Aligarh Journal of Statistics
Vol. 28 (2008), 17-26

ESTIMATION OF RATIOS ON TWO OCCASIONS FOR SEVERAL DOMAINS IN A FINITE POPULATION

T.P. Tripathi, A.K. Das and A. Maity

ABSTRACT

The problem of estimating several domain-ratios on two occasions is considered using partial replacement scheme. The general properties of proposed sampling strategy are studied and the comparison is made with some other strategies. It is advocated that in practice one may retain 25% to 50% of units for the second occasion.

1. INTRODUCTION

Let $U=\{1,2,\cdots,N\}$ be a finite population of N (given) units and y_{hj} and x_{hj} denote the values of characters y and x respectively for $j-$th unit of the population at occasion h. The problem of estimating the population ratio $R_2=\overline{Y}_2 / \overline{X}_2$ on the second occasion has been considered by Tripathi and Sinha (1976), Das (1982) and Chaturvedi and Tripathi (1983) by considering sampling strategies based on partial replacement of units where a portion of the sample on first occasion is retained for observation, in addition to a fresh sample of units drawn on the second occasion.

In many situations of practical importance the estimates are needed not only for the overall ratio R_2 but for ratios at the level of sub-populations as well for which sampling frames are not available. For example, one may need to estimate ratio of female working-force to male working-force for a district as a whole and also separately for different economic/social groups in the district. Similarly the estimates of the ratio of total value of production to total number of workers employed may be needed not only for an industry as a whole but also for different size-groups of that industry. Further the estimates may be needed at two points of time.

In this paper we consider the estimation of ratios $R_{2i}=\overline{Y}_{2i} / \overline{X}_{2i}$ $(i=1,2,\cdots,k)$ for k domains on the second occasion.

2. PROPOSED CLASS OF ESTIMATORS

Let the finite population $U=\{1,2,\cdots,N\}$ be partitioned into k domains on the second occasion. Let the $i-$th domain D_i contain N_{hi} units on the $h-$th occasion $(h=1,2)$ such that $\sum_{i=1}^{k} N_{hi} = N$. Let y_{hij} and x_{hij} denote values of the variates y and x for the $j-$th unit in the $i-$th domain D_i at occasion h; and

$$\overline{Y}_{hi} = \frac{1}{N_{hi}} \sum_{j=1}^{N_{hi}} y_{hij} , \quad \overline{X}_{hi} = \frac{1}{N_{hi}} \sum_{j=1}^{N_{hi}} x_{hij} , \quad (h=1,2) .$$

Let $R_{2i} = \overline{Y}_{2i} / \overline{X}_{2i}$ be the ratio of two means for D_i on the second occasion. For estimation of the domain ratio R_{2i} we propose the following procedure of sample selection.

On the first occasion a sample S_1 of size n is drawn from U using $SRSWOR$. On the second occasion the sample $S_2 = (S_{2m}, S_{2u})$ consists of S_{2m} (matched part) of m units from S_1 retained randomly using $SRSWOR$ and S_{2u} (unmatched part) of u units drawn 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 the $i-$th domain D_i such that

$$\sum_{i=1}^{k} m_{2i} = m \text{ and } \sum_{i=1}^{k} u_{2i} = u .$$

Let

$$\overline{y}_{mhi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} , \quad \overline{x}_{mhi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} x_{hij} , \quad \overline{y}_{n1i} = \frac{1}{n_{1i}} \sum_{j=1}^{n_{1i}} y_{1ij}$$

$$\overline{x}_{n1i} = \frac{1}{n_{1i}} \sum_{j=1}^{n_{1i}} x_{1ij} , \quad \overline{y}_{u2i} = \frac{1}{u_{2i}} \sum_{j=1}^{u_{2i}} y_{2ij} , \quad \overline{x}_{u2i} = \frac{1}{u_{2i}} \sum_{j=1}^{u_{2i}} x_{2ij} , \tag{2.1}$$

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 the $i-$th domain D_i on the first occasion.

Now , for estimating R_{2i} the proposed estimator is given by

$$\hat{R}_{2i} = w\hat{R}_{mi} + (1-w)\,\hat{R}_{u_{2i}} , \tag{2.2}$$

where

$$\hat{R}_{mi} = \frac{\overline{y}_{m2i} - \theta_{1i}\,(\overline{y}_{m1i} - \overline{y}_{n1i})}{\overline{x}_{m2i} - \theta_{2i}\,(\overline{x}_{m1i} - \overline{x}_{n1i})}, \quad \hat{R}_{u2i} = \frac{\overline{y}_{u2i}}{\overline{x}_{u2i}}$$

w is a suitably chosen weight and θ_{1i} and θ_{2i} are some constants not depending on the sampling design.

Following Tripathi (1988) and assuming that m and u are large, we have

$$V(\hat{R}_{u2i}) = \left(\frac{1}{u} - \frac{1}{N}\right) A_i \quad \text{and} \quad V(\hat{R}_{mi}) = \left(\frac{1}{m} - \frac{1}{N}\right) A_i - \left(\frac{1}{m} - \frac{1}{n}\right) B_i , \qquad (2.3)$$

where

$$A_i = \frac{N}{N_{2i}} R_{2i}^2 \, [C_{2yi}^2 + C_{2xi}^2 - 2\, C_{2yxi}]$$

and

$$B_i = \frac{N}{N_{1i}} R_{2i}^2 \, \frac{1}{\overline{Y}_{2i}^2} \left[2 \frac{N_i^*}{N_{2i}} \underline{\theta}' \, \underline{\alpha} - \underline{\theta}' \, G \, \underline{\theta} \right] \text{ with}$$

$$\underline{\theta} = (\theta_{1i}, \, \theta_{2i})', \qquad G = \begin{pmatrix} \sigma_{1yi}^2 & -R_{2i}\sigma_{1yxi} \\ -R_{2i}\sigma_{1yxi} & R_{2i}^2 \sigma_{1xi}^2 \end{pmatrix}$$

and

$$\underline{\alpha} = \begin{pmatrix} \sigma_{12yi}^* - (\overline{Y}_{1i}^* - \overline{Y}_{1i})(R_{2i}\overline{X}_{2i}^* - \overline{Y}_{2i}^*) - R_{2i}\sigma_{1y2x}^* \\ R_{2i}^2 \sigma_{12xi}^* + R_{2i}(R_{2i}\overline{X}_{2i}^* - \overline{Y}_{2i}^*)(\overline{X}_{1i}^* - \overline{X}_{1i}) - R_{2i}\sigma_{1x2y}^* \end{pmatrix}$$

N_i^* is the number of units in D_i common to both the occasions and

$$\overline{Y}_{hi}^* = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} y_{hij} \, , \quad \overline{X}_{hi}^* = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} x_{hij}$$

$$\sigma_{12yi}^* = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} (y_{1ij} - \overline{Y}_{1i}^*)(y_{2ij} - \overline{Y}_{2i}^*)$$

$$\sigma_{12xi}^* = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} (x_{1ij} - \overline{X}_{1i}^*)(x_{2ij} - \overline{X}_{2i}^*)$$

$$\sigma_{1y2x}^* = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} (y_{1ij} - \overline{Y}_{1i}^*)(x_{2ij} - \overline{X}_{2i}^*)$$

$$\sigma^*_{1x2y} = \frac{1}{N_i^*} \sum_{j=1}^{N_i^*} (x_{1ij} - \overline{X}^*_{1i})(y_{2ij} - \overline{Y}^*_{2i})$$

$$\sigma^2_{hyi} = \frac{1}{N_{hi}} \sum_{j=1}^{N_{hi}} (y_{hij} - \overline{Y}_{hi})^2$$

$$C^2_{2yi} = \frac{\sigma^2_{2yi}}{\overline{Y}^2_{2i}}$$

$$\sigma^2_{hxi} = \frac{1}{N_{hi}} \sum_{j=1}^{N_{hi}} (x_{hij} - \overline{X}_{hi})^2$$

$$C^2_{2xi} = \frac{\sigma^2_{2xi}}{\overline{X}^2_{2i}}$$

$$\sigma_{hyxi} = \frac{1}{N_{hij}} \sum_{j=1}^{N_{hi}} (y_{hij} - \overline{Y}_{hi})(x_{hij} - \overline{X}_{hi})$$

$$C_{2yxi} = \frac{\sigma_{2yxi}}{\overline{Y}_{2i}\overline{X}_{2i}} .$$

The value of $\underline{\theta}$ which minimizes $V(\hat{R}_{mi})$ is given by

$$\underline{\theta}_{opt} = \frac{N_i^*}{N_{2i}} G^{-1} \underline{\alpha} \tag{2.4}$$

and the resulting minimal variance is given by

$$V^*(\hat{R}_{mi}) = \left(\frac{1}{n} - \frac{1}{N}\right) A_i + \left(\frac{1}{m} - \frac{1}{n}\right) Q_i , \tag{2.5}$$

where

$$Q_i = A_i - B_i^* \quad \text{and} \quad B_i^* = \frac{N N_i^{*2}}{N_{1i} N_{2i}^2} \frac{1}{\overline{Y}^2_{2i}} R^2_{2i}\, \underline{\alpha}' G^{-1} \underline{\alpha} .$$

To our order of approximation for large samples, optimum w in (2.2) which minimizes the *mse* $M(\hat{R}_{2i})$ is given by $w_0 = \dfrac{M(\hat{R}_{u2i})}{M(\hat{R}_{u2i}) + M(\hat{R}_{mi})}$ and the resulting optimum *mse* is given by $V^*(\hat{R}_{2i}) = \dfrac{M(\hat{R}_{u2i}) M(\hat{R}_{mi})}{M(\hat{R}_{u2i}) + M(\hat{R}_{mi})}$.

Using optimum choice of w in (2.2) and substituting values of $V(\hat{R}_{u2i})$ and $V^*(\hat{R}_{mi})$ from (2.3) and (2.5), we get

$$V^*(\hat{R}_{2i}) = \frac{\frac{1}{m}(Q_i - \lambda f A_i + \lambda B_i^*)(1 - f' + \lambda f)\frac{A_i}{n' - m}}{\frac{1}{m}(Q_i - \lambda f A_i + \lambda B_i^*) + (1 - f' + \lambda f)\frac{A_i}{n' - m}}, \tag{2.6}$$

where, $n' = m + u$ is the sample size on the second occasion, $f = n/N$, $f' = n'/N$ are the sampling fractions and $\lambda = m/n$ is the matched proportion.

Ignoring $1/N$, we have

$$V^*(\hat{R}_{2i}) = \frac{\left(\frac{Q_i}{m} + \frac{B_i^*}{n}\right)\frac{A_i}{n' - m}}{\left(\frac{Q_i}{m} + \frac{B_i^*}{n}\right) + \frac{A_i}{n' - m}} \tag{2.7}$$

In case arbitrary values of $\underline{\theta}$, in (2.3) are used and $n = n'$, it is found that

$$V(\hat{R}_{2i}) = \frac{1}{n}\left(\frac{(1 - f + \lambda f) A_i [(1 - \lambda f) - (1 - \lambda) B_i / A_i]}{(1 - 2\lambda f + 2\lambda^2 f) - (1 - \lambda)^2 B_i / A_i}\right) \tag{2.8}$$

The above expression is similar to that obtained by Tripathi (1988) for estimation of $\overline{Y}_{2i}$.

The expression of w_0 involves some unknown quantities. In practice w_0 can be assessed through available census data or pilot survey data or data provided by the sample at hand.

3. OPTIMUM MATCHING POLICY

It may be shown that A_i, $B_i^* \geq 0$. In case $A_i > B_i^* > 0$, the minimization of $V^*(\hat{R}_{2i})$ in (2.7) w.r.t. m $(0 < m < n)$ yields the optimum value of m given by the solution of the equation

$$B_i^{*2}\frac{m^2}{n^2} + 2 Q_i B_i^* \frac{m}{n} + (Q_i^2 - Q_i A_i) = 0 \tag{3.1}$$

which implies (and conversely)

$$\frac{Q_i}{m} + \frac{B_i^*}{n} = \frac{A_i}{n' - m}. \tag{3.2}$$

Using (3.2) $V^*(\hat{R}_{2i})$ in (2.7) reduces to

$$V_0^*(\hat{R}_{2i}) = \frac{A_i}{(n + n' - 2m_0)}, \tag{3.3}$$

where m_0, the solution of the equation (3.2), is given by

$$m_0 = n\left(\frac{\sqrt{A_i - B_i^*}}{\sqrt{A_i} + \sqrt{A_i - B_i^*}}\right). \tag{3.4}$$

Thus the optimum matched proportion and the resulting minimum variance are given by

$$\lambda_0^{*(i)} = \frac{m_0}{n} = \frac{\left(1 - \frac{B_i^*}{A_i}\right)^{1/2}}{1 + \left(1 - \frac{B_i^*}{A_i}\right)^{1/2}} \tag{3.5}$$

and

$$V_0^*(\hat{R}_{2i}) = \frac{A_i\left(\sqrt{A_i} + \sqrt{A_i - B_i^*}\right)}{(n+n')\sqrt{A_i} - (n+n')\sqrt{A_i - B_i^*}}. \tag{3.6}$$

In case sample sizes are same (n) at both the occasions, (3.6) reduces to

$$V_0^*(\hat{R}_{2i}) = \frac{1}{2n}\sqrt{A_i}\left(\sqrt{A_i} + \sqrt{A_i - B_i^*}\right). \tag{3.7}$$

Let $B_i > 0$. In case arbitrary values of $\underline{\theta}$ are used it is found from (2.8) that the optimum λ is given by

$$\lambda_0^{(i)} = \frac{\left(1 - \frac{B_i}{A_i}\right)^{1/2}}{1 + \left(1 - \frac{B_i}{A_i}\right)^{1/2}} \tag{3.8}$$

and the resulting minimal variance is

$$V_0(\hat{R}_{2i}) = \frac{1}{2n}A_i\left[1 - f + \left(1 - \frac{B_i}{A_i}\right)^{1/2}\right] \tag{3.9}$$

Determination the sign of B_i

The constants θ_{1i} and θ_{2i} in the definition of B_i in (2.3) are in the hands of the statistician and can be chosen such that $B_i > 0$, the good-guessed values of other quantities involved being used, as obtained from a census data or past sample survey data or through a pilot survey. In case it is not possible to obtain $B_i > 0$, one should use complete replacement policy or complete matching policy. From (2.4) and definition of B_i we note that $B_i = \frac{N}{N_{1i}}\left(\frac{R_{2i}}{\bar{Y}_{2i}}\right)^2\left(\frac{N^*}{N_{2i}}\right)^2 \underline{\alpha}' G^{-1}\underline{\alpha} > 0$.

Thus if the choice of $\underline{\theta}$ in B_i is given by (2.4), with good-guessed values of G and $\underline{\alpha}$, it is most likely to yield $B_i > 0$.

4. COMPARISON WITH OTHER ESTIMATORS AND DISCUSSION

Let the sampling strategy proposed in section 2 based on partial matching $(0 < \lambda < 1)$ be denoted by $S_{(\lambda)}$ and the strategies based on complete matching $(\lambda = 1,\ m = n)$ and complete replacement $(\lambda = 0,\ u = n)$ be denoted by $S_{(1)}$ and $S_{(0)}$ respectively.

If n_{2i} and n'_{2i} be the number of units in $\{S_{2m}\}_{m=n}$ and $\{S_{2u}\}_{u=n}$ respectively coming from $i-$th domain D_i. The estimators based on $S_{(1)}$ and $S_{(0)}$ for R_{2i} may be defined by

$$\hat{R}_{2i(1)} = \frac{\frac{1}{n_{2i}}\sum_{j=1}^{n_{2i}} y_{2ij}}{\frac{1}{n_{2i}}\sum_{j=1}^{n_{2i}} x_{2ij}}, \qquad \hat{R}_{2i(0)} = \frac{\frac{1}{n'_{2i}}\sum_{j=1}^{n'_{2i}} y_{2ij}}{\frac{1}{n'_{2i}}\sum_{j=1}^{n'_{2i}} x_{2ij}} \text{ respectively.}$$

For large n, variances of $\hat{R}_{2i(1)}$ and $\hat{R}_{2i(0)}$ are given by

$$V(\hat{R}_{2i(1)}) = V(\hat{R}_{2i(0)}) = \left(\frac{1}{n} - \frac{1}{N}\right)A_i \tag{4.1}$$

which is same as $V(\hat{R}_{2i})$ in (2.8) with $\lambda = 1$ or $\lambda = 0$.

If the sampling fraction f is ignored, from (2.8) and (4.1) we find that the sampling strategy $T_\lambda = (S_{(\lambda)}, \hat{R}_{2i})$ would be better than both the strategies $T_1 = (S_{(1)}, \hat{R}_{2i(1)})$ and $T_0 = (S_{(0)}, \hat{R}_{2i(0)})$ provided $B_i > 0$. It may be noted from (2.3), that $B_i > 0$ provided $\underline{\theta}$ is chosen so as to satisfy $\underline{\theta}'\underline{\alpha} > \frac{N_{2i}}{2N_i^*}\underline{\theta}' G\underline{\theta}$, for example the values satisfying (2.4).

However, in case $\underline{\theta}$ is such that $B_i < 0$, the strategies T_0 and T_1 would be better than T_λ.

It may be shown that $\lambda_0^{*(i)}$ given by (3.5) and $\lambda_0^{(i)}$ given by (3.8) with $B_i > 0$ can never exceed $1/2$.

From (2.8), (3.9) and the definition of A_i it follows that the procedure based on matching and unmatching $(0 < \lambda < 1)$ would be better than those of complete matching $(\lambda = 1)$ and complete replacement $(\lambda = 0)$ in which case $V(\hat{R}_{2i}) = \frac{1}{n} A_i$.

Table 4.1 gives the values of optimum matching $\lambda_0^{(i)}$, in (3.8), and the percent relative efficiency (PRE) of $\lambda_0^{(i)}$ over the choices $\lambda = 0, 1, 1/4, 1/2$ as obtained by $PRE = \frac{V(\hat{R}_{2i})}{V_0(\hat{R}_{2i})} \times 100$ for $f = 0$ in (2.8) and (3.9) and various values of $\frac{B_i}{A_i}$.

Table 4.1: *PRE* of optimum matching over other matching

$\frac{B_i}{A_i}$	Optimum percent matched $100\lambda_0^{(i)}$	Percent relative efficiency of $\lambda_0^{(i)}$ over		
		$\lambda = 0$ or $\lambda = 1$	$\lambda = \frac{1}{4}$	$\lambda = \frac{1}{2}$
0.1	48.7	102.6	100.6	100.0
0.2	47.2	105.6	101.1	100.0
0.3	45.6	108.9	101.5	100.0
0.4	43.6	112.7	101.8	100.2
0.5	41.4	117.2	101.9	100.4
0.6	38.7	122.5	101.7	100.9
0.7	35.4	129.2	101.2	101.8
0.8	30.9	138.2	100.5	103.6
0.9	24.0	151.9	100.0	107.8

From Table 4.1, it is noted that, optimum matching proportion is a monotonically decreasing function of B_i / A_i in $(0, 1)$. The *PRE* of optimum matching $(\lambda_0^{(i)})$ over $\lambda = \frac{1}{2}$ increases monotonically with 0< $0 < \frac{B_i}{A_i} < 1$.

However, *PRE* over $\lambda = \frac{1}{4}$ monotonically increases for $0 < \frac{B_i}{A_i} \le \frac{1}{2}$ and monotonically decreases in the range $\frac{1}{2} < \frac{B_i}{A_i} < 1$.

From Table 4.1, it is observed that the *PRE* of optimum matching over no matching or complete matching increases monotonically for $0 < \frac{B_i}{A_i} < 1$.

It may be advocated that one need not worry about optimum matching policy. If good guessed values of $\frac{B_i}{A_i}$ are available then one may suggest to use the strategy $S_{(\lambda)}$ with $\lambda = 0.5$ for $0 < \frac{B_i}{A_i} \le 0.6$. In all other situations including those in which even approximate values of $\frac{B_i}{A_i}$ (> 0) are not available one may use $S_{(\lambda)}$ with $\lambda = 0.25$ as the resulting loss in precision compared to optimum matching would be negligible.

For estimating the over-all population ratio $R_2 = \frac{\bar{Y}_2}{\bar{X}_2}$, $(\bar{X}_2 \ne 0)$ on the second occasion, one may define an estimator as

$$\hat{R}_2 = \alpha \hat{R}_{2m} + (1-\alpha)\hat{R}_{2u},$$

where $\hat{R}_{2m} = \frac{\bar{y}_{2m} - \theta_1 (\bar{y}_{1m} - \bar{y}_{1n})}{\bar{x}_{2m} - \theta_2 (\bar{x}_{1m} - \bar{x}_{1n})}$ and $\hat{R}_{2u} = \frac{\bar{y}_{2u}}{\bar{x}_{2u}}$ are the estimators based on matched and unmatched parts respectively. The estimator $\hat{R}_2$ is a special case of the estimator discussed by Chaturvedi and Tripathi (1983) and hence all the results related to it, including optimum matching proportion etc. follow immediately. It is found in this case as well that in practice one may choose matched proportion as $\lambda = 0.25$ without any significant loss of precision. Thus in practice for the simultaneous estimation of R_2 and $\underline{R}_2 = (R_{21}, R_{22}, \cdots, R_{2k})'$ the choice of $\lambda = 0.25$ is a suitable one provided $B_i > 0$.

The results for the designs (other than *SRSWOR* at both the occasions) would be similar to those presented in this manuscript provided the variance (*mse*) structures are as given in (2.6), (2.7) and (2.8).

Acknowledgement

The authors are thankful to the Referee for his helpful comments.

REFERENCES

Chaturvedi, D.K. and Tripathi, T.P. (1983): Estimation of population ratio on two occasions using multivariate auxiliary information. *J. Indian Statist. Assoc.*, **21**,113-120.

Das, A.K. (1982): Estimation of population ratio on two occasions. *J. Indian Soc. Agricultural Statist.*, **34**, 1-9.

Tripathi, T.P. and Sinha, S.K.P. (1976): Estimation of ratio on successive occasions. *Symposium on Recent Developments in Survey Methodology*, Indian Statistical Institute, Kolkata.

Tripathi, T.P. (1988): Estimation for domains in sampling on two occasions. *Sankhyā B*, **50**, 103-110.

Received : 16-05-2005

Revised : 14-05-2007

T.P. Tripathi
Maths-Stats Division,
Indian Statistical Institute,
Kolkata, India.

A.K. Das and A. Maity
B.C.K.V., Mohanpur, Nadia-741252, India.
e-mail: akdas_bckv@yahoo.co.in