International Journal of Statistics and Management System
Vol. 4 No. 2 (July-December, 2019)

Some estimators of population product on two occasions*

Amelia Victoria García Luengo† and Inmaculada Oña Casado‡

Received: 23rd March 2018 Revised: 14th April 2018 Accepted: 10th September 2018

Abstract

The problem of estimation of the population product for the current occasion based on the samples selected over two occasions has been considered. This estimator is obtained suitably combining two independent estimates of the population product. One estimate is derived from retained sample using a double-sampling estimator for the product of two means employing information on one auxiliary variable in the first occasion, while the other is an ordinary product estimate derived from the new sample. The expression for optimum estimator, its mean square error, the optimum matched proportion, the gain in efficiency of the proposed estimate over the direct estimate using no information gathered on the first occasion, have been computed. An empirical study is made to study the performance of the proposed strategy.

1 Introduction. The problem of sampling on two successive occasions with a partial replacement of sampling units was first considered by Jessen [8] in the analysis of a survey which collected farm data. Later, the theory was extended by Patterson [12], Rao and Graham [14], Gupta [7], Das [4], Chaturvedi and Tripathi [2], Singh et al. [17], Singh and Priyanka [16], Singh and Vishwakarma [18] among others.

Key words and phrases: Successive sampling, population product estimator, matching fraction, gain in efficiency.

†*Mailing Address:* Department of Statistics and Applied Mathematics, University of Almería, Spain. *E–mail:* amgarcia@ual.es

‡*Mailing Address:* Department of Statistics and Applied Mathematics, University of Almería, Spain. *E–mail:* iocasado@ual.es

However, in many of such repetitive surveys, the estimate of the population ratio and product of two characters for the most recent occasion may be of practical interest. The theory of estimation of the population ratio of two characters over two occasions has been considered by Rao [13], Rao and Pereira [15], Okafor [11], García and Artés [5] among others.

Tripathi and Sinha [19] and Das [4] derived estimators of the population ratio for the most recent occasion when sampling with partial replacement of units. These estimates are obtained by suitably combining two independent estimates of the population ratio. One estimate is derived from retained (matched) sample using the regression method of estimation; while the other is an ordinary ratio estimate derived from the new (unmatched) sample.

Okafor and Arnab [10] also gave some estimators of the population ratio when sampling is done with partial replacement of units. In their case, the estimate of the population total of the character y on the recent occasion is first obtained by a suitable combination of two independent estimates of the population totals from the matched and unmatched samples. The estimate of the population total of x is similarly obtained. These two estimates of the population totals of y and x are then used to derive the estimate of the population ratio.

Further, García [6] presented some sampling strategies for estimating, by a linear estimate, the population product of two characters over two occasions.

In this paper is derived an estimator of the population product for the most recent occasion when sampling with partial replacement of units. This estimator is obtained suitably combining two independent estimates of the population product. One estimate is derived from retained (matched) sample using a double-sampling estimator for the product of two means employing information on one auxiliary variable in the first occasion (Khare [9]); while the other is an ordinary product estimate derived from the new (unmatched) sample. An empirical study is carried out to support the proposed strategy.

2 The sampling scheme and Proposed general estimator

Let Y_{1j}, Y_{2j} and X_{1j} be the value of j-th unit of the characters y_1, y_2 and x_1 respectively. The population means of the main characters y_1, y_2 and the auxiliary character x_1, be denoted by $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{X}_1$, respectively. In case when $\bar{X}_1$ is unknown the use of two phase sampling scheme is proposed. Let $\bar{x}'_1$ denote the sample mean of the auxiliary character, x_1 based on first phase sample of size n' and $\bar{y}_1$, $\bar{y}_2$ and $\bar{x}_1$ denote the sample means based on second phase sample of size n of the characters y_1, y_2, and x_1 respectively.

Suppose that the samples are of size n on both occasions, we use a simple random sampling and the size of the population N is sufficiently great for the factor of correction be ignored.

Let a simple random sample of size n be selected on the first occasion from a universe of size N. Information about auxiliary variable x_1 is available for the first occasion. Let y_1 and y_2 be the variable under study on the second occasion. When selecting the second sample, we assume that $m = pn$ $(0 < p < 1)$ of the units of the selected sample on the first occasion are retained for the second occasion (matched sample) and the remaining $u = n - m = qn$ $(q = 1 - p)$ units are replaced by a new selection from the universe $N - m$ left after omitting the m units.

We denote

$$C_i^2 = \frac{S_i^2}{\bar{Y}_i^2}, \qquad S_i^2 = \frac{\sum_{j=1}^{N}(Y_{ij} - \bar{Y}_i)^2}{N}, \qquad i = 1, 2$$

$$C_0^2 = \frac{S_0^2}{\bar{X}_1^2} \qquad S_0^2 = \frac{\sum_{j=1}^{N}(X_{ij} - \bar{X}_1)^2}{N}$$

The correlation coefficients between (y_1, y_2), (y_1, x_1) and (y_2, x_1) are denoted by ρ_0, ρ_1 y ρ_2 respectively.

$P_2 = \bar{Y}_1\bar{Y}_2$ the population product on the second occasion.

$\widehat{P}_2 = \bar{y}_1\bar{y}_2$ the estimator of the population product on the second occasion.

$\widehat{P}_{2m}$, the estimator of the population product on the second occasion based on the matched sample of m units.

$\widehat{P}_{2u}$, the estimator of the population product on the second occasion based on the unmatched sample of u units.

Hence, we construct an estimate of the population product on the second occasion, $\widehat{P}_2'$, by combining the two independent estimates $\widehat{P}_{2m}'$ and $\widehat{P}_{2u}$, with weights Q and $1-Q$. Thus

$$\widehat{P}_2' = Q\widehat{P}_{2u} + (1 - Q)\widehat{P}_{2m}'$$

For the unmatched portion a direct estimate of the population product, based on the u sampling units is used.

For the matched portion an estimate improved of the direct estimate, may be obtained using a double sampling estimate suggested by Khare [9].

$$\widehat{P}'_{2m} = g(w,u), \quad w = \bar{y}_1\bar{y}_2, \quad u = \frac{\bar{x}'_1}{\bar{x}_1}.$$

The function $g(w,u)$ is such that

$$g(P_2,1) = P_2, \qquad g_1(P_2,1) = 1$$

and satisfy the following conditions:

1. Whatever be the sample chosen, (w,u) assumes values in bounded closed convex subset D_1, of two dimensional real space containing the point $(P_2,1)$.
2. The function $g(w,u)$, and its first and second order partial derivatives exit and are continuous and bounded in D_1.

Under the conditions specified for the function $g(w,u)$ and expanding the function $g(w,u)$ about the point $(P_2,1)$ (Khare [9]), with the help of Taylor´s series up to second order derivatives, we have the expressions of bias (B)

$$\mathrm{B}(\widehat{P}'_{2m}) = P_2\frac{f}{m}\rho_0C_1C_2 + \left(\frac{f}{m} - \frac{f'}{n}\right)C_0^2\left(g_2(P_2,1) + \frac{1}{2}g_{22}(P_2,1)\right) +$$

$$+\frac{1}{2}\left(P_2^2\frac{f}{m}W_1g_{11}(P_2,1) - 2P_2\left(\frac{f}{m} - \frac{f'}{n}\right)W_2C_0g_{12}(P_2,1)\right)$$

where

$$g_1(w,u) = \frac{\partial}{\partial w}g(w,u); \quad g_2(w,u) = \frac{\partial}{\partial u}g(w,u); \quad g_{12}(w,u) = \frac{\partial^2}{\partial u\partial w}g(w,u)$$

$$g_{11}(w,u) = \frac{\partial^2}{\partial w^2}g(w,u); \quad g_{22}(w,u) = \frac{\partial^2}{\partial u^2}g(w,u)$$

$$W_1 = C_1^2 + C_2^2 + 2\rho_0C_1C_2; \qquad W_2 = \rho_1C_1 + \rho_2C_2$$

$$f = \frac{N-m}{N}; \qquad f' = \frac{N-n}{N}$$

Following Cochran [3], the first order approximation to the mean square error (M.S.E) of $\widehat{P'}_{2m}$, assuming an infinite population, is given by

$$\text{MSE }(\widehat{P}'_{2m}) = P_2^2\left[\frac{f}{m}W_1 - \left(\frac{f}{m} - \frac{f'}{n}\right)W_2^2\right]$$

The expression for mean square error of $\widehat{P}'_2$ is

$$MSE\left(\widehat{P}_2'\right)_{min} = \frac{MSE\left(\widehat{P}_{2m}'\right) MSE\left(\widehat{P}_{2u}\right)}{MSE\left(\widehat{P}_{2m}'\right) + MSE\left(\widehat{P}_{2u}\right)} =$$

$$= \frac{P_2^2}{n}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)\frac{1-qZ}{1-q^2 Z} \tag{2.1}$$

where

$$Z = \frac{(\rho_1 C_1 + \rho_2 C_2)^2}{(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)}$$

Also, the mean square error of $\widehat{P}_{2u}$, is given by

$$MSE(\widehat{P}_{2u}) = \frac{P_2^2}{u}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)$$

The best estimate of the P_2 on the second occasion is obtained by using the values of Q that minimize the mean square error (M.S.E.) of $\widehat{P}_2'$,

$$Q_{opt} = \frac{MSE\left(\widehat{P}_{2u}\right)}{MSE\left(\widehat{P}_{2u}\right) + MSE\left(\widehat{P}_{2m}'\right)} = \frac{p}{1 + (1-p)^2 Z}$$

So,

$$MSE(\widehat{P}_{2m}') = \frac{P_2^2}{m}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)[1 - qZ]; \qquad Z = \frac{(\rho_1 C_1 + \rho_2 C_2)^2}{(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)}$$

If, however, the direct estimate $\widehat{P}_{2m}$ based on the m samplings units its mean square error would be

$$MSE(\widehat{P}_{2m}) = \frac{P_2^2}{m}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)$$

We obtain that $\widehat{P}_{2m}'$ is more efficient than $\widehat{P}_{2m}$ if

$$Z = \frac{(\rho_1 C_1 + \rho_2 C_2)^2}{(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)} \geq 0$$

Equating to zero the derivative of $MSE(\widehat{P}_2')$, (2.1), with respect to q, we find that the mean square error of $\widehat{P}_2'$ will have its minimum value if we choose

$$\widehat{q}_{opt} = \frac{1 - \sqrt{1-Z}}{Z}; \qquad Z = \frac{(\rho_1 C_1 + \rho_2 C_2)^2}{(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)}$$

and

$$MSE_{opt}(\widehat{P}_2') = \frac{P_2^2}{n}\left(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2\right)\frac{1 - \widehat{q}_{opt} Z}{1 - \widehat{q}_{opt}^2 Z}$$

3 Comparison with other estimators The simple estimator of the population product on the current occasion $\widehat{P}_2$ is based exclusively on the n sampling units for the second occasion, using no information gathered on the first occasion, its mean square error is given by

$$MSE(\widehat{P}_2) = \frac{P_2^2}{n}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)$$

We can compute the gain in precision G of the combined estimate $\widehat{P}'_2$, obtained by using a double-sampling estimator for the product of two means employing information on one auxiliary variable in the first occasion, over the direct estimate, $\widehat{P}_2$

$$G = \frac{MSE(\widehat{P}_2) - MSE_{min}(\widehat{P}'_2)}{MSE_{min}(\widehat{P}'_2)} = \frac{Zp(1-p)}{1-(1-p)Z}; \quad Z = \frac{(\rho_1 C_1 + \rho_2 C_2)^2}{(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2)}$$

Necessarily $p \leq 1$. If $p = 1$ (perfect matching) or $p = 0$ (no matching), the gain is zero. For other p $(0 < p < 1)$, there will be positive gain if $Z \geq 0$.

3.1 Special case When

$$\rho_1 = \rho = \rho_2 \qquad C_1 = C_2 = C$$

we have

$$MSE(\widehat{P}'_{2m(1)}) = \frac{P_2^2}{m} 2C^2(1+\rho_0)\left[1 - q\frac{2\rho^2}{1+\rho_0}\right]$$

and

$$MSE_{min}\left(\widehat{P}'_{2(1)}\right) = \frac{P_2^2}{n} 2C^2(1+\rho_0)\frac{1-qZ}{1-q^2Z}; \qquad Z = \frac{2\rho^2}{1+\rho_0}$$

The optimum q is given by

$$\widehat{q}_{opt(1)} = \frac{1 - \sqrt{1 - \frac{2\rho^2}{1+\rho_0}}}{\frac{2\rho^2}{1+\rho_0}}$$

The gain in precision, G_1, of the combined estimate, $\widehat{P}'_{2(1)}$, over the direct estimate is given by

$$G_1 = \frac{MSE(\widehat{P}_2) - MSE_{min}(\widehat{P}'_{2(1)})}{MSE_{min}(\widehat{P}'_{2(1)})} = \frac{p\,(1-p)\left(\frac{2\rho^2}{1+\rho_0}\right)}{1-(1-p)\left(\frac{2\rho^2}{1+\rho_0}\right)}$$

For other p $(0 < p < 1)$, there will be positive gain if

$$\frac{2\rho^2}{1+\rho_0} \geq 0$$

We conclude that the gain in precision of the combined estimate, $\widehat{P}'_{2(1)}$, over the direct estimate, $\widehat{P}_2$, increase with increasing ρ value (larger dependence between the auxiliary variable x_1 and the variables under study y_1 and y_2), and decreasing ρ_0 (smaller correlation between y_1 and y_2). Figure 1 show the optimum matching fraction (decreasing when larger dependence between the auxiliary variable x_1 and the variables under study y_1 and y_2) and gain in precision of the combined estimate.

Figure 1: Gain in precision and optimum matching fraction of ρ_0 against ρ_0

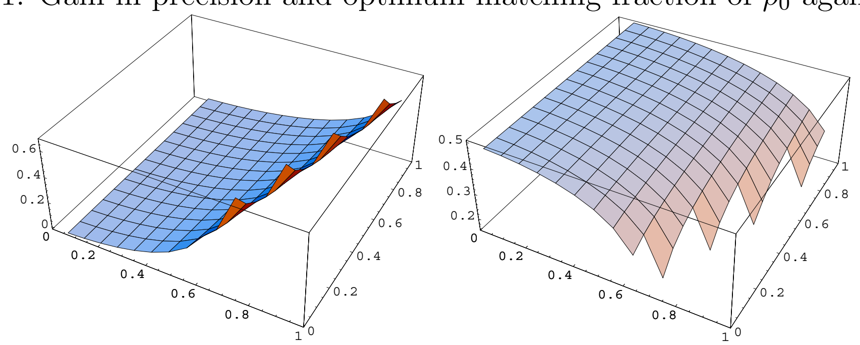


Table 1 show the gain in precision, G_1 of the estimate proposed, $\widehat{P}'_{2(1)}$, over the direct.

Table 1: Gain in precision, G_1

		$\rho = 0.3$			$\rho = 0.6$			$\rho = 0.9$		
$\rho_0 \downarrow$	$q \rightarrow$	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
0.3		3.1	3.7	3.3	12.3	16.6	16.1	25.8	40	51
0.6		2.3	2.9	2.5	9.5	12.5	11.6	22.1	33	38
0.9		2.1	2.6	2.2	6.9	8.8	7.9	18.6	26.9	28.8

3.2 Particular case

When

$$\rho_1 = \rho_2 = \rho_0 = \rho \qquad C_1 = C_2 = C$$

we have

$$MSE(\widehat{P}'_{2m(2)}) = \frac{P_2^2}{m} 2C^2(1+\rho)\left[1 - q\frac{2\rho^2}{1+\rho}\right]$$

and

$$MSE_{min}\left(\widehat{P}'_{2(2)}\right) = \frac{P_2^2}{n} 2C^2(1+\rho)\frac{1-qZ}{1-q^2Z}; \qquad Z = \frac{2\rho^2}{1+\rho}$$

The optimum q is given by

$$\widehat{q}_{opt(2)} = \frac{1 - \sqrt{1 - \frac{2\rho^2}{1+\rho}}}{\frac{2\rho^2}{1+\rho}}$$

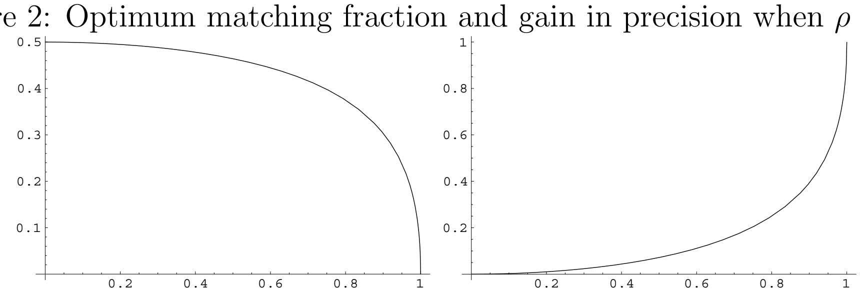
The gain in precision, G_2, of the combined estimate, $\widehat{P}'_{2(2)}$, over the direct estimate is given by

$$G_2 = \frac{MSE(\widehat{P}_2) - MSE_{min}(\widehat{P}'_{2(2)})}{MSE_{min}(\widehat{P}'_{2(2)})} = \frac{p\,(1-p)\left(\frac{2\rho^2}{1+\rho}\right)}{1-(1-p)\left(\frac{2\rho^2}{1+\rho}\right)}$$

Figure 2 show for a series of values of ρ the optimum that should be matched and the gain in precision compared with no matching. The best percentage to match never exceeds 50% and decrease steadily as ρ increases. The greatest attainable gain in precision is 100% when $\rho = 1$. Unless ρ is high, the gains are modest.

Although the optimum percentage to match varies with the correlation coefficient, only a single percentage can be used in practice for all items in a survey.

Figure 2: Optimum matching fraction and gain in precision when $\rho = \rho_0$



4 Empirical study We have used the data collected in a survey on healthy habits and fitness level to assess the optimal operation of the proposed method. This study was carried out over a population of fourteen-year-old schoolchildren in Almería schools during April and June, 1998. (Casimiro [1])

In order to achieve the targets of the study, we have considered the estimation of the endomorphic component (y, one of the multiple variables which affect the survey) at the second occasion, taking as auxiliary variables body mass index (x_1), the add of fold (x_2), the maximum volume of oxygen (x_3) and the arm maintained flexion (x_4) from the first occasion.

The sampling data regarding the number of schoolchildren and the parameters obtained from the two occasions were as follows:

First Occasion (April' 98): Large sample $n = 337$, among the 2681 schoolchildren conforming the population.

Second Occasion (June' 98): Matched sample $m = 135$, unmatched sample $u = 202$.

$$\hat{C}_1 = 0.420 \qquad \hat{\rho}_1 = 0.710$$

$$\hat{C}_0 = 0.173 \qquad \hat{\rho}_0 = 0.730$$

$$\hat{C}_2 = 0.417 \qquad \hat{\rho}_2 = 0.773$$

From these data we can state that

$$MSE_{min}\left(\widehat{P}_2'\right) = 0.80\frac{P_2^2}{n}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2) < \frac{P_2^2}{n}(C_1^2 + C_2^2 + 2\rho_0 C_1 C_2) = MSE\left(\widehat{P}_2\right)$$

which means a gain in precision of 25% of the estimator $\widehat{P}_2'$ over the usual estimator $\widehat{P}_2$. The best percentage to match never exceeds 50% and we have calculated the optimum matching fraction $\widehat{p}_{opt} = 37.82\%$.

In the special case

$$\rho_1 = \rho_2 = 0.773 \qquad \rho_0 = 0.73 \qquad C_1 = C_2 = C$$

we have

$$MSE_{min}\left(\widehat{P}_{2(1)}'\right) = 0.78\frac{P_2^2}{n}2C^2(1+\rho_0) < \frac{P_2^2}{n}2C^2(1+\rho_0)$$

which means a gain in precision of 28% of the estimator $\widehat{P}_{2(1)}'$ over the usual estimator $\widehat{P}_2$ and the optimum matching fraction $\widehat{p}_{opt(1)} = 35.76\%$.

In the particular case

$$\rho_1 = \rho_2 = \rho_0 = 0.773 \qquad C_1 = C_2 = C$$

we have

$$MSE_{min}\left(\widehat{P}'_{2(2)}\right) = 0.79\frac{P_2^2}{n}2C^2(1+\rho_0) < \frac{P_2^2}{n}2C^2(1+\rho_0)$$

which means a gain in precision of 27% of the estimator $\widehat{P}'_{2(2)}$ over the usual estimator $\widehat{P}_2$ and the optimum matching fraction $\widehat{p}_{opt(2)} = 36.98\%$.

So, the estimator $\widehat{P}'_{2(1)}$ is more accurate than the estimators $\widehat{P}'_2$ and $\widehat{P}'_{2(2)}$.

Acknowledgement. The authors wishes to express his gratitude to Casimiro for the use of his numerical data.

References

[1] Casimiro, A. J. (1999). Comparación, evolución y relación de hábitos saludables y nivel de condición física–salud en escolares, entre final de Educación Primaria (12 años) y final de Educación Secundaria Obligatoria (16 años), *Tesis Doctoral*, Universidad de Granada.

[2] Chaturvedi, D. K. and Tripathi T.P. (1983). Estimation of population ratio on two occasions using multivariate auxiliary information. *Journal of the Indian Statistical Association* **21** 113–120.

[3] Cochran, W. G. (1977). Sampling Techniques, Third Edition. *John Wiley* & *Sons*, New York.

[4] Das, K. (1982). Estimation of population ratio on two occasions. *Journal of the Indian Society of Agricultural Statistics.* **34 (2)** 1–9.

[5] García, A. V. and Artés, E. M. (2002). Improvement on estimating of current population ratio in successive sampling. *Brazilian Journal of Probability and Statistics.* **16 (2)** 107–122.

[6] García, A. V. (2008). Estimation of current population product in successive sampling. *Pakistan Journal of Statistics.* **24 (2)** 87–98.

[7] Gupta, P. C. (1979). Some estimation problems in sampling using auxiliary information. *Ph. D. thesis IARS*, New Delhi.

[8] Jessen, R. J. (1942). Statistical Investigation of a Sample Survey for Obtaining Farm Facts. *Iowa Agricultural Experiment Statistical Research Bulletin*. **304**.

[9] Khare, B. B. (1991). Determination of sample sizes for a class of two phase sampling estimators for ratio and product of two population means using auxiliary character. *Metron: Revista Internazionale di Statistica*. **49** 185–197.

[10] Okafor, F. C. and Arnab, R. (1987). Some strategies of two-stage sampling for estimating population ratios over two occasions. *Australian Journal of Statistics*. **29 (2)** 128–142.

[11] Okafor, F. C. (1992). The theory and application of sampling over two occasions for the estimation of current population ratio. *Statistica*. **1** 137–147.

[12] Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*. **B12** 241–255.

[13] Rao, J. N. K. (1957). Double ratio estimate in forest surveys. *Journal of the Indian Society of Agricultural Statistics*. **9** 191–204.

[14] Rao, J. N. K. and Graham, J. E. (1964). Rotation design for sampling on repeated occasions. *JASA*. **59** 492–509.

[15] Rao, J. N. K. and Pereira, N. P. (1968). On double ratio estimators. *Sankhya*. **30 A** 83–90.

[16] Singh, G. N. and Priyanka, K. (2007). Estimation of population mean at current occasion in successive sampling under a super-population model. *Model Assisted Statistics and Applications*. **2** 189–200.

[17] Singh, V. K., Singh, G. N. and Shukla, D. (1991). An efficient family of ratio cum difference-type estimators in successive sampling over two occasions *Journal of Scientific Research* **41** 149–159.

[18] Singh, H. P. and Vishwakarma, G. K. (2007). A general class of estimators in successive sampling. *Metron: Revista Internazionale di Statistica*. **LXV (2)** 201–227.

[19] Tripathi, T. P. and Sinha, S. K. P. (1976). Estimation of ratio on successive occasions. *Proceedings of Conference. Recent Developments in Survey Sampling*, Calcutta.