on sampling over two occasions. *Sankhya*, A, **31**, 463-472.

[5] Raj, Des, 1965. Sampling over two occasions with probability proportionate to size. *Ann. Math. Statist.* **36**, 327-330.

[6] Rao, J.N.K., 1973. On double sampling for stratification and analytical surveys. *Biometrika*, **60**, 125-133.

[7] Rao, J.N.K., Hartley, H.O. and Cochran, W.G., 1962. On a simple procedure of unequal probability sampling without replacement. *Jour. Roy. Statist. Soc.*, B, **24**, 482-491.