Sankhyā : The Indian Journal of Statistics
1988, Volume **50**, Series B, Pt. 1, pp. 103–110.

ESTIMATION FOR DOMAINS IN SAMPLING ON TWO OCCASIONS

By T. P. TRIPATHI

Indian Statistical Institute

SUMMARY. For a strategy of sampling to estimate domain and over-all population means, on a current occasion using sampled data on the current and a previous occasion, optimal matching policies are investigated vis-a-vis specific competitive alternatives.

1. Introduction

In practice, the problem of estimating domain means, totals and proportions etc. on various occasions assumes considerable importance. The problem of estimating sub-population parameters on successive occasions arises mainly due to

(i) change in the value, for various units in the population (hence in the domains), of the character(s) under study from one point of time to the other, or

(ii) change in the number of units in domains, or

(iii) change in the domain structure from one occasion to the other i.e. entry (or exit) of some units into (or from) a domain, or

(iv) any two or more changes listed above in (i) to (iii).

In this paper we consider, for a finite population, the problem of simultaneous estimation of K-domain means and the over-all mean on a current occasion using a partial replacement scheme.

2. Proposed sampling strategy

Let $U = \{1, \ldots, N\}$ be a finite population of N units and y a real variable defined on U. Let $\{D_1, \ldots, D_k\}$ be a partition of U on the second occasion into domains D_i consisting of N_{hi} unknown units on occasion h ($h = 1, 2$), $\sum_{i=1}^{k} N_{hi} = N$. Let y_{hij} denote the value of y for j-th unit in the domain D_i and

$$\bar{Y}_{hi} = \sum_{j=1}^{N_{hi}} y_{hij}/N_{hi}$$

the mean of D_i on occasion $h = 1, 2$.

AMS (1980) *subject classification :* 62D05.

Key words : Domains of study, Scheme of partial replacement, Sampling on two occasions, Estimation of means.

For the simultaneous estimation of domain means $\bar{Y}_{2i}$ $(i = 1, \ldots, k)$ and the population mean on the second occasion, we propose the following sampling procedure.

On the first occasion an SRSWOR S_1 of size n is drawn from U. On the second occasion, the sample $S_2 = (S_{2m}, S_{2u})$ is drawn with S_{2m} (matched) of m units from S_1 as an SRSWOR and S_{2u} (unmatched or replaced) of u units independently from U as another SRSWOR.

Let $m_{2i}\,(\neq 0)$ and $u_{2i}\,(\neq 0)$ be the number of units in S_{2m} and S_{2u} respectively coming from D_i ; $\sum_{i=1}^{k} m_{2i} = m$, $\sum_{i=1}^{k} u_{2i} = u$.

$$\text{Let} \qquad \bar{y}_{m_{hi}} = \sum_{j=1}^{m_{hi}} y_{hij}/m_{hi},\ h = 1, 2,\ \bar{y}_{u_{2i}} = \sum_{j=1}^{u_{2i}} y_{hij}/u_{2i} \qquad \ldots \quad (2.1)$$

$$\bar{y}_{n_{1i}} = \sum_{j=1}^{n_{1i}} y_{1ij}/n_{1i}\ ;$$

where $n_{1i}\,(\neq 0)$ and $m_{1i}\,(\neq 0)$ are the number of units in S_1 and S_{2m} respectively which belonged to D_i on the first occasion.

For estimating $\bar{Y}_{2i}$, we propose an estimator

$$\bar{y}_{2i} = w\bar{y}_{mi} + (1-w)\bar{y}_{u_{2i}} \qquad \ldots \quad (2.2)$$

where $\bar{y}_{mi} = \bar{y}_{m_{2i}} - (\bar{y}_{m_{1i}} - \bar{y}_{n_{1i}})$ with w as a suitably chosen weight.

To derive variance expressions we define

$$d_{hij} = \begin{cases} 1, & \text{if } j\text{-th unit falls in } D_i \text{ on occasion } h, \quad h = 1, 2 \\ 0, & \text{otherwise} \end{cases}$$

and $t_{hij} = d_{hij}\, y_{hij}$ giving

$$\bar{y}_{m_{2i}} = \bar{t}_{2im}/\bar{d}_{2im}; \qquad \bar{y}_{u_{2i}} = \bar{t}_{2iu}/\bar{d}_{2iu}$$

$$\bar{y}_{m_{1i}} = \bar{t}_{1im}/\bar{d}_{1im}; \qquad \bar{y}_{n_{1i}} = \bar{t}_{1in}/\bar{d}_{1in}$$

where $\bar{t}_{him} = \sum_{j=1}^{m} d_{hij}\, y_{hij}/m; \quad \bar{d}_{him} = \sum_{j=1}^{m} d_{hij}/m,\ h = 1, 2$

$$\bar{t}_{2iu} = \sum_{j=1}^{u} d_{2ij}\, y_{2ij}/u; \quad \bar{d}_{2iu} = \sum_{j=1}^{u} d_{2ij}/u$$

$$\bar{t}_{1in} = \sum_{j=1}^{n} d_{1ij}\, y_{1ij}/n; \quad \bar{d}_{1in} = \sum_{j=1}^{n} d_{1ij}/n.$$

Replacing $N/(N-1)$ by unity, to the terms of order m^{-1} and u^{-1} we get

$$V(\bar{y}_{u2i}) = \left(\frac{1}{u} - \frac{1}{N}\right) A_i$$

$$V(\bar{y}_{mi}) = \left(\frac{1}{m} - \frac{1}{N}\right) A_i - \left(\frac{1}{m} - \frac{1}{n}\right) B_i \qquad \ldots \quad (2.3)$$

where $A_i = (N/N_{2i})\,\sigma_{2i}^2$

$$B_i = \frac{2NN_i^*}{N_{1i}N_{2i}}[\rho_i\sigma_{1i}^*\sigma_{2i}^* + (\bar{Y}_{1i}^* - \bar{Y}_{1i})(\bar{Y}_{2i}^* - \bar{Y}_{2i})] - \frac{N}{N_{1i}}\sigma_{1i}^2$$

with N_i^*: number of i-th domain units common to the first and second occasions.

$$\bar{Y}_{hi}^* = \sum_{j}^{N_i^*} y_{hij}/N_i^*; \quad \sigma_{hi}^{*2} = \sum_{j=1}^{N_i^*} (y_{hij} - \bar{Y}_{hi}^*)^2/N_i^*$$

$$\sigma_{12i}^* = \sum_{j=1}^{N_i^*} (y_{1ij} - \bar{Y}_{1i}^*)(y_{2ij} - \bar{Y}_{2i}^*)/N_i^*; \quad \rho_i = \sigma_{12i}^*/\sigma_{1i}^*\sigma_{2i}^*$$

$$\sigma_{hi}^2 = \sum_{j=1}^{N_{hi}} (y_{hij} - \bar{Y}_{hi})^2/N_{hi}.$$

Assuming that the sample size is same on both the occasions, i.e. $m+u=n$, and using optimum weight w, which minimizes the variance of $\bar{y}_{2i}$, we find, using (2.3), that

$$V(\bar{y}_{2i}) = \left(\frac{1}{n}\right) \frac{(1-f+\lambda f)A_i[(1-\lambda f) - (1-\lambda)B_i/A_i]}{(1-2\lambda f + 2\lambda^2 f) - (1-\lambda)^2 B_i/A_i} \qquad \ldots \quad (2.4)$$

where $f = n/N$ is the sampling fraction and $\lambda = m/n$ is the matched proportion.

3. Optimum matched proportion

For the optimal matching proportion, viz.,

$$\lambda_0^{(i)} = \frac{[1-B_i/A_i]^{1/2}}{1+[1-B_i/A_i]^{1/2}} \qquad \ldots \quad (3.1)$$

the minimal resulting variance is

$$V_0(\bar{y}_{2i}) = \frac{1}{2n} A_i[1-f+(1-B_i/A_i)^{1/2}] \qquad \ldots \quad (3.2)$$

In case the structure of the domain D_i remains unchanged on the two occasions, we may assume that

$$N_{hi} = N_i^*; \quad \bar{Y}_{hi} = \bar{Y}_{hi}^*, \sigma_{hi} = \sigma_{hi}^* \text{ for each } h = 1, 2. \qquad \ldots \quad (3.3)$$

Further assuming that

$$\sigma_{1i}^* = \sigma_{2i}^* = \sigma_i^* \text{ (say)} \qquad \ldots \quad (3.4)$$

(3.1) and (3.2) reduce to

$$\lambda_0^{(i)} = \frac{\sqrt{2}\sqrt{1-\rho_i}}{1+\sqrt{2}\sqrt{1-\rho_i}} \qquad \ldots \quad (3.5)$$

$$V_0(\bar{y}_{2i}) = \frac{1}{2n}\frac{N}{N_i^*}\sigma_i^{*2}[1-f+\sqrt{2}\sqrt{1-\rho_i}]. \qquad \ldots \quad (3.6)$$

It may be seen from (2.3) that $A_i - B_i \geqslant 0$ giving $B_i/A_i \leqslant 1$. However B_i/A_i may either be positive or negative. In case $B_i/A_i > 0$, the $\lambda_0^{(i)}$ in (3.1) cannot exceed $\frac{1}{2}$. Thus for $\rho_i \geqslant \frac{1}{2}$ in (3.5), we have $\lambda_0^{(i)} \leqslant \frac{1}{2}$.

Noting that ρ_i is the correlation coefficient between the first and second occasion values on those units of the domain D_i which are common to both the occasions, one may obtain, from (3.5), the optimum matched proportions $\lambda_0^{(i)}$ for estimating $\bar{Y}_{2i}$ corresponding to a specified value of ρ_i in $(-1, 1)$. In practice, simultaneous estimation of all domain means $\{\bar{Y}_{2i}\}$, $i = 1, \ldots, k$ may be required, in which case a compromise has to be made between various values $\lambda_0^{(1)}, \ldots, \lambda_0^{(k)}$. In case $\rho_1, \ldots, \rho_k$ do not differ appreciably from each other, same optimum matched proportion can be used for estimation of $\bar{Y}_{21}, \ldots, \bar{Y}_{2k}$. However, if ρ_i differ considerably from each other (especially if $\rho_i < 0$ for some i and $\rho_i > 0$ for other), a suitable compromise may be difficult and the above sampling strategy may not be suitable for simultaneous estimation of $\bar{Y}_{2i}$ for each $i = 1, \ldots, k$. The results in the next section provide guidelines in case either $\rho_i \leqslant \frac{1}{2}$ or $\rho_i > \frac{1}{2}$ for each or for some $i = 1, \ldots, k$.

4. Comparisons and discussion

Let the sampling scheme proposed in Section 2 based on partial matching $(0 < \lambda < 1)$ be denoted by $S_{(\lambda)}$ and the schemes based on complete matching $(\lambda = 1, u = 0)$ and complete replacement $(\lambda = 0, u = n)$ be denoted by $S_{(1)}$ and $S_{(0)}$ respectively. If n_{2i} and n_{2i}^* denote the number of units in $S_2 = \{S_{2m}\}_{m=n}$ and $S_2 = \{S_{2u}\}_{u=n}$ respectively coming from D_i, the estimators for $\bar{Y}_{2i}$ based on $S_{(1)}$ and $S_{(0)}$ may be taken as

$$\bar{y}_{2i(1)} = \frac{1}{n_{2i}}\sum_{j=1}^{n_{2i}} y_{2ij} = \bar{t}_{2in}/\bar{d}_{2in} \text{ and } \bar{y}_{2i(0)} = \frac{1}{n_{2i}^*}\sum_{j=1}^{n_{2i}^*} y_{2ij} = \bar{t}_{2in}/\bar{d}_{2in}$$

respectively with

$$V(\bar{y}_{2i(1)}) = V(\bar{y}_{2i(0)}) = \left(\frac{1}{n} - \frac{1}{N}\right) A_i \qquad \ldots \quad (4.1)$$

which is same as $V(\bar{y}_{2i})$ in (2.4) with $\lambda = 1$ or $\lambda = 0$.

If the sampling fraction f is ignored, from (2.4) and (4.1) we find that the sampling strategy $T_\lambda = (S_{(\lambda)}, \bar{y}_{2i})$ would be better than both of the strategies $T_1 = (S_{(1)}, \bar{y}_{2i(1)})$ and $T_0 = (S_{(0)}, \bar{y}_{2i(0)})$ provided $B_i > 0$. Thus if (3.3) and (3.4) are satisfied and $\rho_i > \frac{1}{2}$, for each $i = 1, \ldots, k$, one may use a suitable $\lambda_0^{(i)} \epsilon (0, 0.5)$ for simultaneous estimation of $(\bar{Y}_{21}, \ldots, \bar{Y}_{2k})$. In case $B_i \leqslant 0$, or in particular $\rho_i \leqslant \frac{1}{2}$, it would be advisable to use either of the strategies T_1 and T_0.

Table 4.1 gives the values of optimum matchings in (3.1) and the percent relative gain in efficiency (PRGE) of optimum matching over no matching, complete matching, 25% matching, 50% matching and 75% matching ($\lambda = 0$, 1, 1/4, 1/2 and 3/4 respectively) as obtained from

$$PRGE = \frac{V(\bar{y}_{2i}) - V_0(\bar{y}_{2i})}{V_0(\bar{y}_{2i})} \times 100$$

for $f = 0$ in (2.4) and (3.2) and for various values of B_i/A_i.

We note that, $\lambda_0^{(i)}$ is a monotonically decreasing function of B_i/A_i in (0, 1). The PRGE over $\lambda = \frac{1}{2}$ and $\lambda = \frac{3}{4}$ increases monotonically with $0 > B_i/A_i < 1$. However PRGE over $\lambda = \frac{1}{4}$ increases monotonically for $0 < B_i/A_i \leqslant \frac{1}{2}$ and decreases monotonically for $\frac{1}{2} < B_i/A_i < 1$.

TABLE 4.1. PRGE OF OPTIMUM MATCHING OVER OTHER MATCHINGS

B_i/A_i	optimum percent matched $100\,\lambda_0^{(i)}$	percent relative gain in efficiency of $\lambda_0^{(i)}$ over			
		$\lambda = 0$ or $\lambda = 1$	$\lambda = 1/4$	$\lambda = 1/2$	$\lambda = 3/4$
0.1	48.7	2.6	0.6	0	0.7
0.2	47.2	5.6	1.1	0	1.6
0.3	45.6	8.9	1.5	0	2.7
0.4	43.6	12.7	1.8	0.2	4.0
0.5	41.4	17.2	1.9	0.4	5.8
0.6	38.7	22.5	1.7	0.9	8.2
0.7	35.4	29.2	1.2	1.8	11.5
0.8	30.9	38.2	0.5	3.6	16.4
0.9	24.0	51.9	0	7.8	24.8

From Table 4.1 we observe that the PRGE of optimum matching over 50% matching is almost negligible for $0 \leqslant B_i/A_i \leqslant 0.6$ and moderate for $0.6 < B_i/A_i < 0.9$. Further the PRGE of optimum matching over 25% matching is quite small for $0 \leqslant B_i/A_i < 1$. Thus, if $0 < B_i/A_i < 1$ or $\frac{1}{2} < \rho_i < 1$ under (3.3) and (3.4), 25% percent of the units in the sample S_1 may be retained randomly in the sample S_2 for simultaneous estimation of all the domain means $\overline{Y}_{2i}$, $i = 1, \ldots, k$.

However, for $0 \leqslant B_i/A_i \leqslant 0.6$ or $0.5 \leqslant \rho_i \leqslant 0.8$ under (3.3) and (3.4), it may be preferable to retain 50% of the units in S_1 for S_2. As indicated earlier, if $B_i \leqslant 0$ or $\rho_i \leqslant \frac{1}{2}$ under (3.3) and (3.4), one may either retain all the units in S_1 for observing on the second occasion or draw a completely fresh sample.

Thus it is interesting to note that for the estimation of domain means $\overline{Y}_{2i}$ for some i or all $i = 1, \ldots, k$ on the second occasion, one need not bother, for optimum proportion to be matched and may either retain all the units or no unit in case $B_i \leqslant 0$, and retain 25% units, in case $0 < B_i/A_i < 1$.

It may also be shown that for $-\frac{1}{2} \leqslant B_i/A_i < 0$, i.e. $0.25 \leqslant \rho_i < 0.5$ under (3.3) and (3.4), the gain in efficiency by using no matching or 100% matching over 25% matching is not appreciable. Thus in practice if $\rho_i \geqslant 1/2$ for most of the domains and $\frac{1}{4} \leqslant \rho_i < 1/2$ for remaining ones, we may in fact adopt the proposed sampling strategy $T_\lambda = (S_{(\lambda)}, \bar{y}_{2i})$ with $\lambda = m/n = 0.25$. However, if $\rho_i < 1/2$ for most of $i = 1, \ldots, k$, it would be advisable to use either of the strategies

$$T_1 = (S_{(1)}, \bar{y}_{2i(1)}) \text{ and } T_0 = (S_{(0)}, \bar{y}_{2i(0)}).$$

5. Concluding remarks

In most situations of practical importance one is interested in not only estimating the domain means $\overline{Y}_{2i}$ $(i = 1, \ldots, k)$ but also the over-all mean $\overline{Y}_2 = \sum_{j=1}^{N} y_{2j}/N$.

Based on the sampling procedure given in Section 2, an unbiased estimator for $\overline{Y}_2$ may be taken as

$$\bar{y}_2^* = w^* \bar{y}_{2m}^* + (1-w^*) \bar{y}_{2u} \qquad \ldots \quad (5.1)$$

with

$$\bar{y}^*_{2m} = \bar{y}_{2m} - (\bar{y}_{1m} - \bar{y}_{1n})$$

$$\bar{y}_{2u} = \frac{1}{u} \sum_{j \epsilon s_{2u}} y_{2j} \; ; \; \bar{y}_{hm} = \frac{1}{m} \sum_{j \epsilon s_{2m}} y_{hj} \; (h = 1, 2) \; ; \; \bar{y}_{1n} = \frac{1}{n} \sum_{j \epsilon s_1} y_{1j}$$

where w^* is a suitably chosen weight.

The formulae for $V(\bar{y}_{2u})$, $V(\bar{y}^*_{2m})$, $V(\bar{y}^*_2)$ with optimum weight w^*, optimum value of matched proportion $\lambda = m/n$ and the resulting optimum variance $V_0(\bar{y}^*_2)$ are, then, given by (2.3), (2.4), (3.1) and (3.2) respectively with A_i and B_i replaced by A and B respectively, where

$$A = S^2_{2y} \; ; \; B = 2\rho \; S_{1y} \; S_{2y} - S^2_{1y} \qquad \ldots \quad (5.2)$$

with $$S^2_{hy} = \sum_{j=1}^{N} (y_{hj} - \overline{Y}_h)^2/(N-1), \; h = 1, 2 \; ;$$

ρ being the correlation coefficient between y-values on the two occasions. It may be noted that $B/A \leqslant 1$. It follows that discussion in Section 4 with B_i/A_i replaced by B/A is valid for $\bar{y}^*_2$ as well and same λ may be used for simultaneous estimation of $\overline{Y}_2$ and $\overline{Y}_{2i}$ $(i = 1, \ldots, k)$.

Ignoring the sampling fraction $f = n/N$ and assuming that $S_{1y} = S_{2y}$, we get, using optimum weights w^*,

$$V(\bar{y}^*_2) = \frac{1}{n} A. \; \frac{1+(1-2\rho)\,(1-\lambda)}{1+(1-2\rho)\,(1-\lambda)^2} \qquad \ldots \quad (5.3)$$

which is same as the formula obtained by Raj (1965).

Denoting $\bar{y}^*_2$ by $\bar{y}'_2$ in case regression estimator is used in place of $\bar{y}^*_{2m}$ in (5.1), we have, from Cochran (1977, p. 347),

$$V(\bar{y}'_2) = \frac{1}{n} A \, \frac{1-(1-\lambda)\rho^2}{1-(1-\lambda)^2\rho^2}. \qquad \ldots \quad (5.4)$$

From (5.3) and (5.4), the relative gain in efficiency of $\bar{y}'_2$ over $\bar{y}^*_2$, for same λ, is given by

$$RGE(\bar{y}'_2 \,|\, \bar{y}^*_2) = \frac{V(\bar{y}^*_2) - V(\bar{y}'_2)}{V(\bar{y}'_2)} = \frac{\lambda(1-\lambda)\,(1-\rho)^2}{[1-(1-\lambda)\rho^2][1+(1-2\rho)(1-\lambda)^2]} \quad \ldots \quad (5.5)$$

which monotonically decreases with $\rho (0 < \rho \leqslant 1)$ for a given λ.

In case corresponding optimum matched proportions are used in $\bar{y}^*_2$ and $\bar{y}'_2$, the relative gain in efficiency is given by

$$RGE\,(\bar{y}'_{02} \,|\, \bar{y}^*_{02}) = \frac{V_0(\bar{y}^*_2) - V_0(\bar{y}'_2)}{V_0(\bar{y}'_2)} = \frac{\sqrt{2(1-\rho)} - \sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} \qquad \ldots \quad (5.6)$$

which monotonically decreases with $\rho (0.5 \leqslant \rho \leqslant 1)$.

Table 5.1 gives the percent relative gains in efficiency of $\bar{y}_2'$ over $\bar{y}_2^*$ for $\lambda = 0.25$ and $\lambda = 0.40$; for corresponding optimum λ in both and for optimum λ in $\bar{y}_2'$ but $\lambda = 0.25$ in $\bar{y}_2^*$ when $\rho = 0.5$ (0.1) 0.9.

TABLE 5.1. PERCENT RELATIVE GAIN IN EFFICIENCY OF $\bar{y}_2'$ OVER $\bar{y}_2^*$

ρ	100 RGE $(\bar{y}_2' \mid \bar{y}_2^*)$		100 RGE $(\bar{y}_{02}' \mid \bar{y}_{02}^*)$	100 RGE $(\bar{y}_{02}' \mid \bar{y}_2^*)$
	$\lambda = 0.25$	$\lambda = 0.40$		$\lambda = 0.25$
0.5	5.8	7.1	7.1	7.1
0.6	4.6	5.3	5.2	6.4
0.7	3.4	3.6	3.5	5.4
0.8	2.2	2.0	2.3	3.8
0.9	0.9	0.7	0.8	1.3

Table 5.1 reveals that the relative gain in efficiency of $\bar{y}_2'$ over $\bar{y}_2^*$ is not appreciable, especially when $\rho \geqslant 0.7$. Noting the fact that formulae for variance of $\bar{y}_2^*$ is exact while for that of $\bar{y}_2'$ is valid only for large m (because of the use of regression estimator in place of $\bar{y}_{2m}^*$), the use of $\bar{y}_2^*$ over $\bar{y}_2'$ may be recommended especially when ρ is high and the situation demands a small matched proportion, say $\lambda = 0.25$, which fares well in many situations.

Thus based on the discussions in Sections 4 and 5, one may use, in practice, the estimators $\bar{y}_2^*$ and $\{\bar{y}_{2i}\}$ for the simultaneous estimation of $\bar{Y}_2$ and $\{\bar{Y}_{2i}\}$ $(i=1, \ldots, k)$ respectively especially when $0 < B/A < 1$ and $0 < B_i/A_i < 1$ (or alternatively $\frac{1}{2} < \rho < 1$ and $\frac{1}{2} < \rho_i < 1$) in which case one may use $\lambda = 0.25$ without bothering for optimum λ as the resulting loss in efficiency is not appreciable. In other cases the sample S_1 of the first occasion may either be completely replaced or retained.

Acknowledgement. The author wishes to thank the referee whose suggestions have led to an improvement of the paper.

References

COCHRAN, W. G. (1977): *Sampling Techniques*, Third Edition, John Wiley.

RAJ, D. (1965): On a method of using multi-auxiliary information in sample surveys. *J. Amer. Stat. Assoc.* **60**, 270-277.

Paper received : *December, 1986.*

Revised : *July, 1987.*