J. Ind. Soc. Agril. Statist.
59(3), 2005 : 187–191

Two-phase Sampling in Multiple Frame Surveys

Anil Rai, B.C. Saxena and Vijay Bindal
Indian Agricultural Statistics Research Institute, New Delhi
(Received : March, 2000)

SUMMARY

In the present paper an attempt has been made to estimate the domain sizes with minimum variance with the help of first phase samples of two-phase sampling design. Based on the estimated domain size of the overlapping domain from the first phase sample, an estimator of population total has been proposed. The relative efficiency of the proposed estimator for a given cost of survey has been examined with respect to usual Hartley estimator of unknown domain sizes. The performance of the proposed estimator is found to be better than usual estimator for most of the parametric combinations.

Key words: Two-phase sampling, Optimum value, Overlapping frames.

1. INTRODUCTION

Let N_A and N_B be the sizes of two overlapping frames A and B such that these taken together cover the entire population. Let n_A and n_B denotes the independent random samples from frame A and B respectively. In this case there are three mutually exclusive domains i.e. domain (a) of size N_a, domain (b) of size N_b and domain (ab) of size N_{ab}. In case domain sizes are known, the estimator of the population total, say Y is

$$\hat{Y} = N_a \bar{y}_a + N_{ab}(p\bar{y}_{ab} + q\bar{y}_{ba}) + N_b \bar{y}_b \qquad (1.1)$$

where, $\bar{y}_a, \bar{y}_{ab}$ are means of domains (a) and (ab) based on sample from A frame. Similarly, $\bar{y}_b, \bar{y}_{ba}$ are the sample means based on sample from frame B corresponding to domains (b) and (ab) respectively. Here, p and q denote weights (p+q = 1) attached to the sample estimates of the domain (ab) obtained from samples of the frame A and B respectively. The variance of the above estimator after ignoring f.p.c. is given by

$$V(\hat{Y}) = \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + p^2\sigma_{ab}^2\alpha\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta) + q^2\sigma_{ba}^2\beta\right] \qquad (1.2)$$

where, σ_a^2, σ_b^2 and σ_{ab}^2 are domain variances, $\alpha = N_{ab}/N_A$ and $\beta = N_{ab}/N_B$.

The optimum value of p i.e. p_o is obtained subject to the following fixed cost function

$$C = C_A n_A + C_B n_B \qquad (1.3)$$

where C is the total cost, C_A and C_B are the costs of sampling in frame A and B respectively such that $C_A > C_B$.

The problems of dual frame surveys in which one frame is complete but expensive and the other is incomplete but cheaper for sampling has been considered by Hartley (1962, 1974), Cochran (1964, 1965). Steinberg used this technique in case of rare elements in the population whereas Lund (1968) suggested an alternative approach of estimation. The practical application of dual frame surveys has been demonstrated by Vogel (1975), Serrurier and Phillips (1976), Armstrong (1979) etc. The problem of non response in multiple frame surveys using double sampling has been considered by Rao (1968). Saxena *et al.* (1984, 1985, 1986) considered the theory of multiple frame for two stage sampling design as well as multiple character studies. Recently, Skinner and Rao (1996) used maximum likelihood technique in multiple frame surveys.

In case of unknown domain sizes the estimator of the population total takes the form

$$\hat{Y}_H = \frac{N_A}{n_A}(y_a + p_H y_{ab}) + \frac{N_B}{n_B}(y_b + q_H y_{ba}) \qquad (1.4)$$

where p_H and q_H are weights such that $p_H + q_H = 1$. y_a and y_b are sample totals for domains (a) and (b) respectively. The variance expression of the above estimator is

$$V(\hat{Y}_H) = \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + p_H^2\sigma_{ab}^2\alpha + \alpha(1-\alpha)(\overline{Y}_a - p_H\overline{Y}_{ab})^2\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta) + q_H^2\sigma_{ba}^2\beta + \beta(1-\beta)(\overline{Y}_b - q_H\overline{Y}_{ab})^2\right] \qquad (1.5)$$

Here, $\overline{Y}_a$, $\overline{Y}_b$ and $\overline{Y}_{ab}$ are population means of respective domains. The optimum value of p_H i.e. p_{H_0} and the cost function (1.3) can be obtained by solving a bi-quadratic

$$\sqrt{\frac{C_B}{C_A}}\left[\frac{(1-\alpha)\sigma_a^2 + p_H^2\alpha\sigma_{ab}^2 + \alpha(1-\alpha)(\overline{Y}_a - p_H\overline{Y}_{ab})^2}{(1-\beta)\sigma_b^2 + q_H^2\beta\sigma_{ab}^2 + \beta(1-\beta)(\overline{Y}_b - q_H\overline{Y}_{ab})^2}\right]^{1/2} = \left[\frac{(\sigma_{ab}^2 + (1-\alpha)\overline{Y}_{ab}^2)p_H - (1-\alpha)\overline{Y}_a\overline{Y}_{ab}}{(\sigma_{ab}^2 + (1-\beta)\overline{Y}_{ab}^2)q_H - (1-\beta)\overline{Y}_b\overline{Y}_{ab}}\right] \qquad (1.6)$$

In the present paper an attempt has been made to estimate the domain sizes with minimum variance with the help of first phase samples of two phase sampling design. Based on the estimated domain size of the overlapping domain from the first phase samples of both the frames i.e. A and B, an estimator of population total has been proposed. The relative efficiency of the proposed estimator for a given cost of survey has been examined with respect to usual estimator of unknown domain sizes.

2. ESTIMATION OF DOMAIN SIZES

Let simple random first phase sample from frame A i.e. n'_A consists of n'_a units from domain (a) and n'_{ab} units from overlapping domain (ab). Similarly n'_B consists of n'_b units from domain (b) and n'_{ba} units from overlapping domain (ab). The second phase simple random samples which were drawn from the first phase samples of the respective frames are denoted by n_A and n_B. Further, these second phase samples also consists of units from different domains as in the case of first phase samples.

Mathematically

$$n'_A = n'_a + n'_{ab} \qquad n'_B = n'_b + n'_{ba}$$

and

$$n_A = n_a + n_{ab} \qquad n_B = n_b + n_{ba}$$

The estimator of overlapping domain based on first phase samples of both the frames can be obtained as

$$\hat{N}'_{ab} = \frac{n'_{ab}}{n'_A}N_A \text{ and } \hat{N}'_{ba} = \frac{n'_{ba}}{n'_B}N_B$$

respectively. A common weighted estimator of overlapping domain size and its variance can be written as

$$\tilde{N}_{ab} = p^*\hat{N}'_{ab} + q^*\hat{N}'_{ba} \qquad (2.1)$$

and

$$V(\tilde{N}_{ab}) = p^{*2}\frac{N_A^2}{n'_A}\alpha(1-\alpha) + q^{*2}\frac{N_B^2}{n'_B}\beta(1-\beta) \qquad (2.2)$$

where p^* and q^* are weights such that $p^* + q^* = 1$. The optimum value of p^* i.e. p_o^* can be obtained by minimizing the above variance

$$p_o^* = \frac{N_B^2\beta(1-\beta)/n'_B}{N_A^2\alpha(1-\alpha)/n'_A + N_B^2\beta(1-\beta)/n'_B} = \frac{(N_B - \tilde{N}_{ab})n'_A}{(N_A - \tilde{N}_{ab})n'_B + (N_B - \tilde{N}_{ab})n'_A} \qquad (2.3)$$

Substituting the optimum value of p^* we arrive at a quadratic in $\tilde{N}_{ab}$ of the form

$$\tilde{N}_{ab}^2(n'_A + n'_B) - \tilde{N}_{ab}[n'_A(N_B + \hat{N}'_{ba}) + n'_B(N_A + \hat{N}'_{ab})] + (N_A n'_B \hat{N}'_{ba} + N_B n'_A \hat{N}'_{ab}) = 0$$

Thus $\tilde{N}_{ab}$ can be obtained.

The domain sizes of (a) and (b) can be written as

$$\tilde{N}_a = N_A - \tilde{N}_{ab}^*$$

$$\tilde{N}_b = N_B - \tilde{N}_{ab}^* \qquad (2.4)$$

where $\tilde{N}_{ab}^*$ is optimum estimated size of overlapping domain.

3. THE PROPOSED ESTIMATOR

The estimator of population total based on two phase sampling design, when domain sizes were estimated with the help of first phase sample (say, $\hat{Y}_{HD}$) can be written as

$$\hat{Y}_{HD} = \tilde{N}_a\bar{y}_a + \tilde{N}_b\bar{y}_b + \tilde{N}^*_{ab}[w\bar{y}_{ab} + (1-w)\bar{y}_{ba}] \quad (3.1)$$

Here, w is the weight attached to the estimator of mean of the overlapping domain based on sample of frame A. Its variance can be written as

$$V(\hat{Y}_{HD}) = \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + w^2\sigma_{ab}^2\alpha\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta) + (1-w)^2\sigma_{ab}^2\beta\right] + \left[p_o^{*2}\frac{N_A^2}{n'_A}\alpha(1-\alpha) + q_o^{*2}\frac{N_B^2}{n'_B}\beta(1-\beta)\right] + \left[(\bar{Y}_a - w\bar{Y}_{ab})^2 + \{\bar{Y}_b - (1-w)\bar{Y}_{ab}\}^2\right] \quad (3.2)$$

A simple linear function in case of two phase sampling is

$$C = C'_A n'_A + C_A n_A + C'_B n'_B + C_B n_B \quad (3.3)$$

where, C'_A and C'_B are per unit cost of sampling in the first phase. The expression for the optimum value of w under the above cost function takes a complicated form. For simplicity, a realistic assumption viz. the per unit cost ratio of the second phase sampling with respect to first phase sampling has been treated as same and greater than one in both the frames. Mathematically

$$\frac{C_A n_A}{C'_A n'_A} = \frac{C_B n_B}{C'_B n'_B} = \delta \quad \text{(say)} \quad (3.4)$$

Under this assumption the variance expression and cost function reduces to

$$V(\hat{Y}_{HD}) = \frac{N_A^2}{n_A}\left[\sigma_a^2(1-\alpha) + w^2\sigma_{ab}^2\alpha + p_o^{*2}\,\Delta_A\,\psi\alpha(1-\alpha)\right] + \frac{N_B^2}{n_B}\left[\sigma_b^2(1-\beta) + (1-w)^2\sigma_{ab}^2\beta + q_o^{*2}\Delta_B\,\psi\,\beta(1-\beta)\right] \quad (3.5)$$

and

$$C = \Delta[C_A n_A + C_B n_B] \text{ respectively} \quad (3.6)$$

where

$$\psi = (\bar{Y}_a - w\,\bar{Y}_{ab})^2 + \{\bar{Y}_b - (1-w)\bar{Y}_{ab}\}^2$$

$$\Delta = \left(1 + \frac{1}{\delta}\right) \qquad \Delta_A = \frac{\delta C'_A}{C_A} \qquad \Delta_B = \frac{\delta C'_B}{C_B}$$

Now the optimum value of w i.e. w_o can be obtained with respect to the reduced cost function resulting in a bi-quadratic equation of the form

$$\sqrt{\frac{C_A}{C_B}}\frac{\left[(1-\beta)\sigma_b^2 + \beta w'^2\sigma_{ab}^2\right]^{1/2}}{\left[(1-\alpha)\sigma_a^2 + \alpha w^2\sigma_{ab}^2\right]^{1/2}} = \frac{w'\sigma_{ab}^2 - q_o^{*2}\Delta_B(1-\beta)\bar{Y}_{ab}\{(\bar{Y}_a - \bar{Y}_b) + (w'-w)\bar{Y}_{ab}\}}{w\sigma_{ab}^2 - p_o^{*2}\Delta_A(1-\alpha)\bar{Y}_{ab}\{(\bar{Y}_a - \bar{Y}_b) + (w'-w)\bar{Y}_{ab}\}} \quad (3.7)$$

where $w + w' = 1$. The optimum sample sizes n_{ADo} and n_{BDo} from both the frames A and B respectively are

$$n_{ADo} = N_A K_{Do}\sqrt{\frac{(1-\alpha)\sigma_a^2 + \alpha w_o^2\sigma_{ab}^2 + p_o^{*2}\Delta_A\psi\alpha(1-\alpha)}{C_A}}$$

$$n_{BDo} = N_B K_{Do}\sqrt{\frac{(1-\beta)\sigma_b^2 + \beta w_o'^2\sigma_{ab}^2 + q_o^{*2}\Delta_B\psi\beta(1-\beta)}{C_B}}$$

where K_{Do} is a constant that can be determined.

Let, $\rho = \frac{C_A}{C_B}$ $\quad \rho_1 = \frac{C_A}{C'_A}$ $\quad \rho_2 = \frac{C_B}{C'_B}$ $\quad \phi_1 = \frac{\sigma_a}{\sigma_{ab}}$

$\phi_2 = \frac{\sigma_b}{\sigma_{ab}}$ $\quad \phi_3 = \frac{\sigma_{ab}}{\bar{y}_{ab}}$ $\quad \phi_4 = \frac{\sigma_a}{\bar{y}_a}$ $\quad \phi_5 = \frac{\sigma_b}{\bar{y}_b}$

$R_N = \frac{N_A}{N_B}$ and $R_n = \frac{n_A}{n_B}$

The ratio of the variances given by the equations (1.5) and (3.5) after substitution is

$$\frac{V(\hat{Y}_H)}{V(\hat{Y}_{HD})}$$

$$=\frac{K_{Do}}{K_{Ho}}\frac{\sqrt{(1-\alpha)\phi_1^2+p_{Ho}^2\alpha+\alpha(1-\alpha)\left(\frac{\phi_1\phi_3}{\phi_4}-p_{Ho}\right)^2\Big/\phi_3^2}}{\sqrt{(1-\alpha)\phi_1^2+\alpha w_o^2+p_o^{*2}\Delta_A\psi'\beta(1-\beta)/\phi_3^2}}$$

$$\frac{+\frac{\beta}{\alpha\sqrt{\rho}}\sqrt{(1-\beta)\phi_2^2+q_{Ho}^2\beta+\beta(1-\beta)\left(\frac{\phi_2\phi_3}{\phi_5}-q_{Ho}\right)^2\Big/\phi_3^2}}{+\frac{\beta}{\alpha\sqrt{\rho}}\sqrt{(1-\beta)\phi_2^2+\beta w_o'^2+q_o^{*2}\psi'\beta(1-\beta)/\phi_3^2}}$$

where

$$\psi'=\left(\frac{\phi_1\phi_3}{\phi_4}-w\right)^2+\left(\frac{\phi_2\phi_3}{\phi_5}-w'\right)^2$$

The ratio K_{Do}/K_{Ho} can be obtained with the help of the cost function (3.6). So, the above ratio can be written as

$$\frac{V(\hat{Y}_H)}{V(\hat{Y}_{HD})}$$

$$=\frac{1}{\Delta}\left[\frac{\sqrt{(1-\alpha)\phi_1^2+p_{Ho}^2\alpha+\alpha(1-\alpha)\left(\frac{\phi_1\phi_3}{\phi_4}-p_{Ho}\right)^2\Big/\phi_3^2}}{\sqrt{(1-\alpha)\phi_1^2+\alpha w_o^2+p_o^{*2}\Delta_A\psi'\beta(1-\beta)/\phi_3^2}}\right.$$

$$\left.\frac{+\frac{\beta}{\alpha\sqrt{\rho}}\sqrt{(1-\beta)\phi_2^2+q_{Ho}^2\beta+\beta(1-\beta)\left(\frac{\phi_2\phi_3}{\phi_5}-q_{Ho}\right)^2\Big/\phi_3^2}}{+\frac{\beta}{\alpha\sqrt{\rho}}\sqrt{(1-\beta)\phi_2^2+\beta w_o'^2+q_o^{*2}\psi'\beta(1-\beta)/\phi_3^2}}\right]^{1/2} \quad (3.8)$$

The above ratio is the relative efficiency of the proposed estimator with respect to Hartley's estimator of unknown domain sizes. It can be seen that theoretically comparison of these estimators is not possible. Hence, in the following section the relative efficiency has been compared empirically.

4. THE RELATIVE EFFICIENCY

The relative efficiency of the estimator of the population total based on two phase sampling approach i.e. $\hat{Y}_{HD}$, when the domain sizes are not known is calculated by assigning realistic values to the parameter involved in the equation (3.8). The values assigned to different parameters are

α = 0.10, 0.20, 0.30, 0.40, 0.50
β = 0.20, 0.40, 0.60, 0.80
δ = 2.00, 4.00, 6.00, 8.00, 10.0
ρ = 3.00, 9.00, 12.00
ρ_1 = 2.00, 4.00, 12.00
ρ_2 = 2.00, 4.00, 12.00
ϕ_1 = 0.25, 0.50, 1.00, 2.00, 4.00
ϕ_2 = 0.25, 0.50, 1.00, 2.00, 4.00
ϕ_3 = 0.05, 0.10, 0.15, 0.20, 0.25
ϕ_4 = 0.05, 0.10, 0.15, 0.20, 0.25
ϕ_5 = 0.05, 0.10, 0.15, 0.20, 0.25
R_N = 1.00, 3.00, 5.00
R_n = 1.00, 0.33, 0.20

The relative efficiency of the proposed estimator for estimation of the population total in case of unknown domain sizes were calculated and for some of the parametric combinations are presented in Table 1. This can be observed that the relative efficiency of the proposed estimator is greater than one in most of the feasible parametric combinations. It is seen that the relative efficiency increases when $N_A > N_B$ i. e. the size of frame A is greater than frame B. It is also observed that efficiency is higher when coefficient of variation of domain 'a' is larger and that of overlapping domain 'ab' is small. Further, it is seen that if C.V. of domain 'a' is small and C.V. of domain 'ab' is large, the efficiency is slightly of lower order. One of the limitations in this empirical investigation has been that for some of the parametric combinations no feasible solutions in either of the two bi-quadratics are available in the admissible range of weights. Hence, for those few cases it is not possible to find out the relative efficiencies.

Table 1. Relative efficiency of the proposed two phase sampling estimator with respect to usual estimator when domain sizes are unknown

$\phi_1 = 0.50, \quad \phi_2 = 0.50, \quad \rho = 3, \quad \rho_1 = \rho_2 = 12, \quad \delta = 6$

	N_A/N_B	1			3			5		
	n_A/n_B	1	.33	.20	1	.33	.20	1	.33	.20
ϕ_4	ϕ_5	$\phi_3 = .05$								
.05	.05	1.14	1.36	1.62	1.49	1.49	1.49	1.49	1.49	1.49
	.15	1.14	1.36	1.44	1.36	1.40	1.41	1.40	1.41	1.41
	.25	1.11	1.32	1.42	1.34	1.39	1.39	1.39	1.40	1.40
.15	.05	1.87	2.07	2.20	2.09	2.12	2.12	2.11	2.12	2.12
	.15	1.42	1.74	1.97	1.96	2.04	2.05	2.04	2.05	2.05
	.25	1.43	1.75	1.85	1.93	2.02	2.04	2.02	2.04	2.04
.25	.05	1.89	2.13	2.25	2.19	2.22	2.22	2.22	2.22	2.22
	.15	1.44	1.78	2.03	2.05	2.14	2.16	2.14	2.16	2.16
	.25	1.46	1.80	2.00	2.03	2.13	2.14	2.12	2.15	2.15
		$\phi_3 = .15$								
.05	.05	1.22	1.48	1.58	1.67	1.73	1.74	1.73	1.74	1.74
	.15	1.27	1.52	1.61	1.68	1.73	1.74	1.73	1.74	1.74
	.25	1.26	1.50	1.60	1.68	1.73	1.74	1.73	1.74	1.74
.15	.05	-	0.98	1.01	1.03	1.73	1.74	1.73	1.74	1.74
	.15	0.91	1.00	1.02	1.03	1.73	1.74	1.73	1.74	1.74
	.25	-	0.98	1.01	1.03	1.73	1.74	1.73	1.74	1.74
		$\phi_3 = .25$								
.05	.05	1.22	1.50	1.63	1.72	1.73	1.74	1.73	1.74	1.74
	.15	-	1.57	1.68	1.70	1.73	1.74	1.73	1.74	1.74
	.25	-	1.57	1.67	1.69	1.73	1.74	1.73	1.74	1.74
.15	.05	0.91	1.06	1.00	1.04	1.73	1.74	1.73	1.74	1.74
	.15	1.08	1.14	1.16	1.14	1.73	1.74	1.73	1.74	1.74
	.25	1.06	1.13	1.16	1.11	1.73	1.74	1.73	1.74	1.74

REFERENCES

Armstrong, B. (1979). Test of multiple frame sampling techniques for agricultural surveys. *Survey Methodology,* **5(2)**, 178 - 199.

Cochran, R.S. (1964). Multiple frame surveys. *Proc. Amer. Statist. Assoc. Soc. Statist. Sec.*, Chicago, Illinos.

Cochran, R.S. (1965). Theory and applications of multiple frame surveys. *Ph.D. Dessertation*, Iowa State University, Iowa.

Hartley, H.O. (1962). Multiple frame surveys. *Proc. Amer. Statist. Assoc. Soc. Statist. Sec.*, 203 - 208.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhya*, **C36**, 99 - 118.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proc. Amer. Statist. Assoc.Soc. Statis. Sec.*, 282 - 288.

Rao, J.N.K. (1968). Some non response sampling theory when the frame contains an unknown amount of duplication. *J. Amer. Statist. Assoc.*, **63(321)**, 87 - 90.

Serrurier, D. and Phillips, J.E. (1976). Double frame Ontario pilot hog surveys. *Survey Methodology*, **2(2)**.

Saxena, B.C., Narain, P. and Srivastava, A.K. (1984). Multiple frame surveys in two stage sampling. *Sankhya*, **B46**, 75 - 82.

Saxena, B.C., Narain, P. and Srivastava, A.K. (1985). Estimation of total for two characters in multiple frame surveys. *Survey Methodology Statistics*, Canada, **11(2)**, 119 - 132.

Saxena, B.C., Narain, P. and Srivastava, A.K. (1986). Robustness of Hartley's estimator for multiple frame surveys. *J. Ind. Soc. Agril. Statist.*, **38(3)**, 370 - 378.

Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *J. Amer. Statist. Assoc.*, **91**, 349 - 356.

Vogel, F.A. (1975). Surveys with overlapping frames - Problems in applications. Presented at the *135th Annual Meeting of American Statistical Association*, Atlanta, Georgia.